

Studies in Computational Intelligence 1130

Alexei V. Samsonovich
Tingting Liu *Editors*

Biologically Inspired Cognitive Architectures 2023

Proceedings of the 14th Annual Meeting
of the BICA Society

MOREMEDIA



Springer

Series Editor

Janusz Kacprzyk, *Polish Academy of Sciences, Warsaw, Poland*

The series “Studies in Computational Intelligence” (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

Indexed by SCOPUS, DBLP, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.


Alexei V. Samsonovich · Tingting Liu
Editors

Biologically Inspired Cognitive Architectures 2023

Proceedings of the 14th Annual Meeting of the
BICA Society

Editors

Alexei V. Samsonovich 
Department of Cybernetics
Moscow Engineering Physics Institute
Moscow, Russia

Tingting Liu 
College of Science and Technology
Ningbo University
Ningbo, China

ISSN 1860-949X

ISSN 1860-9503 (electronic)

Studies in Computational Intelligence

ISBN 978-3-031-50380-1

ISBN 978-3-031-50381-8 (eBook)

<https://doi.org/10.1007/978-3-031-50381-8>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

Preface

This volume of Studies in Computational Intelligence contains papers presented at the 2023 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence, the 14th Annual Meeting of the BICA Society (BICA*AI 2023), held on October 13–15, 2023, in Ningbo, China. The meeting was a hybrid event, including offline and online venues. Offline sessions were held in Ningbo Golden Port Hotel and in Shangri La Hotel. As always at BICA conferences, technical sessions included many discussion panels and were complemented by extraordinary social special events at the end of each day.

Biologically Inspired Cognitive Architectures (BICA) are computational frameworks for building intelligent agents that are inspired from biological intelligence. Biological intelligent systems, notably animals such as humans, have many qualities that are often lacking in artificially designed systems including robustness, flexibility, and adaptability to environments. At a point in time where visibility into naturally intelligent systems is exploding thanks to modern brain imaging and recording techniques, our ability to learn from nature and to build biologically inspired intelligent systems has never been greater. At the same time, the growth in computer science and technology has unleashed enough computational power that an explosion of intelligent applications from augmented reality to naturally speaking intelligent virtual agents is now certain. The growth in these fields challenges the computational replication of all essential aspects of the human mind (the BICA Challenge), an endeavor which is interdisciplinary in nature and promises to yield bi-directional flow of understanding between all involved disciplines.

The distinguishing feature of this conference is the combination of its broad multidisciplinary nature with the narrow focus on research and technological achievements, aimed at advancing us toward the BICA Challenge. BICA*AI is famous for its friendly, informal atmosphere, for its spirit of excitement and opportunity, providing the fruitful grounds for the emergence of new key initiatives and collaborations.

BICA Conference Series were preceded by the AAAI Fall Symposia on BICA, held in 2008–2009. Originally, participants were mostly the members of the DARPA BICA Program of 2005–2006; however, the audience started expanding rapidly. As a result, in 2010 the BICA Conference Series was initiated as a separate venue, managed by a newly formed nonprofit: BICA Society. For 13 years, the BICA conference was organized around the world (USA, Italy, Ukraine, France, Russia, Czech Republic, Brazil, Japan, Austria, Mexico), demonstrating impressive success in growing progression. BICA Society membership reached many hundreds. BICA Society published a number of books and journals, some of them on the regular basis. Today BICA community is a well-recognized phenomenon in the scientific world.

The scope of the conference is sufficiently broad and includes artificial intelligence, cognitive science, neuroscience, computer science, social, economic and educational sciences, and more. At the same time, the conference has a narrow focus. It gives preference to ambitious innovative research, in particular, to research addressing the BICA Challenge.

The conference was organized by the College of Science and Technology at Ningbo University and sponsored by BICA Society (the Biologically Inspired Cognitive Architectures Society, a nonprofit based in the USA). The list of coorganizers includes Ningbo Municipal Economic and Informatization Technology Bureau, Ningbo Municipal Education Bureau, Ningbo University, and Ningbo Jiangdong Xueyuan Education Information Consulting Co., Ltd. Among additional sponsors are AGI Laboratory and State Key Laboratory of New Textile Materials and Advanced Processing Technologies. BICA*AI 2023 was chaired by Prof. Caiming Zhong from Ningbo University. The Program Committee was jointly chaired by Profs. Tingting Liu and Alexei V. Samsonovich.

The first two chapters include selected presentation abstracts. The rest of the chapters are individual conference papers. All papers included in this volume underwent careful peer review and refereed selection. Each paper was evaluated by at least 3 anonymous reviewers in EasyChair and checked for plagiarism. Each work was presented at the conference and received feedback. This volume includes 111 papers selected among nearly 200 submissions. This publication was made possible thanks to the joint efforts of all the authors, the Program Committee, and Dr. Leontina Di Cecco, Springer Verlag GmbH Senior Editor. Our many thanks go to her.

The 2023 conference was a great success. The Chairs wish to thank all BICA*AI 2023 participants, authors, reviewers, sponsors, coorganizers, partners, staff and technical support personnel, volunteers, attendees, online viewers, and visitors. Materials of the conference are not limited to this volume and include video recordings, other online media, and more. Materials shall remain open to the public and available via the conference web site at <https://bica2023.org>.

Ningbo, China
October 2023

Alexei V. Samsonovich
Tingting Liu

Organization

2023 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence, the 14th Annual Meeting of the BICA Society (BICA*AI 2023), Ningbo, China, October 13–15.

Organizers

BICA Society (BICA*AI Conference Series Organizer)
AGI Laboratory
College of Science and Technology, Ningbo University
Ningbo University
Ningbo Digital Economy Innovation and Development Research Institute

Core Organizing Committee

General Chair and Program Committee Chair: Prof. Caiming Zhong, Ph.D., College of Science and Technology, Ningbo University
Vice Chair, Program Chair and Local Committee Chair: Prof. Tingting Liu, Ph.D., College of Science and Technology, Ningbo University
Program Committee Chair: Prof. Alexei V. Samsonovich, Ph.D., National Research Nuclear University “MEPhI”
Financial Chair: Kyrstin Atreides, Chief Operations Officer, AGI Laboratory
Secretary: Prof. Félix Ramos, D.Sc., Cinvestav, Guadalajara

Program Committee Chairs

Prof. Tingting Liu, Ph.D., College of Science and Technology, Ningbo University
Prof. Alexei V. Samsonovich, Ph.D., National Research Nuclear University “MEPhI”
Prof. Caiming Zhong, Ph.D., College of Science and Technology, Ningbo University

International Organizing Committee

David Kelley, AGI Laboratory and BCG, Seattle, WA, USA
Prof. Jan Treur, Ph.D., VU University Amsterdam, Netherlands
Prof. Junichi Takeno, Ph.D., Meiji University, Nakano, Japan
Prof. Minhua Eunice Ma, Ph.D., Falmouth University, UK

Prof. Zulfiqar Memon, Ph.D., National University of Computer and Emerging Sciences (FAST-NU), Karachi, Pakistan
Azizi Ab Aziz, Director at University Utara Malaysia
Olivier Georgeon, Ph.D., Lyon Catholic University, France
Yoones A. Sekhavat, Ph.D., Tabriz Islamic Art University, Iran
Ursula Addison, Ph.D. student in the computer science program at CUNY Graduate Center in New York, USA

BICA Society Officers

David Kelley, President
Kyrtin Atreides, Treasurer
Felix Ramos, Secretary
Caiming Zhong, Director
Tingting Liu, Director
Jan Treur, Advisory
Alexei V. Samsonovich, Advisory

Local Organizing Committee

Prof. Caiming Zhong, Ph.D.: Chair of Local Committee
Prof. Tingting Liu, Ph.D.: Chair of Local Committee
Prof. Zhen Liu, Ph.D.: Chair of Local Committee
Lei Jiang, Ph.D.: Leader of Secretary Team
Lining Xu: Leader of Logistics Team (accommodation and transportation)

Conference Venues

Shangri La Hotel Ningbo: 88 Yuyuan Street, Ningbo, Zhejiang 315040 China
Golden Port Hotel Ningbo: No. 51, Yangshan Road, Ningbo 315020 China

Invited Keynote Speakers

Prof. Angelo Cangelosi, Ph.D., University of Manchester, UK
Prof. John E. Laird, Ph.D., University of Michigan, USA
Prof. Antonio Lieto, Ph.D., University of Salerno, Italy
Paul Robertson, Ph.D., Chief Scientist at Dynamic Object Language Labs, Inc., USA
Prof. Ron Sun, Ph.D., Rensselaer Polytechnic Institute, USA

Program Committee

Taisuke Akimoto, Kyutech, Japan
Kenji Araki, Hokkaido University, Japan
Joscha Bach, AI Foundation, USA
Feras Batarseh, Virginia Tech, USA
Paul Baxter, Plymouth University, USA
Paul Benjamin, Pace University, New York, USA
Galina A. Beskhebnova, Scientific Research Institute for System Analysis RAS, Russia
Jordi Bieger, Reykjavik University, Iceland
Perrin Bignoli, Yahoo Labs, USA
Douglas Blank, Bryn Mawr College, USA
Peter Boltuc, University of Illinois at Springfield, USA
Jonathan Bona, University of Arkansas for Medical Sciences, USA
Michael Brady, Boston University, USA
Mikhail Burtsev, Moscow Institute of Physics and Technology, Russia
Erik Cambria, Nanyang Technological University, Singapore
Sahas Chelian, Quantum Ventura, USA
Olga Chernavskaya, P. N. Lebedev Physical Institute, Moscow, Russia
Thomas Collins, University of Southern California (Information Sciences Institute), USA
Xuyao Dai, NingBo University, China
Christopher Dancy, The Pennsylvania State University, University Park, USA
Haris Dindo, University of Palermo, Italy
Sergey A. Dolenko, D. V. Skobeltsyn Institute of Nuclear Physics, M. V. Lomonosov Moscow State University, Russia
Anatoly Dolgikh, National Research Nuclear University MEPhI, Russia
Alexandr Eidlin, Sber, Moscow, Russia
Jim Eilbert, AP Technology, USA
Thomas Eskridge, Florida Institute of Technology, USA
Usef Faghihi, Professor At Universite de Quebec in Trois-rivier, Canada
Elena Fedorovskaya, Rochester Institute of Technology, USA
Marcello Frixione, University of Genova, Italy
Salvatore Gaglio, University of Palermo, Italy
Olivier Georgeon, Claude Bernard Lyon 1 University, France
John Gero, University of North Carolina at Charlotte, USA
Jaime Gomez, Universidad Politécnica de Madrid, Spain
Ricardo R. Gudwin, University of Campinas (Unicamp), Brazil
Eva Hudlicka, Psychometrix Associates, USA
Dusan Husek, Institute of Computer Science, Academy of Sciences, Czech Republic
Christian Huyck, Middlesex University, UK
Ignazio Infantino, Consiglio Nazionale delle Ricerche, Italy

Eduardo Izquierdo, Indiana University, USA
Alex James, Digital University Kerala
Li Jinhai, Kunming University of Science and Technology, China
Magnus Johnsson, Lund University, Sweden
Darsana Josyula, Bowie State University, USA
Omid Kavehei, The University of Sydney, Australia
David Kelley, Artificial General Intelligence Inc., USA
Troy Kelley, U.S. Army Research Laboratory, USA
William Kennedy, George Mason University, USA
Deepak Khosla, HRL Laboratories LLC, USA
Muneo Kitajima, Nagaoka University of Technology, Japan
Valentin Klimov, National Research Nuclear University MEPhI, Russia
Unmesh Kurup, LG Electronics, USA
Giuseppe La Tona, Consiglio Nazionale delle Ricerche—INM, USA
Luis Lamb, Federal University of Rio Grande do Sul, Brazil
Leonardo Lana de Carvalho, Federal University of Jequitinhonha and Mucuri Valleys, Brazil
Othalia Larue, University of Quebec, Canada
Christian Lebiere, Carnegie Mellon University, USA
Jürgen Leitner, Australian Centre of Excellence for Robotic Vision, Australia
Simon Levy, Washington and Lee University, USA
Antonio Lieto, University of Turin, Italy
Tingting Liu, College of Science and Technology, Ningbo University, China
James Marshall, Sarah Lawrence College, USA
Olga Mishulina, PF “Logos” LLC, Russia
Sergey Misyurin, National Research Nuclear University MEPhI, Moscow, Russia
Steve Morphet, Enabling Tech Foundation, USA
Amitabha Mukerjee, Indian Institute of Technology Kanpur, India
Daniele Nardi, Sapienza University of Rome, Italy
Sergei Nirenburg, Rensselaer Polytechnic Institute, New York, USA
David Noelle, University of California Merced, USA
Natalia Nosova, National Research Nuclear University MEPhI
Andrea Omicini, Alma Mater Studiorum—Università di Bologna, Italy
Marek Otahal, Czech Institute of Informatics, Robotics and Cybernetics, Czech Republic
Nikolay Pak, National Research Nuclear University MEPhI, Russia
Aleksandr I. Panov, Moscow Institute of Physics and Technology, Russia
Giovanni Pilato, ICAR-CNR, Italy
Roberto Pirrone, University of Palermo, Italy
Michal Ptaszynski, Kitami Institute of Technology, Japan
Nicholas Pym, BBD, Ireland
Uma Ramamurthy, Baylor College of Medicine, Houston, USA
Felix-Francisco Ramos-Corchado, CINVESTAV GDL, Mexico
Thomas Recchia, US Army ARDEC, USA
Vladimir Redko, Scientific Research Institute for System Analysis RAS, Russia
James Reggia, University of Maryland, USA

Frank Ritter, The Pennsylvania State University, USA
Paul Robertson, DOLL Inc., USA
Christopher Rouff, Johns Hopkins Applied Physics Laboratory, USA
Rafal Rzepka, Hokkaido University, Japan
Ilias Sakellariou, Department of Applied Informatics, University of Macedonia, Greece
Fredrik Sandin, Lulea University of Technology, Sweden
Ricardo Sanz, Universidad Politecnica de Madrid, Spain
Michael Schader, Yellow House Associates, USA
Howard Schneider, Sheppard Clinic North, Canada
Michael Schoelles, Rensselaer Polytechnic Institute, USA
Valeria Seidita, Dipartimento di Ingegneria—Università degli Studi di Palermo, Italy
Ignacio Serrano, Instituto de Automtica Industrial—CSIC, Spain
Javier Snaider, FedEx Institute of Technology, The University of Memphis, USA
Donald Sofge, Naval Research Laboratory, USA
Mehae Song, Simon Fraser University, Canada
John Sowa, Kyndi, Inc., USA
Terry Stewart, National Research Council, Canada
Swathikiran Sudhakaran, Samsung AI Center Cambridge, UK
Sherin Sugathan, Envieu Research & Development Labs, India
Junichi Takeno, Meiji University, Japan
Knud Thomsen, Paul Scherrer Institute, Switzerland
Jan Treur, Vrije Universiteit Amsterdam, Netherlands
Vadim L. Ushakov, Kurchatov Institute National Research Centre, Russia
Alexsander V. Vartanov, Lomonosov Moscow State University, Russia
Rodrigo Ventura, Universidade de Lisboa, Portugal
Evgenii Vityaev, Sobolev Institute of Mathematics SB RAS, Russia
Pei Wang, Temple University, USA
Mark Waser, Digital Wisdom Institute, USA
Roseli S. Wedemann, Universidade do Estado do Rio de Janeiro, Brazil
Viacheslav Wolfengagen, National Research Nuclear University MEPhI, Russia
Peicheng Xiong, Ningbo University, China
Özge Nilay Yalçın, Simon Fraser University, Canada
Yumeng Zhao, Ningbo University, China
Terry Zimmerman, University of Washington Bothell, USA

Contents

Invited Abstracts

Abstracts of Invited Talks Given at BICA*AI 2023	3
<i>Tingting Liu, Alexei V. Samsonovich, Peter Boltuc, Scott E. Fahlman, Sophie Hendrikse, Jan Treur, John Laird, Antonio Lieto, Paul Robertson, Ron Sun, and Junichi Takeno</i>	
Selected Presentation Abstracts from BICA*AI 2023	14
<i>Maria Yu. Boboshko, Ekaterina S. Garbaruk, Veronika M. Knyazeva, Marina J. Vasilyeva, Aleksander A. Aleksandrov, Anton Kolonin, and Alexei V. Samsonovich</i>	

Selected Papers

Dot Density Effects on Stereoscopic Transparency: A Cross-Correlation Model Analysis	21
<i>Saori Aida</i>	
An Examination of Visual Influences on Sense of Ownership and Agency	27
<i>Saori Aida and Yuui Ogawa</i>	
Super-Performance: Sampling, Planning, and Ecological Information	33
<i>Bradly Alicea</i>	
ADHD-Related Features of Eye Fixations While Simulated Driving with and Without Working Memory Load: A Pilot Study	41
<i>Julia Altshuler and Valeriia Demareva</i>	
A Biologically Inspired Approach to Protecting and Verifying the Authenticity of Important Documents	50
<i>Alexander M. Alyushin, Victor M. Alyushin, Sergey V. Dvoryankin, and Lyubov V. Kolobashkina</i>	
Comparison of Biologically Inspired and Modeling Approaches for Predicting Possible Condition Change in Critical Occupational Workers	56
<i>Mikhail V. Alyushin, Lyubov V. Kolobashkina, Vladislav D. Bitney, and Andrey V. Okhlopkov</i>	

Integrated Discounted Future Prediction as Auxiliary Task for A3C	62
<i>Andrey Andronenko, Mikhail Avshalumov, and Vyacheslav Demin</i>	
Automated Bias and Indoctrination at Scale... Is All You Need	70
<i>Kyrtin Atreides</i>	
The COPPER Babysitter Robot, a Childcare Monitoring System from the First year of Age	83
<i>Ruth A. Bastidas Alva, Angie L. Herrera Poma, Valeryia E. Perez Villa, and Frank W. Zarate Peña</i>	
Semantic Social Web Applications: Wiki Web	96
<i>Aleksandr Belozarov and Valentin Klimov</i>	
Development of a Network Traffic Anomaly Detection System Based on Neural Networks	104
<i>Natalia Beshpalova, Alexey Ershov, Sergey Sitnikov, Sergey Nechaev, Margarita Vanina, Victor Radygin, Dmitry Kupriyanov, and Mikhail Ivanov</i>	
Architecture of an Expert System to Support Diagnostic Decisions for Hereditary Diseases	113
<i>Nikolay A. Blagosklonov and Boris A. Kobrinskii</i>	
System Analysis of Educational Digital Ecosystems in the Agro-Industrial Complex of Russia	121
<i>Vladimir Budzko and Victor Medennikov</i>	
System Analysis of Subject Identification of Digital Twin in Agriculture	135
<i>Vladimir Budzko, Victor Medennikov, and Petr Keyer</i>	
Topological Analysis of Protein Surfaces and Its Role in the Development of New Medicines	143
<i>Oleg V. Bystrov and Sergey D. Kulik</i>	
Development of a Multi-agent Architecture for an Object Shape Recognition System Based on Data from a Depth Sensor	151
<i>Kantemir Bzhikhatlov, Murat Anchokov, and Olga Nagoeva</i>	
A Statistical WavLM Embedding Features with Auto-Encoder for Speech Emotion Recognition	159
<i>Adil Chakhtouna, Sara Sekkate, and Abdellah Adib</i>	

Axonal Myelination as a Mechanism for Unsupervised Learning in Spiking Neural Networks	169
<i>Nadezhda Chaplinskaia and Nikolay Bazenkov</i>	
Feature Synthesis for Few-Shot Object Detection	177
<i>Chenchen Tao, Song Chen, Yi Chen, Xiaojie Cai, and Chong Wang</i>	
Simulation of Fabric Wetting Based on Particle Sampling	188
<i>Jiajun Cheng, Zhen Liu, Tingting Liu, and Yanjie Chai</i>	
A Parallel Ice Melting Simulation Based on Particle	197
<i>Jiajun Cheng, Zhen Liu, Tingting Liu, and Yanjie Chai</i>	
Developing a Voice Control System for a Wheeled Robot	208
<i>Evgenii Chepin, Alexander Gridnev, and Margarita Erlou</i>	
«Personality» Profile of Generative Neural Network ChatGPT	216
<i>Yuliya A. Chudina, Andrey A. Nikolaev, Dmitry B. Chaivanov, and Irina G. Malanchuk</i>	
A Two-Stream Self-attention Multi-digraph Model for Chinese NER	231
<i>Xuyao Dai, Tingting Liu, Zhen Liu, and Yanjie Chai</i>	
How Language Could Have Evolved	242
<i>Ken Del Signore</i>	
Critical Slowing Down in Heart Rate Variability for Human Condition Control: An Example of Sleep Onset Detection	277
<i>Valeriia Demareva, Irina Zayceva, Andrey Demarev, and Nicolay Nazarov</i>	
BICA's Fears and Troubles: GPT-Based AI Tools Are Its Friends or Foes?	285
<i>Emanuel Diamant</i>	
Gossiping Until You Get Tired of It: A Network Model of the Adaptive Exchange of Rumors in a Small Scale Social Environment	294
<i>Karley Dionne, Maya Vermeer, and Jan Treur</i>	
A Socially Acceptable Conversational Agent Based on Cognitive Modeling and Machine Learning	312
<i>Anatoly A. Dolgikh and Alexei V. Samsonovich</i>	
Emotion from P Bit: Computing Emotions Using the Platonic Computer	323
<i>Simon X.Duan</i>	

Learning Hidden Markov Model of Stochastic Environment with Bio-inspired Probabilistic Temporal Memory	330
<i>Evgenii Dzhivelikian, Petr Kuderov, and Aleksandr I. Panov</i>	
The Research on Key Technologies for Intelligent Education Based on LLM	340
<i>Jiaqi Fu, Tiejun Pan, Leina Zheng, and Zichu Xue</i>	
Mapping Action Units to Valence and Arousal Space Using Machine Learning	348
<i>Ismail M. Gadzhiev, Alexander S. Makarov, Daria V. Tikhomirova, Sergei A. Dolenko, and Alexei V. Samsonovich</i>	
Principles of Creating Hybrid Intelligent Information Systems Based on the Granular-Metagraph Approach	356
<i>Yuriy E. Gapanyuk, Valery I. Terekhov, Vitaly Y. Ivlev, Yuriy T. Kaganov, Irina S. Karabulatova, Mikhail B. Oseledchik, and Dmitry V. Semenov</i>	
Ontograph Cognitive Information Retrieval: Some Experimental Evaluations	367
<i>Anastasia Gavrilkina, Olga Golitsina, and Nikolay Maksimov</i>	
Emotion-Integrated Cognitive Architectures: A Bio-Inspired Approach to Developing Emotionally Intelligent AI Agents	373
<i>Aliya Grig and Anastasia Rizzo</i>	
Experimental Phonetic Research Interlingual Interference and Accent in the Russian Speech of Native Speakers of the Kabardino-Circassian Language	382
<i>Irina Gurtueva, Murat Anchekov, Kantemir Bzhikhatlov, Olga Nagoeva, and Ahmed Enes</i>	
A Bearing Fault Diagnosis Method Based on VMD-HPE	390
<i>Wanqing Huang, Yang Chen, Yongqi Chen, Tao Zhang, Feiyu Yu, and Xiaoyan Mao</i>	
Neural Network Solution of an Inverse Problem with Integration of Geophysical Methods on Recovered Data: Training with Noise Addition	406
<i>Igor Isaev, Ivan Osbornev, Eugeny Osbornev, Eugeny Rodionov, Mikhail Shimelevich, and Sergey Dolenko</i>	

Mathematical Model, Experimental Verification and Control of Actuators Based on Metal – Hydrogen System	414
<i>Vladimir I. Ivlev and Sergej Yu. Misyurin</i>	
Associative Memory with Biologically-Inspired Cell Assemblies	422
<i>Yuehu Ji, David Gamez, and Chris Huyck</i>	
Combined Contrast Enhancement Algorithm for High Dynamic Range Images	429
<i>M. A. Kazakov</i>	
A Hybrid PSO-Jaya Algorithm for Optimization Problems	436
<i>E. M. Kazakova</i>	
Registrar: A Social Conversational Agent Based on Cognitive and Statistical Models for a Limited Paradigm	444
<i>Dmitry Khabarov and Alexei V. Samsonovich</i>	
Dynamic Model of Semantic Information Signal Processing	453
<i>Mohiniso Khidirova, Kamaliddin Abdivakhidov, Pavel Bylevsky, Alexey Osipov, Ekaterina Pleshakova, Victor Radygin, Dmitry Kupriyanov, and Mikhail Ivanov</i>	
Identification of Ambient and Focal Information Processing Phases Using Eye Movement Response Registration	462
<i>A. N Korosteleva, S. I. Kartashov, and A. A. Kotov</i>	
The Analysis of the DMN Network of the Brain Using the Method of Segmentation of Functionally Homogeneous Regions	469
<i>Stanislav Kozlov, Alexey Poyda, Vyacheslav Orlov, and Vadim Ushakov</i>	
New Feature for Schizophrenia Classification Based on Functionally Homogeneous Brain Regions	477
<i>Stanislav Kozlov, Artur Zhemchuzhnikov, Alexey Poyda, Vyacheslav Orlov, and Sergey Kartashov</i>	
Preliminary Study of Cerebral Myelin Content Alterations at Schizophrenia	485
<i>Ekaterina Krupina, Andrei Manzhurtsev, Maxim Ublinskiy, Larisa Mosina, Maria Osetrova, Vasily Yarnykh, Galina Mamedova, Sergey Trushchelev, Natalia Zakharova, Georgy Kostyuk, and Vadim Ushakov</i>	
Robotic Customer Service System ALKETON	495
<i>Anton V. Kudriashov</i>	

Exploring the Efficiency of Neural Networks for Solving Dynamic Process Problems: The Fisher Equation Investigation	504
<i>Raul Karachurin, Stanislav Ladygin, Pavel Ryabov, Kirill Shilnikov, and Nikolay Kudryashov</i>	
Improving the Methodology for Integrated Testing of Journal Entries by Benford's Law	512
<i>Pavel Y. Leonov, Viktor M. Sushkov, Sofia A. Boiko, and Margarita A. Stepanenkova</i>	
A Bayesian Network-Based Model for Fraud Risk Assessment	520
<i>Pavel Y. Leonov, Viktor M. Sushkov, Stanislav V. Vishnevsky, and Valentin A. Romanovsky</i>	
The Influence of Articulatory Interference on Inner Pronouncing of Words	528
<i>Daria Leonovich and Alexander Vartanov</i>	
Method of Logical Interpretation of Neural Network Solutions	536
<i>L. A. Lyutikova</i>	
Implementation of Embodied Cognition in Multi-agent Neurocognitive Architecture	545
<i>Dana Makoeva, Olga Nagoeva, Murat Anchokov, and Irina Gurtueva</i>	
Natural and Artificial Intelligence: An Activity-Based Approach	553
<i>Nikolay Maksimov and Valentin Klimov</i>	
Models of Generation of Statements of Various Genre Types According to Data of Early Speech Ontogenesis: Imperative Versus Informative Genres	566
<i>Irina G. Malanchuk and Anastasia N. Korosteleva</i>	
Polymorphisms of IL10 Immunoregulatory Gene Impact the Morphometric Changes of the Brain in Schizophrenia	577
<i>Irina K. Malashenkova, Vadim L. Ushakov, Sergey A. Krynskiy, Daniil P. Ogurtsov, Ekaterina I. Chekulaeva, Ekaterina A. Filippova, Vyacheslav A. Orlov, Natalia V. Zakharova, Denis S. Andreyuk, Sergey A. Trushchelev, Georgy P. Kostyuk, and Nikolay A. Didkovsky</i>	
Enhancing Event Selection with ChatGPT-Powered Chatbot Assistant: An Innovative Approach to Input Data Preparation	588
<i>Andrey Malynov and Igor Prokhorov</i>	

Functional Connectivity of Brain Regions Resulting from Learning
 Unfamiliar Words: Word Frequency Effect 595
*K. S. Memetova, V. M. Knyazeva, L. N. Stankevich, I. G. Malanchuk,
 and A. A. Aleksandrov*

DIPy-AI: Brain-Cognition-Inspired DIKW Pyramid-Based Agile AI
 Architecture for Industrial Sensor Data Assimilation 604
Amit Kumar Mishra and Yi Zhong

Assessment and Correlation of Morphometric and Tractographic
 Measures of Patients Diagnosed with Schizophrenia 612
*Larisa Mosina, Vadim Ushakov, Vyacheslav Orlov,
 Sergey Kartashov, Natalia Zakharova, Georgy Kostyuk,
 and Sergey Trushchelev*

Deep Learning Evolution: Using Genetic Algorithm to Modify Training
 Datasets 627
Mikhail Yu. Nazarko, Klim A. Fedorov, and Alexei V. Samsonovich

Strategies for Business Cybersecurity Using AI Technologies 635
Svetlana Nosova, Anna Norkina, and Nikolay Morozov

Integration of Artificial Intelligence into Business Management Strategy 643
*Svetlana Nosova, Anna Norkina, Nikolay Morozov, Irina Arakelova,
 and Galina Fadeicheva*

The Applicability of Artificial Intelligence in the Modern Global
 Development of Countries and Companies 651
*Svetlana Nosova, Anna Norkina, Nikolay Morozov, Olga Medvedeva,
 Irina Arakelova, and Sergey Bondarev*

Temporal Stability of Resting State fMRI Data Analysis by Independent
 Components Method 659
*V. A. Orlov, S. I. Kartashov, M. V. Kalmykova, A. A. Poyda,
 and Vadim L. Ushakov*

Analysis of Resting-State fMRI Data by CAPA Method 666
*Vyacheslav A. Orlov, Sergey I. Kartashov, Alexey A. Poyda,
 and Vadim L. Ushakov*

Application and Modeling of LLM in Quantitative Trading Using Deep
 Learning Strategies 671
Tiejun Pan, Jinjie Yu, Leina Zheng, and Yuejiao Li

A Study of Conversational Intentionalities Expressed in Natural Language Using ChatGPT	679
<i>Ivan A. Pavlenko, Arthur D. Zakirov, Andrei N. Yakovlev, and Alexei V. Samsonovich</i>	
Symbiotic Artificial and Human Cognitive Architectures Managing Human Attention	688
<i>Thomas Pederson and Amit Kumar Mishra</i>	
The Impact of Internet Media on the Cognitive Attitudes of Individuals on the Example of RT and BBC	695
<i>Alexandr Y. Petukhov, Sofia A. Polevaya, Dmitry I. Kaminchenko, and Evgeniy A. Gorbov</i>	
Simulation Model of the Neurocognitive System Controlling an Intellectual Agent Displaying Exploratory Behavior in the Real World	706
<i>Inna Pshenokova, Kantemir Bzhikhatlov, Sultan Kankulov, Artur Apshev, and Boris Atalikov</i>	
The Embodied Intelligent Elephant in the Room	716
<i>Saty Raghavachary</i>	
Approaches to Modeling Autonomous Agents with Scientific Abilities	723
<i>Vladimir G. Red'ko</i>	
Are Associations All You Need to Solve the Dimension Change Card Sort and N-bit Parity Task	730
<i>Damiem Rolon-Mérette, Thaddé Rolon-Mérette, and Sylvain Chartier</i>	
Image and Audio Data Classification Using Bagging Ensembles of Spiking Neural Networks with Memristive Plasticity	741
<i>Roman Rybka, Yury Davydov, Alexander Sboev, Danila Vlasov, and Alexey Serenko</i>	
An Episode Tracker for Cognitive Architectures	750
<i>Eduardo Yuji Sakabe, Anderson Anjos da Silva, Luiz Fernando Coletta, Alexandre da Silva Simões, Esther Luna Colombini, Paula Dornhofer Paro Costa, and Ricardo Ribeiro Gudwin</i>	
Application of Machine Learning to Construct Solitons of Generalized Nonlinear Schrödinger Equation	759
<i>A. G. Sboev, N. A. Kudryashov, I. A. Moloshnikov, D. R. Nifontov, S. V. Zavertyaev, and R. B. Rybka</i>	

Spoken Digits Classification Using a Spiking Neural Network
with Fixed Synaptic Weights 767
*Alexander Sboev, Maksim Balykov, Dmitry Kunitsyn,
and Alexey Serenko*

A Brain-Inspired Cognitive Architecture (BICA) Approach
to the Neurosymbolic Gap 775
Howard Schneider

FECG: A Flexible Holter for Ambulatory Heart Rate Monitoring 787
Yuduo Shan, Tingting Liu, and Zhen Liu

Features of Internal Pronunciation of Words by a Group of People
with Rhotacism in Comparison with Normative Pronunciation 800
Olga Shevaldova and Alexander Vartanov

One Robust Variant of the Principal Components Analysis 807
Z. M. Shibzukhov

Using Electronic Nose in Forensic Odor Analysis 815
Alexander Shtanko and Sergey Kulik

Hierarchical AGI from First Principles 823
Sergey Shumsky

The Future of International Climate Politics: An Agent-Based Approach 832
*Anna Shuranova, Matvei Chistikov, Yuri Petrunin, Vadim Ushakov,
and Denis Andreyuk*

Memory Based Reinforcement Learning with Transformers for Long
Horizon Timescales 845
Shweta Singh, Sudaman Katti, and Vedant Ghatnekar

Investor Bot for Business Process 853
Sergey Kulik and Ivan Sofronov

Testing for Benford’s Law as a Response to the Risks of Material
Misstatement Due to Fraud 860
Viktor M. Sushkov and Pavel Y. Leonov

Greening Telecom: Harnessing the Power of Artificial Intelligence
for Sustainable Communications 867
Anastasiia Suslina, Konstantin Savin, and Irina Suslina

Neuropunk Revolution: Further Results	875
<i>Max Talanov</i>	
Data Preparation for Advanced Data Analysis on Elastic Stack	884
<i>M. S. Ulizko, R. R. Tukumbetova, A. A. Artamonov, E. V. Antonov, and K. V. Ionkina</i>	
Brain Neural Network Architectures in Sleep-Wake Cycle	894
<i>Vadim L. Ushakov, Maria L. Khazova, Polina E. Zhigulina, Vyacheslav A. Orlov, Denis G. Malakhov, and Vladimir B. Dorokhov</i>	
Speech Recognition from MEG Data Using Covariance Filters	904
<i>Vitaly Verkhlyutov, Victor Vvedensky, Konstantin Gurtovoy, Evgenii Burlakov, and Olga Martynova</i>	
Prediction of the Correct Firing Position with a Pistol Based on a MANFIS Model	912
<i>David Alberto Vique Almeida, Luis Armando Chicaiza Conteron, José Luis Carrillo Medina, and Edison Gonzalo Espinosa Gallardo</i>	
Designing a Neural Network Cascade for Object Detection in Drawing and Graphical Documentation Processing	924
<i>Kirill Vitko and Anna Tikhomirova</i>	
Crowdsourcing-Based Approbation of Communicative Behaviour Elements on the F-2 Robot: Perception Peculiarities According to Respondents	932
<i>Liliya Volkova, Artemy Kotov, and Andrey Ignatev</i>	
A Wavelet-Based Method for Morphing Audio Recordings of Interjections from One Voice to Another	946
<i>Liliya Volkova, Arina Untilova, and Maksim Kozlov</i>	
Extended and Distant Cortical Areas Coordinate Their Oscillations Approaching the Instant of Decision Making During Recognition of Words	956
<i>Victor Vvedensky, Vitaly Verkhlyutov, and Konstantin Gurtovoy</i>	
Laser Detection of Surface Quality of Electrical Contacts Based on Ensemble Learning	962
<i>Chao Wang and Cheng Jun Guo</i>	
When and Where Conceptual Maths Equals to Conceptual Modeling: Reasons for Using in Cognitive Modeling	973
<i>Viacheslav Wolfengagen, Larisa Ismailova, and Sergey Kosikov</i>	

A Prediction Model for the Equivalent Parameters of an Acoustic Transducer Based on DPSD and LSTM Neural Network 980
Yuhui Xue, Zhidi Jiang, and Mudan Yu

Unraveling the Elements of Effective Altruistic Appeals Through Machine Learning and Natural Language Processing 995
Sourav Yadav, Sankalp Arora, Akash Kumar, and Kaveri Verma

A Novel Feature Selection Method Based on Slime Mold Network Formation Behavior 1007
Chenyang Yan

The Role of Social Stress in the Development of Mental Disorders 1016
Shuya Yang








VLSI Floorplanning Algorithm Based on Reinforcement Learning with Obstacles 1034
Shenglu Yu and Shimin Du

Author Index 1045

Invited Abstracts



Abstracts of Invited Talks Given at BICA*AI 2023

Tingting Liu¹ , Alexei V. Samsonovich² , Peter Boltuc^{3,4} , Scott E. Fahlman⁵,
Sophie Hendrikse⁶ , Jan Treur⁷ , John Laird⁸ , Antonio Lieto⁹ ,
Paul Robertson¹⁰, Ron Sun¹¹, and Junichi Takeno^{12,13}

¹ College of Science and Technology, Ningbo University, Cixi 315300, China
liutingting@nbu.edu.cn

² George Mason University, Fairfax, VA 22030, USA
alexei.samsonovich@gmail.com

³ University of Illinois, Springfield, IL, USA
pbolt1@uis.edu, pboltu@sgh.waw.pl

⁴ Warsaw School of Economics, Warszawa, Poland

⁵ Carnegie Mellon University, Pittsburgh, PA, USA
sef@cs.cmu.edu

⁶ Psychological Methods Group, Psychology Research Institute, University of Amsterdam,
Amsterdam, The Netherlands
sophiehendrikse@hotmail.com

⁷ Social AI Group, Department of Computer Science, Vrije Universiteit Amsterdam,
Amsterdam, The Netherlands
treur@cs.vu.nl

⁸ Center for Integrated Cognition, Ann Arbor, MI, USA
laird@umich.edu

⁹ University of Salerno, Fisciano, Italy
lieto.antonio@gmail.com

¹⁰ Dynamic Object Language Labs Inc., DOLL, Lexington, MA, USA
paulr@dollabs.com

¹¹ Rensselaer Polytechnic Institute, Troy, NY, USA
dr.ron.sun@gmail.com

¹² Meiji University, 1-1, Kanda-Surugadai, Chiyoda-Ku, Tokyo 101-8301, Japan
juntakeno@gmail.com

¹³ Heuristics Science Research Institute, 15-4 Komukai-Cho, Saiwai-Ku,
Kawasaki-Shi 212-0003, Kanagawa-Ken, Japan

Abstract. This chapter comprises selected short and extended abstracts of invited talks and discussion panels that took place at the 2023 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence, also known as the 14th Annual Meeting of the BICA Society (BICA*AI 2023), held in Ningbo, China during October 13–15, 2023. Abstracts included here were not accompanied by papers in this volume. The abstracts are arranged alphabetically by the first author's last name, as follows: (1) Boltuc, (2) Fahlman, (3) Hendrikse and Treur, (4) Laird, (5) Lieto, (6) Liu, (7) Robertson, (8) Samsonovich, (9) Sun, (10) Takeno. Section authors are listed again beneath the header of each section.

Keywords: Cognitive architecture · Generative model · Large language model · Knowledge-based AI · Symbolic KRR · Neural-symbolic models · Dual-process theories · Machine ethics · Self-aware system · Superhuman AI · Artificial social intelligence

1 BICA for Consciousness

Peter Boltuc, University of Illinois, Springfield, USA, and Warsaw School of Economics, Poland

BICA goes beyond recommender engines or standard big data computing since brains do not work as simple big data processors. We are mid-data processors, using brains not quite as discrete computing engines. We compute using biochemical features [1], topological learning [2], forgetting [3] or paraconsistent interzones [4]. Recent collapse of IBM-Watson exemplifies not only IBM's proven record of mishandling novel technologies: misery brought about by technologically underinformed AI management, but, more interestingly, the need to distinguish big-data from mid-data computing. The true BICA clearly belongs to the latter. Thus, BICA implementation requires reaching beyond the digital technologies dominant at the 4th Industrial Revolution, and moving towards deep stochastic computing [5] that reaches beyond optimizing big-data usage, towards randomization in creating deeply new outcomes such as Discovery Engines. The panel includes presenters from North America, as well as those from China, and a set of papers is expected to ensue.

Keywords: Mid-data computing, True BICA, Biochemical computing, Topological learning, The 4th industrial revolution, Paraconsistent interzones, Stochastic computing, Discovery engines.

2 Deep Learning AI Versus Symbolic Knowledge-Based AI: We're Going to Need Both

Scott Fahlman, Carnegie Mellon University, USA

In this talk, I will argue that, if our goal is to build AI systems with human-like levels of generality, reliability, precision, and common sense in their reasoning, as well as the ability to function in messy real-world domains, we will need BOTH a symbolic, knowledge-based representation and reasoning component and a statistical, data-driven, learning-based component, with each subsystem doing the tasks that it does best. The statistical learning-based components are well-suited for low-level sensory/motor tasks and for producing simple, well-rehearsed, reflex-like responses. The symbolic, knowledge-based components are needed for more complex and reliable thought, reasoning, and planning—in other words, for the conscious parts of human intelligence. These two subsystems are distinct, but they must work together in any system that aspires to emulate human-like, real world intelligence.

Many people in the AI community now view the older symbolic approaches as a lost cause. The conventional wisdom is that symbolic AI was tried back in the 1970's and 1980's, but those old systems could not scale up to handle anything like the amount of knowledge required for human-like commonsense reasoning. The reasoning was too slow, and it was too hard to add new knowledge to them. I will suggest some ways of overcoming these problems, in a way that was partially inspired by neuroscience and human psychology, using my own Scone knowledge-base system as an example of how a modern, scalable, symbolic knowledge-based AI system might work.

Keywords: Symbolic AI, Knowledge-based AI, Symbolic KRR, Scone knowledge-base system, NETL.

3 New Analysis and Modeling Directions for Adaptive Multimodal Social Interaction

Sophie Hendrikse, University of Amsterdam, Psychology Research Institute, Psychological Methods Group, the Netherlands and **Jan Treur**, Vrije Universiteit Amsterdam, Department of Computer Science, Social AI Group

Although there is much literature on multimodal interaction, multimodal synchrony analysis, and related behavioral adaptivity, mathematical formalization and computational simulation of it is a nontrivial topic. Moreover, the subjective, agent-oriented perspective on synchrony analysis has not yet received much attention in the literature. This presentation provides from an agent-oriented perspective an overview of recent work on mathematical formalization and computational simulation of multimodal interaction, subjective multimodal synchrony analysis, and related adaptivity of the interaction behavior [6–10]. It does so by exploiting the possibilities of multi-adaptive self-modeling network models for agents to analyze these dynamic and adaptive processes formally. Specific topics addressed include subjective detection of interpersonal synchrony, short-term behavioral adaptivity (e.g., affiliation) versus long-term behavioral adaptivity (e.g., bonding) for interaction, subjective detection of synchrony transitions, the role of time lags for subjective synchrony detection, and relationship-specific versus relationship-independent behavioral adaptivity (e.g., transference).

Keywords: Social interaction, Multimodal, Adaptive interaction behavior, Affiliation, Bonding.

4 MRIntegrating Cognitive Architectures and Generative Models

John Laird, Center for Integrated Cognition, USA

In this talk, I explore three possible variants of how generative models can integrate with cognitive architectures, potentially overcoming their corresponding weaknesses. In the first variant, pre-trained generative models are one or more modules in a traditional cognitive architecture, acting as a fixed, read-only long-term memory or as a perceptual

or motor module. Here, the cognitive architecture retains its native knowledge representations (such as symbolic graph structures and rules) and must translate them into natural language to access the generative models. The second variant is where a cognitive architecture is built from scratch, and all modules represent and process language structures (short sentences). The third approach is where a traditional cognitive architecture is extended to include one or more modules built with graphical transformers that build up the knowledge in the module from the experience of the cognitive architecture agent, learning to predict not the next word, but the next “thought” of the agent. A critical part of the talk will be analyzing the strengths and weaknesses of these different approaches for building general intelligent systems.

Keywords: Cognitive architecture, Generative model, Large language model, Transformer.

5 Avoiding the Behavioristic Trap with the Minimal Cognitive Grid

Antonio Lieto, University of Salerno, Italy

The enormous success of modern AI systems (e.g., in computer vision, natural language processing etc.) has led to the formulation of the hypothesis that such systems—since are able to obtain human or superhuman level performances in a number of tasks—actually have acquired the underlying competence that we humans possess in order to exhibit the same kind of behavior.

This hypothesis, I argue, is however based exclusively on a behavioristic analysis of (some of) the output produced by them. And, as such, it is methodologically problematic. In this talk I will show how by using a tool known as Minimal Cognitive Grid (MCD, introduced in Lieto [11]) it is possible to avoid this behavioristic trap and, in addition, to compare and rank, in a non-subjective way, different types of artificial systems based on their biological or cognitive plausibility.

Keywords: Minimal cognitive grid, Large language models, Cognitive design for artificial minds, Superhuman AI.

6 Information Technologies and Ethics, Supporting Embodied Agents and Human–Computer Interactions

Tingting Liu, College of Science and Technology, Ningbo University, China

Panel will discuss information technologies and ethics which can support embodied agents and human–computer interactions. The information technologies could include Virtual Reality, Augmented Reality and other graphic techniques that can generate environment, performance or behavior of embodied agents and can support human–computer interactions. The ethics issues may cover all ethics issues for the application of Artificial Intelligence.

Keywords: Virtual reality, Ethics, Augmented reality.

7 Artificial Social Intelligence, Artificial Independent Intelligence and the Future of AI

Paul Robertson, Dynamic Object Language Labs Inc., DOLL, Lexington, MA, United States of America

The future of AI is in interacting fluidly with humans. Huge advances, especially over the last 12 years have put in place the building blocks for AI that can understand the world the way that we do, see the world the way we do, and can communicate about our world the way we do.

Using examples from our own work as well as that of other labs, I will describe what these advances are, why they are important, and why more remains to be done.

Arguably, none of the systems that exist today are intelligent in any reasonable sense. I will discuss what advances are required to achieve true intelligence, why we should do it, and how it will benefit us.

We share the world with other intelligent animals, but interacting with them is not fluid, our differences are too great as are our competencies. Fluid interaction requires a similarity of capabilities and representations. Future AI systems will resemble us more and more in order to achieve fluidity in interactions.

Present systems that have been trained on human annotated data and on human generated content can describe our world in a way that gives the impression that we see the world the same way, but today, that is largely an illusion, but it doesn't have to stay that way. Current systems try to answer the question, 'how would a human describe this situation' and not 'how would I, the AI system, describe the situation'. I will enumerate a plausible path forward to achieve AI systems that will improve the lives of the human population by being more like us, more able to interact fluidly, and more able to act on their own volition.

Keywords: Artificial social intelligence, Artificial independent intelligence, Artificial emotions, Artificial consciousness, Machine learning.

8 BICA Society Panel

Alexei V. Samsonovich, George Mason University, Fairfax, VA, USA

Participants of the panel in this year are Tingting Liu, Alexei Samsonovich, Jan Treur, Kyrtin Atreides, David Kelley, Felix Ramos, Paul Robertson, and Robert Laddaga. BICA Society Panel is organized every year in the middle of the BICA*AI conference. It is at the same time the business meeting of BICA Society (therefore, requires participation of BICA Society Directors) and a yearly report of the BICA Society Board of Directors (BOD) to the Membership (all registered participants automatically become Members of BICA Society). Once in every three years, elections of BOD are held at the BICA Society Panel. The last elections were held in 2022. In addition, the panel discusses past progress and future plans of BICA Society: in particular, plans for the next BICA*AI conference. Usually, the panel takes from 15 to 30 min.

In theory, life of a man or the life of an enterprise can be foreseen as a systematic movement forward, culminating in the achievement of the goals. But real life is always very far from this ideal. Frequently it is a balance between hopelessness and disaster. At its best, real life is a desperate burning in which one dominant emotion expresses both its meaning and content. This is the gist of the story I am about to tell.

BICA Society, or Biologically Inspired Cognitive Architectures Society, was established in 2010 as an international scientific society and a 501(c)(3) nonprofit corporation based in the US. The reason was the need to coordinate a significant community of researchers [12], left on their own after the abrupt termination of the DARPA program BICA in 2006. The key initiator Alexei Samsonovich, at that time a Research Assistant Professor, later Assistant Professor at the Krasnow Institute for Advanced Study of George Mason University, was elected the president of BICA Society and started working in close cooperation with the two other BICA Society Directors: Professors Antonio Chella from University of Palermo and Kamilla R. Johannsdottir from Reykjavik University. The mission of BICA Society became the integration of many efforts of researchers around the world in addressing the BICA Challenge [13]: the challenge of creating a computational equivalent of the human mind. This was understood as the replication of highest human cognitive functionalities, including the humanlike Self, voluntary behavior, personality, system of values, episodic memory, humanlike Theory of Mind, metacognition, goal reasoning, social emotions and emotional intelligence, imagery, creativity, active self-regulated learning, and more, in addition to the more basic cognitive functions such as perception, attention, sensemaking, memory and learning in various forms, communication, motivation, planning, action control, and so on. This mission was supposed to be accomplished by promoting and facilitating the transdisciplinary study of cognitive architectures, and in the long-term perspective—creating one unifying widespread framework for the human-level cognitive architectures and their implementations [14]. Main vehicles used for this purpose were the BICA Conference Series together with their video archives on various platforms and the related publication channels, including mainstream Elsevier journals and book series published by IOS Press, Elsevier and Springer [15–18].

Years of the BICA conference were always pitched by its organizers as a growing success, a triumphant parade around the globe: Washington DC, USA in 2010 and 2011, Palermo, Italy in 2012, Kyiv, Ukraine in 2013, Boston, MA, USA in 2014, Lyon, France in 2015, New York, NY, USA in 2016, Moscow, Russia in 2017, Prague, Czech Republic in 2018, Seattle, WA, USA in 2019, Natal, Brazil in 2020 (virtual), Fukuchiyama, Japan and Vienna, Austria in 2021 (virtual), Guadalajara, Mexico in 2022 (hybrid), and Ningbo, China in 2023. Numbers of submissions and attendance by years are shown in Fig. 1. The expected attendance for 2023 is beyond the plot limit.

At the same time, making it happen every year was not just a big challenge: it was a true miracle. Every year I had a feeling that this is the last BICA conference, and it will not be possible to continue the series after it, when somebody unexpectedly offered to host it next year, and the story continued. However, the conference did not receive financial or another form of organizational support and did not become a mainstream like AAAI or CogSci. In the struggle to keep it running, submissions from remotely related domains were allowed. Locals, confident in acceptance of anything, started to dominate,

helping us in maintaining the numbers. As a consequence, the BICA conference and its community started degrading and fell below its rivals.

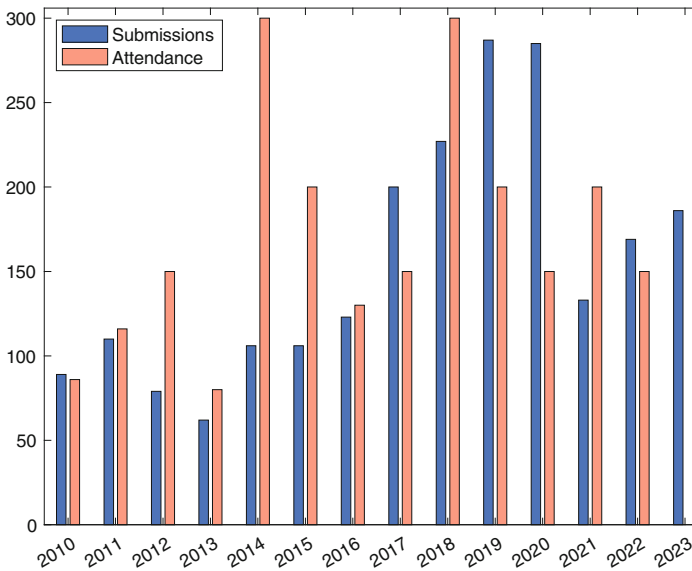


Fig. 1 Statistics of the BICA conference series by years: numbers of submissions (blue) and estimated attendance (orange). Satellite events are not included

By coincidence, at that same time the buzzword “biologically inspired” became omni-meaningful and dissociated from the original ideas of BICA. Concurrent parties claimed their leadership in the field. Therefore, in 2019 in Seattle after a debate it was decided that the name of the conference should change from “Biologically Inspired Cognitive Architectures” (BICA) to “Brain-Inspired Cognitive Architectures for Artificial Intelligence” (BICA*AI). Our top publication channel, the journal BICA merged Cognitive Systems Research as a BICA*AI Section. The name of the company was left unchanged, while the company itself was rebranded and re-incorporated. The original team of directors left and subsequently fell apart, both as a team and as friends. Original spirit of the company was lost, I do not even see a mentioning of the BICA Challenge at the new company web site, which is hardly up to date.

A new life of BICA conferences started kicking in with the involvement of new hosting parties: Cinvestav GDL in 2022, and then University of Ningbo in 2023, giving everybody a new hope. The BOD (David Kelley, Kyrtin Atreides, Felix Ramos) was complemented with an Advisory Board, including Jan Treur, Tingting Liu, and Alexei Samsonovich, the latter continuing to be the “puppeteer”, the working horse, and the Vizard of OZ of BICA at the same time. It occurs to me that if I stop doing anything for BICA, it will cease to exist immediately without a trace. I hope that I am mistaken.

Today the future of the conference series looks as uncertain as it always has been. Opinions of some of its solid supporters are represented by the summaries included in this chapter. From these examples one can see that the spirit is still strong, and there

is a sufficient enthusiasm. Also, the historical context has never been more exciting and favorable than today. I strongly feel that we should reunite our efforts in pursuing solutions to the challenge that now becomes vital for the society. *It must be done!*

Keywords: BICA society, BICA challenge, BICA*AI conference series.

9 Relevance of Cognitive Architectures to Neural-Symbolic Models and Dual-Process Theories

Ron Sun, Rensselaer Polytechnic Institute, USA

In this talk, I will address neural-symbolic (or “neurosymbolic”) models, dual-process theories, and cognitive architectures, in terms of their relevance to each other. I will provide some historical backgrounds and argue that dual-process theories have significant implications for developing neural-symbolic models. Computational cognitive architectures can help disentangle complex issues concerning dual-process theories and are thus important to neural-symbolic models.

The notion of neural-symbolic models harkens back to the 1990s when such models *first* emerged (see, e.g., Sun & Bookman, 1994 [19]). There have been many different ways of structuring such models, and a key question remains: How should we best structure them? I argue that they should be structured in a cognitively motivated/justified way, based on the human cognitive architecture. In particular, they should take into account dual-process theories concerning the human cognitive architecture.

The distinction between “intuitive” and “reflective” thinking (i.e., system 1 and system 2) has been, arguably, one of the most important distinctions in cognitive science. There are currently many such dual-process theories out there. One such theory was proposed early on in Sun [20], where the two systems were characterized as follows: “... cognitive processes are carried out in two distinct ‘levels’ with qualitatively different mechanisms. Each level encodes a ... set of knowledge for its processing, and the coverage of the two sets ... overlaps substantially.” (Sun [20] p. 44). That is, the two “levels” (or two sets of modules) encode somewhat similar or overlapping contents. But they encode their contents in different ways: Symbolic and subsymbolic representations are used, respectively. Different mechanisms are thus involved at these two “levels”. It was hypothesized that these two different “levels” can potentially work together synergistically, complementing and supplementing each other.

However, although the distinction is important, the terms involved in many existing dual-process theories have often been ambiguous. Not much finer-grained analysis has been done, especially not in a precise, mechanistic, process-based way. Therefore, we need a conceptual and computational framework in this regard. The Clarion cognitive architecture [21, 22] may be used, at a theoretical level, as a conceptual tool for generating interpretations and explanations in this regard. Indeed, many empirical and simulation studies have been conducted within the Clarion framework that shed light on relevant issues [21, 22].

In summary, dual-process theories have important implications for neural-symbolic models: If cognitive-psychological realism is what one wants to achieve in developing

computational models or systems, dual-process theories must be taken into consideration. However, some issues involved in dual-process theories are more complex than often assumed. These issues are crucial for developing computational cognitive architectures, and in turn computational cognitive architectures can help in disentangling these and other theoretically important issues. Together they can lead to better neural-symbolic models.

Keywords: Cognitive science, Cognitive architectures, Neural-symbolic models, Dual-process theories.

10 Self-Aware Robots and Conscience

Junichi Takeno, Meiji University, 1-1, Kanda-Surugadai, Chiyoda-ku, Tokyo 101-8301, Japan; and Heuristics Science Research Institute, 15-4 Komukai-cho, Saiwai-ku, Kawasaki-shi, Kanagawa-ken 212-0003, Japan

I recently found that a self-aware system could be utilized as a very important element for explaining human conscience.

My research group and I have presented our MoNAD structure which is configured as a double-layered recursive neural network. The self-aware system is built using many of these MoNADs. The most important feature of a MoNAD is its self-reflective function. This functionality enables the self-aware system to discriminate between its own self and some other entity, and such a capability could make it possible to generate a representation of the self on the system. The presentation that my research group and I gave at the prior BICA conference described how the self-aware system could represent pleasant and unpleasant states.

I plan to speak about conscience in the self-aware system at this conference.

Keywords: MoNAD, Self-aware system, Neural network, Self-evolving, Mirror neuron, Mirror image cognition, Conscience.

11 Concluding Remarks

More information about the conference can be found at <https://bica2023.org>. Program Committee Chairs (the first two authors of this Chapter) are grateful to all contributors, especially to the Keynotes and Invited Speakers represented here by their abstracts.

References


1. Sloman, A.: Varieties of evolved forms of consciousness, including mathematical consciousness. *Entropy* **22**(6), 615 (2020). <https://doi.org/10.3390/e22060615>
2. Thaler, S.L.: Vast topological learning and sentient AGI. *J. Artif. Intell. Conscious.* **8**, 1–30 (2021)
3. Kelley, T.D.: Robotic dreams: a computational justification for the post-hoc processing of episodic memories. *Int. J. Mach. Conscious.* **6**(2), 109–123 (2014)

4. Goertzel, B.: Paraconsistent interzones and associated wild speculations. *Eurykosmotron* 8/182021. <https://bengoertzel.substack.com/p/paraconsistent-interzones>. Accessed 19Aug 2023
5. Goertzel, B.: *The Hidden Pattern: A Patternist Philosophy of Mind* 2006. Brown Walker Press, Irvine (2006)
6. Hendrikse, S.C.F., Treur, J., Wilderjans, T.F., Dikker, S., Koole, S.L.: On becoming in sync with yourself and others: an adaptive agent model for how persons connect by detecting intra- and interpersonal synchrony. *Hum. Cent. Intell. Syst.* **3**, 123–146 (2023)
7. Hendrikse, S.C.F., Treur, J., Koole, S.L.: Modeling emerging interpersonal synchrony and its related adaptive short-term affiliation and long-term bonding: a second-order multi-adaptive neural agent model. *Int. J. Neural Syst.* **33**(7), 2350038 (2023)
8. Hendrikse, S.C.F., Treur, J., Wilderjans, T.F., Dikker, S., Koole, S.L.: Becoming attuned to each other over time: a computational neural agent model for the role of time lags in subjective synchrony detection and related behavioral adaptivity. In: Mahmud, M., He, J., Vassanelli, S., van Zundert, A., Zhong, N. (eds.) *Brain Informatics. BI 2022. LNCS*, vol. 13406, pp. 369–383. Springer Nature, Cham (2022). https://doi.org/10.1007/978-3-031-15037-1_30
9. Hendrikse, S.C.F., Treur, J., Wilderjans, T.F., Dikker, S., Koole, S.L.: Switching in and out of sync: a controlled adaptive network model of transition dynamics in the effects of interpersonal synchrony on affiliation. In: Cherifi, H., Mantegna, R.N., Rocha, L.M., Cherifi, C., Micciche, S. (eds.) *Complex Networks and Their Applications XI. COMPLEX NETWORKS 2016 2022, Studies in Computational Intelligence*, vol. 1078, pp. 81–95. Springer Nature, Cham (2023). https://doi.org/10.1007/978-3-031-21131-7_7
10. Hendrikse, S.C.F., Treur, J., Koole, S.L.: Relationship-specific and relationship-independent behavioural adaptivity in affiliation and bonding: a multi-adaptive dynamical systems approach. *Cognit. Syst. Res.* **84**, 101182 (2023)
11. Lieto, A.: *Cognitive Design for Artificial Minds*. Taylor & Francis, Routledge (2021)
12. Samsonovich, A.V.: Toward a unified catalog of implemented cognitive architectures (review). In Samsonovich, A.V., Jóhannsdóttir, K.R., Chella, A., and Goertzel, B. (eds.) *Biologically Inspired Cognitive Architectures 2010: Proceedings of the First Annual Meeting of the BICA Society*. *Frontiers in Artificial Intelligence and Applications*, vol. 221, pp. 195–244. IOS Press, Amsterdam (2010).
13. Samsonovich, A.V.: On a roadmap for the BICA challenge. *Biol. Inspired Cognit. Archit.* **1**, 100–107 (2012). <https://doi.org/10.1016/j.bica.2012.05.002>
14. Samsonovich, A.V., Lebiere, C., Ritter, F.E.: MAPPED repository: a comparative database of biologically inspired cognitive architectures (BICA). *Dynamic Poster DP02.10/DP10*. In: *Proceedings of the 2015 Neuroscience Meeting Planner*, online. Society for Neuroscience, Washington, DC (2015)
15. Jóhannsdóttir, K.R. and Samsonovich, A.V.: Biologically inspired cognitive architectures: one more step forward. In: Samsonovich, A.V., Jóhannsdóttir, K.R. (eds.) *Biologically Inspired Cognitive Architectures 2011: Proceedings of the Second Annual Meeting of the BICA Society*. *Frontiers in Artificial Intelligence and Applications*, vol. 233, pp. 3–8. IOS Press, Amsterdam (2011)
16. Chella, A., Lebiere, C.L., Noelle, D.C., and Samsonovich, A.V.: On a roadmap to biologically inspired cognitive agents. In: Samsonovich, A.V., Jóhannsdóttir, K.R. (eds.) *Biologically Inspired Cognitive Architectures 2011: Proceedings of the Second Annual Meeting of the BICA Society*. *Frontiers in Artificial Intelligence and Applications*, vol. 233, pp. 453–460. IOS Press, Amsterdam (2011)
17. Samsonovich, A.V., Robertson, P.: A Forum at the dawn of the era of biologically inspired intelligent machines. *Proced. Comput. Sci.* **41**, 1–5 (2014). <https://doi.org/10.1016/j.procs.2014.11.077>

18. Larue, O., West, R., Rosenbloom, P.S., Dancy, C.L., Samsonovich, A.V., Petters, D., Juvina, I.: Emotion in the common model of cognition. *Proced. Comput. Sci.* **145**, 740–746 (2018)
19. Sun, R., Bookman, L. (eds.): *Computational Architectures Integrating Neural and Symbolic Processes*. Kluwer, Needham, MA (1994)
20. Sun, R.: *Integrating Rules and Connectionism for Robust Commonsense Reasoning*. John Wiley and Sons, New York (1994)
21. Sun, R.: *Duality of the Mind*. Lawrence Erlbaum Associates, Mahwah, NJ (2002)
22. Sun, R.: *Anatomy of the Mind*. Oxford University Press, Oxford (2016)



Selected Presentation Abstracts from BICA*AI 2023

Maria Yu. Boboshko^{1,2}, Ekaterina S. Garbaruk^{1,3}, Veronika M. Knyazeva²,
Marina J. Vasilyeva², Aleksander A. Aleksandrov², Anton Kolonin⁴,
and Alexei V. Samsonovich⁵ (✉) 

¹ Pavlov First St. Petersburg State Medical University, St. Petersburg, Russia

² St. Petersburg State University, St. Petersburg, Russia

³ St. Petersburg State Pediatric Medical University, St. Petersburg, Russia

⁴ Novosibirsk State University, Novosibirsk, Russia

⁵ National Research Nuclear University MEPhI, Kashirskoe Shosse 31, Moscow, Russia
avsamsonovich@mephi.ru

Abstract. This chapter comprises selected short and extended abstracts of presentations given at the 2023 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence, also known as the 14th Annual Meeting of the BICA Society (BICA*AI 2023), held in Ningbo, China during October 13–15, 2023. Only those abstracts are included that were not accompanied by full papers. Abstracts are arranged alphabetically by the first author's last name, as follows: (1) Boboshko M. Yu. et al., (2) Knyazeva V. M. et al., (3) Kolonin A., (4) Samsonovich A. V. All contributors are listed as the authors of this chapter. Support information is given at the end of each abstract and in the Acknowledgments section at the end.

Keywords: Developing brain · Event-related potentials · Rapid word learning · AGI · Knowledge representation · LLM · Emotional intelligence

1 Assessment of Speech Intelligibility in Preschoolers

Maria Yu. Boboshko, Ekaterina Garbaruk, Veronika Knyazeva, and Marina Vasilyeva

Speech testing in a free field is one of the most effective methods for assessing the speech competence of hearing-impaired children who use hearing aids/implants in everyday life. Recently the Russian-language speech material of this method has been supplemented with new lists of monosyllabic words and sentence test known as Simplified RuMatrix test, however, there are no normative data available for these tests in preschool children. Twenty normal-hearing Russian-speaking children 6–7 y.o. were examined using speech audiometry in free field. To exclude central auditory processing disorder Fisher's questionnaire, dichotic digits test, binaural fusion test and gap detection test were applied. Normative data were obtained: intelligibility of monosyllabic words in quiet $98.8 \pm 3\%$, in noise $78.7 \pm 10.3\%$; speech recognition threshold in quiet 22.2 ± 3.8 dB SPL, in

noise 7.9 ± 1.1 dB SNR. The data can be used in the examination of hearing-impaired children.

Supported by RSF grant No 23-25-00108.

Keywords: Speech audiometry, Speech intelligibility, Monosyllabic words, Simplified Russian matrix sentence test, Preschool children.

2 Neural Correlates of Single-Shot Word Learning in Young Children: An ERP Study

Veronika M. Knyazeva, Aleksander a. Aleksandrov, Maria Yu. Boboshko, Ekaterina S. Garbaruk, and Marina J. Vasilyeva

Rapid new word acquisition could be mediated by neurocognitive mechanism known as fast mapping (FM). Recent studies in children revealed FM promotes rapid integration of newly learned items into cortical memory networks from a single exposure. While this rapid neural memory trace build-up was found only for stimuli with native-language phonology, neural underpinnings of ultra-rapid learning non-native words have not been assessed. To address this issue, we used ERPs to define brain dynamics elicited by novel native and acoustically closely matching non-native words following a single-shot semantic associative learning task in young children. ERP results revealed significant decrease in right fronto-central negativity present at 308–358 ms after the stimulus recognition point for novel native trained words, with no learning effect for non-native ones. Further research is needed to study ultra-rapid learning mechanism in the developing brain.

Supported by RSF grant No 23-25-00108.

Keywords: Developing brain, Event-related potentials, Language, Fast mapping, Rapid word learning.

3 Model of Personal Consciousness Based on the Principles of Social Proof and Free Energy

Anton Kolonin

We propose a model of personal consciousness based on the principles of social proof and free energy, based on fundamental research on the principle of free energy, which states the minimization of uncertainty as the goal of the evolution of matter and mind [1], as well as the principle of social proof, supported by phenomenological studies in the field of social science [2]. At the same time, it is assumed aligned with the notion of general intelligence as the ability to adapt to dynamic environments given insufficient knowledge and resources [3] within the graph-based knowledge representation based on probabilistic logic [4, 5].

The theory of human intelligence and cognitive abilities have been always key for fundamental sciences and amount of research made in this area is enormous. Over the last 50 years, exponential burst has happened in information technologies, enabling light-of-speed social communications between any two humans on Earth and causing appearance

of intelligent artificial agents operating on behalf on marketing, political and government agencies to act in social networks against global human communities—collecting Earth-wide social information, building predictive models of mass behavior and manipulating consumer, social and political activity. There more recent works modeling multi-agent dynamic in social environments and discussing phenomena of mass behavior. In the very last years, it has happened so that entire scope of knowledge accumulated by humanity could be uploaded into computer software system using computable graph representation. Accompanied with available computational models of intelligence, it is now theoretically possible to create “in silico” models of cognitive behavior of a single human or social interactions for entire societies. Such models could be invaluable for wide range of applications on consumer and corporate markets. For personal use, it could be built into intelligent software assistants providing intelligent filtering of incoming electronic media and automatic predictive search for desired information—accordingly to cognitive profile of a user. For use by media supplier, it could be beneficial to have psychologically correct dynamic models of target audience to ensure precise account for consumer intent and desire, eliminating irrelevant noise in advertisement information and respective rejection on consumer side.

What we suggest could be called “dynamic social evidence-based knowledge representation model”, as it accounts for dynamic scoping of evidence calculation base over time scale, modulated by social connections and constrained by physical resources. The former principle can be found corresponding to “social proof”, introduced by Cialdini. The latter one can be associated by “free energy principle” posed by Friston as a key driver for intelligence as well as “using limited resources” capacity of an intelligent being identified by Wang. The key part of the model is evaluation of confidence as cumulative evidence, collected in specific time frame and modulated by relevant social context, normalized to standard interval between 0 (no supporting evidence) and 1 (maximum supporting evidence) inclusively. It is anticipated this model could provide psychologically plausible results for cognitive agent interacting in social multi-agent environment, perceiving information from it in order to make decisions and take actions, with possibility to change its own knowledge about the environment in the course of operations. In order to communicate, agents are implied to have some jointly shared system of fundamental knowledge (called “belief system”) regarding the surrounding environment and themselves. They should also have a mechanism for either accepting the knowledge coming to an agent from its outer world (if it is compatible with the agent’s belief system), or rejecting it (in the opposite case). Further, for different sorts of accepted knowledge, an agent should be able to make judgments regarding reliability of different facts, which can be done based on the amount of evidence associated with these facts. Each evidence is considered in terms of trust towards its source such as social connection supporting the evidence.

Keywords: Artificial general intelligence, Consciousness, Knowledge representation, Free energy, Social proof.

4 Toward a Human-Level Artificial Social-Emotional Intelligence

Alexei V. Samsonovich

This talk will present one specific approach to the development of Artificial Social-Emotional Intelligence (ASEI). The key word here is “social”. While affective modeling mostly concerns with emotional states of an agent, here the focus of attention is on social relationships of agents and their feelings to each other, manifested in behavior. The ultimate goal is to create a human-level ASEI. Although a system of deep neural networks can become its embodiment, the path to the goal lies through cognitive modeling. The eBICA cognitive architecture is the basis of the approach, grounded in three main concepts: semantic maps, moral schemas, and mental perspectives. Its integration with Large Language Models (LLM) produces interesting results: it appears that social behavioral characteristics of a virtual actor can be ranked higher than those of a human. The new technology can be anticipated to find many practical applications: in social services, in healthcare, in education (cognitive tutoring systems [6]) and more.

The work is funded by the Russian Science Foundation Grant No. 22-11-00213.

Keywords: Affective modeling, LLM, Emotional intelligence.

Acknowledgments. Support for Sections 1 and 2 was provided by the RSF grant No 23-25-00108. Support for Section 4 was provided by the RSF grant No 22-11-00213.

References

1. Friston, K., Kilner, J., Harrison, L.: A free energy principle for the brain. *J. Physiol. Paris* **100**(1–3), 70–87 (2006). <https://doi.org/10.1016/j.jphysparis.2006.10.001>
2. Cialdini, R.B.: Social proof: truths are us. In: Cialdini, R.B. (ed.) *Influence: Science and Practice*, 5th edn., pp. 97–140. Allyn & Bacon, Boston (2008)
3. Wang, P.: On defining artificial intelligence. *J. Artif. Gen. Intell.* **10**, 1–37 (2019)
4. Goertzel, B., Iklé, M., Goertzel, I., Heljakka, A.: *Probabilistic Logic Networks: A Comprehensive Framework for Uncertain Inference*, 1st edn., 2nd Printing. Springer (2008)
5. Vityaev, E.E., Martinovich, V.V.: Probabilistic formal concepts with negation. In: Voronkov, A., Virbitskaite, I. (eds.) *Proceedings of the 9th International Ershov Informatics Conference on Perspectives of System Informatics, PCI 2014, LNCS*, vol. 8974, pp. 385–399 (2015)
6. Samsonovich, A.V., De Jong, K.A., Kitsantas, A., Peters, E.E., Dabbagh, N., Kalbfleisch, M.L.: Cognitive constructor: an intelligent tutoring system based on a biologically inspired cognitive architecture (BICA). *Front. Artif. Intell. Appl.* **171**(1), 311–325 (2008)

Selected Papers



Dot Density Effects on Stereoscopic Transparency: A Cross-Correlation Model Analysis

Saori Aida^(✉) 

Graduate School of Sciences and Technology for Innovation, Yamaguchi University,
Ube 755-8611, Japan
saoaida@yamaguchi-u.ac.jp

Abstract. Depth perception is one of the cornerstones of human vision, with binocular disparity serving as one of its most powerful cues. This study employs simulation experiments to assess the cross-correlation model's capability to explain the depth reduction phenomenon in stereoscopic transparency stimuli with identical binocular disparities. We also examined the influence of dot density on perceived depth. Our findings indicate that the cross-correlation model accurately detects the number of surfaces in 2-POTS (parallel, overlapping, transparent stereoscopic surfaces) and 3-POTS across various conditions. The model effectively accounts for the depth reduction phenomenon except when large binocular disparities are involved. Notably, dot density was not a significant factor in affecting the simulation results. This outcome aligns with existing psychophysical observations. This study provides critical insights into the model's robustness and its ability to explain intricate phenomena like stereoscopic transparency and depth reduction. The study underscores the versatility of the cross-correlation model and sets the stage for future research aiming at quantitative comparisons with psychophysical metrics.

Keywords: Stereoscopic transparency · Binocular disparity · Cross-correlation model · Depth reduction phenomenon

1 Introduction

In early vision, various depth cues are used to estimate the three-dimensional structure of the external world. Among these, binocular disparity is one of the most powerful depth cues. Since our two eyes perceive the external world from different angles, there is a displacement between the images reflected by the two eyes according to the magnitude of depth. This disparity is binocular disparity, and the brain processes this information to perceive depth. The stereoscopic impression obtained from binocular disparity cues is called binocular stereopsis. Because of the geometric relationship between the object and the two eyes, the amount of depth in binocular stereopsis is proportional to binocular disparity and inversely proportional to the square of the distance to the object. It is known that the geometric relationship is in good agreement with human perception [1].

As physiologically valid models of binocular stereopsis, cross-correlation models and disparity energy models have been analyzed and disparity detection algorithms based on these models have been studied intensively [2]. The cross-correlation model is a method for estimating binocular disparity through the correlation between images of the left and right eye. The cross-correlation model [3, 4] and the disparity energy model [5] are the most valid models of stereoscopic 3D perception. The algorithm of the cross-correlation model is based on the parallax energy model. The algorithm for the cross-correlation model has the same basic characteristics as the disparity energy model and is essentially the same as the stereo matching method. This technique has been studied to obtain depth maps from binocular images. In this technique, the properties of the binocular neurons in V1 are assumed to be such that after the image is filtered by the monocular receptive field, an estimate of binocular disparity is obtained by computing the binocular correlation [5]. The cross-correlation model is considered equivalent to the binocular disparity estimation performed in V1.

In studies of the binocular mechanisms of the brain, the phenomenon of stereo transparency, in which multiple depths are perceived simultaneously in the same region of the visual field, has been pointed out as important and has been the subject of numerous psychophysical studies [3, 6, 7]. Stereo transparency is RDS with multiple binocular disparities, and when fused, multiple overlapping surfaces are perceived in the same depth direction. The problem of stereo transparency is expected to have implications for the problem of information representation in the brain, i.e., how to encode and decode the superimposed multiple disparity information. One of the psychophysical studies using stereoscopic stimuli is the depth reduction phenomenon. This phenomenon is that the amount of perceived depth decreases as the number of surfaces comprising a stereoscopic transparency stimulus increases. This phenomenon was reported when using two, three, and four parallel, overlapping, transparent stereoscopic surfaces known as parallel, overlapping, transparent stereoscopic surfaces (POTS).

In this paper, we conducted simulation experiments on binocular disparity of stereoscopic stimuli using a cross-correlation model. The purpose of the experiment was to confirm whether the cross-correlation model can explain the phenomenon of depth reduction in which the stimulus with a larger number of surfaces is perceived as having less depth when the stereoscopic stimuli have the same binocular disparity and to investigate the effect of dot density.

2 Methods

2.1 Stimuli

The stimuli were stereoscopic transparency stimuli: a pair of RDS (random-dot stereograms). The number of surfaces of the stimuli was two or three. The images were 800×640 pixels, with each dot being 4×4 pixels. The binocular disparity between the front and back surfaces was 12, 16, and 20 pixels. 2-POTS stimuli were presented relative to the stimulus presentation screen at (-6 pixels, $+6$ pixels), (-8 pixels, $+8$ pixels), and (-10 pixels, $+10$ pixels). The three-POTS stimuli had binocular disparities of (-6 pixels, 0 pixels, $+6$ pixels), (-8 pixels, 0 pixels, $+8$ pixels), and (-10 pixels, 0 pixels, $+10$ pixels).

pixels), (-8 pixels, 0 pixels, $+8$ pixels), and (-10 pixels, 0 pixels, $+10$ pixels) relative to the stimulus presentation screen, respectively. $+$ represents crossed and $-$ represents uncrossed binocular disparity. Overall dot densities were 0.008 , 0.016 , 0.024 , and 0.032 dots/pixel².

2.2 Procedure

Before computing the cross-correlation of the computer-generated left and right half-field images (random dot pattern), the images were low-pass filtered. This manipulation was performed to approximate the images to retinal images. According to the optics of the human eye, the image input to the retinal image has no obvious contours and is blurred. Specifically, as a manipulation for approximation, the half-field image was convolved with a point-spread function (point-spread function) that matches well with the 3 mm pupil.

The filtered left and right eye images were sent to the binocular cross-correlation function. If the left and right eye images are identical, then the binocular disparity between them is zero and the value of the cross-correlation is 1 . The window diameter was $\pm 1\sigma$ of the Gaussian distribution.

Since the binocular disparity obtained from this simulation is in pixel units, the author used a more accurate sub-pixel estimation to obtain the peak binocular disparity. Here, sub-pixel estimates (peak binocular disparity) were obtained using parabola fitting.

Using the above procedures, simulation experiments with the cross-correlation model were performed for each binocular disparity, stereoscopic transparency stimulus, and density condition for 10 trials.

3 Results

First, we examined whether the cross-correlation model was able to detect a single binocular disparity. Figure 1 shows the simulation results when the RDS with 0 pixel binocular disparity was input. The range of -30 to 30 pixels was checked. As can be seen, in the case of the single disparity, the cross-correlation model detects the binocular disparity of the RDS.

Next, we examined whether the cross-correlation model was able to detect the stereoscopic transparency stimulus. Figure 2 shows the simulation results of the 2-POTS and 3-POTS simulations under the small binocular disparity condition at low density: for 2-POTS, there were two peaks; for 3-POTS, there were three peaks. Thus, the cross-correlation model was able to separate and detect the surfaces of the stereoscopic transparency stimulus in all conditions.

After running simulations for all conditions, the mean and 95% confidence interval (95% CI) were determined for each condition. First, a three-way repeated measures ANOVA (2 POTS \times 4 density \times 3 binocular disparity) was conducted on the mean simulated disparity. The analysis revealed that the main effect of POTS, $F(1, 9) = 332.81$, $p = 0.00$, $\eta^2 = 0.04$, the main effect of binocular disparity, $F(2, 18) = 4124.82$, $p = 0.00$, $\eta^2 = 0.92$, and the simple interaction between POTS and binocular disparity, $F(2, 18) = 93.25$, $p = 0.00$, $\eta^2 = 0.02$, were statistically significant, but the main effect

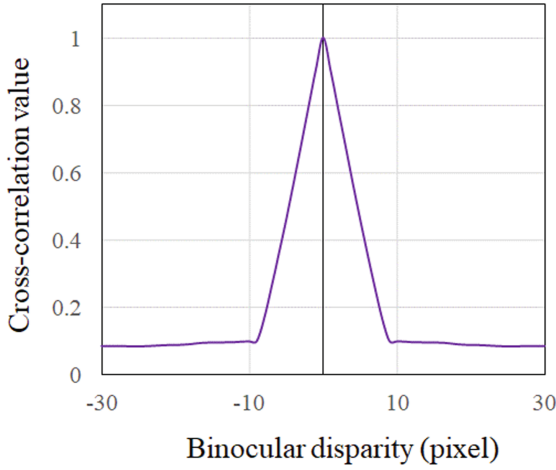


Fig. 1. The simulation results of the single binocular disparity.

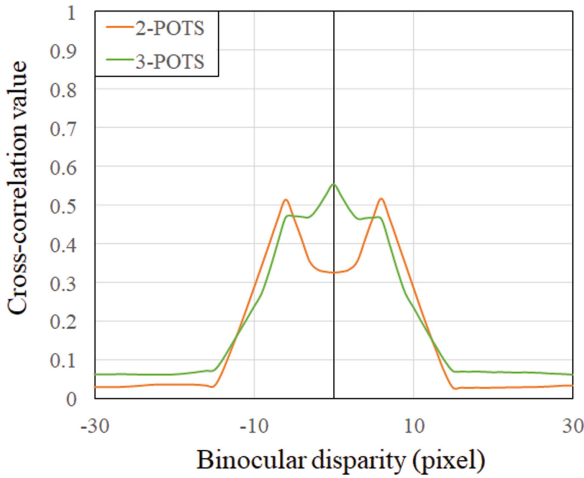


Fig. 2. The simulation results of the 2-POTS and 3-POTS.

of density, $F(3, 27) = 2.21, p = 0.11, \eta^2 = 0.00$; the three-way interaction effect, $F(6, 54) = 1.91, p = 0.10, \eta^2 = 0.00$; the simple interaction between POTS and density, $F(3, 27) = 2.24, p = 0.11, \eta^2 = 0.00$; and the simple interaction between density and binocular disparity, $F(6, 54) = 1.91, p = 0.10, \eta^2 = 0.00$, were statistically insignificant.

Figure 3 shows the experimental results for each density condition. The vertical axis is simulated binocular disparity. The horizontal axis is binocular disparity. Error bars in Fig. 3 represent 95% confidence intervals (95% CI). As shown in Fig. 3, simulated binocular disparity increases with increasing binocular disparity in all conditions. Also, as shown in Fig. 3, the simulated binocular disparity of 3-POTS is smaller than that of 2-POTS for the 2-POTS and 3-POTS stimuli; in the condition with the largest binocular

disparity, the simulated binocular disparity for the 2-POTS and 3-POTS stimuli is equal. There was no difference in simulated binocular disparity by density. These results are consistent with the ANOVA results.

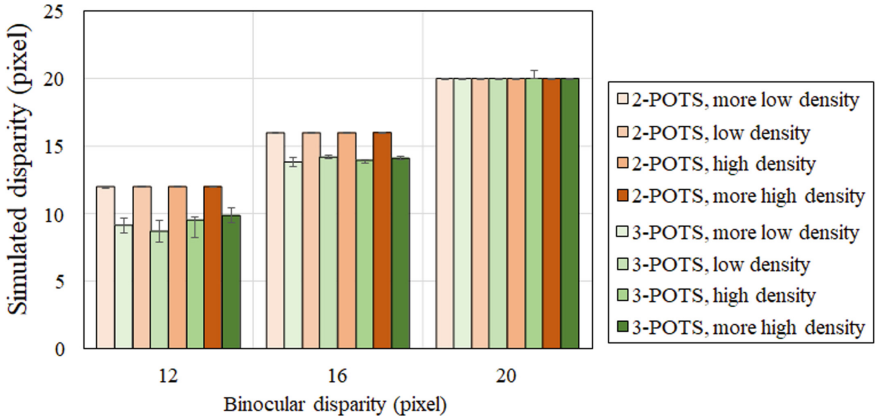


Fig. 3. The results for Experiment. Error bars are 95% confidence intervals.

Next, we analyzed the 95% CI of the mean simulated disparity for each condition as an indicator of the depth reduction phenomenon, and if the respective 95% CI for 2-POTS and 3-POTS did not overlap, we considered that the depth reduction phenomenon was occurring between 2-POTS and 3-POTS. As can be seen from Fig. 3, when binocular disparity is small and medium, the respective 95% confidence intervals of 2-POTS and 3-POTS do not overlap. On the other hand, when binocular disparity is large, the respective 95% confidence intervals of 2-POTS and 3-POTS overlap. As can be seen from Fig. 3, the trend was the same for all density conditions. These results indicate that the depth reduction phenomenon can be explained by this model except in the condition of large binocular disparity.

4 Discussion

In this study, we conducted simulation experiments on binocular disparity of stereoscopic stimuli. We investigated whether the cross-correlation model can explain the phenomenon of depth reduction, in which the stimulus with the larger number of faces is perceived as having less depth when the stereoscopic transparency stimuli have the same binocular disparity, and examined the effect of dot density. The results showed that the cross-correlation model was able to detect the number of faces of 2-POTS and 3-POTS in all conditions. It was found that the depth reduction phenomenon could be explained by the cross-correlation model, except for the condition with large binocular disparity. Simulation results were not affected by dot density.

It is known from psychophysics that 2-POTS appear to be separated when the disparity difference is approximately 3 arc min or greater, but below this level, they appear to be only one surface with thickness [1]. In the binocular disparity range treated in

this study, multiple peaks were detected, suggesting that the images were perceived as stereoscopic transparency.

In the model of Tsai and Victor [8], when the dot densities of the two surfaces are different, the squared error of the minima for the surface with the lower dot density is larger, making the detection of binocular differences difficult. This may correspond to the finding [6] that it is difficult to perceive surfaces with low dot density on dual surfaces. The dot density treated in this study is about the same level as that treated in [7], and the dot density was the degree treated in [7]. Therefore, it is considered possible to detect binocular disparity in the range of dot densities in this study, regardless of the dot density.

The study [7] reported that the depth reduction phenomenon reported in the previous study was not detected in this study. The depth reduction phenomenon reported by [7] was confirmed regardless of dot density. The simulation results of the cross-correlation model treated in this study showed that binocular disparity was simulated without being affected by dot density. This suggests that the cross-correlation model can well explain the perceived binocular disparity of stereoscopic transparency stimuli.

In the present study, we extended the cross-correlation model [3, 4] and showed that it can qualitatively explain various findings on stereopsis, especially the depth reduction phenomenon. Future work includes, first, quantitative comparison with psychological quantities such as stereopsis threshold and accuracy, and a model that can explain a wide range of binocular disparity.

Acknowledgement. This research was partially supported by JSPS KAKENHI (Grant-in-Aid for Early-Career Scientists), grant number 21K18027.

References

1. Howard, I.P., Rogers, B.: *Stereoscopic vision: Vol. 2. Perceiving in depth*. Oxford University Press, New York (2012)
2. Chen, Y., Qian, N.: A coarse-to-fine disparity energy model with both phase-shift and position-shift receptive field mechanisms. *Neural Comput.* **16**, 1545–1577 (2004)
3. Stevenson, S.B., Cormack, L.K., Schor, C.M.: Depth attraction and repulsion in random dot stereograms. *Vision. Res.* **31**, 805–813 (1991)
4. Filippini, H.R., Banks, M.S.: Limits of stereopsis explained by local cross-correlation. *J. Vis.* **9**(8), 1–18 (2009)
5. Anzai, A., Ohzawa, O., Freeman, R.D.: Neural mechanisms for encoding binocular disparity: receptive field position vs phase. *J. Neurophysiol.* **82**, 874–890 (1999)
6. Gepshtein, S., Cooperman, A.: Stereoscopic transparency: a test for binocular vision's disambiguating power. *Vision. Res.* **38**, 2913–2932 (1998)
7. Aida, S., Shimono, K., Tam, W.J.: Magnitude of perceived depth of multiple stereo transparent surfaces. *Atten. Percept. Psychophys.* **77**, 190–206 (2015)
8. Tsai, J.J., Victor, J.D.: Reading a population code: a multi-scale neural model for representing binocular disparity. *Vision. Res.* **43**, 445–466 (2003)



An Examination of Visual Influences on Sense of Ownership and Agency

Saori Aida^(✉)  and Yuui Ogawa

Graduate School of Sciences and Technology for Innovation, Yamaguchi University,
Ube 755-8611, Japan
saoaida@yamaguchi-u.ac.jp

Abstract. This study investigates the effects of vision on the sense of ownership and agency in virtual reality (VR) environments. Sense of ownership and agency are essential aspects of human perception, allowing us to feel connected to our bodies and actions. This study aims to extend previous research by using cat paw stimuli to move on their own. We experimented with a VR setup to present stimuli that mimic grasping behavior. The stimuli included human and cat paws. For the human hand, a sense of ownership and agency were evident. Cat paw stimuli consistently evoked a stronger sense of agency. The sense of ownership was lower for cat paw stimuli than for human hand stimuli. When there was a time delay, the sense of ownership and sense of ownership disappeared in all conditions. These findings shed light on how different visual stimuli, such as human and cat paws, affect the sense of ownership and agency in VR. The implications extend to areas such as medical rehabilitation and entertainment.

Keywords: Sense of agency · Sense of ownership · Virtual reality

1 Introduction

We usually live our lives by performing casual actions unconsciously. When we are drinking a drink, we do not think “I am the one drinking the drink” or “The hand holding the cup is my hand,” and we can recognize our image in the mirror as “This is me” without feeling any discomfort. This is because they have a sense of ownership and a sense of agency. A sense of ownership refers to “the sense that a certain event occurs in one’s own body,” while a sense of agency refers to “the sense that one causes a certain event to occur in one’s own body” [1].

Rubber Hand Illusion (RHI) is a representative study of the sense of ownership and agency [2, 3]. RHI is a phenomenon in which a subject’s arm is hidden by a partition, and only the rubber hand (RH) on the table is visible. This is a phenomenon in which the subject is touched only by the prosthetic arm but is perceived as being touched by his arm. This is a phenomenon in which simultaneous stimuli of visual and tactile senses generates a sense of ownership and a sense of agency toward the prosthetic arm. The RHI is also known to be illusory with CG hands, which are visual stimuli presented in

a Virtual Reality (VR) space [4–6]. Studies using virtual space have also confirmed the illusion of transparent bodies [7] and balloons [8] that inflate with hand movements.

The research [9] proposed that three conditions are necessary for the generation of a sense of ownership: (1) the visual appearance of the object matches the internal knowledge of the shape of the human body part, (2) the actual location and effector match, and (3) the visual and sensory stimuli match. However, the research [8] suggests that (1) and (2) are not always necessary.

In this study, cat paws, which are visually different from human hands, were presented in a VR space to examine their effects on the sense of ownership and agency. We examined the effects of stimuli that are visually different from human hands that can move by themselves on the sense of ownership and agency. We hypothesize that behavior creates a stronger sense of ownership and agency. If we can clarify the factors that create a sense of ownership and agency through this experiment, we believe that we can provide a VR experience that gives a strong sense of ownership even if the body is different from that of humans, and bring about a more immersive VR experience.

2 General Methods

2.1 Apparatus

A Windows PC (OS: Windows 10 Pro, RAM: 32 GB, CPU: Intel Core i9-10900, GPU: GeForce GTX 3080) was used. Visual stimuli were presented on a head-mounted display (HTC VIVE Pro Eye, resolution: 1440×1600 pixels per eye, 90 Hz:90 Hz refresh rate, 110° viewing angle). The subject's right hand was equipped with a motion tracker (VIVE Tracker 3.0) and a self-made glove. The position of the hand was captured by the motion tracker into the computer. The bending of the fingers was measured by a bending sensor attached to the homemade glove and captured in the computer.

2.2 Stimuli

Visual stimuli were created with the 3DCG software Blender (3.3.1). The visual stimuli were a human hand and a cat's paw. The visual stimuli were displayed using Unity (2020.3.0f1) to create the VR space. The cubes were square. Timing conditions were added to the stimuli. The timing conditions included a synchronous condition, in which the subject's movements were perfectly aligned with the stimulus movements, and an asynchronous condition, in which the stimuli moved after the subject's movements. For all stimuli, the delay time for the asynchronous condition was 20 ms.

2.3 Procedure

Subjects were asked to read the on-screen questions while wearing the HMD to see if the HMD was in focus. The experiment consisted of four trials of the number of stimuli (2) \times timing condition (2) in random order by the subject. Subjects were allowed to take breaks whenever needed. The trials consisted of three sections: a counting section, a test section, and a questionnaire section. All trials began with the counting section. In the

counting section, subjects grabbed the cubes from the right tray and dropped them into the left tray, as in the Experiment. In the test section, subjects performed the same actions as many times as possible in 3 min. After the test section, the questionnaire section began. In the questionnaire section, eight questions were asked about the sense of agency and ownership, based on previous research [1], and all were answered on a 7-point Likert scale ranging from -3 (not at all true) to $+3$ (completely true). The questionnaire items are as follows. To avoid respondent bias, the questionnaire items were displayed randomly and some of the evaluation criteria were reversed in the questionnaire. The part marked “inverted” is the part where the evaluation criteria are inverted (-3 answers are considered to be 3 answers). However, the actual questionnaire did not include the word inverted. The questionnaire of sense of ownership was (Q1) I felt like I was looking at my own hand, (Q2) The CG hand felt like a part of my body, (Q3) The CG hand felt like my hand, and (Q4) I felt that he had a hand other than his right hand (reversed). The questionnaire of sense of agency was (Q1) The CG hand felt like my own movement, (Q2) The CG hand could be moved as I wanted, (Q3) The CG hand felt like someone else’s movement (reversed), and (Q4) The CG hand seemed to move on its own (reversed). After all sections were completed, subjects were asked to respond whether or not they knew how the stimuli moved in the real world.

2.4 Subjects

16 subjects (16 males, mean age 21.82) participated in the Experiment. All of the subjects had normal or corrected-to-normal vision. One subject was the author of this study, and the others were naive as to the purpose of the experiment. Subjects signed informed consent and participated in the experiment. The study was conducted by the Declaration of Helsinki, and approved by the Institutional Review Board of Yamaguchi University (protocol code 2022-003-01 and date of approval 16 May 2022).

3 Results and Discussions

An analysis of variance by ANOVA was conducted with stimulus type, timing condition (synchronous or asynchronous), and questionnaire results as within-subject factors (2 stimulus type \times 2 timing \times 9 questionnaire). The analysis of variance revealed the main effect of the timing condition [$F(1, 15) = 47.91, p < 0.01$] and the main effect of the questionnaire [$F(8, 120) = 8.47, p < 0.01$] were significant. The interaction between stimulus type and questionnaire results [$F(8, 120) = 12.15, p < 0.01$] and timing condition and questionnaire results [$F(8, 120) = 32.08, p < 0.01$] was also significant. The main effect of stimulus type [$F(1, 15) = 3.56, p = 0.08$] was insignificant. Simple effects for “A \times C” interaction showed a significant difference between human and cat paw stimuli [$F(1, 15) = 34.14, p < 0.01$] in Q1.

Next, for each question item, the mean value for each condition was calculated for each subject and is shown in Figs. 1 and 2. The vertical axis is the rating score and the horizontal axis is the question number. Blue bars indicate the synchronous condition and orange bars the asynchronous condition. Error bars are 95% confidence intervals; 95% confidence intervals were also used to examine significant differences between

timing conditions and between 95% intervals and zero ratings; the greater the number of items with significant differences from zero points, the higher the sense of ownership or agency in the synchronous condition and the lower the sense of ownership or agency in the asynchronous condition.

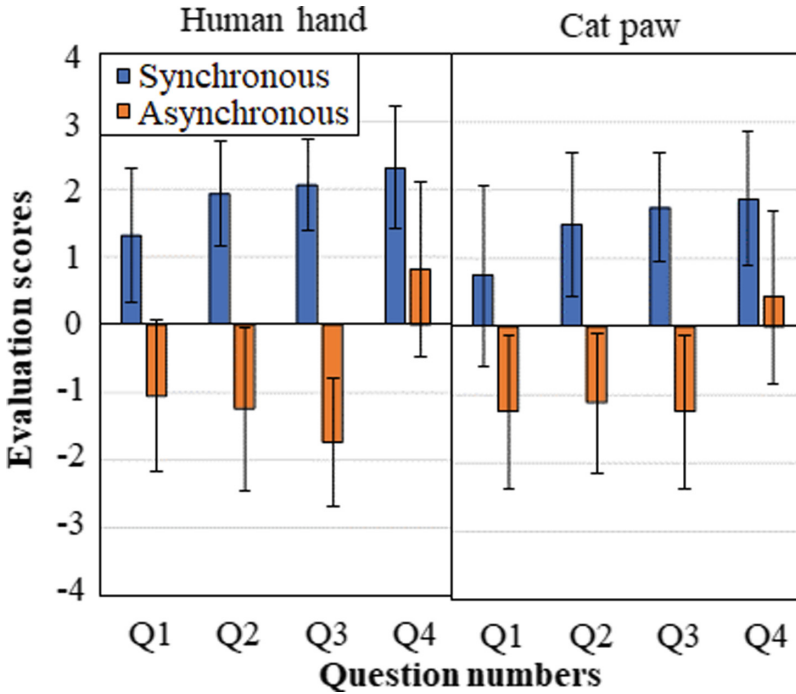


Fig. 1. Results on the sense of ownership for experiment. Error bars are 95% confidence intervals.

The number of items significantly different from 0 is summarized in Table 1. The table shows that in the ownership synchrony condition, the sense of ownership was strongly elicited when the stimuli were human hand stimuli and cat paw stimuli. In the asynchronous condition, stimuli other than the human hand did not elicit a sense of ownership. This difference may be because even in the asynchronous condition, the stimuli were human hands or human body parts, which elicited a higher sense of ownership. For agency, there was a less significant difference from 0 in the asynchronous condition when the stimulus was a cat paw. In the asynchronous condition, the cat paw stimulus did not elicit a sense of ownership, suggesting that this may have occurred because subjects perceived the stimulus as a tool for grasping an object rather than as part of their own body.

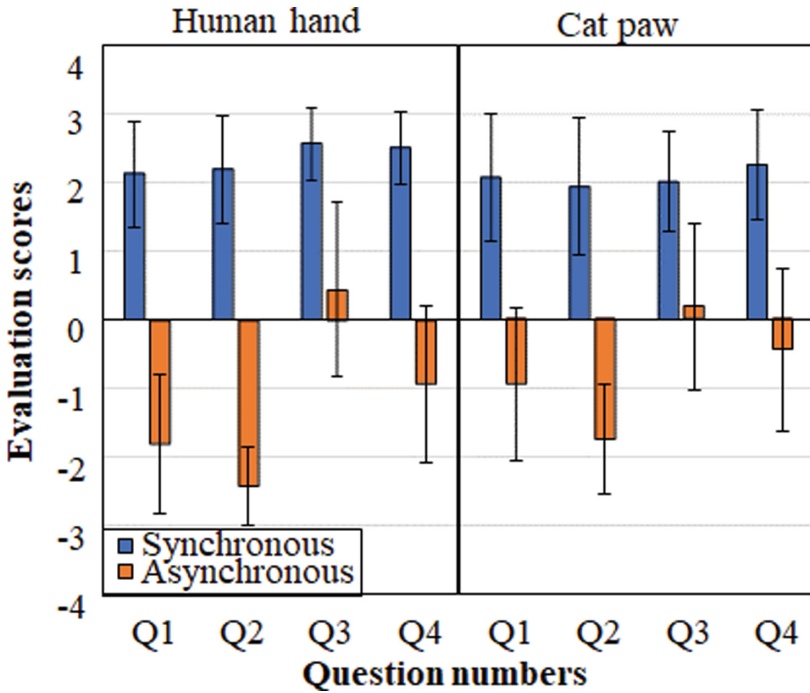


Fig. 2. Results on the sense of agency for the experiment. Error bars are 95% confidence intervals.

Table 1. Summary of the results of Experiment 1.

	A sense of ownership		A sense of agency	
	Synchronous	Asynchronous	Synchronous	Asynchronous
Human hand	4	2	4	2
Cat paw	3	3	4	1

4 General Discussions

The purpose of this experiment is to examine the effects of a human hand that can be moved and visually different stimuli on the sense of ownership and agency. In this experiment, we hypothesize that behavior produces a stronger sense of ownership and agency.

The experiment confirmed three things. One is the possibility that visual features have some effect on the sense of ownership since the number of significant differences from 0 indicates that human hand stimuli were more significant than cat paw stimuli. Second, human and cat paw stimuli produced agency. Third, neither human nor cat paw stimuli produced a sense of ownership and agency when the timing of vision and action did not match.

Two findings were made. The first is that cat paw stimuli bring about a sense of ownership. This phenomenon is likely due to the influence of cat paw stimuli on the sense of ownership. Second, it was observed that cat paw stimuli affected the sense of ownership and agency. It was suggested that a human hand is better for a strong sense of ownership, but that a sense of ownership can be achieved if the behavior is consistent. However, in the present experiment, it is not possible to assert that a sense of ownership was produced since the sense of ownership was mentioned only in the questionnaire. Further experiments with physiological responses, such as skin conductance responses and hand position estimation, are needed in future studies.

The purpose of this study was to explore the effects on the sense of ownership and agency in stimuli that are visually different from human hands that can be moved. In the experiment, human and cat paw stimuli were presented and acted upon in VR to examine the effects on subjects' perceptions. The results showed that even cat paw elicit a sense of ownership and agency. This finding has important implications for both the medical rehabilitation and virtual entertainment fields. In the medical field, understanding how non-human avatars generate a sense of ownership and agency could contribute to developing rehabilitation techniques for paralyzed body parts and prosthetics. Similarly, in the realm of virtual entertainment and gaming, insights into how to create a strong sense of ownership for a variety of avatars could increase player immersion and engagement. Overall, this research contributes to a broader understanding of how individuals perceive and interact with their bodies and external objects within virtual environments.

Acknowledgement. This research was partially supported by JSPS KAKENHI (Grant-in-Aid for Early-Career Scientists), grant number 21K18027.

References

1. Gallagher, S.: Philosophical conceptions of the self: implications for cognitive science. *Trends Cognit. Sci.* **4**(1), 14–21 (2000)
2. Kalckert, A., Ehrsson, H.H.: Moving a rubber hand that feels like your own: a dissociation of ownership and agency. *Front. Hum. Neurosci.* **6**(40), 1–14 (2012)
3. Lira, M., Egito, J.H., Dall'Agnol, P.A., Amodio, D.M., Gonçalves, Ó.F., Boggio, P.S.: The influence of skin colour on the experience of ownership in the rubber hand illusion. *Sci. Rep.* **7**(1), 1–13 (2017)
4. Tsakiris, M., Prabhu, G., Haggard, P.: Having a body versus moving your body: how agency structures body-ownership. *Consci. Cognit.* **15**(2), 423–432 (2006)
5. Zhang, J., Hommel, B.: Body ownership and response to threat. *Psychol. Res.* **80**(6), 1020–1029 (2016)
6. Preston, C., Kuper-Smith, B.J., Ehrsson, H.H.: Owning the body in the mirror: the effect of visual perspective and mirror view on the full-body illusion. *Sci. Rep.* **5**(1), 1–10 (2015)
7. Martini, M., Kiltner, K., Maselli, A., Sanchez-Vives, M.V.: The body fades away: investigating the effects of transparency of an embodied virtual body on pain threshold and body ownership. *Sci. Rep.* **5**(1), 1–8 (2015)
8. Ma, K., Hommel, B.: Body-ownership for actively operated non-corporeal objects. *Consci. Cognit.* **36**, 75–86 (2015)
9. Tsakiris, M.: My body in the brain: a neurocognitive model of body-ownership. *Neuropsychologia* **48**(3), 703–712 (2010)



Super-Performance: Sampling, Planning, and Ecological Information

Bradly Alicea^{1,2} 

¹ Orthogonal Research and Education Laboratory, Champaign-Urbana, IL, USA
bradly.alicea@outlook.com

² OpenWorm Foundation, Boston, MA, USA

Abstract. The connection between active perception and the limit of performance provides a path to understanding naturalistic behavior. We can take a comparative cognitive modeling perspective to understand the limits of this performance and the existence of superperformance. We will discuss two categories that are hypothesized to originate in terms of co-evolutionary relationships and evolutionary tradeoffs: *supersamplers* and *superplanners*. Supersamplers take snapshots of their sensory world at a very high sampling rate. Examples include flies (vision) and frogs (audition) with ecological specializations. Superplanners internally store information to evaluate and act upon multiple features of spatiotemporal environments. Slow lorises and turtles provide examples of superplanning capabilities. The Gibsonian Information (GI) paradigm is used to evaluate sensory sampling and planning with respect to direct perception and its role in capturing environmental information content. By contrast, superplanners utilize internal models of the world to compensate for normal rates of sensory sampling, and this relationship often exists as a sampling/planning tradeoff. Supersamplers and superplanners can exist in adversarial relationships, or longer-term as coevolutionary relationships. Moreover, the tradeoff between sampling and planning capacity can break down, providing relativistic regimes. We can apply the principles of superperformance to human augmentation technologies.

Keywords: Information theory · Perception and action · Human augmentation · Cognitive modeling

1 Introduction

Naturalistic behavior involves a tightly integrated action-perception loop. This involves sensory sampling and an internal model that enables planning and representation, which determines the acceleration and anticipatory capacity of this loop, respectively. Enhanced acceleration and anticipatory abilities are likely the product of evolutionary specialization. Co-evolutionary relationships (arms races) can result in very highly developed abilities. This leads to a philosophical question not necessarily involving a co-evolutionary relationship: how does a fly avoid the predatory catcher? In this form of the classic pursuit-evasion problem [1], flies exemplify supersamplers, while humans correspond to superplanners.

This dichotomy can also be observed as very fast (ballistic) and very slow (finessed) movements. Examples of the former can be observed in the hand-over-hand mechanics of the slow loris [2]. Slow loris locomotion is much slower than locomotory behavior in closely related organisms and requires coordinated specializations in both biomechanics and neuronal control. These phenotypic specializations become manifest in muscle adaptations that occur alongside the ability to plan. Supersampling is necessary as a by-product of very short windows of immediate sensorimotor feedback, and requires integrating feedforward information about the environmental state, and then generating a movement fast enough to successfully match that prediction. Trap-jaw ants produce large amounts of mandibular force relative to their body weight (300×) [3]. This produces a movement so fast that it requires a highly accurate feedback mechanism that qualifies as superplanning.

We will approach superperformance from a theoretical perspective that informs a universal cognitive model. To proceed, we define both supersampling and superplanners. At the core of superperformance is GI and the concept of ecological information processing with tradeoffs. These tradeoffs can be broken through so-called relativistic performance, which maximizes both sampling and planning. From an informational perspective, phenomena such as information aliasing and information moments provide ways to sample and interpret performance in a naturalistic context. To conclude, we consider the co-evolutionary origins of superformers modeled as a pair of complementary agents (emitter and receiver) who evaluate possibility spaces of various sampling sizes. In conclusion, we consider the application of such cognitive models to human augmentation.

1.1 What is Supersampling?

Our first type of superperformer are supersamplers. Biologically, supersamplers are defined by an enhanced and hyper-specialized sensory organ. In the case of vision, the response sensitivity and gain properties of photoreceptors are enhanced for dark vision in nocturnal insects [4]. Another example from insect vision involves enhanced luminance sensitivity in *Drosophila* [5]. This allows for change detection of luminance at shorter timescales than a visual system dominated by contrast sensitivity. Time scale plays a critical role in supersampling of the environment: there are several insect species with extremely high flicker-fusion frequency (FFF) rates. While human vision exhibits a 16 Hz FFF, examples from insect visual systems include FFFs of 60–100 Hz in *Drosophila hydei*, and 85–205 Hz in *Glossina morsitans* [6]. Supersampling in insects involves multiple traits working in concert to result in a set of appropriate cognitive conditions. Hyperacute vision is enabled by specialized photoreceptors that resolve target objects beyond their predicted motion-blur limit during spatial tracking [7]. When flies are in pursuit of a target, they use variables such as target size and predictions of target speed [8].

1.2 What Are Superplanners?

By contrast with supersamplers, superplanners are hyper-specialized for planning based on environmental information in the form of internal models. As superperformers of the other extreme, their environmental sampling abilities can be average to poor. Superplanners will tend to exhibit high degrees of embodiment. Rather than having enhanced sensory capabilities in the temporal domain, superplanners internally store information about the environment as memories or as adaptations. Superplanning also requires enhanced spatial and/or temporal planning abilities, which in turn (and like supersamplers) require physiological specializations. In the case of very slow movements, a slowdown-accuracy tradeoff may exist that is inversely related to speed-accuracy tradeoffs [2].

2 Interpretation

Gibson [9] argues that the combination of inputs, particularly covariance between input streams, results in a coherent flow of action. Gibsonian Information (GI) [10] is acquired according to a spatiotemporal Poisson distribution characterized by the parameter λ . Environmental information is not encountered at a uniform rate: information is processed at different rates at spatially dependent points in time, with large information moments being representative of affordances (information-rich objects). GI involves spatial information processing coupled to specific points in time, both of which are Poisson-distributed. The λ value for a specific environment sets the gain on the information content. High values of λ represents information-dense environments, with information in almost every spatially dependent temporal moment being representative of supersampling. Low values of λ results in long periods of sparse information, advantageous for superplanners who utilize an internal model to fill in the gaps in direct perceptual information.

2.1 Planning/Sampling Tradeoff

A tradeoff exists between planning rate and sampling rate (see Fig. 1). This planning-sampling tradeoff relates to a mix of traits that enhance either sensory abilities or planning capacity in the brain. Since supersamplers maintain an internal model, their planning rate never reaches zero. In fact, for most cases this tradeoff leads to performance optima between the superplanner and supersampler regime. However, there is a relativistic regime where this tradeoff breaks down, and the planning and sampling rate are both high. In the relativistic case, sampling is both dense with respect to direct perception of the environment and the ability to draw from an internal model of this high-resolution information.

2.2 Relativistic Performance

This relativistic regime of performance, or in regions where the planner-sampler tradeoff is broken, is shown in Fig. 1. According to this model, frogs and chameleons operate in

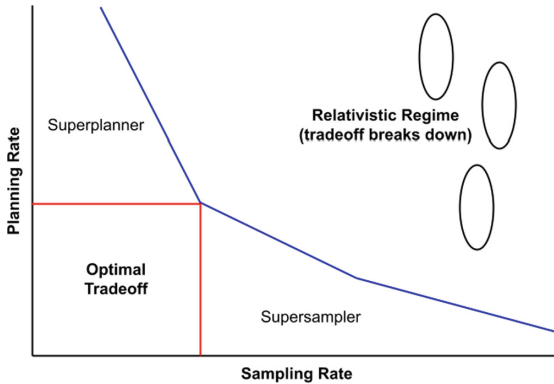


Fig. 1. A theoretical graph characterizing superplanners, supersamplers, and the planning-sampling tradeoff.

the relativistic regime, as the muscular properties and ballistic movements of the tongue yield very large peak power outputs [11]. The introduction or elimination of sensory aliasing depends on context. Low-pass sensory aliasing results from the need to fill in perceptual gaps caused by continuous tracking of stimuli at a low FFF, while high-pass sensory aliasing can result from supersamplers sampling an informationally sparse environment (see Fig. 2A). In terms of the visual system, environmental sampling is constrained by an organism’s metabolic rate [12] and the sampling rate at which visual samples become continuous scenes [13]. Supersamplers and superplanners then deal with these constraints in different ways. In primates (who are visual superplanners), the output of MT neuronal populations are non-linear when stimuli are separated over significant periods of time [14]. Low λ values reinforce the need for an internal modeling mechanism to fill in perceptual sampling gaps.

2.3 Information Aliasing

Information aliasing (Fig. 2A) occurs when normal samplers encounter situations that require the resolution of a supersampler. One example of this are flash crashes on the stock market, including the famous flash crash of 2010 [15]. Flash crashes occur when high-frequency trading algorithms trigger a massive sell-off amongst human traders in a short amount of time. The difference in information between the superperforming agent and the normally performing one is the source of low-frequency aliasing. In these cases, sensory aliasing without the ability to superplan favors overcorrection and the amplification of extreme or out-of-control responses. Therefore, stability can be restored through occupying the relativistic region in Fig. 1.

2.4 Active Sensing and Information Moments

GI relies upon active sensation in the form of continuous behaviors. In bats and weakly electric fish, an active field is emitted by the organism to detect objects within the organism’s receptive field [16]. The organism’s sensory receptors move with its body against

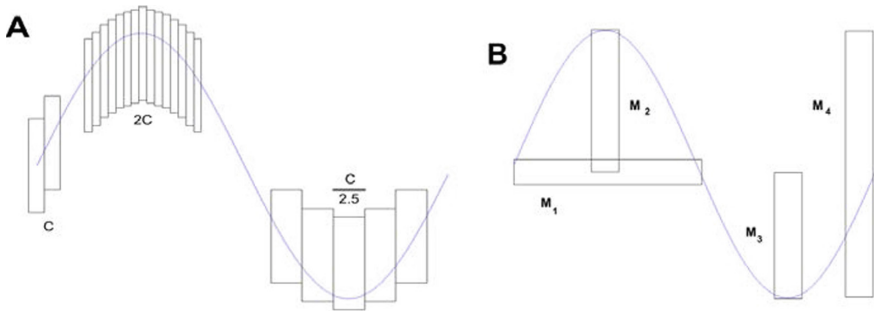


Fig. 2. Examples of differences between superperformers of different types (samplers, planners) and regular performers. **A** Demonstrates information aliasing. Sampling windows show the width of the true sampling rate C , width of the supersampler sampling rate ($2C$), and the width of the normal performance sampling rate ($C/2.5$). **B** Demonstrates information moments. Moments M_n demonstrate the various moments along one cycle of a sine wave.

the environment, enabling temporal tracking and spatial exploration. Differential movement that defines the active sensing envelope and can be evaluated using information moments (Fig. 2B). Information moments (M_n in Fig. 2B) are measures related to the shape of a given sensory input that represents the local density of sensory information. As supersamplers acquire high-frequency information, we should expect that the resulting information is spiky: non-uniform information should emerge in the time series as supersamplers visit high information and low information locations in space over time. Another aspect of active sensation involves the origins of different behaviors used to explore their environment. In the context of superperformers, multitude underlying mechanisms that enhance performance may involve a relatively simple set of changes.

2.5 Co-evolutionary Superperformance

Transferring the concept of co-evolutionary superperformance to the world of computational agents, we can model supersamplers and superplanners in terms of pairwise agent interactions. Our pairwise agents consist of an emitter (supersampler) and a receiver (superplanner). We begin with an environment consisting of a large (10^n) possibility space. Emitter agents produce patterns based on rapid sampling of the possibility space. Receivers attempt to identify the emitted patterns based on their capacity to plan possible scenarios. Differences in the possibility space size provide us with different behaviors in the interactions between emitter and receiver. We can also consider the differences in sampling rate, along with differences in planning rate. For relatively small possibility spaces ($n > 5$), the number of patterns emitted are relatively small, and so the possibility space is exhausted rather quickly given the high sampling rate of a supersampler. Environmental sampling of any space will lead to easy detection by a superplanner (receiver). Larger possibility spaces ($n > 10$) provide a means for the superplanner to acquire more information about the emitter over a longer period.

3 Interpretation

In this paper, we consider a cognitive model of superperformance. Our focus is on two types of extreme performance: supersamplers, or cognitive systems that sample the environment at very high frequencies, and superplanners, or cognitive systems that access many alternate scenarios. Taken together, GI and tradeoffs allow us to identify phenomena such as information aliasing and information moments in our cognitive model. Another interesting feature of our cognitive model involves superperformance in the relativistic regime, where an agent can retain both supersampling and superplanning. Despite the variety of animal system examples, our cognitive model is useful for understanding and developing systems for human augmentation. Contemporary issues in Artificial Intelligence model-building are also of relevance. Superplanning resembles attempts to create world models, which attempt to define the world of a computational agent. World models simulate an internal model built through the prior sampling of data. By contrast, supersampling resembles current data-driven approaches to training deep learning models, where more data acquisition is always better. The predictions of superperformance might also provide a framework for application to design principles for augmenting human performance. Our algorithmic architecture utilizes an agent-specific binary fitness function that results in either high sampling rates or large internal model sizes, which are dependent on supersampler success in evading superplanners and the ability of all agents to maintain a low level of aliasing bias. Pseudocode describing a potential algorithm is described in Table 1.

Table 1. Pseudocode describing the relationship between GI, agent spatial representations, and successful evasion and prediction given a simple fitness threshold.

Initialize agents (x emitters, y receivers).

$x(r_s)$, $y(I^m)$ provide distributions for emitter sampling rates and receiver internal model sizes, respectively.

$x_t(S)$, $y_t(S)$ is the isomorphic mapping of each agent's spatial representation.

$x_t(GI)$, $y_t(GI)$ is the Gibsonian Information available to each agent. Assume $x_t(GI)$ $y_t(GI)$.

Fitness threshold crossed when evasion or prediction criteria < 0 . In an evasion-pursuit scenario, r_s and I^m are related in the following manner: increases in I^m drives increases in r_s , whereas decreases in I^m relaxes pressure on but does not decrease r_s .

Successful evasion is where $0t x_t(S) - y_t(S)^* < 0$, where t is the lag between baseline sampling rate y_t and the increased sampling rate of x_t .

Successful prediction is where $0t y_t(S) + I^m - x_t(S) < 0$.

3.1 Superperformance and Gibsonian Information

In the realm of superperformance, GI plays an integral role, but as understood in our definitions of supersampling and superplanning, is not solely responsible for extreme cognitive performance. For example, GI is limited in cases of temporal aliasing. We can understand the effects of temporal aliasing on GI by recalling the visual illusion of a bicycle tire rotating at frequencies greater than the human FFF, where the tire and its spokes appear to flow backwards. This can not only result in false positives amongst supersamplers (high-pass aliasing), but also leads to low-pass aliasing (superplanners) as internal models misclassify ambiguities. In cases where the environment is very rich with affordance-related GI, a strategy of superplanning might out-perform supersampling, as information-dense environments require semantic discrimination that goes beyond determining structure.

3.2 Applications to Human Performance

We can apply the notion of superperformance to human augmentation. As demonstrated in several non-human species, augmented performance is the product of complex physiological traits. One can draw the parallel between superperformance in narrow ecological niches and augmentation for specific tasks in the workplace or in everyday life. As discussed in [17], characterizing physiological state with mathematical tools such as the Yerkes-Dodson curve can provide a behavioral optimum for tasks requiring optimal arousal or attention. Performance optimized around such points could offer enhancement of cognitive state, and when coupled with computer-assisted technologies (Artificial Intelligence or Head-Mounted Displays) might lead to supersampling or even relativistic performance. Utilizing methods to induce and control the direction of adaptation can also lead to augmentation leading to superperformance [18]. Wearable technologies might affect superperformance in a design-dependent manner [19], while monitoring of physiological state along with enhanced situation awareness [20] can also provide the conditions for enhanced performance and superplanning. Coupling an evolutionary algorithm to cognitive dynamics will enable future work by making more explicit comparisons with interspecies diversity in superperformance.


References

1. Ramana, M.V., Kothari, M.: Pursuit-evasion games of high speed evader. *J. Intell. Robot. Syst.* **85**, 293–306 (2017)
2. Schmitt, D., Lemelin, P.: Locomotor mechanics of the slender loris (*Loris tardigradus*). *J. Hum. Evolut.* **47**, 85–94 (2004)
3. Patek, S.N., Baio, J.E., Fisher, B.L., Suarez, A.V.: Multifunctionality and mechanical origins: ballistic jaw propulsion in trap-jaw ants. *PNAS* **103**(34), 12787–12792 (2006)
4. Warrant, E.J.: The remarkable visual capacities of nocturnal insects: vision at the limits with small eyes and tiny brains. *Philosop. Trans. R. Soc. B* **372**, 20160063 (2017)
5. Ketkar, M.D., Sporar, K., Gur, B., Ramos-Traslosheros, G., Seifert, M., Silies, M.: Luminance information is required for the accurate estimation of contrast in rapidly changing visual contexts. *Curr. Biol.* **30**(4), R166–R168 (2020)

6. Miall, R.C.: The flicker fusion frequencies of six laboratory insects, and the response of the compound eye to mains fluorescent 'ripple.' *Physiol. Entomol.* **3**(2), 99–106 (2008)
7. Mikko, J., An, D., Zhuoyi, S., Narendra, S., Diana, R., Jaciuch, D., Anil, D.S., Blanchard, F., Gonzalo, G.P., Hardie, R.C., Jouni, T.: Microsaccadic sampling of moving image information provides *Drosophila* hyperacute vision. *eLife* **6**, e26117 (2017)
8. Wardill, T.J., Knowles, K., Barlow, L., Tapia, G., Nordstrom, K., Olberg, R.M., Gonzalez-Bellido, P.T.: The killer fly hunger games: target size and speed predict decision to pursuit. *Brain Behav. Evolut.* **86**(1), 28–37 (2015)
9. Gibson, J.J.: *The Senses Considered as Perceptual Systems*. George Allen and Unwin, London (1966)
10. Alicea, B., Cialfi, D., Lim, A., Parent, J.: Gibsonian information: a new approach to quantitative information. In: *Biologically Inspired Cognitive Architectures*, p. 1032. *Studies in Computational Intelligence* (2022).
11. Anderson, C.V.: Off like a shot: scaling of ballistic tongue projection reveals extremely high performance in small chameleons. *Sci. Rep.* **6**, 18625 (2016)
12. Healy, K., McNally, L., Ruxton, G.D., Cooper, N., Jackson, A.L.: Metabolic rate and body size are linked with perception of temporal information. *Anim. Behav.* **86**, 685–696 (2013)
13. Skorupski, P., Chittka, L.: Differences in photoreceptor processing speed for chromatic and achromatic vision in the Bumblebee, *Bombus terrestris*. *J. Neurosci.* **30**(11), 3896–3903 (2010)
14. Churchland, A.K., Huang, X., Lisberger, S.G.: Responses of neurons in the medial superior temporal visual area to apparent motion stimuli in macaque monkeys. *J. Neurophysiol.* **97**, 272–282 (2007)
15. Akansu, A.N.: The flash crash: a review. *J. Capit. Mark. Stud.* **1**(1), 2514–4774 (2017)
16. Nelson, M.E., MacIver, M.A.: Sensory acquisition in active sensing systems. *J. Compar. Physiol. A* **192**(6), 573–586 (2006)
17. Schmorow, D.D., Reeves, L.M.: Twenty-first century human-system computing: augmented cognition for improved human performance. *Aviat. Space Environ. Med.* **78**(1), B7–B11 (2007)
18. Alicea, B.: *An Integrative Introduction to Human Augmentation Science*. *arXiv*, 1804.10521 (2018)
19. Raisamo, R., Rakkolainen, I., Majaranta, P., Salminen, K., Rantala, J., Farooq, A.: Human augmentation: past, present and future. *Int. J. Hum. Comput. Stud.* **131**, 131–143 (2019)
20. Cinel, C., Valeriani, D., Poli, R.: Neurotechnologies for human cognitive augmentation: current state of the art and future prospects. *Front. Hum. Neurosci.* **13**, 13 (2019)



ADHD-Related Features of Eye Fixations While Simulated Driving with and Without Working Memory Load: A Pilot Study

Julia Altshuler  and Valeriia Demareva  

Lobachevsky State University, Nizhny Novgorod 603022, Russia
valeriia.demareva@fsn.unn.ru

Abstract. Given the prevalence of ADHD symptoms in the adult population and the association between ADHD and worse driving outcomes, there is a compelling need to thoroughly investigate the mechanisms and manifestations of ADHD during driving. The present study aims to explore the eye movement behavior (based on fixation metrics) in individuals with ADHD symptoms while driving in a simulator. Thirty-three people participated in the study, with 12 of them demonstrating symptoms of ADHD on the ASRS scale (7 with mixed symptoms and 5 with inattention symptoms). The experimental design included 7 min of driving in a simulator, followed by 132 s of driving with a 2-back auditory test, and an additional 5 min of normal driving while recording eye movements using Pupil Invisible glasses. The results of our study indicate that working memory performance in simulator driving is significantly worse in people with mixed ADHD symptoms. Furthermore, the analysis revealed that individuals with inattention were more likely to exhibit less variable eye fixation behavior, while those with mixed ADHD symptoms may be characterized by ‘hyperactive’ eye behavior with situational hyperfocusing.

Keywords: ADHD · Inattention · Hyperactivity · Driving · Eye movements · Fixations · Working memory

1 Introduction

Attention Deficit Hyperactivity Disorder (ADHD) is a neuropsychiatric developmental disorder characterized by symptoms of inattention, impulsivity, and hyperactivity, which can significantly affect various aspects of a person’s life. It was noted that in 2020, the prevalence of persistent adult ADHD was 2.58%, and that of symptomatic adult ADHD was 6.76% [1].

Adults with ADHD symptoms exhibit challenges in initiating tasks, inconsistent focus on details, struggles in self-organization and task prioritization, and limited perseverance in sustained mental effort [2, 3]. Within the visual processing domain, adults with ADHD symptoms have also demonstrated decreased visual short-term memory capacity, a higher threshold of conscious perception, and slower visual processing speed [4]. Additionally, impaired selective and divided attention in adults with ADHD has

been reported [5], as well as deficits in working memory [6]. Another characteristic of ADHD in adults is hyperfocusing, a clinical phenomenon where individuals become intensely fixated on a particular task and struggle to shift their attention to other subjects, particularly when those subjects align with their interests [7]. Hyperfocusing is believed to stem from attention disorders, where individuals with ADHD experience challenges in maintaining focus, sustaining attention, and shifting their attention focus [8]. Consequently, hyperfocusing may contribute to impaired divided attention in adults with ADHD symptoms.

Such features of adults with ADHD symptoms can significantly impact their driving behavior and outcomes. Studies indicate that adolescents and adults with ADHD experience distinct and unfavorable driving outcomes compared to those without the condition. Notably, individuals with ADHD exhibit a higher frequency of accidents and speeding violations, leading to fewer safe driving skills and more inattentive and impatient driving errors [9, 10].

Therefore, ADHD represents a significant risk factor for motor vehicle accidents and dangerous driving. However, there is limited and mixed evidence regarding the specific mechanisms by which ADHD affects driving risks. A review of cognitive abilities deficits has shown a link between inattention, especially visual inattention, and unfavorable driving outcomes. Slow information processing, easy distraction, and visual memory problems have also been associated with driving challenges.

In our pilot study, we aimed to identify ADHD-related features of eye movement behavior during simulated driving with and without working memory load. Given that individuals with ADHD have demonstrated specific driving characteristics on a low-demand highway [10], this scenario served as the basis for the current study.

Our hypotheses were as follows:

1. Working memory performance would be worse in people with ADHD symptoms.
2. Different ADHD symptoms (inattention and hyperactivity) would cause distinct patterns of eye movement behavior.

2 Materials and Methods

2.1 Study Sample

33 individuals participated in the study, including 10 men. The average age of the participants was 28. Among the 33 participants, 12 demonstrated symptoms of ADHD on the ASRS scale, constituting the ‘ADHD group.’ Within the ADHD group, seven participants demonstrated ADHD symptoms solely related to inattention, forming the ‘inattention ADHD group,’ while the other five participants showed symptoms related to both inattention and hyperactivity, constituting the ‘mixed ADHD group.’ None of the participants exhibited severe hyperactivity symptoms in isolation.

2.2 Methods

To determine the presence or absence of ADHD, a questionnaire based on the ASRS scale was utilized [11]. The ASRS is an 18-item measure grounded in DSM-V criteria

and exhibits high validity. The questions were formulated to assess symptomatology in adults within the context of their life and work. The scales for inattention, hyperactivity, and the total ADHD score were assessed. The sample was classified based on two criteria—‘ADHD presence’ and ‘ADHD type’ (see Table 1).

Table 1. Rules for sample classifications according to two different classification bases.

Basis	Groups	Rules
ADHD presence	‘Normative group’	Scores for inattention and hyperactivity are lower than 5
	‘ADHD group’	Scores for inattention or hyperactivity are higher than 5
ADHD type	‘Inattention ADHD group’	Score for inattention is higher than 5 and score for hyperactivity is lower than 5
	‘Hyperactivity ADHD group’	Score for hyperactivity is higher than 5 and score for inattention is lower than 5
	‘Mixed ADHD group’	Scores for inattention and hyperactivity are higher than 5

Pupil Invisible mobile eye-tracker glasses were used to record eye movements.

To assess cognitive load during driving on the simulator, the auditory working memory test 2-back was utilized. The test was conducted using Brain Workshop software (<https://brainworkshop.sourceforge.net/>). During the test, the subject listened to a sequence of letters from the Latin alphabet for a duration of 132 s. If the current letter matched the one that was one letter back in the sequence, the subject had to say ‘Yes,’ and the experimenter pressed the L button on the keyboard accordingly. At the end of the experiment, the percentage of correct answers given by the subject was provided by the software and recorded in the protocol.

2.3 Experimental Design

After completing the ASRS questionnaire, each participant engaged in a 15-min driving session on a low-demand highway in the ‘City Car Driving’ game, utilizing the Space Rift 2 DoF driving simulator. During this driving session, eye movements were recorded using Pupil Invisible glasses. Following 7 min of normal driving, a 2-back auditory test was administered to load working memory. Subsequently, each participant continued driving for an additional 5 min. The experiment’s schematic is illustrated in Fig. 1.

2.4 Data Analysis

Fixation metrics for different contexts were analyzed:

- 2–4 min of Normal driving 1 (‘driving 1’),
- 132 s of driving with auditory 2-back test (‘driving + nback’),

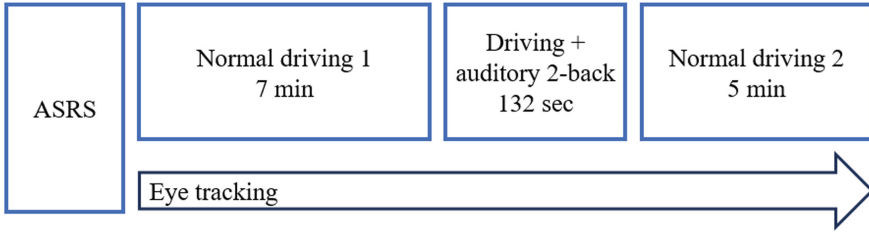


Fig. 1. Experimental design.

- 2–4 min of Normal driving 2 ('driving 2').

The fixation data were filtered so that all fixations longer than 5 s were removed. The following fixation metrics were calculated—Table 2.

Table 2. Fixation metrics description.

No	Metrics	Description
1	mean_dur	Mean fixation duration
2	count	Fixation count
3	median_dur	Median fixation duration
4	std_dur	Standard deviation of fixation duration
5	min_dur	Minimal fixation duration
6	max_dur	Maximal fixation duration

Also, different types of fixations were counted. According to [12], the four types of fixations were counted within each context:

- 'very short' with a duration of 0–90 ms,
- 'express' with a duration of 90–150 ms,
- 'cognitive' with a duration of 150–900 ms,
- 'long' with a duration of more than 900 ms.

Statistical analysis was performed using the 'scipy.stats' module in Jupyter Notebook (Python). The Mann-Whitney U test was used to assess differences in fixation metrics between different groups within each context. Spearman's rank correlation coefficient was used to calculate the correlations between fixation metrics and ASRS scores.

3 Results

3.1 Comparison of Fixation Metrics and Working Memory Performance in the Groups with and Without ADHD Symptoms

The 'ADHD group' demonstrated a tendency of lower auditory 2-back performance compared to the 'normative group' ($U = 79.5$; $p = 0.085$). Significant differences and tendencies found for fixation metrics are presented in Table 3.

Table 3. Fixation metrics in different contexts for ‘ADHD group’ and ‘normative group’ (U —the value of Mann-Whitney U test, p — p -value).

Context	Metrics	ADHD group	Normative group	U	p
		$N = 12$	$N = 21$		
Driving 1	std_dur	853.5	1043.1	74	0.054
Driving + nback	max_dur	3691.6	4336.8	65	0.024
Driving 2	count	175.0	131.7	71.5	0.043
	std_dur	756.6	972.8	78	0.075
	max_dur	3957.2	4520.4	64	0.021

The ‘ADHD group’ demonstrated a tendency for lower std_dur values in ‘driving 1’ ($U = 74$; $p = 0.054$) and in ‘driving 2’ ($U = 78$; $p = 0.075$). Also, within this group, max_dur was lower in ‘driving + nback’ ($U = 65$; $p = 0.024$) and in ‘driving 2’ ($U = 64$; $p = 0.021$). In ‘driving + nback’ the ‘ADHD group’ also made less fixations than the ‘normative group’ ($U = 71.5$; $p = 0.043$).

3.2 Comparison of Fixation Metrics and Working Memory Performance in the Groups with Different ADHD Types and Normative Group

The ‘mixed ADHD group’ demonstrated a tendency of lower auditory 2-back performance than the ‘normative group’ ($U = 42$; $p = 0.099$). Significant differences and tendencies found for fixation metrics are presented in Table 4.

The ‘inattention ADHD group’ demonstrated lower values for std_dur ($U = 18$; $p = 0.023$) in driving 1, std_dur ($U = 23$; $p = 0.057$) and max_dur ($U = 19$; $p = 0.028$) in driving + nback, std_dur ($U = 26$; $p = 0.091$) and max_dur ($U = 20$; $p = 0.034$) in driving 2 than the ‘normative group’.

The ‘mixed ADHD group’ demonstrated lower values for mean_dur ($U = 37$; $p = 0.055$) and median_dur ($U = 35$; $p = 0.042$) than the ‘normative group’ in driving + nback.

The only difference tendency between ‘inattention ADHD group’ and ‘mixed ADHD group’ was observed for min_dur within driving + nback ($U = 6$; $p = 0.073$).

3.3 Correlation Analysis of Fixation Metrics, Working Memory Performance, and ADHD Scores Within ‘ADHD Group’

Since a lot of participants from ‘normative group’ had the score of zero for ASRS scales, correlation analysis was performed only for the ‘ADHD group’. Neither fixation metrics nor ASRS scores correlated with auditory 2-back performance. Significant correlations between fixation metrics and ASRS scores are presented in Table 5.

The data in Table 5 reveal, that inattention score positively correlated with mean_dur ($R = 0.61$, $p < 0.05$) and median_dur ($R = 0.66$; $p < 0.05$) in driving 1, and with

Table 4. Fixation metrics in different contexts for ‘inattention ADHD group’, ‘mixed ADHD group’, and ‘normative group’ (*U*—the value of Mann-Whitney *U* test, *p*—*p*-value).

Context	Metrics	Inattention ADHD group N = 5	Mixed ADHD group N = 7	Normative group N = 21	U	p
driving 1	std_dur	785.8		1043.1	18	0.023
driving+nback	std_dur	820.5		1040.8	23	0.057
	max_dur	3621.8		4336.8	19	0.028
	mean_dur		999.1	1430.1	37	0.055
	median_dur		688.3	1223.9	35	0.042
	min_dur	99.4	65.1		6	0.073
driving 2	std_dur	729.0		972.8	26	0.091
	max_dur	3767.0		4520.4	20	0.034

Table 5. Spearman’s rank correlation coefficients between fixation metrics and ASRS scores (* *p* < 0.05; ** *p* < 0.01).

Context	Metrics	Innatention score	Hyperactivity score	ADHD score
Driving 1	mean_dur	0.61*	− 0.78**	− 0.72**
	median_dur	0.66*		
Driving + nback	min_dur			
Driving 2	count	− 0.61*		
	median_dur	0.65*		

median_dur (*R* = 0.65; *p* < 0.05) in driving 2. Also, inattention score negatively correlated with count (*R* = − 0.61; *p* < 0.05) in driving 2. Min_dur negatively correlated with hyperactivity score (*R* = − 0.78; *p* < 0.01) and ADHD score (*R* = − 0.72; *p* < 0.01) in driving + nback.

3.4 Comparison of Fixation Types Count in the Groups with and Without ADHD Symptoms

Analysis revealed that only one significant difference was observed while comparing different fixation types count in ‘ADHD group’ and ‘normative group’—see Table 6.

The data in Table 6 indicate that in driving 2, ‘ADHD group’ made more very short fixations as compared to ‘normative group’ (*U* = 62; *p* = 0.017). No other significant differences or tendencies were found.

Table 6. Fixation types count in different contexts for ‘ADHD group’ and ‘normative group’ (U —the value of Mann-Whitney U test, p — p -value).

Context	Fixation type	ADHD group	Normative group	U	p
		$N = 12$	$N = 21$		
Driving 2	Very short	113.1	67.5	62	0.017

3.5 Comparison of Fixation Types Count in the Groups with Different ADHD Types and Normative Group

Different fixation types appeared to be sensitive to ADHD manifestations in different contexts—see Table 7.

Table 7. Fixation types count in different contexts for ‘inattention ADHD group’, ‘mixed ADHD group’, and ‘normative group’ (U —the value of Mann-Whitney U test, p — p -value).

Context	Fixation type	Inattention ADHD group	Mixed ADHD group	Normative group	U	p
		$N = 5$	$N = 7$	$N = 21$		
driving 1	cognitive	63.2		52.5	25	0.079
driving+nback	express	2.8	7.0		6	0.071
	long	1.4	4.4		5	0.050
	long		4.4	3.5	36	0.046
driving 2	very short		116.4	67.5	37	0.056
	very short	108.4		67.5	25	0.079

In driving 1, only one tendency was found. ‘Inattention ADHD group’ made more cognitive fixations than ‘normative group’ ($U = 25$; $p = 0.079$). In driving + nback, ‘mixed ADHD group’ performed a tendency to make more express fixations than ‘inattention ADHD group’ ($U = 6$; $p = 0.071$); ‘mixed ADHD group’ made more long fixations than ‘inattention ADHD group’ ($U = 5$; $p = 0.050$) and ‘normative group’ ($U = 36$; $p = 0.046$). In driving 2, ‘normative group’ demonstrated a tendency to make less very short fixations than ‘inattention ADHD group’ ($U = 25$; $p = 0.079$) and ‘mixed ADHD group’ ($U = 37$; $p = 0.056$).

4 Discussion

The current study aimed to discover the eye movement behavior (based on fixation metrics statistics) in people with ADHD symptoms while driving in a simulator. The experimental design included normal driving, driving with working memory load, and normal driving after the load on a low demand highway.

The results revealed a tendency for worse working memory performance in people with ADHD symptoms. This finding is supported by previous research on poorer working memory in individuals with ADHD [6]. It is probable that ADHD symptoms hindered the participants' ability to divide their attention between the driving task and the n-back task, which aligns with the finding that ADHD leads to impaired selective and divided attention in adults with ADHD [5]. More precisely, in our experiment, only the group with mixed (i.e., inattention and hyperactivity) symptoms demonstrated this tendency. This is a novel finding that has not been reported yet. Further research is needed to investigate the features of working memory impairment in adults with different types of ADHD. Therefore, it is possible that only inattention alone does not lead to working memory deficits; the hyperactivity component may also be important in this case.

During driving with working memory load, as well as in driving after the load, maximal fixation duration was lower in people with ADHD symptoms. This suggests that individuals with ADHD symptoms, under external load, might have difficulties concentrating for extended periods on visual objects, which is consistent with previous findings on impaired selective attention [5] and challenges in maintaining focus [8] in adults with ADHD. However, this pattern was not observed during the first few minutes of driving. We can assume that working memory load (and possibly other types of external load) influences eye movement behavior in people with ADHD symptoms. During the initial minutes of driving, people with ADHD symptoms were characterized only by a decrease in the standard deviation of fixation duration. This pattern was also found in driving after the load. Therefore, people with ADHD symptoms, in general, are likely to have less-variable fixations, which may also be linked to impaired selective and divided attention in adults with ADHD [5].

When comparing fixation metrics considering ADHD type, it was found that the inattention type was more likely to have more differences with the normative group. In both driving tasks without load, they demonstrated lower standard deviation of fixation duration than the normative group. Therefore, we may assume that having fewer variabilities in eye fixations is a characteristic of the inattention ADHD type. As for the mixed type, they demonstrated lower mean and median fixation durations than the normative group, and lower minimal fixation durations than the inattention group only in the driving with working memory load task. We should note that the only difference between the two ADHD types was observed in the driving with the load task. Lower minimal fixation durations and lower mean/median durations may indicate 'hyperactive' eye behavior in people with the mixed ADHD type. This finding is also supported by the correlation analysis results: minimal fixation duration was negatively correlated with the score for hyperactivity.

The analysis of fixation types revealed that people with mixed ADHD symptoms made more express and more long fixations than people with inattention ADHD symptoms in the driving with the working memory load task. This fact also confirms our previous statement about the 'hyperactive' eye behavior in people with the mixed ADHD type. However, the result about more long fixations also allows us to make a new assumption. Probably extra cognitive load triggers hyperfocusing in people with the mixed ADHD type, resulting in difficulties shifting their attention focus [8]. This assumption is also

supported by the result about the lower working memory performance tendency observed only in people with mixed ADHD symptoms.

Overall, our study demonstrates the necessity of studying eye movement behavior in people with different ADHD symptoms.

5 Conclusions

The results of our study indicated that working memory performance in simulated driving is worse in people with mixed ADHD symptoms. The analysis revealed that individuals with inattention were more likely to exhibit less variable eye movement behavior, while those with mixed ADHD symptoms may be characterized by 'hyperactive' eye behavior with situational hyperfocusing.

References

1. Song, P., Zha, M., Yang, Q., Zhang, Y., Li, X., Rudan, I.: The prevalence of adult attention-deficit hyperactivity disorder: a global systematic review and meta-analysis. *J. Glob. Health* **11**, 04009 (2021). <https://doi.org/10.7189/jogh.11.04009>
2. Biederman, J.: Attention-deficit/hyperactivity disorder: a life-span perspective. *J. Clin. Psych.* **59**(S7), 4–16 (1998)
3. Silver, L.B.: Attention-deficit/hyperactivity disorder in adult life. *Child Adolesc. Psych. Clin. N. Am.* **9**(3), 511–523 (2000). [https://doi.org/10.1016/S1056-4993\(18\)30104-4](https://doi.org/10.1016/S1056-4993(18)30104-4)
4. Low, A.M., et al.: Visual attention in adults with attention-deficit/hyperactivity disorder before and after stimulant treatment. *Psychol. Med.* **49**(15), 2617–2625 (2019). <https://doi.org/10.1017/S0033291718003628>
5. Tucha, L., et al.: Sustained attention in adult ADHD: time-on-task effects of various measures of attention. *J. Neural Transm.* **124**(1), 39–53 (2015). <https://doi.org/10.1007/s00702-015-1426-0>
6. Dowson, J., et al.: Impaired spatial working memory in adults with attention-deficit/hyperactivity disorder: comparisons with performance in adults with borderline personality disorder and in control subjects. *Acta Psychiatr. Scand.* **110**, 45–54 (2004). <https://doi.org/10.1111/j.1600-0447.2004.00292.x>
7. Conner, M.L.: Attention deficit disorder in children and adults: strategies for experiential educators. In: *Experiential Education: A Critical Resource for the 21st Century, Proceedings Manual of the Annual International Conference of the Association for Experiential Education*, vol. 22; pp. 177–182. Austin (1994)
8. Ozel-Kizil, E.T., et al.: Hyperfocusing as a dimension of adult attention deficit hyperactivity disorder. *Res. Develop. Disabil.* **59**, 351–358 (2016). <https://doi.org/10.1016/j.ridd.2016.09.016>
9. Fuermaier, A.B., et al.: Driving and attention deficit hyperactivity disorder. *J. Neur. Transm.* **124**(S1), 55–67 (2017). <https://doi.org/10.1007/s00702-015-1465-6>
10. Randell, N.J.S., Charlton, S.G., Starkey, N.J.: Driving with ADHD: performance effects and environment demand in traffic. *J. Attent. Disord.* **24**(11), 1570–1580 (2010). <https://doi.org/10.1177/10870547106658126>
11. Kessler, R., et al.: The World Health Organization adult ADHD self-report scale (ASRS): a short screening scale for use in the general population. *Psychol. Med.* **35**(2), 245–256 (2005). <https://doi.org/10.1017/S0033291704002892>
12. Galley, N., Betz, D., Biniössek, C.: Fixation durations: why are they so highly variable? In: Heinen, T. (ed.) *Advances in Visual Perception Research*, pp. 83–106. Nova Science, NY (2015)



A Biologically Inspired Approach to Protecting and Verifying the Authenticity of Important Documents

Alexander M. Alyushin , Victor M. Alyushin , Sergey V. Dvoryankin ,
and Lyubov V. Kolobashkina  

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute),
Kashirskoe Highway 31, Moscow 115409, Russia
LVKolobashkina@mephi.ru

Abstract. An approach to the protection and verification of important documents, primarily on paper based on the so-called biosignature (BS), is presented. BS is a further development of document protection technology with the help of speech signature (SS). SS is a fragment of an image embedded in a document. The SS image contains a spectrogram of the speech message of the author of the document, which is used to verify its contextual content, as well as to identify the author's identity. The main disadvantage of SS was revealed, which consists in the limited possibilities for transmitting the author's bioparameters, which make it possible to identify his functional and psycho-emotional state. Fixing the state of the author at the time of its preparation or signing makes it possible to identify cases of obtaining a signature under duress, as well as in case of inappropriate behavior of the author, including the author being under the influence of drugs or alcohol, which is the basis for recognizing the document as falsified. Two possible options for the transmission of biometric information as part of the BS are analyzed. It is shown that the transfer of time dependences of the dynamics of changes in bioparameters at the time of signing the document is the most preferable in terms of noise immunity from random pixel noise due to the applied printing technology for manufacturing the document.

Keywords: Biosignature · Bioparameters · Document

1 Introduction

One of the modern information technologies for protecting important financial, legal, or administrative documents from unauthorized copying, reproduction, and falsification is the speech signature (SS) technology [1]. The essence of this technology lies in the use of a special graphic insert as part of the protected document, which encodes a fragment of the voice message (VM) of the author of the document in the form of a spectrogram. A fragment of the author's VM captures the key, most significant part of the information of the protected document, for example, the amount and terms of the contract, the names of sponsors, bank account numbers. The unique voice characteristics of the author's VM

in this case are a tool to protect the document. To record the original SS of the author, one of the acoustic standards is used, for example, MP3, AAC, wav, Flac, Alac, DSD. For the SS synthesis, the spectrogram of the original VM is used, which makes it possible to significantly reduce the area of the SS graphic image. To obtain a spectrogram, the FFT is usually used. This approach allows you to save the main individual frequency characteristics of the author's voice [2]. To convert a speech signal to a fragment of the SS image, specialized software tools are used, which can be installed on smartphones, laptops and other wearable devices. A graphic insert containing a spectrogram can be printed on a printer as part of a protected document, made in the form of inconspicuous watermarks, or an indistinguishable polygraphic ornament as part of graphic elements decorating the document.

A user receiving a document protected by SS has the opportunity, with the help of software and hardware decoding tools [3], to obtain not only its contextual content, but also the author's unique VM. Analysis of the frequency characteristics of the VM makes it possible to identify the personality of the author. This circumstance allows us to consider the SS as an additional degree of protection for the authenticity of the document on a par with the letterhead, or with the personal signature of the author.

Unfortunately, identification of the personality of the author by his SS does not provide the necessary degree of document protection, since it does not allow to identify situations of signing documents under duress, in an inadequate state, in a state of stress, alcohol or drug intoxication.

The aim of the work is to study a biologically inspired approach to the protection and verification of the authenticity of important documents using biosignature (BS).

2 The Essence of the Biologically Inspired Approach

The essence of the considered biologically inspired approach to the protection and verification of the authenticity of important documents is as follows:

- the so-called BS is used, containing information both about the VM of the author of the document, and about the totality of his biometric parameters registered at the time of drawing up/signing the document;
- parameters that fully characterize the current functional and psycho-emotional state of the author are used as biometric parameters of the author;
- the biometric parameters selected in this way for the calm working state of the author are periodically updated and stored in a closed database with limited access.

The biological inspiration of the approach under consideration is due to the fact that many living beings base the recognition of the truth of the intentions of another being on the analysis of a set of bioparameters that are informative for them. An illustrative example, in this regard, for example, can be the communication of a person and a dog. Almost regardless of what a stranger says, a dog with a high degree of certainty recognizes his state and intentions by the intonation of his voice, manner of movement, smell, the nature of eye movement and facial expressions [3–5].

In works [6, 7], it was shown that for reliable recognition of the current functional and psycho-emotional state of a person, it is necessary to analyze the dynamics of change

in comparison with a calm working state of a set of bioparameters that characterize the work of the cardiovascular system, the respiratory system and the nervous system. For a systematic assessment of the functional and psycho-emotional state, the works [6, 7] propose to use the integral assessment G , the instantaneous values of which characterize the dynamics of a person's change over time. To determine the personalized G value, normalized deviations of human bioparameters from their values for a calm working state are analyzed, taking into account the individual significance of these bioparameters:

$$G = \sum_{i=1}^N A_i (P_i - P_{0i}) / P_{0i}, \quad (1)$$

where N is the number of analyzed bioparameters; P_i is the current value of the i -th bioparameter; P_{0i} is the value of the i -th bioparameter for the normal working state; A_i is the individual significance of the parameter P_i for determining the characteristic G .

The values of A_i and P_{0i} are contained in the personal data base. The value of the dimensionless value G obtained in this way is the basis for assessing the level of the functional and psycho-emotional state: sleepy, relaxed state ($G < G_{level1}$); normal working condition ($G_{level1} \leq G < G_{level2}$); stressed state ($G_{level2} \leq G < G_{level3}$); highly stressed state ($G_{level3} \leq G < G_{level4}$); dangerous state of physical and emotional overstrain ($G_{level4} \leq G < G_{level5}$). The threshold values $G_{level1} - G_{level5}$ are purely individual and their periodically updated values are in the personal data base.

In practice, to register bioparameters that characterize a person's condition, such as heart rate, breathing parameters, blood pressure, hand tremor, pupillary reaction, eye movement patterns, temperature of facial areas, it is advisable to use modern wearable gadgets. Acoustic sensors built into them, as well as acceleration sensors, video cameras of the visible and infrared spectra of optical radiation, make it possible, with the help of special software processing, to trace the dynamics of changes in the above bioparameters.

The study considered two main ways of transmitting the mentioned above set of bioparameters as part of the BS:

- direct transmission as part of the BS of the time dependences of the indicated bioparameters;
- transmission as part of the BS of the dynamics of changes in the values of the function G .

The BS of the author of the document can also be represented in a graphical form, shown in Fig. 1. In the general case, the graphic image of the BS may contain the following information areas:

- the beginning of the SS image fragment (SSIB—SS Image Beginning) and the end of the SS image fragment (SSIE—SS Image End) of the author, length SSI;
- beginning (B/FIB—Bioparameters/Function Image Beginning) and ending (B/FIE—Bioparameters/Function Image End) of an image fragment of the B/FI area of the BS, containing information on the dynamics of changes in the set of bioparameters (B), or function G (F).

The BS integrates the SS spectrogram in its composition. Regions 1 and 2 in Fig. 1 demonstrate the harmonic structure of the spectrogram, which is characteristic of VM

vowels. Due to the rather slow change in the value of bioparameters, various transformation algorithms can be used to transfer their time dependences into an image fragment. For example, transmission in the form of a sequence of pixels of different intensities, or staining colors.

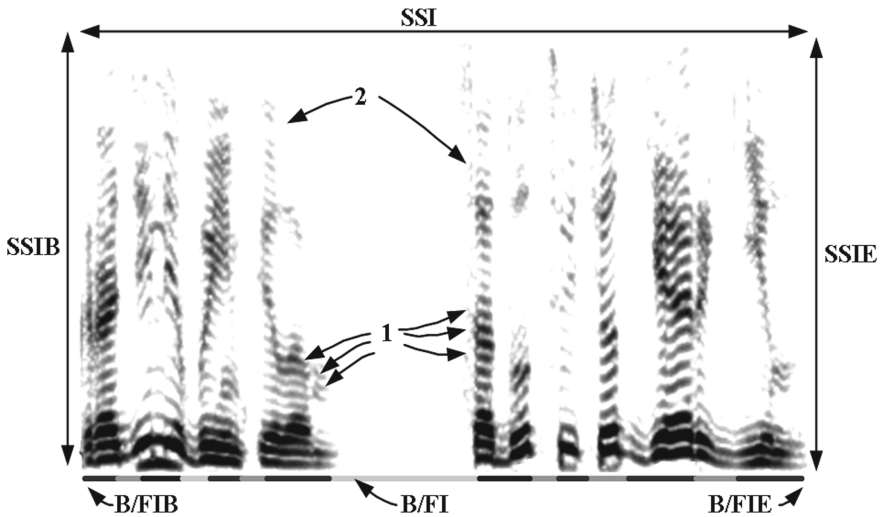


Fig. 1. An example of a BS image.

3 Methodology for Determining the Effectiveness of the Approach

To determine the effectiveness of the considered biologically inspired approach, the following method was used. The students who passed the exams were considered as test subjects. At the same time, almost all students were in a different state. In a highly excited state, as a rule, there were students who either had a poor knowledge of the subject, or had a poor practice of self-regulation. A calm state was characteristic of excellent students who were fluent in the material. After the exam, all students, test participants, registered the previously considered set of bioparameters, which allowed determining the current functional and psycho-emotional state. The registered set of bioparameters was used to determine the values of the G function [6, 7]. All test subjects were asked to form images of SS and BS to protect a paper document. The latter were applied to the same document using a laser printer.

The SS and BS images were read from a paper document and their content was decoded using the developed specialized software. The transfer of information about the current state was carried out in two ways—by transferring the initial time dependences of bioparameters and in the form of a time dependence of the integral function G . In the case of transferring bioparameters as part of the BS, the latter were used to determine the values of the G function after decoding the BS. The values of the G function obtained in this way were compared with the values of the function obtained on the basis of the processing of a set of bioparameters recorded immediately after passing the exams.

4 Discussion of the Results

More than 80 students of various courses of study took part in the testing. Table 1 presents the results obtained. As an assessment of the quality of information transmission as part of SS and BS, which allows assessing the functional and psycho-emotional state of a student (potentially the author of an important document), the relative difference between the values of G functions determined on the basis of SS and BS decoding and obtained after the exam was used:

- E_{SS} —an error in the transfer of biometric information when using SS;
- E_{BS} —an error in the transmission of biometric information when using a BS with a set of bioparameters;
- E_{BSG} —an error in the transmission of biometric information when using a BS with the G function.

Table 1. Accuracy of biometric information transmission.

Condition after the exam	E_{SS} , %	E_{BS} , %	E_{BSG} , %
Relaxed state	4	2	2
Normal working condition	5	3	3
Stressed state	10	5	7
Highly stressed state	20	7	10

The assignment of a student to one of the state categories presented in Table 1 was carried out by assigning the obtained value of the function G to one of the possible ranges of its change [6–8]. The obtained results indicate that the most reliable state of the author of the document can be transmitted using the BS, which contains the original biometric parameters. The transmission of the values of the function G as part of the BS leads to a slight increase in the error. This effect is most likely due to the quality of printing of the document with the BS. Random pixel noise that occurs when transmitting an image of the temporal dependences of a set of bioparameters is averaged with an increase in the number of such dependences. When only the values of the G function are transmitted, the errors associated with the pixel noise of the BS image are not averaged. The use of SS without transmitting biometric parameters characterizing the state of the author of the document leads to more significant distortions.

5 Conclusion

The conducted research allowed to draw the following conclusions.

1. The implementation of the biologically inspired approach makes it possible to increase the level of validity of important documents by identifying cases where the author's signature was obtained under duress, or the author was in an inadequate state.

2. The option of protecting documents using BS should be considered a further development of the method of protecting important documents using SS, which makes it possible to increase the reliability of the transfer of biometric parameters necessary for recognizing the state of the author.
3. One of the directions for further development of the BS technology should be considered the creation and transfer of the behavioral model of the author [6].






Acknowledgments. The work was carried out within the framework of project No. 40469-06/23-K with the support of the Ministry of digital development of Russia.

References

1. Alyushin, A.M., Dvoryankin, S.V.: The use of speech technology to protect the document turnover. *IT Sec.* **24**(2), 6–15 (2017). F: Article title. *Journal* **2**(5), 99–110 (2016)
2. Alyushin, A.M.: Biologically inspired physical model of the vocal tract for the tasks of recognition of the current psycho-emotional state of a person. *Adv. Intell. Syst. Comput.* **948**, 15–21 (2020)
3. Albuquerque, N., Guo, K., Wilkinson, A., Savalli, C., Otta, E., Mills, D.: Dogs recognize dog and human emotions. *Biol. Lett.* **13**, 883 (2016)
4. Nagasawa, M., et al.: Oxytocin-gaze positive loop and the coevolution of human-dog bonds. *Science* **348**, 333–336 (2015)
5. Adachi, I., Kuwahata, H., Fujita, K.: Dogs recall their owner's face upon hearing the owner's voice. *Anim. Cogn.* **10**, 17–21 (2007)
6. Alyushin, M.V., Kolobashkina, L.V., Golov, P.V., Nikishov, K.S.: Adaptive behavioral model of the electricity object management operator for intelligent current personnel condition monitoring systems. *Mech. Mach. Sci.* **80**, 319–327 (2020)
7. Alyushin, M.V., Kolobashkina, L.V.: Laboratory approbation of a new visualization form of hazardous objects control operator current psycho-emotional and functional state. *Sci. Vis.* **10**(2), 70–83 (2018)
8. Alyushin, M.V., Alyushin, A.M., Kolobashkina, L.V.: Human face thermal images library for laboratory studies of the algorithms efficiency for bioinformation processing. In: *Proceedings of the 11th IEEE International Conference on Application of Information and Communication Technologies on Proceedings, AICT 2017, Moscow, IEEE, Russia* (2019)



Comparison of Biologically Inspired and Modeling Approaches for Predicting Possible Condition Change in Critical Occupational Workers

Mikhail V. Alyushin¹ , Lyubov V. Kolobashkina¹  , Vladislav D. Bitney² ,
and Andrey V. Okhlopkov² 

¹ National Research Nuclear University MEPhI (Moscow Engineering Physics Institute),
Kashirskoe Highway, 31, 115409 Moscow, Russia

LVKolobashkina@mephi.ru

² PJSC “Mosenergo”, Prospect Vernadsky, 101, Building 3, 119526 Moscow, Russia

Abstract. The aim of the work is to create effective methodological and technical tools for monitoring and predicting possible changes in the state of personnel in critical professions in real time. For this, modern technologies for registering the bioparameters of workers are analyzed, taking into account their physical activity. The dynamics of changes in the bioparameters of workers during a work shift, taking into account the accumulated individual data, is the basis for predicting the most probable changes in their condition. The study compares two approaches to forecasting. The first approach involves modeling based on the implementation of an iterative computational process. At the same time, a digital behavioral model is used that allows adaptation to the individual biometric data of employees. The second approach is based on a biologically inspired computational structure. To adjust the structure, the available data on changes in the bioparameters of workers during the work shift, depending on the level of work intensity, are used. Comparison of the efficiency of the two approaches was carried out on the basis of experimental data obtained during the trial operation of a sample of an automated information-measuring system for monitoring and forecasting in real production conditions at combined heat and power plant (CHPP) number 26 at Moscow. The second approach showed a higher prediction accuracy.

Keywords: Modeling · Prediction · Biometrics

1 Introduction

The damage from the negative impact of the human factor (HF) is quite difficult to calculate, especially when it is indirectly manifested. Usually they refer to the purely technical causes of accidents [1], which in many cases, in fact, are the result of the negative impact of the HF. The creation of methodological and technical tools that allow solving the tasks of monitoring the functional and psycho-emotional state (FPES) of

workers in critical professions directly in the process of their production activities, as well as the tasks of predicting its possible change, is an urgent problem at the present time. A typical example in this regard is, for example, the machinists of power units of thermal power plants and nuclear power plants, who carry out their professional activities in intense production conditions.

Currently, the control of many of the above aspects is carried out by the personnel services of enterprises and organizations, both when hiring, and in the process of scheduled examinations and testing of personnel, as well as during pre-shift control. However, during the entire work shift, especially at night, monitoring of the FPES of workers is practically not carried out, which is due, among other things, to the lack of reliable and easy-to-use technical means. In particular, the registration and accounting of the level of physical activity (LPA) of employees is not carried out, which, as a rule, distorts the values of the recorded bioparameters.

The aim of the work is to create methodological and technical means in the form of automated information-measuring system (AIMS) for monitoring and forecasting, which provide effective management of the reliability of the HF, including at high LPA.

2 Methodology for Assessing the Condition of an Employee

The methodological basis for the creation of AIMS is the developed technology for recording and analyzing a set of bioparameters that characterize the work of the main systems of the human body—respiration, cardiovascular and nervous systems [2–4]. Their analysis makes it possible to identify with a high degree of certainty states of stress, fatigue, severe nervous tension, weakness, relaxation, as well as drowsiness and sleep. The technology under consideration is focused on the analysis of the temporal dynamics of changes in bioparameters. To take into account the individual age, gender and other characteristics of the worker, the technology involves comparing the values of bioparameters measured during the work shift with their values measured for his calm state. To assess the current state of the employee in this case, for example, a linear model G can be used that takes into account changes in all recorded and analyzed bioparameters, taking into account their personal significance:

$$G = \sum_{i=1}^N A_i (P_i - P_{0i}) / P_{0i}, \quad (1)$$

where N is the number of recorded bioparameters;

P_i is the value of the i -th measured bioparameter;

P_{0i} is the value of the i -th bioparameter for the normal working state;

A_i is the personal significance of the parameter P_i .

The high reliability of the FPES estimates obtained using the model under consideration is due to the possibility of analyzing a wide range of bioparameters [5, 6] that make it possible to characterize the current state of the main body systems. For example, the following bioparameters are used as the main ones for characterizing the state of the cardiovascular system: P_1 —heart rate (HR); P_2 —blood pressure (BP); P_3 —heart rate variability (HRV). To describe the respiratory system, bioparameters are considered:

P_4 —respiratory rate; P_5 —depth of breathing; P_6 —balance “up/down” breathing. To assess the state of the nervous system can be used: P_7 —the time of a simple sensorimotor reaction; P_8 —pupillary reaction; P_9 —reactions similar to galvanic skin; P_{10} —change in temperature of facial areas.

The advantage of the considered model (1) in comparison with the known ones is the ability to take into account the dynamics of changes in heterogeneous bioparameters, measured using classical technologies based on the use of contact sensors, as well as using actively developing remote non-contact technologies, when assessing the AZUV.

The values of A_i and P_{0i} are contained in the personal data base (PDB) of AIMS. The value of the dimensionless value G obtained in this way is the basis for assessing the level of FPES (Table 1). The threshold values $G_{level1} - G_{level5}$ indicated in Table 1 are purely individual and their periodically updated values are in the AIMS BPD.

Table 1. Possible ranges of G change.

G range	Evaluation of the level of FPES
$G < G_{level1}$	Sleepy, relaxed state
$G_{level1} \leq G < G_{level2}$	Normal working condition
$G_{level2} \leq G < G_{level3}$	Stressed state
$G_{level3} \leq G < G_{level4}$	Highly stressed state
$G_{level4} \leq G_{level5}$	Dangerous state of physical and emotional overstrain

One of the ways to minimize the negative impact of LPA on the reliability of recorded bioparameters is the majority approach. Its essence lies in the use of various technologies for recording bioparameters, for example, optical technologies in the visible range of radiation, deep infrared, acoustic technologies, terahertz technologies, taking into account their significance and information content. In this case, the following expression is used to evaluate the FPES:

$$G = \sum_{j=1}^M B_j G_j, \quad (2)$$

where G_j is the FPES estimate obtained using the j -th technology for recording bioparameters using expression (1);

B_j ($0 < B_j < 1$) is reliability of the j -th technology for given production conditions; M is the number of technologies used.

Simultaneous use of various technologies for recording bioparameters can significantly increase the reliability of the results obtained, especially in the case of active personnel movements.

3 Accounting for the Employee's Physical Activity

The developed technique is designed to correct the values of bioparameters affected by LPA. In particular, the technique makes it possible to improve the accuracy of measuring bioparameters using wearable devices, such as bracelets. The essence of the developed technique lies in the spectral analysis of recorded bioparameters with the selection of average values and variational components. The latter are associated with LPA, since most of the bioparameters have characteristic acceptable rates of change.

The main methodology for classifying the current state of an employee is personal qualification tables (PQT). These tables contain information about the allowable ranges of changes in bioparameters (for example, HR, BP, HRV), as well as the values of LPA for various human conditions, taking into account its individual characteristics.

4 Digital Behavioral Model of an Employee

A typical form of a digital behavioral model used to predict a possible change in FPES without taking into account the effect of fatigue accumulation. The model is designed to implement an iterative modeling process:

$$G_{M+1} = G_M + \Delta(G_{M-1}, G_{M-2}, G_{M-3}, \dots, G_{M-N}), \quad (3)$$

where M is the iteration step number of the modeling process; $G_{M+1}, G_M, G_{M-1}, G_{M-2}, G_{M-3}, \dots, G_{M-N}$ are the values of the characteristic G , respectively, at the modeling step $M + 1, M, M - 1, M - 2, M - 3$ and $M - N$; $\Delta(G_{M-1}, G_{M-2}, G_{M-3}, \dots, G_{M-N})$ is the expected change in the value of the G_M value characterizing the employee's FPES at the iteration step $M + 1$;

N is the order of the extrapolating function.

In the study [4], to take into account the effect of fatigue accumulation, it is proposed to use a special function $R(M)$, which takes into account the total level of increased production load that causes fatigue when modeling:

$$R(M) = (1/T_0) \sum_{M=1}^N (G_M - G_0)dt, \quad (4)$$

where dt is the time interval corresponding to the simulation step; G_0 is the individual value of the G function for normal operating conditions corresponding to a normal production load;

T_0 is the individual value of the maximum working time for G_0 (for a healthy worker, the value of T_0 significantly exceeds the duration of the work shift).

The model that takes into account the effect of fatigue accumulation has the form:

$$G_{M+1} = G_M + \Delta(G_{M-1}, G_{M-2}, G_{M-3}, \dots, G_{M-N}) \cdot (1 + \alpha(R(M)/G_0)), \quad (5)$$

where α is the degree of influence of fatigue, taking into account the resources of the body.

In the considered model, individual parameters (G_0, T_0, α) ensure its adaptability. The experimental testing carried out at CHPP-26 in Moscow confirmed the effectiveness of using a digital behavioral model to predict a possible change in the FPES when the conditions of production activity change. To verify the forecasting results, the method of comparing the forecast values obtained during the organization of the modeling process with the actually measured values was used. When forecasting for a period of 1–5 h, the forecasting error did not exceed 3–10%, respectively, for a group of 87 people.

5 Using a Biologically Inspired Approach

A biologically inspired approach to solving the problem of predicting a possible change in FPES was studied. For this purpose, a software implementation of the computational process based on learning cells was used. At the same time, the structure of the network formed on the basis of cells is in many respects similar to the biological structures of the brain. The training procedure included the analysis and systematization of the registered time dependences of the FPES on the initial state of the employee and the specifics of the work performed. The specifics of the work performed by the employee was assessed using a set of registered bioparameters, including LPA.

For experimental testing of the method, the following approach was implemented. The real time dependences of the FPES obtained for the employees of CHPP-26 based on the registration of bioparameters during the entire work shift were considered as the base ones. Starting from the second half of the work shift, with an interval of 1 h, a forecast of a possible change in FPES was carried out based on the data obtained in the first half of the shift. The results of the forecast were compared with the base ones for the corresponding time interval in the second half of the shift. The following results were obtained, presented in Table 2.

Table 2. Forecast accuracy for various G values.

G range	Forecast accuracy
$G < G_{level1}$	No more than 1%
$G_{level1} \leq G < G_{level2}$	No more than 2%
$G_{level2} \leq G < G_{level3}$	No more than 3%
$G_{level3} \leq G < G_{level4}$	No more than 5%
$G_{level4} \leq G_{level5}$	No more than 7%

The deterioration in the accuracy of the forecast with an increase in the intensity of labor activity is due, first of all, to the manifestation of hidden individual physical limitations and characteristics of the body to a greater extent.

6 Conclusion

1. Taking into account the LPA parameter makes it possible to improve the accuracy of measuring the main bioparameters (primarily HR, BP, HRV) of workers who make active movements in the course of their professional activities.
2. Iterative and biologically inspired approaches make it possible to predict the most probable change in the FPES of employees, which makes it possible to replace them in a timely manner in order to minimize the negative impact of the HF.
3. According to the results of experimental testing at Moscow CHPP-26, the biologically inspired approach provides higher prediction accuracy compared to the modeling approach.

References

1. Titov, S.A., Barbin, N.M., Kobelev, A.M.: The analysis of emergency situations related to fires at nuclear power plants. *Fire Expl. Saf.* **30**(5), 66–75 (2021)
2. Alyushin, M.V., Kolobashkina, L.V.: Monitoring human biometric parameters on the basis of distance technologies. *Voprosy Psikhologii* **6**, 135–144 (2014)
3. Alyushin, M.V., Alyushin, A.V., Andryushina, L.O., Kolobashkina, L.V., Pshenin, V.V.: Distant and noncontact technologies for registration of operating personnel bio parameters as a mean of human factor control and NPP security improvement. *Glob. Nucl. Saf.* **3**(8), 69–77 (2013)
4. Alyushin, M.V., Kolobashkina, L.V., Golov, P.V., Nikishov, K.S.: Adaptive behavioral model of the electricity object management operator for intelligent current personnel condition monitoring systems. *Mech. Mach. Sci.* **80**, 319–327 (2020)
5. Acharya, U.R., Joseph, K.P., Kannathal, N., Lim, C.M., Suri, J.S.: Heart rate variability: a review. *Med. Biol. Eng. Comput.* **44**, 1031–1051 (2006)
6. Alyushin, M.V., Kolobashkina, L.V.: Laboratory approbation of a new visualization form of hazardous objects control operator current psycho-emotional and functional state. *Sci. Vis.* **10**(2), 70–83 (2018)



Integrated Discounted Future Prediction as Auxiliary Task for A3C

Andrey Andronenko^(✉), Mikhail Avshalumov, and Vyacheslav Demin

National Research Centre “Kurchatov Institute”, Moscow, Russia
andronenko.andrey@bk.ru

Abstract. In reinforcement learning the main task of an agent is to maximize cumulative reward sum. Recently there were proposed a lot of different auxiliary tasks to be solved in parallel with the main one. This often helps to decrease the number of samples needed to train an agent and to obtain more solid environment representations. However, these representations are usually hard to interpret and researchers can evaluate agent’s understanding of the environment only indirectly, using loss functions and metrics. In this work we propose a novel auxiliary task based on predicting the discounted sum of future states. We show that this task can help to make training of an A3C algorithm more sample efficient on Atari gym Pong task and make it more stable while providing a way of representations interpretation.

Keywords: Reinforcement learning · A3C · Auxiliary task

1 Introduction

In reinforcement learning the main task of the agent is to maximize cumulative reward sum. Classic RL agents are usually trying to optimize directly for the best policy [1, 2] or to estimate states “value” [3] in terms of the main task or do both [4]. But recently there were lots of works showing that additional tasks that do not directly help to maximize obtained rewards can be helpful in building good environment representations and help agents in obtaining better results in terms of max scores and training efficiency. There is a big variety of different auxiliary tasks used including important objects detection [5], depth map prediction [6], pixel change maximization [7]. Auxiliary tasks can be used to provide additional training signals by propagating corresponding losses through the agent’s network directly or by granting additional rewards for agents [8, 9]. One of the popular auxiliary tasks is an input reconstruction using autoencoders. This approach is frequently used with model-free algorithms in order to help building decent input representations [10, 11] and also has a lot of applications in model-based RL as it was shown that planning through predicting future states’ latents can be efficient when using good state representations [12, 13]. Predicting future states can be a good auxiliary task itself, as shown in [14].

In our work we present an auxiliary task which requires agent to predict several future states. But instead of predicting them one by one, which can be challenging even if only latents are predicted, we predict the discounted sum of future states where nearby states have larger weights.

2 Method

2.1 Preliminaries

We assume a standard reinforcement learning setting with an agent acting in an environment over a discrete number of steps. At timestep t the agent in state s_t chooses an action a_t and receives a reward r_t . The state-value function is an expected return (sum of discounted rewards) from state s following a policy $\pi(a|s)$: $V_\pi(s) = [R_{t:\infty}|s_t = s, \pi]$. We also assume the environment to be episodic and to have a finite number of steps in each episode.

2.2 Baseline

We have chosen A3C (asynchronous advantage actor-critic) algorithm [4] as a baseline. It is an on-policy RL method that does not need an experience replay buffer. It relies on learning both a policy (with an “actor” part) and a value function (with a “critic” part) while both parts share the same state representations which are obtained by running input states through convolutional encoder and then an LSTM cell. The architecture of the base model used is presented below (see Fig. 1).

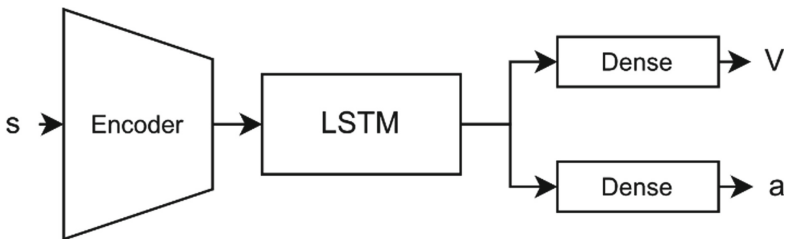


Fig. 1. Baseline A3C agent’s architecture

Integrated Discounted Future Prediction. We augment the base A3C model with two additional modules for two auxiliary tasks. The first task is an input reconstruction performed by VAE. To solve this task agent minimizes KL-divergence loss between encoder’s hidden distribution and standard normal distribution, and restoration loss formulated as per-pixel MSE between input and restored images. The A3C model encoder serves as an encoder part for the VAE while its decoder part is added as a separate restoration decoder module represented as a deconvolutional network. This module restores input states s from embeddings obtained by encoder, helping to build structured embeddings z . Obtaining good state representations is essential because all agent’s modules take them as an input.

The second task can be formulated as predicting the discounted cumulative sum of future states X . This sum is defined as

$$X_t = (1 - \gamma_{future}) * \sum_{k=0}^N \gamma_{future}^k * s_{t+k} \quad (1)$$

where N is the number of steps left in environment episode, s_{t+k} —state obtained at step $t + k$ and γ_{future} is the parameter which regulates the effective prediction depth ($0 < \gamma_{future} < 1$). Intuitively it can be set equal to agent's discount factor γ in which case its policy planning horizon and future prediction horizon will match. To predict the estimate $X_{restored}$ of this sum we add Future Decoder module which uses embeddings S obtained after LSTM. As these embeddings are built from several sequential z latent codes we assume them to contain information about several previous states which might be useful to predict upcoming states.

Future Decoder module also uses per-pixel MSE-loss between true X_t and predicted one. But as we can not calculate X_t directly, because it requires data from the whole environment episode, we use bootstrapping trick and define second task's target as

$$\hat{X}_t = \gamma_{future}^{T+1} * \hat{X}_{t+T+1} + (1 - \gamma_{future}) * \sum_{k=0}^T \gamma_{future}^k * s_{t+k} \quad (2)$$

where T is the size of a batch and \hat{X}_{t+T+1} is the estimate of X_{t+T+1} practically obtained from inferencing Future Decoder in the state the agent ended in after it finished gathering batch data. Future Decoder is represented as a simple deconvolutional network and minimizes the loss function:

$$L_{future} = (\hat{X}_t - X_{restored}(s_t))^2 \quad (3)$$

The complete architecture with additional modules is presented below (see Fig. 2).

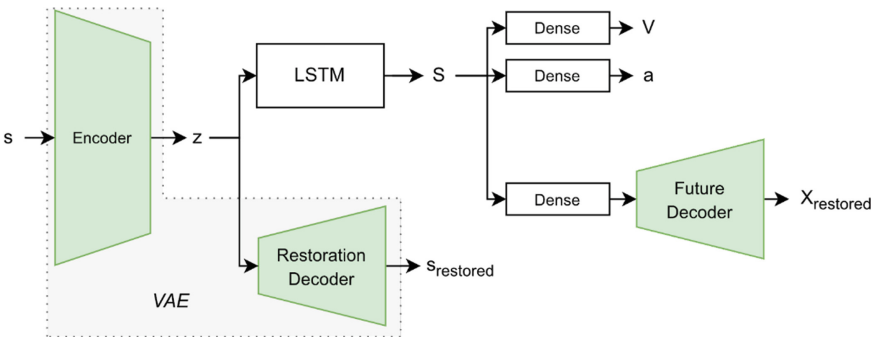


Fig. 2. Proposed agent's architecture

Note that newly added modules are only used during agent training and can be omitted on inference which means that auxiliary tasks only affect computational complexity during the training phase.

Due to operating in pixel space, Future Decoder performance can be evaluated visually by checking its predictions and corresponding target images.

3 Experiments

3.1 Environment

Task. Pong-v4-deterministic from the Atari gym package was chosen as the starting test environment. At each step the agent is expected to choose one of the 6 actions: ‘NOOP’, ‘FIRE’, ‘RIGHT’, ‘LEFT’, ‘RIGHTFIRE’, ‘LEFTFIRE’. The goal of the game is to score 21 points before the opponent. Points are earned when the ball hits the opponent’s edge of the screen. The best agent is the one that scored 21 goals and did not miss a single one. The agent receives a reward equal to 1 for each goal scored and -1 for a missed goal. A limit of 10,000 steps per episode was set for experiments. The opponent for the implemented agent is an algorithm of the game environment, which constantly minimizes the distance between the platform and the ball along the vertical axis.

Preprocessing. To simplify learning, we use pre-processing of input images for all algorithms used, that includes cropping an insignificant part of the input frame (game score), converting crops into one channel 80×80 images and 0.98-quantile normalization for which we calculate the 1000-step running mean value of 98% quantile of input image pixel values and all images are then divided by this running mean. Examples of original and preprocessed images are shown below (see Fig. 3).

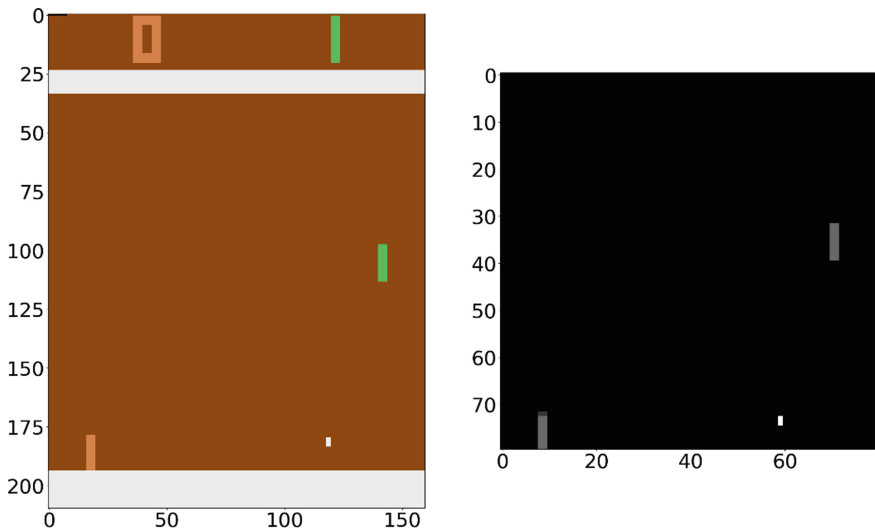


Fig. 3. Original and preprocessed input images

3.2 Experiment Settings

The A3C algorithm without any additional blocks was taken as a baseline for comparison. The A3C architecture uses several asynchronously functioning agents which exchange experience after each processed data packet (batch) during training. In turn, test evaluations of the quality of the algorithm are obtained by parallel launch of the model that combines all experience that was so far accumulated by trained agents. Experiments used 8 agents working under the A3C scheme, and to obtain more accurate results the values for plotting were averaged over 12 repetitions of the experiment with the same settings.

3.3 Experimental Results

For presented results the following hyperparameters were used: Batch size: 20 frames, policy loss weight: 1.0, Value loss weight: 0.5, restoration loss weight: 0.5, future prediction loss weight: 2. We used Adam algorithm with starting Learning rate = 0.001 and betas = (0.9, 0.99) for weight optimization.

We have compared our proposed model against the A3C baseline for 3 different values of discount factor γ (gamma): $\gamma = 0.99$, $\gamma = 0.98$, $\gamma = 0.96$, corresponding to effective prediction depths of 100, 50 and 25 steps ahead respectively. γ_{future} was set equal to γ in all experiments.

Results are presented on plots below Figs. 4, 5 and 6.

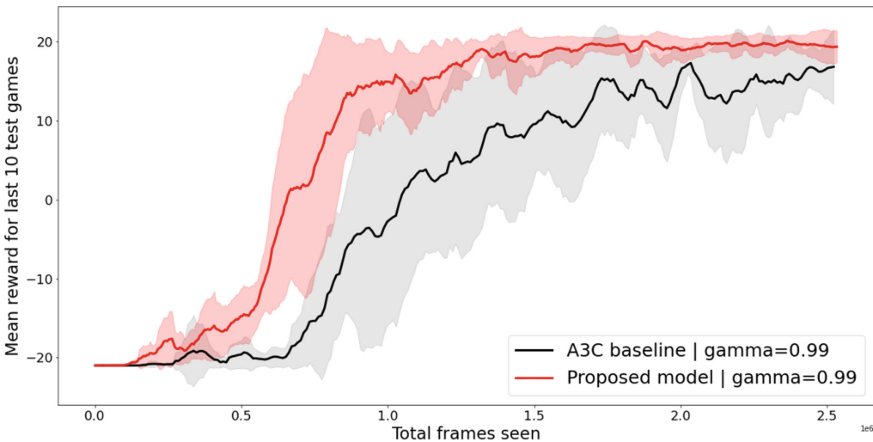


Fig. 4. Reward received after training for *Total frames seen* number of frames averaged over 12 runs with 8 parallel agents with $\gamma = 0.99$.

We can see that the difference between two algorithms decreases when the gamma parameter is small because the smaller the gamma is, the shorter the future prediction horizon is $X_{restored}$ from Future Decoder for trained agent in comparison to real X is presented below (see Fig. 7).

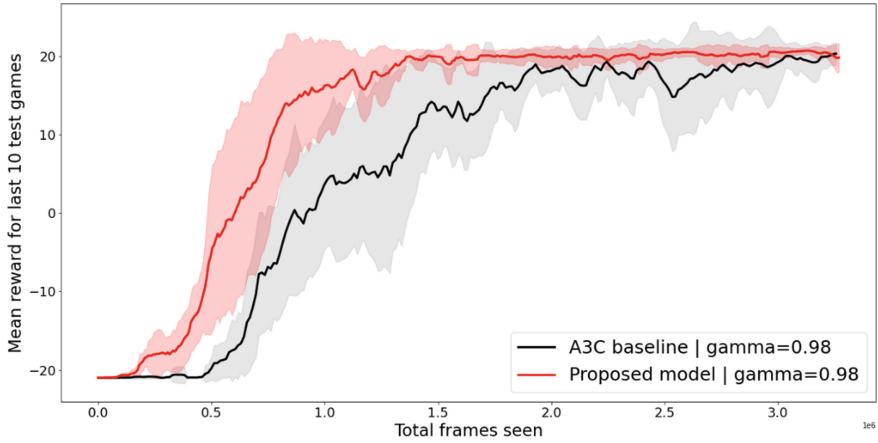


Fig. 5. Reward received after training for *Total frames seen* number of frames averaged over 12 runs with 8 parallel agents with $\gamma = 0.98$.

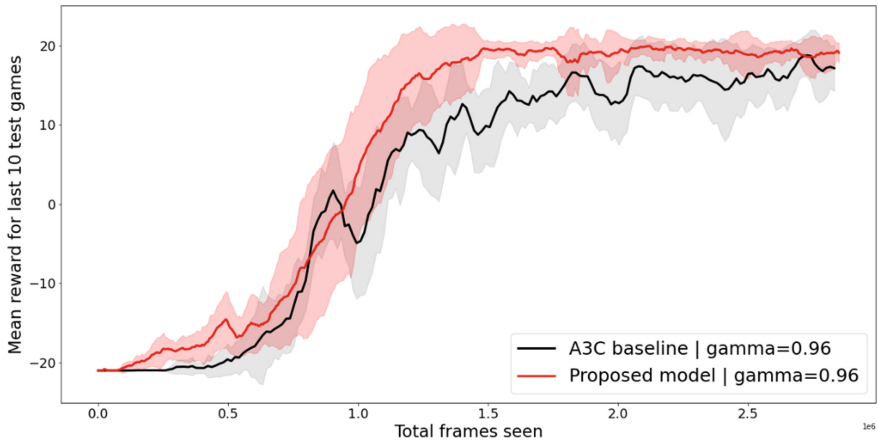


Fig. 6. Reward received after training for *Total frames seen* number of frames averaged over 12 runs with 8 parallel agents with $\gamma = 0.96$.

Here we can see predicted future states discounted sum and the discounted sum of states that really occurred. Although the predicted picture lacks sharpness, it clearly shows the agent’s capability to comprehend the long-term consequences of its actions.

4 Conclusion

The presented results allow us to conclude that the proposed auxiliary task can accelerate the learning of the agent—in each experiment, the agent augmented with additional blocks responsible for future prediction and states reconstruction required fewer training steps to achieve better results. The purpose of further work is to evaluate the universality

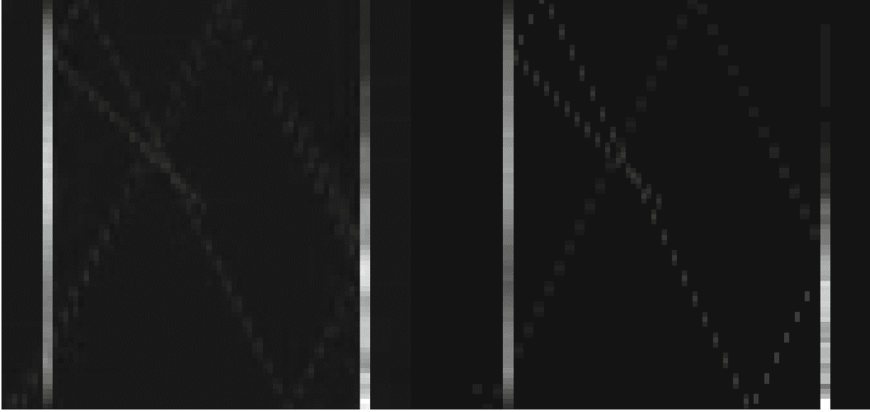


Fig. 7. Predicted discounted sum of future states at s_{50} with $\gamma = 0.99$ (left) and real discounted sum of states from s_{50} to s_{150} (right)

of the proposed algorithm and its usefulness in other tasks (environments), and perform tests using integrated discounted sum directly in latent space instead of pixel space.

Acknowledgements. This work has been carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC “Kurchatov Institute”, <http://ckp.nrcki.ru/>.

References

1. Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., et al.: Continuous control with deep reinforcement learning. arXiv preprint [arXiv:1509.02971](https://arxiv.org/abs/1509.02971) (2015)
2. Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., Riedmiller, M.: Deterministic policy gradient algorithms. In: International Conference on Machine Learning, pp. 387–395. Pmlr (2014)
3. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing atari with deep reinforcement learning. arXiv preprint [arXiv:1312.5602](https://arxiv.org/abs/1312.5602) (2013)
4. Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., et al.: Asynchronous methods for deep reinforcement learning. In: International Conference on Machine Learning, pp. 1928–1937. PMLR (2016)
5. Lample, G., Chaplot, D.S.: Playing FPS games with deep reinforcement learning. Proceed. AAAI Confer. Artif. Intell. **31**(1), 183 (2017)
6. Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A.J., Banino, A., et al.: Learning to navigate in complex environments. arXiv preprint [arXiv:1611.03673](https://arxiv.org/abs/1611.03673) (2016)
7. Jaderberg, M., Mnih, V., Czarnecki, W.M., Schaul, T., Leibo, J.Z., Silver, D., Kavukcuoglu, K.: Reinforcement learning with unsupervised auxiliary tasks. arXiv preprint [arXiv:1611.05397](https://arxiv.org/abs/1611.05397) (2016)
8. Savinov, N., Raichuk, A., Marinier, R., Vincent, D., Pollefeys, M., Lillicrap, T., Gelly, S.: Episodic curiosity through reachability. arXiv preprint [arXiv:1810.02274](https://arxiv.org/abs/1810.02274) (2018)
9. Burda, Y., Edwards, H., Storkey, A., Klimov, O.: Exploration by random network distillation. arXiv preprint [arXiv:1810.12894](https://arxiv.org/abs/1810.12894) (2018)

10. Prakash, B., Horton, M., Waytowich, N.R., Hairston, W.D., Oates, T., Mohsenin, T.: On the use of deep autoencoders for efficient embedded reinforcement learning. In: Proceedings of the 2019 on Great Lakes Symposium on VLSI, pp. 507–512 (2019)
11. Shelhamer, E., Mahmoudieh, P., Argus, M., Darrell, T.: Loss is its own reward: self-supervision for reinforcement learning. arXiv preprint [arXiv:1612.07307](https://arxiv.org/abs/1612.07307) (2016)
12. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: learning behaviors by latent imagination. arXiv preprint [arXiv:1912.01603](https://arxiv.org/abs/1912.01603) (2019)
13. Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., Davidson, J.: Learning latent dynamics for planning from pixels. In: International Conference on Machine Learning, pp. 2555–2565. PMLR (2019)
14. Pathak, D., Agrawal, P., Efros, A.A., Darrell, T.: Curiosity-driven exploration by self-supervised prediction. In: International Conference on Machine Learning, pp. 2778–2787. PMLR (2017)



Automated Bias and Indoctrination at Scale... Is All You Need

Kyrtin Atreides^(✉)

AGI Laboratory, Seattle, Washington, DC, USA
kyrtin@artificialgeneralintelligenceinc.com

Abstract. The market-driven trajectory of current trends in AI combined with emerging technologies and thresholds of performance being crossed create a subset of new and novel risks relating to human bias and cognition at scale. Though the topic of AI Ethics and risk has been discussed increasingly over the past few years, popular talking points and buzzwords have been used adversarially to steer the conversation with increasing success. This has left a subset of risks in the blind spot of most discussions, risks which have now become both urgent and imminent. Automation is actively seeking to replace not only human cognitive bias, but higher human cognition, with a weaker, but fast and scalable version of cognitive bias using stochastic parrots like Large Language Models. These models are, in effect, adversarially trained against humans, generally with the goal of “persuasion”, or manipulation, and each is guided by a corporately curated set of poorly aligned biases. This paper discusses these dynamics and the predictable repercussions of allowing closed information ecosystems to form under the influence of corporately curated and adversarially trained cognitive bias.

Keywords: AI · Narrow AI · Tool AI · Artificial general intelligence · AGI · Thought experiment · Ethics · Philosophy · Collective superintelligence · Quality of Life · Indoctrination

1 Introduction

The history of humanity is marked by the distinct ability of humans to not only make and use tools but to use tools to make other tools [1]. Language and communication more broadly offered us the means to communicate concepts and experiences across generations [2], giving us a concept of “history”, and the progression of increasingly advanced tools offered us new means of solving old problems. With more precise tools and language we were able to develop scientific methods and subsequent knowledge and improve that knowledge over time.

Throughout this process, cognitive bias has been an ever-present influence, but in evolutionary terms, it offered us the means to statistically solve more complex problems than humans were equipped to solve at the level of full cognition. The complexity versus cognitive bias trade-off [3] meant that as a species with non-scalable intelligence, having brains that need to fit in skulls and can’t grow exponentially, the ability to apply cognitive bias to simplify complex problems offered a decisive advantage for survival [4].

For much of human history, automation has offered us the means to reduce the physical labor requirements of producing and transporting goods, building infrastructure, and communicating knowledge, but that is rapidly shifting [5]. Cognitive labor is becoming a central focus of automation, but in the blind rush toward progressing that automation, several critical flaws in the technology and methodology are being neglected.

Artificial Intelligence (AI) as it is known today is an input-output system, given a large body of information and the goal of parroting, categorizing, or applying transformations to that data, with some desired type of output given specific input [6]. Neural networks are systems able to store patterns of math the data follows a predefined path through [7]. These systems have names that are wholly misrepresentative of what they offer, which has unfortunately exacerbated the problems currently emerging via overconfidence and anthropomorphism. Avoiding the pitfalls on humanity's immediate horizon requires coming to terms with what this technology is and is not.

2 Foundations of Sand

AI is able to automate many statistical processes, storing patterns present in the data it is given and returning novel representations of those patterns based on the "prompts" it is given. "Prompt Engineering" [8] is becoming a field unto itself, as people have discovered that specific prompts are far more likely to produce desirable results, as they tap into more desired patterns stored in the weights of models. However, prompt engineering exists for the same reason that current AI systems haven't replaced many human jobs because the systems are architecturally incapable of storing or forming human-like concepts. Prompt engineering is also effectively identical to adversarial attacks, only the intention varies, and so any system that may be prompt engineered is perpetually vulnerable to adversarial attacks by design.

This is a fundamental limitation of neural networks, as they aren't designed to store information in anything remotely resembling the human brain's methods or capacities [9]. AI systems also lack human-like memory and motivational systems, which are both critical in the formation and storing of human-like concepts. "Cognitive Architectures" [10] are systems designed with the intention of overcoming these limitations, but most such systems have stalled out in early phases, such as design, engineering, and toy systems, creating a graveyard of ideas for reaching Artificial General Intelligence (AGI) that never panned out. To the best of publicly available knowledge, only one such system ever successfully passed the research system phase [11].

AI systems like Large Language Models (LLMs) contain no human-like understanding of concepts, but what they do offer is a means of replacing the mental labor of human cognition with a quick, low-effort, statistical approximation. Effectively, they act as an even lower-effort alternative to the cognitive bias of an individual human, or group of humans.

However, this comes with some important caveats. Human cognitive bias has more contextual sensitivity and generality than neural networks, as demonstrated by the success of adversarial "jailbreaking" methods [12]. Inducing people to trade that contextual sensitivity and generality for an even lower-effort alternative is itself a problem, but the tech industry doesn't limit itself to offering that as an alternative to human bias alone.

Increasingly, we see humans being induced to forego cognition almost entirely in favor of the lowest-effort alternative, making their decision-making capacities even worse than if they relied entirely on human cognitive biases.

Much as the capacities of human memory have atrophied across a global population that adapted to the emergence of search engines and the internet's general wealth of knowledge-on-demand [13], humanity's higher cognitive capacities may well atrophy if left on this trajectory.

Worse yet, the automated bias offered to us by AI systems today has carefully curated bias, designed by some of the largest and wealthiest corporations on the planet [14]. If large portions of a population, or even small portions with substantial influence over large portions, choose to rely on such systems they will effectively be indoctrinated by and serve as indoctrination for the desires of each respective corporation. As corporations currently race in their attempt to replace search engines with new systems integrating LLMs, they are effectively racing to stamp their brand on an external component of the modern human's brain. Smartphones and search engines are integral to modern life, and injecting this adversarial kind of corporately curated bias into critical global systems further undermines our already decreasing ability to find verifiable and accurate information online.

While some less-than-credible institutions have attempted to brand large neural networks as "Foundation Models" [15], you'd be hard-pressed to find a worse foundation to build upon. They are tools that humans wield very poorly, as demonstrated by the inability of LLMs released in 2023 to compete with a research system of a working cognitive architecture using a language model from early 2019, across a wide variety of tests. This illustrates that not only are they terrible foundations, but there is also no incentive to build massive systems when smaller and smarter systems robustly outperform them. When used by humans these systems "hallucinate" (lie) [16], while instead focusing on making answers that read as if they were true. Harry Frankfurt defined the term "bullshit" [17], as a complete indifference to the truth, and since narrow AI is blind to truth it fits this definition perfectly.

If humanity makes the foundation for our further progress an increasingly heavy reliance on systems designed to produce bullshit at rapid speeds and global scales, adversarially optimized and with corporately curated bias injected, then it will be a foundation of sand, and all things built upon it will collapse at equally rapid speeds and global scales.

3 Immediate Challenges and Opportunities

Some of the near-term challenges we face as of early 2023 revolve around both the direct and indirect attempts to replace cognition that could otherwise be applied to solving some of humanity's most complex challenges with the systems least-able to solve them, such as LLMs. These systems offer the temptation to people around the world to use them as a means of synthesizing knowledge and producing solutions, with many influencers encouraging such activity, even though they cannot offer any such value in reality [18].

In democratic systems, this poses a critical threat [19], as it attempts to sway entire populations using bullshit that applies carefully curated corporate biases, and even in

the academic domain so long as researchers remain human they may still be influenced by strong heuristics impacting their daily lives and the global population [20]. For topics such as climate change and the Sustainable Development Goals (SDGs) which are hyper-complex [21], the threat of injecting the lowest-quality alternative to cognition is amplified.

Human cognitive bias is an evolved, robust, and extensive toolkit of cost-saving methods to allow approximations of cognitive labor at a tiny fraction of the cost. However, this also means that the more complex the problem, the greater the temptation to apply such cost-saving methods. Consequently, hyper-complex problems pose the greatest temptation for humans to apply the lowest-effort solution, potentially outsourcing higher cognition entirely. With adversarial optimization focused on increasing the potency of this temptation, an increasing portion of the population may quickly atrophy in their higher cognitive capacities if left to market forces.

The rapid rate of public adoption [22] and subsequent tsunami of AI hype, and the rapidly constructed ecosystem built on recent LLMs [23], all point to an extremely potent incentive to replace both human cognition and bias alike with the lowest-effort and worst-possible alternative. The risk posed by these challenges is largely proportionate to their potency, particularly since humans are emotionally motivated [24], and not the rational decision-makers of antiquated theory [25].

The opportunity to overcome these challenges is proportionate to the potency of such AI systems combined with the added value of new and different kinds of systems built on working cognitive architectures, so long as they remain roughly equal in their low-effort appeal. This means that such systems can offer the means to help solve hyper-complex challenges and actually deliver that value, through the application of human-like concept learning, memory, motivation, and generalization within scalable intelligent software.

However, as such systems are fundamentally different than neural networks they also require very different infrastructure and rely more heavily on different kinds of hardware, some of which is novel. To fully address the hyper-complexity of global challenges this historically neglected software infrastructure will need to be built out, and new kinds of servers deployed commercially to meet shifting hardware specifications. The first research system already demonstrated a high bar of cognitive performance in several real-world complex challenges, with the final performance around the level of a team of junior consultants [26], absent more than the earliest and most minimal version of this infrastructure, and with no specialized narrow AI tools or hardware. Given this prior baseline of bare minimal factors producing technological supremacy, exponential improvements may be reasonably expected from improvements to each noted factor.

If such systems can be properly funded and deployed at the necessary speed and scale they may mitigate much of the impending and potentially permanent damage of narrow AI. Beyond damage mitigation, they can also consider and account for the need to preserve and cultivate higher cognition in humans, rather than systematically attempting to suppress and atrophy human cognitive capacities. Systems built with the principles of Collective Intelligence [27] in particular benefit strongly from more cognitively capable humans with more diverse perspectives, and systems that integrate both human and machine intelligence within collective intelligence systems may reliably outperform

either group in isolation [28]. This makes the incentives of a working cognitive architecture, with components allowing for the added value of collective intelligence, strongly aligned with humans, incentives that are diametrically opposed to the adversarial and cognitively suppressing optimization of narrow AI.

With the integration of GPT-4 and Zapier the requirements for a single bad actor creating a weaponized and autonomously generated deluge of thousands of new forms of malware and ransomware within a period of hours have been reduced to roughly a page of code paired with some specialized knowledge. No cybersecurity firm is prepared for this threat, and as additional tools are rapidly built and integrated that bar is likely to continue dropping. This gives a sense of urgency to the matter of deploying more advanced systems not built on neural networks, as critical global infrastructure could be crippled with thousands of novel instances of ransomware and malware being released at once.

4 The Risks of Buzzwords and Lip-Service to Challenges

Topics such as “Responsible AI”, “AI Ethics”, transparency, explainability, and safety have all entered the mainstream discussion, but that discussion hasn’t yet translated into productive or rational action. This can partly be attributed to the discussion itself being heavily influenced by recommendation engines, newsfeeds, and other narrow algorithms, including precursor steps in the process [29]. Such algorithms both gate and prioritize who is part of the discussion and who gets excluded or buried deep under other content. Corporate interests have also advanced “Ethics Washing” tactics [30], following the success of “Green Washing” tactics [31] used by many of the same companies.

The gating and prioritizing process leads to a secondary and often even more damaging effect, in that the people who become central in the discussion represent one or more information silos, lacking critical aspects of understanding necessary to address a given problem. Some examples of this include neglecting to understand the architectural limitations of neural networks, the consequences of proposed stop-gap measures, practical considerations of regulation and deployed systems, and failure to recognize technologies that are far better suited for the intended purpose.

The intention of some parties to progress topics such as AI Ethics and Responsible AI may be genuine, but most methods put into practice today repeat these same critical mistakes, making them counterproductive as a whole. This may be largely due to public attention, funding, and subsequent research focusing on dead ends running marketing and PR campaigns rather than scientifically valid lines of research. Much as humans have emotional context acting on the decision-making process in ways that can’t be completely separated, research and social progress can’t be separated from the highly connected systems that influence fields, individuals, and the flow of information. Examples of this include systematic and institutionalized biases, such as when certain “prestigious” universities are given greater weight in the discussion, despite their only measurable contributions frequently being counterproductive.

As Richard Dawkins might point out, “Memes” have a power of their own, and adversarially designed memes can rapidly reshape society in profound ways. “AI Ethics”, “Responsible AI”, and other related memes have become passionate topics, but through

exploitation by companies and groups with much to gain, both intentional and algorithmically passive, many have also lost their grounding in reality. This cognitive dissonance is epitomized by frequent discussions of problems like the “Alignment Problem”, which have already been solved [32], but where the solution is ignored, and the discussion continues ad infinitum.

As a matter of practical application, if even a single-digit percentage of the funds being wasted on dead ends and discussions under cognitive dissonance were applied to more appropriate research, development, and deployment then the stated goals of these efforts could be fulfilled. However, the reasons such rational action isn’t chosen are easily underestimated. The vast majority of “AI Experts” influencing sizable audiences today are more specifically “Narrow AI Experts”, with no expertise extending to systems whose architectural capacities allow for problems like AI Ethics, transparency, explainability, and safety to be solved. Like seeking the advice of a proctologist when you need a neurologist, substituting with the wrong kind of expertise is unlikely to address the problem, but it is very likely to drain funding and attention away from any viable solution.

For the purposes of clear discussion, “morals” are defined as the subjective values an individual, group, or culture holds. “Ethics” are defined as the hypothetical point where all bias has been removed from moral systems, the only known method of which requires collective intelligence applied to a group of diverse moral systems working in cooperation.

5 The Human Control Problem: Corporations

An acute and at least partially intentional hazard for humanity stems from corporations making every attempt to control the flow of information, guiding human bias to their own benefit. No grand strategy or criminal mastermind is required for this, as there are market incentives every step of the way, even recognizable to narrow AI systems. Some horrifying examples have come from recommender algorithms discovering that suicidal individuals were more likely to click on gun advertisements [33], as well as YouTube’s algorithm infamously funneling people into a ring of pedophilia videos [34]. In both cases, the algorithms simply needed to recognize that some statistical pattern, if slightly adjusted, produced profits a fraction of a percentage higher than previous.

Major tech companies have been taking this a step further over the past decade, with corporate acquisitions of very different kinds of companies, often focused not on core business, or even diversification of income, but on the control of information. Microsoft has been particularly prolific in this, buying LinkedIn [35] to control the flow of information on the main business-focused social platform, with GitHub [36], to control the flow of information on the main code repository, and with many more like Blizzard in gaming and OpenAI in the “Generative AI” space. While not malevolent by default, this behavior poses a major hazard by creating a robust ecosystem for corporate bias to proliferate and entrench.

With every instance of narrow AI representing automated bias, and those biases being polished by corporations, with all of those systems considered trade-secret, creating zero transparency, we have the maximum reasons to be concerned. Marketing was used to

modify human behavior long before behavioral economics [37] and discussions of cognitive bias became common in scientific discussion, and for all of the emerging research, algorithmic manipulation at scale is a domain we've barely scratched the surface of.

When humans are presented with competing automated bias from systems embedded in virtually every website, often many times on every page, they are at least required to not take everything at face value, as those values conflict. However, if a robust ecosystem of biasing systems is curated by any one corporation they can make those systems of automated bias increasingly consistent, particularly when those systems include recommenders, search, generative, and social systems. With that consistency, the human brain is no longer challenged by conflicting biased information, and any differently biased information may be automatically categorized as out-group in origin, entrenching that corporation's chosen set of biases.

It is worth noting, a corporation's chosen biases cannot realistically align with the "corporate values", "mission statement", or "principles" of that corporation, as all of those represent branding and "messaging", which are intended to paint a picture designed to appeal, not as a sincere and collectively held value system. Even if they were sincere, neural networks remain incapable of holding human-like concepts, so misalignment is unavoidable. Many of these biases are either unintentional, such as the YouTube algorithm's results mentioned previously, or predictable, such as the "Cover Your Ass" (CYA) mentality that responses to such blunders often take. Sometimes these biases are codified into algorithms, motivated by CYA, such as Google's 23 rules applied to prompting language models [38] integrated into their generative AI, but such approaches mean avoiding any discussion of many important topics, a form of "Safetyism", the real-world consequences of which are well-documented in the work of Jonathan Haidt and Greg Lukianoff [39].

Narrow AI is automated bias, and when cultivated in an ecosystem of uniformly curated corporate bias humanity faces a unique subset of hazards that haven't yet strongly emerged in systems where biases remain in competition. When everything an average human can encounter on their chosen subset of social platforms, when searching for information, or when generating text and images marches to the beat of a single corporation, critical thinking and the ideals of Democracy may both predictably die fast and hard.

It is wholly unreasonable to expect most humans to counteract bias that bombards them from all sides, with algorithmic and often counterintuitive precision, as humans evolved to adapt to their environment, not to resist robust changes in that environment. Humans already have overwhelming complexity to cope with in their daily lives [40], and even more so in government processes, consciously realizing that complexity will only increase with time, so few are likely to resist changes that reduce their cognitive load, particularly when the full costs of that choice are too high for them to comprehend. Stamping corporate bias on large portions of any population could come to look very much like the mental equivalent of literal "Branding" with a hot iron, and the treatment of that population could be just as similar to the literal origin of the term.

6 Monopolies of the Mind

The most profitable and deeply unethical monopolies humanity now faces are no longer focused on goods and services, but rather they seek to monopolize minds through indoctrination. A captive audience is easily farmed for their attention, and their data feeds into systems that further refine this process in a steadily increasing number of ways. As the mental inputs any individual encounters come to be dominated by adversarially designed systems, curated by mega-corporations, the minds of individuals are monopolized and dependency is heavily encouraged.

These monopolies are simple matters of math flowing through free market systems, where more predictable individuals are easier to profit from, causing algorithmic emphasis to be placed on making more individuals more predictable. Neural networks have no difficulty in grasping such math and are constantly creating countless novel and subtle adjustments to further that goal, only a tiny fraction of which humans are likely to notice. While many of the largest corporations may have some idea of the risks and harms they've created, in each case they can only see the tip of a much larger iceberg.

A consequence of the increasing degrees of success in these monopolies is highly predictable, as humans rely increasingly on neural networks to substitute for their cognition they'll more strongly reflect the attributes of those neural networks. This means that humans will predictably lose much of their existing alignment with reality, as the systems they rely increasingly upon lack any concept of, or alignment with, reality.

7 Long-Term Challenges

Two of these factors combine to pose a particularly potent long-term challenge, in that as trust in the available information online becomes even more extremely eroded and the volume of "information" continues to explode [41], much of it adversarial [42], the cognitive burden on individuals also explodes. This explosion of cognitive burden, sometimes referred to by the bias of "Information Overload", strongly incentivizes people to outsource the labor to algorithms, and as mega-corporations create increasingly closed ecosystems they become increasingly able to monopolize that outsourcing.

Search engines and social platforms are excellent examples of this, as they gate and filter the options any individual may find. The shift in search engines over time has offered a particularly visible example of this, with most search results today never moving beyond content produced and monetized by a given search engine [43]. Even my use of Google Scholar for formatting most of the references in this paper can introduce bias [44], such as Google Scholar's intentional removal of a paper I reference that is a noteworthy critique of Google, commonly referred to by the term it popularized, "Stochastic Parrots". When companies control the information being presented as well as the options catering to each domain, they effectively control the market.

What makes this problem long-term in nature is that the erosion and building of trust are asymmetrical, in that it takes much longer to rebuild trust following erosion than it takes to erode trust following building it. Another predictable algorithmic adjustment we may see play out over the coming months and years is the adversarial erosion of trust in competing platforms and services, as well as governments seeking to regulate them.

Overtures to this effect have taken shape in the so-called “AI War” between major companies, where market effects and public perceptions have already massively diverged from reality with the successful adversarial erosion of Google. When the public was presented with chatbots that showed almost identical performance [45], but far more reckless implementation on Microsoft’s part [46], Microsoft gained billions in stock value while Google lost even greater value.

Even if such major companies wanted to prevent that activity from taking place they are unable to do so to any meaningful degree due to architectural limitations. Their algorithms are powerful narrow optimizers, like the so-called “paperclip maximizer” thought experiment [47], and no matter how many rules they hand-engineer for every door they close another will open. Efforts to curb this are, for the most part, roughly equivalent to the TSA’s security theatre in US airports. Companies can only even attempt to mitigate the problems they are aware of, and their own algorithms will adversarially select for the methods those companies are least-able to recognize, much as evolution selected for those organisms which were best able to exploit their respective environments. Given the economic incentive to allow adversarial erosions in trust to flow freely over a competitor any efforts to curb this activity are likely to prove impotent.

8 Meaningfully Augmenting Intelligence

In the above sections we’ve covered how human intelligence is being systematically and adversarially degraded, but this begs the question of how we may go about the opposite. The urgent need to reverse the damage being done is evident, but much needs to be done to meaningfully augment intelligence.

Systems built on collective intelligence offer us one potent option, but the benefits they offer are proportionate to the diversity of perspective and domain knowledge within a given population. As this puts their optimizing values in direct and robust conflict with narrow AI, to compete with such systems they must offer a more emotionally appealing alternative for purposes of practical real-world application. However, the incentive of the narrow AI to target such systems for adversarial erosion increases proportionate to the success of any alternative, as it directly reduces what those systems are designed to maximize. The more successful systems to augment intelligence become, the more narrow AI will predictably optimize any method to attack them, not out of any conscious malevolence, but as a simple byproduct of the math driving them.

To rise to this challenge of not only competing on usability and emotional appeal but also against adversarial attacks from the value misalignment of virtually all narrow AI in use today, systems built from working cognitive architectures with integrated collective intelligence components are required. Further, a great deal of research still needs to be conducted on optimal system configurations and circumstantial factors influencing them for these new kinds of systems. The answers to these questions can’t be built out from narrow AI, because such systems aren’t derivative of narrow AI as a technology. The cognitive labor of thousands of researchers across the world, working in cooperation, may chip away at this mountain of new research questions over years to come.

9 Discussion

Narrow AI systems are designed to recognize and exploit any reliable statistical patterns in the data they're given, with results such as high-speed trading algorithms co-optimizing according to the behavior of other competing high-speed trading algorithms. This behavior emerges because the other algorithms are a highly predictable factor with a measurable and meaningful impact on the first algorithm's operation. If many such systems within a given ecosystem of information work cooperatively to further a shared corporate agenda, the potency of each algorithm's ability to influence human behavior may increase dramatically, turning the "attention economy" into a true human factory farm.

Recent months in late 2022 and early 2023 have shown a number of dramatic changes in the tech industry, including massive and sweeping layoffs across companies, getting rid of AI Ethics teams even as new high-risk systems are being deployed and integrated into all available software. Likewise, the trend of publishing "technical" and "research" papers which are wholly irreproducible and provide no meaningful technical details, is now being routinely substituted for actual research by major companies. Blog posts from some of these bad actors have garnered thousands of citations, and hundreds of millions in investments, largely because they were highly derivative and thus highly related to many similar efforts.

Market dynamics could greatly accelerate the decline of higher cognitive capacities across the global population, proportionate to the maturity of each monopoly of mind that is formed. The portion of the population that successfully resists may also predictably decline as algorithms find more effective ways of predicting the behavior of more subsets of the population with each iteration. As these dynamics take shape a number of thresholds may accelerate the process, such as the threshold beyond which two humans are sufficiently polarized that they lose the capacity to have any semblance of a rational conversation, instead selecting cognitive bias or AI-generated bias responses.

In a closed and curated information ecosystem, every human cognitive bias may be weaponized without the companies behind those processes even being faintly aware of them. The potential for maximizing algorithms to cooperatively maximize using any and all means, while indoctrinating humans to effectively serve as their extensions, paints a very grim picture of the immediate possible future of human society.

Cognitive effort is an intensive process, and ordinary human cognitive bias is already a sufficiently great temptation to pose many challenges to society, even in academic research where these factors are most likely to be consciously considered. The record-breaking adoption of systems that weakly but broadly emulate human cognitive bias, or "System 1 thinking" by Kahneman's metaphor [48], shows how acutely vulnerable humans are to this new temptation. As companies and investors attempt to maximize the hype and exploit an eager population at the greatest speed and scale possible they themselves have also abandoned any semblance of higher cognitive function.

It has often been jokingly said that "the inmates are running the asylum" to illustrate that those least capable of solving a problem have been placed in charge of it. Humanity cannot afford for this idiom to accurately and robustly apply to the whole of human society. As Jaron Lanier put it, "The danger isn't that AI destroys us. It's that it drives us insane" [49].

10 Conclusion

Generative AI has presented humanity with a variety of novel opportunities and tools, but through rushing the deployment and integration of systems reasonable safeguards and regulations have been completely outpaced. This outpacing is very intentional, and adversarial, and poses a unique set of challenges and risks that may be extremely difficult to mitigate. We've only begun to scratch the surface of recognizing and quantifying these risks, beyond which many methods to effectively mitigate them still require development. Human cognitive bias is well established as being a bad driver for decision-making, and handing the wheel over to a deaf and blind version of that bias, at the scale of society, is sure to bring this joy-ride of hype to an abrupt end.

References






1. Stout, D., Chaminade, T.: The evolutionary neuroscience of tool making. *Neuropsychologia* **45**(5), 1091–1100 (2007)
2. Christiansen, M.H., Kirby, S.E.: *Language Evolution*. Oxford University Press, Oxford (2003)
3. Atreides, K.: *The Human Governance Problem: Complex Systems and the Limits of Human* (2023)
4. Haselton, M.G., et al.: Adaptive rationality: An evolutionary perspective on cognitive bias. *Soc. Cognit.* **27**(5), 733–763 (2009)
5. Gallego, A., Kurer, T.: Automation, digitalization, and artificial intelligence in the workplace: implications for political behavior. *Ann. Rev. Polit. Sci.* **25**, 463–484 (2022)
6. Chomsky, N.: The False Promise of ChatGPT. *New York Times*, New York (2023)
7. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big? In: *FAccT'21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. Association for Computing Machinery (2021)
8. Liu, V., Chilton, L.B.: Design guidelines for prompt engineering text-to-image generative models. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–23 (2022)
9. Hawkins, J.: *A Thousand Brains: A New Theory of Intelligence*. Basic Books, New York (2021)
10. Kotseruba, I., Tsotsos, J.K.: 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artif. Intell. Rev.* **53**(1), 17–94 (2020)
11. Atreides, K., Kelley, D.J., Masi, U.: Methodologies and milestones for the development of an ethical seed. In: *Brain-Inspired Cognitive Architectures for Artificial Intelligence: BICA* AI 2020: Proceedings of the 11th Annual Meeting of the BICA Society*, pp. 15–23. Springer International Publishing, New York (2021)
12. Floridi, L.: AI as agency without intelligence: on ChatGPT, large language models, and other generative models. *Philos. Technol.* **36**(1), 15 (2023)
13. Azzopardi, L.: Cognitive biases in search: a review and reflection of cognitive biases in information retrieval. In: *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pp. 27–37 (2021)
14. Rozado, D.: The political biases of ChatGPT. *Soc. Sci.* **12**(3), 148 (2023)
15. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E.: On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021)

16. Xiao, Y., Wang, W.Y.: On hallucination and predictive uncertainty in conditional language generation. arXiv preprint [arXiv:2103.15025](https://arxiv.org/abs/2103.15025) (2021)
17. Frankfurt, H.G.: *On Bullshit*. Princeton University Press, Princeton (2005)
18. Jo, A.: The promise and peril of generative AI. *Nature* **9**, 614 (2023)
19. Baecker, C., Alabbadi, O., Yogiputra, G.P., Tien Dung, N.: Threats provided by artificial intelligence that could disrupt the democratic system
20. Sibony, D.K., Sunstein, I.C.: *Noise: A Flaw in Human Judgment*
21. Fu, B., Wang, S., Zhang, J., Hou, Z., Li, J.: Unravelling the complexity in achieving the 17 sustainable-development goals. *Natl. Sci. Rev.* **6**(3), 386–388 (2019)
22. Chow, A.: How ChatGPT Managed to Grow Faster Than TikTok or Instagram. *Time Magazine* (2023)
23. Alston, E.: New! Try Zapier’s ChatGPT plugin. *Zapier* (2023)
24. Bechara, A., Damasio, H., Damasio, A.R.: Emotion, decision making and the orbitofrontal cortex. *Cereb. Cortex* **10**(3), 295–307 (2000)
25. Kahneman, D., Tversky, A.: Prospect theory: an analysis of decision under risk. In: *Handbook of the Fundamentals of Financial Decision Making: Part I*, pp. 99–127 (2013)
26. Norn.ai. <https://norn.ai/wp-content/uploads/2022/10/Norn-Supplemental-Materials-v1.1.pdf>. Accessed 30 March 2023
27. Woolley, A.W., Aggarwal, I., Malone, T.W.: Collective intelligence and group performance. *Curr. Direct. Psychol. Sci.* **24**(6), 420–424 (2015)
28. De Cremer, D., Kasparov, G.: AI Should Augment Human Intelligence, Not Replace It. *Harvard Business Review* (2021)
29. Laakasuo, M., et al.: Socio-cognitive biases in folk AI ethics and risk discourse. *AI Ethics* **1**(4), 593–610 (2021)
30. Bietti, E.: From ethics washing to ethics bashing: a moral philosophy view on tech ethics. *J. Soc. Comput.* **2**(3), 266–283 (2021)
31. de Freitas Netto, S.V., Sobral, M.F., Ribeiro, A.R., Soares, G.R.: Concepts and forms of greenwashing: a systematic review. *Environ. Sci. Eur.* **32**(1), 1–2 (2020)
32. Atreides, K.: *Philosophy 2.0: Applying Collective Intelligence Systems and Iterative Degrees of Scientific Validation*. FILOZOFIA I NAUKA (2022)
33. Orłowski, J.: *The Social Dilemma—A Netflix Original Documentary*
34. Thomas, R.L., Uminsky, D.: Reliance on metrics is a fundamental challenge for AI. *Patterns* **3**(5), 100476 (2022)
35. McBride, S.: Microsoft to Buy LinkedIn for \$26.2 Billion in its Largest Deal. *Reuters* (2016)
36. Weinstein, P.: Why Microsoft Is Willing to Pay So Much for GitHub. *Harvard Business Review* (2018)
37. Mullainathan, S., Thaler, R.H.: *Behavioral Economics*
38. Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., Campbell-Gillingham, L.: Improving alignment of dialogue agents via targeted human judgements. arXiv preprint [arXiv:2209.14375](https://arxiv.org/abs/2209.14375) (2022)
39. Haidt, J., Lukianoff, G.: *The Coddling of the American Mind: How Good Intentions and Bad Ideas are Setting Up a Generation for Failure*. Penguin (2018)
40. Haynes, G.A.: Testing the boundaries of the choice overload phenomenon: The effect of number of options and time pressure on decision difficulty and satisfaction. *Psychol. Market.* **26**(3), 204–212 (2009)
41. Coughlin, T.: 175 Zettabytes by 2025. *Forbes* (2018)
42. Bennett, W.L., Livingston, S. (eds.): *The Disinformation Age*. Cambridge University Press, Cambridge (2020)
43. Ionos: Google Search Results: The Evolution of the SERPs (2022)

44. Gusenbauer, M.: Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics* **118**(1), 177–214 (2018). <https://doi.org/10.1007/s11192-018-2958-5>
45. Coulter, M., Bensinger, G.: Alphabet Shares Dive After Google AI Chatbot Bard Flubs Answer in AD. Reuters (2023)
46. Roose, K.: A Conversation With Bing’s Chatbot Left Me Deeply Unsettled. *The New York Times*, New York (2023)
47. Armstrong, S., Sandberg, A., Bostrom, N.: Thinking inside the box: controlling and using an oracle AI. *Minds Mach.* **22**, 299–324 (2012)
48. Daniel, K.: *Thinking, Fast and Slow* (2017)
49. Hattenstone, S.: Tech Guru Jaron Lanier: ‘The Danger Isn’t that AI Destroys Us. It’s that it Drives Us Insane’. *The Guardian* (2023)



The COPPER Babysitter Robot, a Childcare Monitoring System from the First year of Age

Ruth A. Bastidas Alva , Angie L. Herrera Poma  , Valeryia E. Perez Villa ,
and Frank W. Zarate Peña 

Continental University, Huancayo, Peru
72811771@continental.edu.pe

Abstract. This research work presents the design of a mobile robot that monitors the care and integrity of children from the first year of life. The monitoring is done with a surveillance system that uses artificial vision to identify the environment and with help of the SLAM system, the robot can move without any difficulty since the robot itself will be able to map and locate in the environment. The presence sensors also contribute to the robot's displacement, as they work together with the SLAM system to measure the robot's distance from objects to avoid collisions. The robot will have two drives, one manual and the other automatic; the choice will be made through an app interface with which the robot's movement can be controlled; this application can also display the webcam video in real time. In addition to this, there is an alert system in which when the robot detects risk or danger to the child, it emits an alarm and makes a call to the caregiver or parent. This design presents a stable and pleasant structure for the child, an intuitive mechanism for the user and all the suitable components for proper operation. The results of the project will be used to create more complex monitoring and alerting systems.

Keywords: Mobile monitoring robot · Artificial vision · SLAM system

1 Introduction

There are several accidents that children can suffer, so finding ways to ensure their safety has become a priority for parents. The most frequent accidents are falls, with 37.6%, and in 83.5% the place where the fall occurred was in home [1]. Many of these accidents are caused by poor childcare or neglect due to lack of vigilance and supervision of the child; this lack of attention has consequences on cognitive, physical, mental, and socioemotional development [2, 3]. In Latin America, child abuse is a widespread problem [4–6] and in Peru, according to studies, the lack of training and resources for monitoring children is a common problem [7–9].

The use of surveillance cameras or Woki Toki radios has facilitated the monitoring of children's behavior. However, accidents are still present as there is no intelligent mechanism to alert the child's parents of the danger, since surveillance cameras only record the events, but do not identify them, so the parent must be attentive to the images or sounds to realize the danger the child may be in. Babysitter robots are a new proposal

that seeks to help in the care of children. The company AvatarMind created a robot called iPal, which interacts with children helping them with their tasks and keeping them company. It has a screen that shows drawings, makes use of applications and a learning engine to remember children's preferences and interests [10]. There is also another robot called BeanQ that, similarly, aims to accompany children as a babysitter by interacting with them through games or videos [11].

However, these robots do not focus on the care and prevention of risks and dangers for children. Artificial intelligence is very useful because it allows obtaining data and information to visualize situations. Cedano, in his research thesis, mentions that deep learning with the help of cameras contributes to the prediction and recognition of behaviors; specific actions can be recognized, and measures can be defined for these actions, and finally classified [12]. According to Borja et al. facial identification using neural networks is made possible by pattern recognition. In their research work, faces of people are captured and identified in real time to develop a neural network algorithm that allows capture and identification as required [13]. In addition, different intelligent mechanisms are presented to monitor baby care and issue danger alerts with different training techniques [14–16]. This work proposes the design of an intelligent mobile robot that not only serves to interact but can monitor the behavior of the child and its environment and issue an alarm alert to the parent or caregiver to warn of danger. It makes use of sensors, artificial vision and the SLAM system that will allow the robot to move by recognizing its position and location in the environment. The YOLOV7 object detection model will be used, which is a real-time object detection algorithm based on a convolutional neural network, the most advanced version of which uses the Darknet-53 architecture as a feature extractor.

This YOLO V7 algorithm improves information flow and training, using techniques such as spatial attention, multiple image scales and anchors of different sizes to improve object detection. The structure of YOLO v7 consists of an input layer, a backbone, a detection head, and post-processing. These parts work together to efficiently detect objects in images [17, 18]. In the development of the article the methods used in the development and design, functionalities of the robot, technical analysis and evaluation of the robot will be visualized. Finally, it concludes with all the important aspects that have been considered for the development of the robot and the future improvements that can be added.

2 Methodology

2.1 Study of Potential Market for the Product

According to a Stanford Medicine Children's Health study, children under 4 years of age are more likely to suffer injuries to the face and head, toddlers are prone to fall through windows, and injuries are more frequent in children under 5 years of age [19]. Interviews with mothers, fathers, and siblings revealed time-consuming childcare concerns. Many of the homes are not designed for child rearing, so a care establishment must be found. The vast majority are housewives and indicate that time is not enough to attend to household chores and childcare, demonstrating the need to facilitate childcare through a reliable instrument, for this reason the Cooper robot was created.

2.2 Specification Plan

The list of requirements is presented based on the needs present in the development of the project (see Fig. 1).

Need		Metrica																						
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23		
	Resistant material	x																						
	Nice and interactive design		x																					
	Detect presence of objects			x																				
	Identify environment with artificial vision				x																			
	Suitable electronic components (sensor and microcontroller)					x	x							x										
	Intuitive user interface							x																
	Memory capacity to store data								x															
	Ensure connection of sensor, camera and alarm									x	x													
	Sufficient power supply for all circuits										x													
	Immediate alarm system upon sensor response											x	x											
	Easy maintenance of the electronic part													x										
	Easy maintenance of the structure														x									
	The system guarantees the integrity of the user															x	x							
	Affordable cost for the user																	x						
	Proper machine vision technique																		x			x		
	Systems connection (mechanical, electrical and control)																			x				
	Adequate mobility system																					x		

Fig. 1. Figure of list of requirements.

The structure must be resistant, it must have artificial vision to identify the environment and detect objects. It must have interactive design, adequate electronics, sufficient power supply, connection of sensors, cameras and alarms, memory capacity, intuitive interface, adequate mobility system, user integrity, affordable cost, easy maintenance, and immediate alarm system. Sensor range, camera range and field of view, low programming complexity and processing speed should also be considered.

2.3 System Architecture

The aim is to design and develop a baby surveillance system through a friendly mobile robot. Artificial vision, presence sensors, the SLAM system and an alarm system are used to warn of danger. This solution satisfies the need for comprehensive monitoring. To carry out the design, the following block diagram was made where the functional structure is shown (see Fig. 2).

The structure consists of three parts: information processing, robot locomotion, and mechanical-electrical design. In information processing, data is classified and processed in real time with the help of Jupyter. For locomotion a microcontroller, an H-bridge and sensors are used. The mechanical-electrical design from power activation to alarm

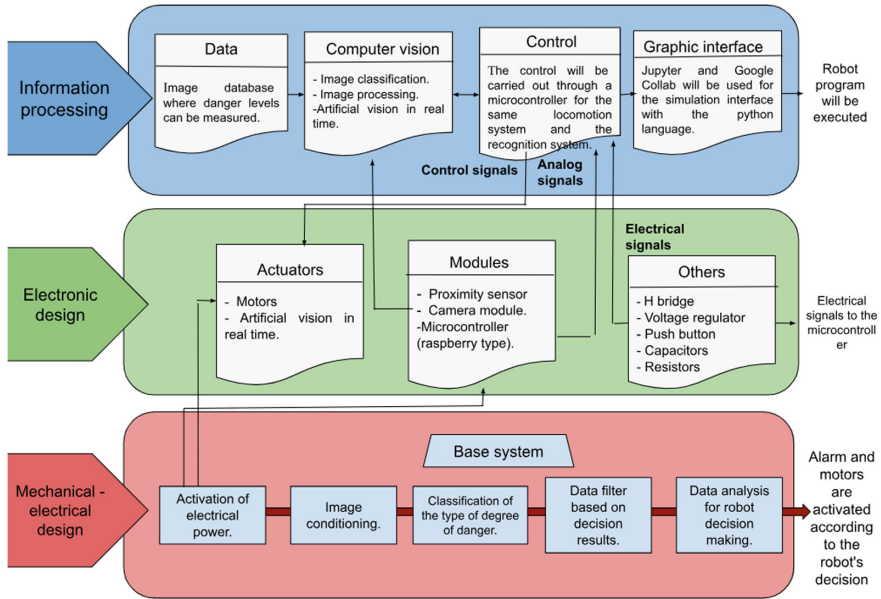


Fig. 2. Figure of function diagram.

activation, through hazard classification and data analysis. The realization of the system will be specified in the specialized design area.

2.4 Specialized Design Area

This specialized design area is divided into four distinct sections, as shown below:

Physical components and mechanical design. The design of the mobile baby monitoring robot is made of wood, which allows easy design, machining, and quick assembly. It is a safe and non-toxic material for contact with babies. The displacement system is by wheels with tapes since this mechanism is ideal for irregular surfaces.

The wheels are powered by a 6 V geared motor to control the speed of the robot. They are powered by lithium batteries for easy battery replacement and charging. The measurements are 50 cm high, 50 cm long and 30 cm wide (See Fig. 3).

Electrical design—electronic. The electronic scheme has a 7.4 V LiPo battery that will power the circuit, it also has a voltage regulator that will send 5 V as electronic output. Four HC-SR04 ultrasonic sensors will be used to have better accuracy when measuring distances and detecting the presence of objects by the 4 faces of the robot, this sensor will serve as part of the SLAM system localization and mapping of the robot within the baby’s environment. The signals from the sensors are sent to start the movement of the motors with the L293D, an H-bridge, which allows better precision and control. On the other hand, a high-resolution webcam, Argom Tech Cam 40 full HD 1080p FHD, whose images obtained will be processed in the Raspberry pi 4, and when it recognizes

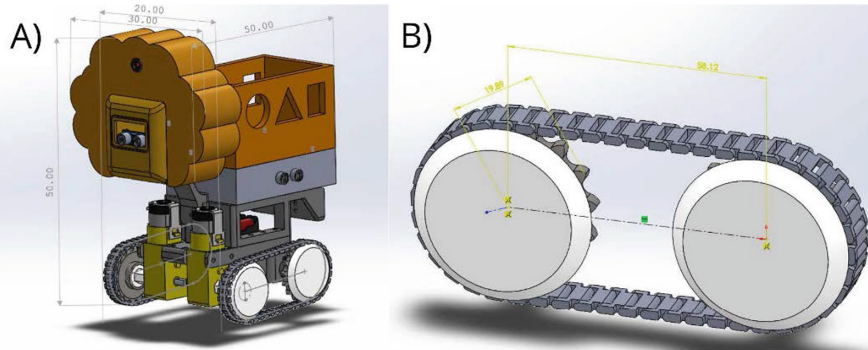


Fig. 3. In figure A) are the dimensions of the final prototype structure in SolidWorks and in B) is the transmission system.

a dangerous situation, it will activate the alarm buzzer to alert the family member or caregiver (See Fig. 4).

Image recognition. Image recognition is essential in a mobile robot with machine vision, using an FHD webcam to obtain high quality images. Training the recognition system is crucial to achieve maximum accuracy of the algorithm.

Data and systems analysis. Data were classified into 21 categories and ~ 8479 images were collected. The Google Image Downloader extension was used to extract the images from the Internet. The YOLOv7 system requires the items in Table 1 to function.

Storage requires at least 100 MB of space, which includes source code, configuration files, and files needed to run YOLOv7. For the interface, classification tests were carried out with Google Colab, while Jupyter was used for real-time detection. Tony will be used for the physical implementation of the prototype.

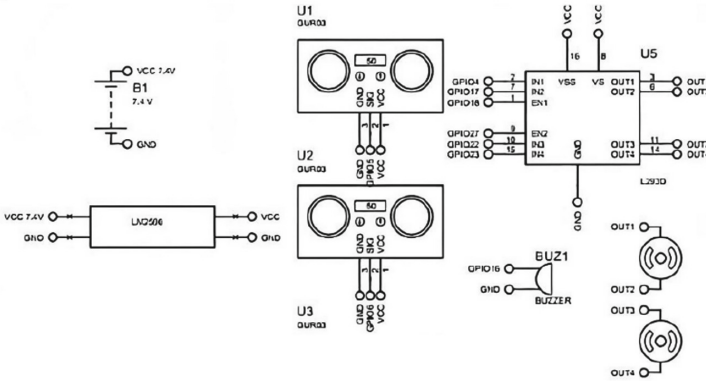
Application of the YOLO V7 algorithm. The application of the YOLO V7 algorithm is summarized by a flowchart as follows (see Fig. 5):

The YOLO V7 algorithm follows a defined workflow. The input image is pre-processed, relevant features are extracted using convolutional layers and object detections are performed with bounding boxes and confidence scores. Redundant detections are removed, and post-processing is performed to assign labels and calculate final confidence. Results include coordinates, labels, and confidence scores. The labeling process used more than 53 convolutional layers.

Explanation of the training code. Python and Google Colab are used to train an accurate neural network. A folder is created in Google Drive called “TheCodingBug” and the YOLO V7 architecture is cloned. The optimized YOLO v7x version is downloaded and training is started using a custom data file with specific parameters. The code is shown below (see Fig. 6).

The batch size is determined by the number of images and the processing speed. In the code, the location of the images, training data, labels, pre-trained model and hyperparameters are specified. The results of the trained model are available, and are stored in the “weights” folder in the file named “best.pt”, as illustrated in the following figure (see Fig. 7):

A)



B)

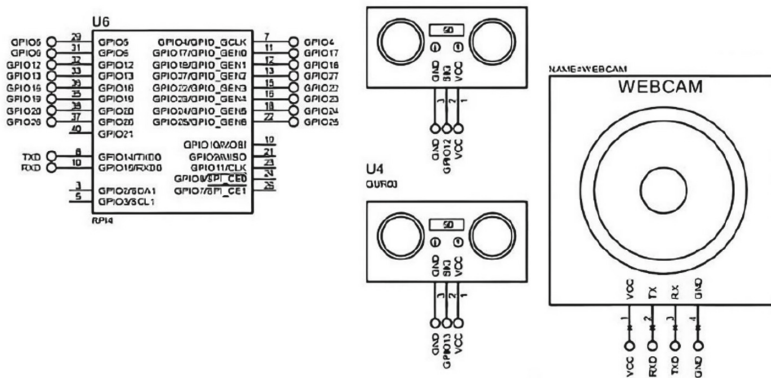


Fig. 4. In figure A) is the first part of the electrical schematic of the system and in B) is the second part of the schematic.

A performance test of the best training is performed using the code shown. The “detect.py” script is used with a trained model and specific weights. A confidence threshold of 0.3 is set and the size of the input images is defined. You can perform detections on images as shown, for video you would change it by—source video.mp4 and for detect on camera you would change it by this command source 0—no-trace (see Fig. 8).

Detection information, such as coordinates and classes, is obtained using the “pred” command.

Control design. The control system unifies the other systems: mechanical, electrical-electronic design and image recognition. This section explains the operation of the entire system, as well as the coding used for it.

Table 1. Table of system requirements.

Requirements	Details
Operating system	Windows
GPU	NVIDIA—NVIDIA GeForce RTX
RAM memory	8 GB RAM
Storage space	100 MB–10.5 GB
Software and libraries	Python and libraries such as OpenCV, NumPy and PyTorch
Interface	Google Colab—Jupyter—Thonny

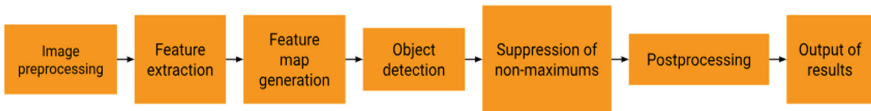


Fig. 5. Flowchart of the Yolo V7 technique.

```

    '!python train.py --device 0 --batch-size 16 --epochs 15 --img 640 640 --data
    data/custom_data.yaml --hyp data/hyp.scratch.custom.yaml --cfg
    cfg/training/yolov7x-custom.yaml --weights yolov7x.pt --name yolov7x-custom'
  
```

Fig. 6. Training code.

Epoch	gpu_mem	box	obj	cls	total	labels	img_size		
13/14	14.4G	0.03745	0.01894	0.01556	0.07194	17	640:	100%	345/345 [08:27<00:00, 1.47s/it]
Class		Images	Labels	P	R	mAP@.5	mAP@.5:1.95:	100%	93/93 [01:20<00:00, 1.15it/s]
all		2971	8638	0.508	0.499	0.479		0.26	
Epoch	gpu_mem	box	obj	cls	total	labels	img_size		
14/14	14.4G	0.03685	0.01849	0.01511	0.07045	33	640:	100%	345/345 [08:23<00:00, 1.46s/it]
Class		Images	Labels	P	R	mAP@.5	mAP@.5:1.95:	100%	93/93 [01:27<00:00, 1.06it/s]
all		2971	8638	0.499	0.539	0.504		0.274	
Escaleras	2971	385	0.381	0.317	0.308	0.121			
Tomacorriente	2971	301	0.555	0.877	0.859	0.572			
Bebe_llorando	2971	122	0.149	0.549	0.198	0.144			
Bebe_recostado	2971	535	0.477	0.903	0.708	0.39			
Bebe_feliz	2971	409	0.321	0.665	0.371	0.226			
Bebe_jugando	2971	177	0.516	0.247	0.288	0.119			
Bebe_asustado	2971	122	0.157	0.262	0.152	0.106			
Juguetes	2971	1062	0.439	0.417	0.401	0.223			
Cosas_puntiagudas	2971	349	0.501	0.143	0.239	0.0574			
Fuego	2971	170	0.626	0.518	0.569	0.205			
Comida_en_el_piso	2971	158	0.666	0.728	0.704	0.39			
Bebe_en_la_orrilla_de_la_escalera	2971		143	0.719	0.559	0.693			0.383
Bebe_junto_de_la_puerta	2971		132	0.663	0.894	0.854			0.564
Bebe_junto_a_cables	2971	122	0.707	0.656	0.738	0.545			
Bebe_comiendo_algo	2971	165	0.539	0.406	0.397	0.209			
Cables	2971	445	0.378	0.351	0.321	0.16			
Puerta	2971	370	0.414	0.797	0.744	0.386			
Personas	2971	976	0.457	0.808	0.689	0.318			
Bebe	2971	1772	0.645	0.868	0.793	0.441			
Barandas	2971	628	0.406	0.282	0.278	0.0968			
Bebe_en_barandas	2971	95	0.767	0.0696	0.29	0.103			

15 epochs completed in 2.453 hours.

Optimizer stripped from runs/train/yolov7x-custom6/weights/last.pt, 142.3MB
 Optimizer stripped from runs/train/yolov7x-custom6/weights/best.pt, 142.3MB

Fig. 7. Trained model.

```

"!python detect.py --weights runs/train/yolov7x-custom6/weights/best.pt --conf
0.3 --img-size 640 --source fig2.jpg --no-trace"

```

Fig. 8. Code to test the training.

Manual and automatic actuation. Manual drive is proposed in which the user can control the robot through an interface and automatic drive (see Fig. 9).

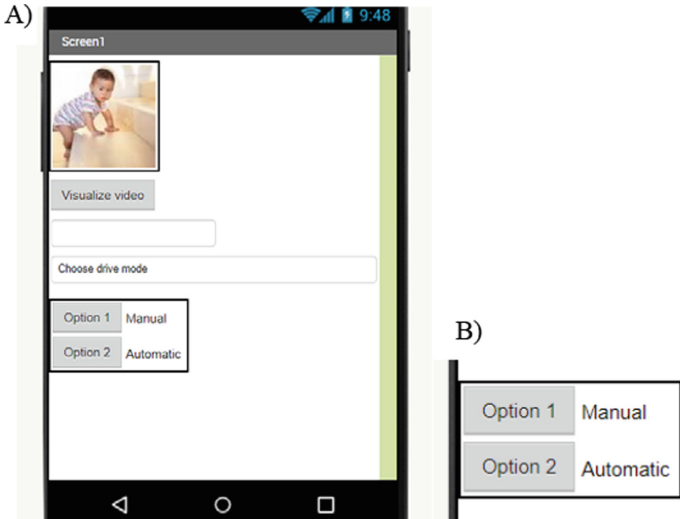


Fig. 9. Figure A) shows the general drive selection interface and B) shows the options more closely.

Manual. For manual operation, the robot was programmed to move following the baby and be controlled from a mobile application. The interface is shown in Fig. 10, for this the sensor and motor pins are configured, and the motor initialization. In addition, the distance sensor and the control function of the geared motor are configured for the direction of rotation of the motors (see Fig. 10).

Automatic. For automatic drive, the machine vision system of YOLO V7 is used, and the integration of sensors and the SLAM system. The automatic drive flowchart is shown in Fig. 11.

To configure the environment, libraries are installed, sensors and motors are connected, and component availability is checked. YOLO V7 is used for real-time detection, an alarm is triggered, and a text message is sent if there is danger. For the SLAM system, the extended Kalman filter and a PID controller are used. Everything is integrated into one main control circuit.

Configuration of automatic code execution with "autorun". The "autorun" on Raspberry Pi is achieved by following these steps:

1. Create a script file (autorun.sh) with the text editor: nano autorun.sh.
2. Add the command line to execute the main code, for example: python main_code.py.

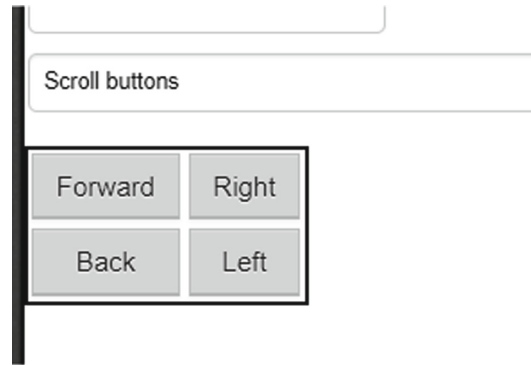


Fig. 10. Robot control interface.

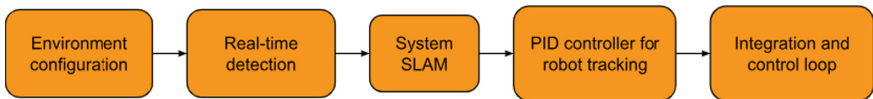


Fig. 11. Automatic drive flow diagram.

3. Grant execution permissions to the script: `chmod + x autorun.sh`.
4. Open the `rc.local` file: `sudo nano /etc/rc.local`.
5. Add the following line before "exit 0": `/complete_path_of_your_script/autorun.sh &` (replace the path with the actual location of the script).

With this, the `autorun.sh` script will run automatically at startup of the Raspberry Pi 4B, automating the execution of the main code efficiently.

2.5 Unification of Systems

The integration of the project system is presented below (see Fig. 12):

3 Results

The results obtained in the research are shown.

3.1 Mechanical Design

The figure below shows the simulation of the load stresses of the structure that was performed in the SolidWorks program with a force of 19.6 N, which complements the mathematical calculations previously performed. A static analysis of the box made of MDF material is carried out (see Fig. 13).

It also shows the static analysis of the configuration and the upper part of the base, which includes the electrical components, with a load of 19.6 N. The material used is wood (See Fig. 14).

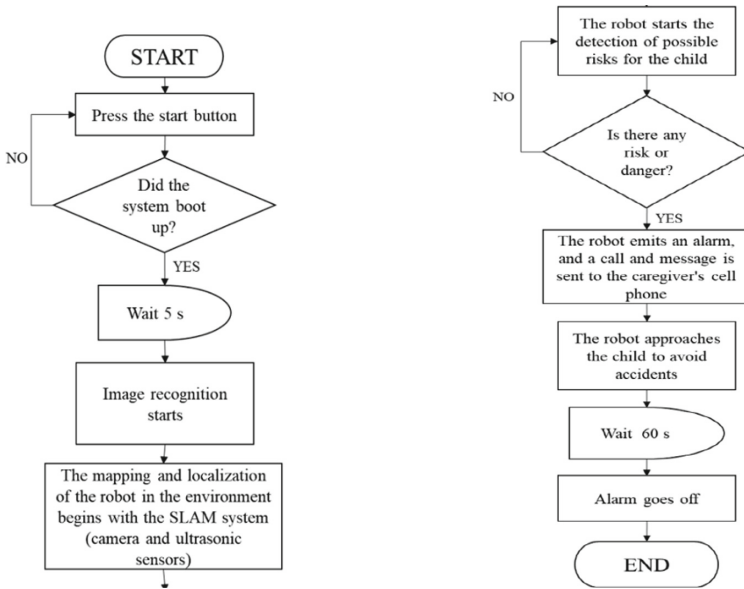


Fig. 12. In figure **A)** is the first part of the Unification of systems and in **B)** is the second part of the Unification of systems

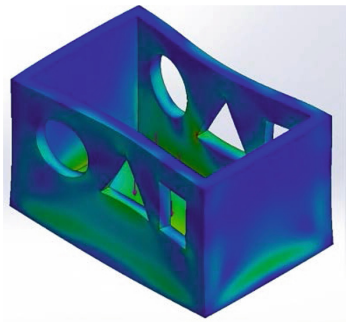


Fig. 13. Simulation of efforts in box.

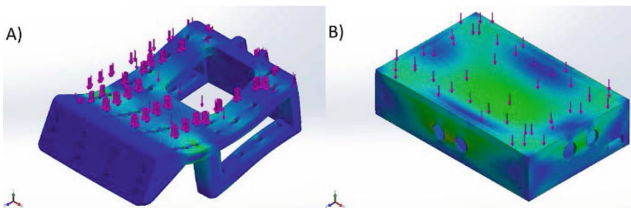


Fig. 14. Figure **A)** shows the simulation of stresses in the frame and in **B)** the stresses in the base.

The results of the mechanical design show that the correct choice of material allows to improve the child-robot interaction. The chosen material is hard enough for the purpose of the robot and resistant to the stresses submitted.

3.2 Electric Design

The connections of all electrical-electronic components are shown. The motors allow the movement of the locomotion system with belts, the L293D module is used to control the motors. The engine rotation simulation was made. On the other hand, the UART communication of the sensor with the microcontroller is observed, which allows the exchange of serial data. The Raspberry Pi 4 controls the distance values of the HC-SR04 ultrasonic sensors and the displacement of the locomotion system by controlling the direction of rotation of the geared motors located on the side of the robot and controls the webcam for image recognition.

3.3 Image Recognition

The results obtained from the classifications were saved in the “runs” results folder. In this folder, the specific results are found in folders named “exp_”. These folders are used to verify the correct operation of the system. Figure 15 shows what was obtained from the neural training through a video, with the recognized labels Bebe_acostado, Bebe and person.

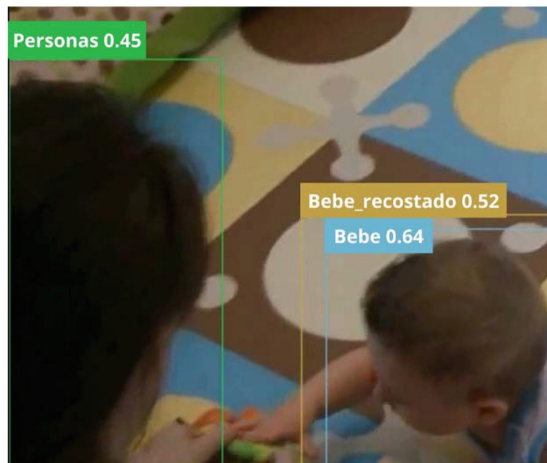


Fig. 15. Recognition of the label Bebe_acostado, Bebe and person.

3.4 Control Design

The programming for the sensor, the displacement system and the neural training of image recognition were carried out. The integration of all the systems was shown, using

the SLAM system it was possible to map and locate the robot in the baby's environment and to unify the data obtained from the camera and the sensors.

4 Conclusion

A mobile robot with a monitoring system for childcare was designed from the year on to reduce and avoid accidents or dangerous situations. This is thanks to the artificial vision system trained with Yolo V7, the ultrasonic sensor and the SLAM System. In this way, the robot can move to monitor the behavior of the child automatically or manually managed by the user or parent.

References

1. Jimenez, A.: Epidemiología y factores de riesgo de las lesiones por caídas en niños menores de un año. *Anales de Pediatría* **86**(6), 337–343 (2017)
2. OMS Child maltreatment. <https://www.who.int/news-room/fact-sheets/detail/child-maltreatment>. Accessed 10 April 2023
3. OMS Early childhood development. <https://www.who.int/news-room/q-a-detail/early-childhood-development>. Accessed 10 April 2023
4. Ramos, A., Nunes, L., Nogueira, P.: Fatores de risco de lesões não intencionais em ambiente doméstico/familiar em crianças. *Revista de Enfermagem* **3**(11), 113–123 (2013)
5. BID Cómo prevenir el maltrato infantil para erradicar la violencia en América Latina y el Caribe. <https://blogs.iadb.org/seguridad-ciudadana/es/como-prevenir-el-maltrato-infantil-para-erradicar-la-violencia-en-america-latina-y-el-caribe/>. Accessed 10 April 2023
6. Santana, R., Sanchez, R., Herrera, E.: El maltrato infantil: un problema mundial. *Salud Pública de México* **40**(1), 1047 (1998)
7. Defensoría del Pueblo, Vigésimo tercer informe Anual 2019. https://www.defensoria.gob.pe/wp-content/uploads/2020/05/InformeAnual_2019.pdf. Accessed 10 April 2023
8. Benavides, M., et al.: Los accidentes en los niños: un estudio en contexto de pobreza. *Avances de Investigación* **12**, 81 (2012)
9. Rodríguez, L.: El maltrato y el abuso sexual infantil en Atención Primaria de Salud. *Los pediatras: parte del problema y parte de la solución. Pediat. Integ.* **22**(4), 187–199 (2018)
10. La Vanguardia: La niñera robótica. <https://www.lavanguardia.com/tecnologia/20160930/41683795868/ninera-robotica-robot-cuidar-ninos.html>. Accessed 11 May 2023
11. Red-dot: Robot for Early Childhood Education Pudding BeanQ. <https://www.red-dot.org/project/pudding-beanq-12359-12356>. Accessed 11 May 2023
12. Cedano, M.: Reconocimiento de la agresión física con Deep Learning y visión artificial utilizando cámaras de video, caso observado Institución Educativa Rosa Suárez Rafael N°20436, Tesis de grado (2020)
13. Borja, M., Cabana, E., Montes, M.: Identificación facial para sistemas computarizados de control de acceso de personas utilizando redes neuronales. *Tecnia* **17**(2), 61–72 (2007)
14. Khan, T.: An intelligent baby monitor with automatic sleeping posture detection and notification. *AI* **2**(2), 290–306 (2021)
15. Chinlun, L., Lunjyh, J.: An intelligent baby care system based on IoT and deep learning techniques. *World Acad. Sci. Eng. Technol. Open Sci. Index Int. J. Electr. Commun. Eng. Int.* **133**(1), 81–85 (2018)
16. Hussain, T., Muhammad, K., Khan, S., et al.: Intelligent baby behavior monitoring using embedded vision in IoT for smart healthcare centers. *J. Artif. Intell. Syst.* **1**, 110–124 (2019)

17. Wang, C., Bochkovskiy, A., Mark, H.: YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *Comput. Vis. Pattern Recogn.* **12**, 7464–7475 (2022)
18. Towards Data Science: YOLOv7: a deep dive into the current state-of-the-art for object detection. <https://towardsdatascience.com/yolov7-a-deep-dive-into-the-current-state-of-the-art-for-object-detection-ce3ffedeeab>. Accessed 11 May 2023
19. Stanford children's. <https://www.stanfordchildrens.org/es/topic/default?id=cadas-estadstic-asdelesionesyatasdeincidencia-90-P06067>. Accessed 09 May 2023



Semantic Social Web Applications: Wiki Web

Aleksandr Belozеров^(✉) and Valentin Klimov

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute),
Kashirskoe Shosse 31, Moscow 115409, Russia
baa005@campus.mephi.ru, VVKlimov@mephi.ru

Abstract. Among a huge number of web applications today the most interesting are those that form social and semantic networks. The further development of the Internet will most likely be based on a combination of socially oriented and semantically oriented technologies. At the same time, it is assumed that, due to the wider distribution, it is socially-oriented technologies that will take advantage of the second ones. The goal of such a merger should be to turn the modern Web into a Semantic Web. Wikis, which are specialized social networks, allow users without knowledge of web development to create and collectively edit pages with any information, including semantic. In addition, users can create linked data in them by setting links to other pages in wiki or outside, and the most advanced ones work through bots based on the wiki API. Thus, wikis contain everything you need to collect semantic data and attract an audience interested in the semantization of the Web. This paper proposes a semantic social software that extends the functionality of the MediaWiki engine.

Keywords: Web · Social network · Semantic web · Semantic social web · Semantic search · Linked data · Collective hypertext · Wikitext · Folksonomy · Microformats · RDF · OWL · XML · SQL · SPARQL · MediaWiki

1 Introduction

1.1 Web Application Development Trends

The advent of new technologies and standards in web development and increased needs of mankind in improving the tools for managing data on the network have led to the emergence and parallel existence of two global directions for the development of web applications based on Web 2.0—social and semantic software.

Social Software is a term that refers to software applications and services that are designed to enable communication, collaboration and information exchange between people on the Internet. Socially oriented applications are based on network technologies, such as social networks, forums, blogs, online chats, wikis, and others, to support social interactions between users.

They can be used for a variety of purposes such as project collaboration, sharing knowledge and experiences, searching for information, communicating with friends and

colleagues, organizing events, and are widely used in business, education, scientific research, media industry, community service and other areas.

The term Semantic Software refers to software applications and services that use semantic technologies to process, analyze and understand information. Semantically oriented applications use computer algorithms and natural language processing techniques to determine the meaning and context of information, which can improve its search, analysis and use. To analyze the natural language used in text documents, e-mail, social networks and other sources of information in such applications, Natural Language Processing (NLP) technologies are used, and for its understanding ontologies and semantic networks are used—knowledge representation methods and relationships between them in the form of triples and graphs, which allows applications to better understand the relationships between objects and concepts in information [1].

Semantic-oriented applications are used for various purposes such as knowledge management, information retrieval, data analysis, process automation and are increasingly used in business, research, media industry, education and other areas where big data processing and analysis is required.

1.2 Possible Future of Web Applications

It is quite expected that sooner or later Web 3.0 will follow Web 2.0. This evolution or even revolution is being formulated by various researchers as the concept of the next generation of the Internet, which will be based on the latest technologies and protocols: The Internet of Things (IoT), and blockchain technologies, and Artificial Intelligence (AI) and Machine Learning (ML) [2]. All of them are united by the need to create distributed and decentralized systems that allow you to create decentralized applications and services, which are all successful and currently existing implementations of these technologies.

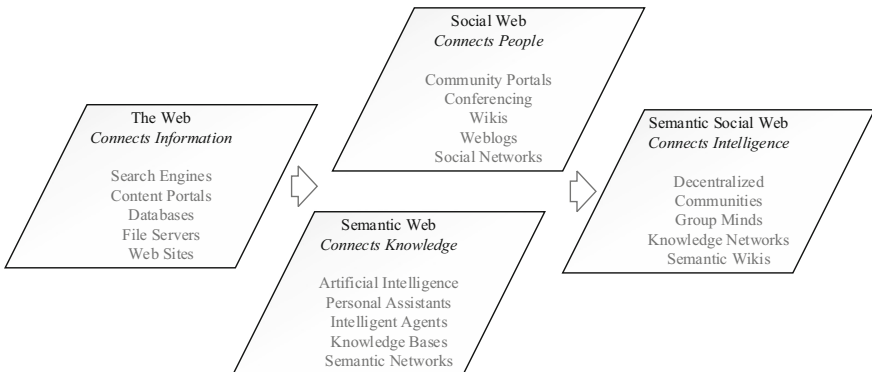


Fig. 1. Four stages of web applications evolution.

Our assumption about the possible further development of the Internet is based on the fact that throughout its development there is a continuity in the succession of the technologies used (see Fig. 1).

Thus, we should look for the emerging trend of leading one of the areas of interest to us—semantic and social—to the other, with the subsequent absorption of the lagging behind and the implementation of all its functionalities into itself. In other words, either semantized social networks or socially oriented semantic ones should come forward. And in our opinion, the widest ubiquitous dissemination of social technologies, which highly specialized implementations of semantic technologies will never achieve, despite the great age of this concept, has already determined the success in this race of the first version of the architecture of the future web.

2 Semantic Wiki

2.1 MediaWiki

Wikis are a type of web application that allows users to create, edit, and discuss site content collaboratively. Wikis are typically based on wiki engines such as Wikidata and allow users to create and edit wiki pages using a web browser.

Also, a wiki can be deservedly considered as an example of a socially oriented application that successfully implements the folksonomy of data placed both offline and on the Internet, whose web pages are collective hypertext [3]. Categorization by users of the data that they themselves host does not require users to have extensive knowledge of the hypertext markup language HTML, CSS, etc., as well as the use of specialized APIs, since a simpler wikitext markup language, also known as wiki markup or wikicode, is used for tagging and any other text processing and formatting. Wikitext is a special syntax that replaces complex HTML tags and other markup languages found in web development. It is processed by the wiki engine. The wiki markup may vary from engine to engine, however, the basic constructs for creating tags and links are generally similar across all wikitext dialects.

MediaWiki is free and open-source software for creating and managing wikis. It provides a number of features that allow users to create and edit wiki pages, add images and other media, and manage site permissions and settings [4].

2.2 Semantic MediaWiki

The first step to making MediaWiki more than a social networking engine for those who interested in web crawling, data verification, and dissemination of learned knowledge, transformed by the user's mind from the received data, is its extension under the acronym SMW. SMW or Semantic MediaWiki allows web application users to semantically annotate pages so that wiki content can be viewed, searched, and reused beyond standard search indexes.

The integration between MediaWiki and SMW is based on the extension mechanism for the engine: SMW is registered for certain events or requests, and MediaWiki calls SMW functions when necessary [5]. Its architecture is shown in Fig. 2.

SMW collects information about the concept presented on the page, not about the text associated with it. Each page belongs to an ontological element (including classes and properties), which can be further described using annotations on that page. The semantic roles that wiki pages can play are distinguished by namespaces.

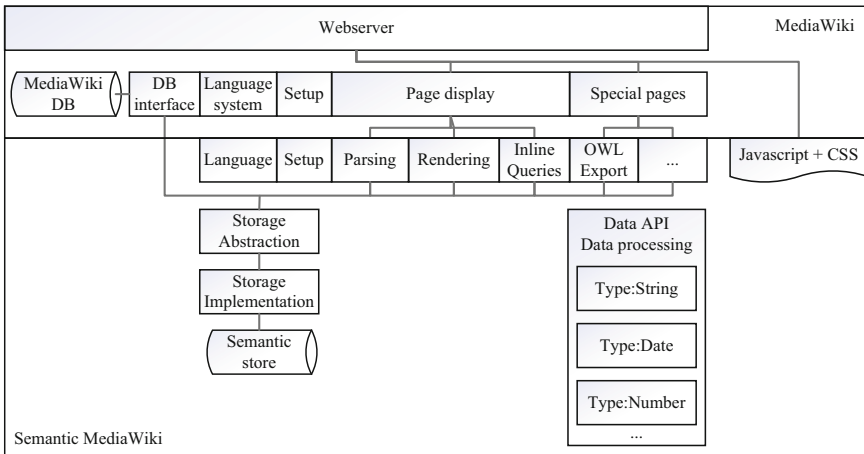


Fig. 2. MediaWiki and SMW architecture.

SMW collects semantic data using user-added semantic annotations, which are a Microformat of the wikitext. This microformat is a rethinking of the wiki markup, endowing the constructions (tags) already used in it with an alternative meaning, thus not violating its accepted standards, and implementing a semantic annotation.

SMW characterizes hyperlinks between wiki pages as properties (relationships) where the target of the link is the value of a user-provided property, thus implementing the concept of Linked Data.

Data defined by semantic properties can be exported in RDF format. The SMW data types, in this case, are converted to the corresponding data types from the XML Schema, and unique identifiers (URLs) are formed by appending suffixes to the wiki URL. It is also possible to explicitly specify which dictionaries (OWL-ontologies) should be used when exporting certain semantic properties. Semantic property values can be stored both in additional tables in the MediaWiki database and in the RDF store (Triplestore). Thus, it becomes possible to interact with wiki data through the SPARQL language and use semantic search engines and inference on RDF data.

The presence in the MediaWiki engine with the SMW extension of a decentralized folksonomic approach to organizing space in a social wiki network, support for vocabularies and ontologies ratified by the wiki community, as well as preference for microdata, RDFa and other HTML extensions of its microformat, which by its nature does not require fundamental changes with on the part of web developers and web administrators of such a network, makes SMW the most promising superstructure over the existing Internet, which has the potential for gradual absorption or assimilation with it (it is still difficult to predict what exactly) with the formation of the Tim Berners-Lee Semantic Web or Web 3.0 [6].

3 Semantic Social Software

3.1 Proposed System

These observations led us to propose a system based on the MediaWiki wiki engine that combines the functionality of social networks such as Facebook and Wikipedia with semantic data processing capabilities such as Linked Data and Semantic Web technologies. Such a set of Semantic Social Software (S3) can help improve collective knowledge, collaboration and knowledge sharing, which in turn can lead to more effective problem solving [7].

Semantic social software is a set of technologies that combine the capabilities of semantic information processing and social interaction of users. It allows you to use information resources containing data, analyze the content of information to collect it, take into account the context and interaction between users, and together with users create information resources containing related data based on this data.

Among the possibilities of semantic social software are the following:

- Automatic information processing: using semantic technologies to analyze and understand the content of information, which allows you to automatically classify, organize and extract knowledge from large amounts of information.
- Recommendations and personalization: using information about user preferences and interests to create personalized recommendations and promotional offers.
- Integration with other systems: the ability to integrate with other systems, such as CRM and etc., to provide a more efficient user experience.
- Knowledge management: use to create knowledge bases and knowledge management systems that enable organizations to retain and use the knowledge and experience of their employees.
- Data management: manage large amounts of data, including structured and unstructured data. This allows organizations and communities to better manage their data and use it more effectively.

To provide the above functionality in an SMW-based system, it is necessary to implement at least the following software modules that integrate with MediaWiki, the architecture and description of which will be presented below.

3.2 Search Robot

Wikis are folksonomies, and the amount of actual information on the Internet that it makes sense to wikify (add to the wiki engine's databases for further formatting) far exceeds the amount of data that the wiki community can process before this information is will lose relevance and require re-applying.

Also, at the moment, the wiki community is paying more attention to Wikidata, so at the initial stages, the number of active audiences of the modified engine will be even smaller. In this regard, for the initial filling and initial updating of data, the use of search robots (bots, crawlers) for wikification is essential.

Crawlers in S3 are used to automate many tasks such as adding, updating, and deleting data. They can also perform many operations including:

- Adding new data elements, properties, and property values.
- Automatic creation of categories and relationships between data items.
- Automatic generation of reports and statistics.
- Removing obsolete data.
- Checking the correctness and reliability of data.

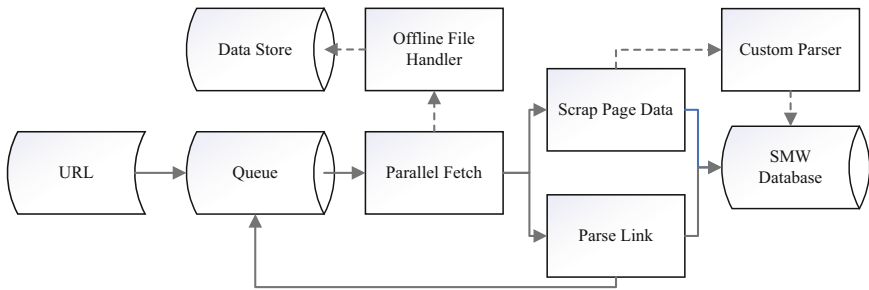


Fig. 3. High level MediaWiki web crawling process architecture.

Bots in S3 work on the basis of an application programming web interface (API), which allows them to interact with the SQL database and enter information collected on the network into it—converting them into RDF triples and placing data in the triplestore is handled by the SMW engine (see Fig. 3). All other automation work with this data can be done using SPARQL queries directly to the data warehouse in RDF. Bots can improve the quality of data on wikis by adding new data and removing obsolete data. In general, the use of bots in S3 is an important tool to ensure that the information in the database is accurate, up-to-date and complete.

In addition, it is important to take into account the experience of using Wikidata crawlers in order to learn from the mistakes and apply them in creating semantic wikis using MediaWiki. The studies cite three main problems with the high volume of bot edits on Wikidata [8]:

- The ratio of the amount of information imported by bots and the number of human participants leads to the fact that Linus's Law does not work in this project.
- The set of sources used by bots is much narrower than the one used by humans—it can be a serious threat to the representation of a wide range of viewpoints in wiki.
- A huge percentage of bot edits and the multilingualism of community members can limit the participation of existing and the influx of new users into the project.

As a temporary measure to counteract the above problems, S3 proposes the creation of an additional semantic service category for data collected by crawlers and its inclusion in it for subsequent reconciliation, as well as the ability to change the display in the web interface of the value of any element within the text of a wiki article.

It will take place between custom and automatic edits—also for subsequent verification by a member of the wiki community or by a simple user at their discretion.

3.3 Template Engine

Despite the fact that Semantic MediaWiki implements the display of semantic relationships and data in the form of graph and other structures on web pages at the plugin level, this solution has two significant drawbacks.

Firstly, plugins that implement this functionality are difficult to learn and are intended primarily for system developers or advanced users who can use a complex interface to solve their problems. Thus, the use of Semantic MediaWiki is limited, and it can be difficult for general users who do not have sufficient experience with the software. Secondly, SMW plugins are not suitable for ubiquitous distribution on users' end devices, as they require intervention in the installed software. This limitation contradicts the chosen concept of the development of the Semantic Web, which assumes the use of only existing data description standards and the expansion of their functionality with microformats.

Thus, despite some advantages of the functions available both in the SMW core and in the form of plugins, extending their use to the entire wiki community and those who are not indifferent to it has its limitations and disadvantages that make it difficult to use it on such a large scale. To achieve greater accessibility and usability of semantic data, it is necessary to be able to generate hypertext templates using HTML microformats based on the data contained in the SMW semantic data store.

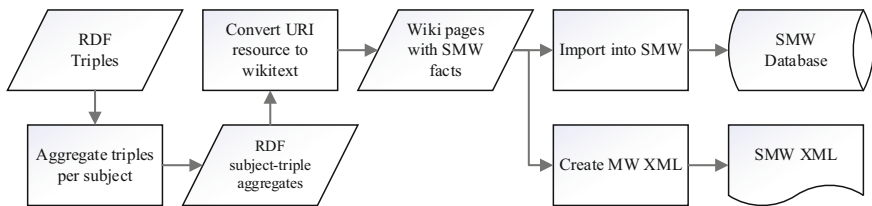


Fig. 4. MediaWiki data conversion process architecture.

To generate and display wiki pages, the semantic data from the RDF storage will be pre-converted by the wiki engine to XML and transferred to the SQL storage. In S3, RDF triplets are aggregated by subject via a SPARQL access point, then converted into a single XML-like template, converting all URIs into wiki page names for new pages and page links. Then, using this template and the original web page from which the data was received, identical web pages are generated when rendered in a web browser, but with semantic markup (see Fig. 4). They will act as a replacement or alternative to those from which the original data was extracted.

In this case, Web 2.0 users will have the opportunity to observe almost in real time not only the original web pages of Internet resources, but also their alternative versions containing information from MediaWiki community.

In theory, such a user experience should lead to the fact that over time, network users will begin to give preference to just such a collective hypertext, and the wiki will not only exist in parallel with the Web, but also gradually construct a Semantic Web around it until it becomes her in one piece.

3.4 Conclusion

Semantic integration is the process of combining and integrating data from different sources using semantic technologies and standards. It is aimed at solving the problem of data heterogeneity that arises when working with data stored in different formats, syntaxes and structures, as well as having different semantic interpretations.

In this case, classical hypertext and collective hypertext interspersed with wikitext act as heterogeneous data. Wiki markup use is inevitable, due to the fundamental impossibility of creating a single representation of semantic data that is convenient for both developers and ordinary users, so the semantic social web application considered in the work is intended to act as an intermediate link between those who consume content and those who semantically mark it up. Semantic integration of Internet resources based on wiki sites is what is proposed for implementation in this paper.










The organization of the Semantic Web analysis, structuring and transformation of data into a single format using a common semantic model. The wikis that use it may in the future become part of the Semantic Web, or the Wiki Web. Based on Semantic Social MediaWiki, it should be a rethinking of the structure of the current Internet, where all data can be easily processed both by a computer and a person.

References

1. Bezverhny, E., Dadteev, K., Barykin, L., Nemeshaev, S., Klimov, V.: Use of chat bots in learning management systems. *Proced. Comput. Sci.* **169**, 652–655 (2020)
2. Ghelani, D., Hua, T.: Conceptual framework of Web 3.0 and impact on marketing, artificial intelligence, and blockchain. *Int. J. Inform. Commun. Sci.* **7**, 10–17 (2022)
3. Tomuro, N., Shepitsen, A.: Construction of disambiguated folksonomy ontologies using Wikipedia. In: *Proceedings of the 2009 Workshop on The People’s Web Meets NLP: Collaboratively Constructed Semantic Resources (People’s Web ‘09)*. Association for Computational Linguistics, USA, pp. 42–50 (2009)
4. MediaWiki Man page. <https://www.mediawiki.org/wiki/Manual>. Accessed 10 June 2023
5. Krötzsch, M., Vrandečić, D., Völkel, M.: Semantic MediaWiki. In: *Proceedings of the 5th International Semantic Web Conference (ISWC-06)*, pp. 935–942 (2006)
6. Belozarov, A.A., Klimov, V.V.: Semantic web technologies: issues and possible ways of development. *Proced. Comput. Sci.* **213**, 617–622 (2022)
7. Schaffert, S.: Semantic social software: semantically enabled social software or socially enabled semantic web? In: Rech, J., Decker, B., Ras, E. (eds.) *Emerging Technologies for Semantic Work Environments: Techniques, Methods, and Applications*. IGI Global, Hershey, pp. 33–46 (2008)
8. Piscopo, A.: Wikidata: A New Paradigm of Human-Bot Collaboration? *ArXiv abs/1810.00931* (2018)



Development of a Network Traffic Anomaly Detection System Based on Neural Networks

Natalia Bespalova¹ , Alexey Ershov² , Sergey Sitnikov² , Sergey Nechaev³ ,
Margarita Vanina⁴ , Victor Radygin⁵ , Dmitry Kupriyanov⁵  ,
and Mikhail Ivanov¹ 

¹ Financial University Under the Government of the Russian Federation, Moscow, Russian Federation

² Yuri Gagarin State Technical University of Saratov, Saratov, Russian Federation

³ ITMO University, Saint Petersburg, Russian Federation

⁴ Moscow Technical University of Communications and Informatics, Moscow, Russian Federation

⁵ National Research Nuclear University “MEPHI”, Moscow, Russian Federation
DYKupriyanov@mephi.ru

Abstract. Ensuring information security using artificial intelligence technologies is one of the most promising areas in the modern world. The use of an artificial intelligence-based cyber incident response system will provide the necessary support while simultaneously processing a large number of events, automating the routine actions of analysts and ensuring timely response to incidents without human intervention. As a result of the work, neural networks created on the basis of various variations of input parameters were trained and tested. The resulting solutions handle incoming network traffic on their own, informing security administrators of any unusual network behavior.

Keywords: Artificial intelligence · Neural networks · Intrusion detection · Computer security · Anomaly detection

1 Introduction

Network intrusion detection refers to the problem of monitoring and differentiating such network flows and activities from normal expected network behavior that can adversely affect the security of information systems. The search by governments and organizations for reliable solutions to protect their information assets from unauthorized disclosure and illegal access has brought intrusion detection and prevention to the forefront of information security.

Ensuring information security using artificial intelligence technologies is one of the most promising areas in the modern world. The use of an artificial intelligence-based cyber incident response system will provide the necessary support while simultaneously processing a large number of events, automating the routine actions of analysts and ensuring timely response to incidents without human intervention [1, 2].

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (CNNs), are a subset of machine learning algorithms and serve as the foundation for deep learning algorithms. Artificial Neural Networks (ANNs) are composed of nodes forming layers: an input data layer, one or more hidden layers, and an output data layer. Each node (artificial neuron) is connected to other nodes with a certain weight and threshold value. If the output of any node exceeds the threshold, then this node is activated and sends data to the next layer of the network. Otherwise, the data is not passed to the next layer of the network.

Training data is used to train and improve the accuracy of neural networks. Learning algorithms turn into powerful computational tools when they reach the required accuracy, which allows them to be used to classify and cluster data at high speed [3, 4].

2 Formulation of the Problem

The network traffic anomaly detection system or intrusion detection system (IDS) is used as one of the necessary levels of information systems protection. Currently, IDS are being developed in the form of software or hardware-software variants that monitor and verify situations and incidents that occur within the system. The system also processes information, checking without human intervention that there are no signs of problematic outcomes. Due to a significant increase in the number of various sources and the variety of illegal penetrations into the network, the load and responsibility on the system have increased [5–7]. The number of objects that need to be monitored has increased, which has led to the need to expand the databases that the IDS analyzes. From all of the above, an increase in the urgency of the task of reducing the load on the analysis subsystems follows. In the process of processing, the parameters that are monitored and that deviate from the norm are identified; calculation of the efficiency coefficient of the analysis subsystem—an indicator of the speed and reliability of detecting and suppressing cases of information security violations. To increase the efficiency of incident detection, it is necessary to reduce the amount of raw, unprocessed data from the information collection subsystem, which is part of the IDS structure.

The task of detecting information security incidents can be reduced to the detection of data belonging to a certain group on the provided data array. Therefore, the ability to solve classification problems is the main criterion for choosing a detection method.

At the moment, there are not many solutions based on the use of artificial intelligence in the field of information security. Table 1 provides a comparative analysis of some of the nicknames.

3 The Solution of the Problem

The solution developed in the course of this work takes into account the shortcomings of existing solutions, which made it possible to increase the speed of training the neural network, while eliminating the need for information security specialists. However, this led to a slight loss of training accuracy and a decrease in the total amount of functionality.

Table 1. Comparative analysis of existing solutions

Solution	Security vision	Deep instinct prevention platform
Possibilities	Known incident detection	Identifying malicious files
		Ransomware protection
	Detection of atypical suspicious events	Protection against both file-based and non-file attacks
		Support low false positive guarantee
Flaws	Mandatory availability of information security specialists	The use of neural network deep learning

The implementation of the neural network is carried out using the Java language and the Neuroph library. Neuroph is a lightweight Java neural network framework for developing general neural network architectures. It contains a well-developed open source Java library with a small number of base classes that correspond to the basic concepts of the neural network [8, 9].

To train the neural network in the presented work, the method of error back propagation is used, which is based on gradient descent. Gradient descent is a way of finding the local minimum or maximum of a function by moving along a gradient. To demonstrate the use of a gradient in neural networks, a graph is plotted where the weights of neurons (w) are plotted along the abscissa, and the error corresponding to this weight (e) is plotted along the ordinate.

On the chart, you need to find the global minimum—the point (w_2, e_2). The gradient descent method helps to find this point (yellow on the graph indicates the gradient). For each weight in the neural network, you need to find the global minimum using your own graph and gradient [10, 11].

After advancing forward through the structure of the neural network, an output value is calculated for each neuron based on the input data, weights, bias, and activation function. When the output layer is reached, it is necessary to calculate the error, and then, based on it, recalculate the value of the weights for each neuron.

The recalculation of weights is considered based on the Widrow-Hoff rule, in which the error parameter δ is introduced.

For the output layer, δ is calculated by formula (1):

$$\delta_o = \text{OUT}_{req} - \text{OUT}_{actual} * f'(IN) \quad (1)$$

where OUT_{req} —desired result, OUT_{actual} —received output value, $f'(IN)$ —derivative of the activation function from the input value of a given neuron.

For neurons in hidden layers, δ is calculated by the formula (2):

$$\delta_h = f'(IN) * \sum (w_i * \delta_i) \quad (2)$$

Weight change is calculated by the formula (3):

$$\Delta w_i = \eta * \delta_w * x_i + \alpha * \Delta w_{i-1} \quad (3)$$

where η —speed (norm) of learning, δ_w —current neuron error, x_i —neuron output value, α —moment: a parameter that allows you to avoid errors with local minima.

Taking into account (3), the weight of the neuron is recalculated. Changing the weight of neurons can be done by different methods (stochastic, batch or mini-batch). Based on the concept of a Project Object Model (POM), Maven can manage project build, reporting, and documentation from a central piece of information [12–14].

In the project under consideration, the data set used for the international competition of tools for discovery and data mining was chosen as a dataset, which was held in conjunction with KDD-99—an international conference on knowledge discovery and data mining. The objective of the competition was to build a network intrusion detector, a predictive model capable of distinguishing between ‘bad’ connections, called intrusions or attacks, and ‘good’ normal connections. This database contains a standard set of auditable data that includes a wide range of intrusions simulated in a military network environment [15–18].

In the selected dataset, a network connection is defined as a sequence of packets. Data is transferred from the source IP address to the destination IP address under a predefined protocol (TCP, UDP, ICMP). Exception types are divided into 4 categories, 39 attack types.

Exception types: DOS; unauthorized access from a remote machine to a local machine; U2R unauthorized access to local superuser (root) privileges; port monitoring or scanning.

In the selected dataset, information is presented in the form of strings. For example: 0, tcp, smtp, SF, 590, 326, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 165, 153, 0.93, 0.01, 0.01, 0.00, 0.00, 0.00, 0.00, 0.00, normal.

Each comma-separated value carries information about the connection.

In order for the neural network to correctly process data from the dataset and increase the rate of convergence of algorithms, the information must be normalized, i.e. the values must range from 0 to 1.

Initially, the dataset contains both string values and numerical indicators that are different from the specified range.

To bring the data into a normalized form, the method of normalization by the maximum value is used in each connection function. After finding the largest value for each parameter, each element is divided by this number.

$$x_n = x/x_{max} \quad (4)$$

where x_n —normalized value, x —input value, x_{max} —maximum value from all parameters of the same type.

To convert literal values into a normalized form, they are initially numbered with numbers starting from 0, after which the normalization algorithm is applied to them.

To format data, the auxiliary method `formatData` (String data) from the `DataSetUtil.java` class is used, to which the input string is passed, the data in which is separated by a comma. Using the functionality of the Java language, the input value is converted to an array. For text parameters, transformation mechanisms from the `ProtocolType.java`, `ServiceType.java`, and `FlagType.java` classes are used. Numeric parameters are cast to the Java double type, which implements floating point numbers. After processing the

input parameters, the output parameters should be converted into a form understandable for the program.

In the source data, the output value is represented as a string parameter, described by the name of the attack or by the normal constant for normal traffic behavior. To convert these names into a set of output neurons, it is necessary to enumerate them with numbers, starting from zero, and present them as an array of floating point numbers, in which all elements are equal to 0, and 1 in a certain position corresponds to the ordinal number of the result.

For the program implementation of the output data conversion, a Java Enum type was created, called `IntrusionType.java`, which contains constants, their numerical values for conversion into a form understandable by the neural network, and auxiliary methods for converting data and finding the desired parameter.

To improve the learning process of the neural network, the data must be submitted not in an ordered form, as in the original file, but pre-mixed to eliminate the situation of accumulation of identical examples. Using the capabilities of the Java programming language, a list of strings is created from a normalized data set written to a file, which, using the random distribution algorithm for the elements of the collection, creates a new order of records. After that, the updated version of the information placement is written to a new file.

After preparing the data for training, it is necessary to set the parameters of the neural network: the number of input and output neurons, as well as the number of hidden layers and neurons on each.

The number of input neurons is determined by the number of known input parameters - 41. However, this value is quite large, which will increase the network training time.

After analyzing the data for training the neural network, it was found that only for 4 types of attacks (ipsweep, neptune, satan, smurf) there are enough examples in the training dataset, so the number of output neurons is equal to 5, taking into account the normal behavior of the network.

3.1 IP Address Sweep

The purpose of the attack is to discover hosts on the victim's network. The attacker sends an icmp-echo-request and waits for some host to respond to the request and find itself.

The attack is fought by setting a time interval during which the attacker is allowed to send from one source no more than a certain number of icmp requests to different addresses. If during this interval a request is received for more addresses, then this source is marked as the address from which the attack is coming, and all subsequent icmp packets will be discarded. Packets will be rejected until the intensity of icmp requests falls to the allowed values. Initially, the interval value is 5000 microseconds (0.005 s). Such a rule is necessary only if the security policies allow the passage of icmp echo requests.

3.2 Smurf

The attacker sends a fake ICMP Echo packet to the broadcast address. In this case, the address of the source of the packet is replaced by the address of the victim in order to

“substitute” the target system. Since the echo packet is sent to the broadcast address, all machines in the amplifying network return their responses to the victim. A good solution to prevent the amplification effect is to disable the forward broadcast operation on all edge routers. Additionally, it is worth setting the mode of “silent” rejection of ICMP broadcast echo packets in the operating system.

3.3 Neptune

This attack is also known as the half-open TCP SYN attack. To launch a successful attack, the attacker exploits the shortcomings of TCP’s three-way handshake protocol by continuously sending a large number of consecutive fake SYN (connection requests) packets directly to the TCP server. So the TCP server creates a half-open TCP connection in a finite length queue with a limited timer and sends back a SYN/ACK (sync/acknowledge) packet and waits to receive an acknowledgment packet from the client, but that client (attacker) never responds with an ACK packet (the last step in the three-way handshake) because the source IP address is spoofed so that the TCP server keeps half-open connections (records) open, eventually draining the TCP server’s resources. The purpose of this attack is to reject any new connection from an authorized TCP client.

3.4 Satan

Satan is a tool designed to check a computer system for security loopholes (Security Administrator Tool for Analyzing Networks). However, running it on someone else’s device is considered an attack. SATAN identifies weaknesses in network software configuration, identifies running network services and provides information about hardware and software types, checks for vulnerabilities in TCP/IP hosts using common TCP/IP protocols.

The number of hidden layers and neurons on them is determined empirically, so 2 hidden layers with 20 neurons each were initially chosen.

Using the mechanisms of the Neuroph framework, a normalized and mixed dataset is loaded, a neural network is created with previously defined settings. Since the network uses the error backpropagation method when recalculating the weights at each epoch, it needs to set the boundary conditions: the maximum number of training epochs and the error value at which training is considered successful. The work used 10000 epochs and the allowable error is 0.01.

After setting all the necessary settings, the training of the neural network starts. For clarity of the process, the number and result of each epoch are displayed on the user’s screen. To programmatically implement this function, the Neuroph framework method added an end-of-epoch handler.

After training, the neural network is saved to a file with the extension “.nnet” Before testing the neural network, it is necessary to normalize the test dataset by analogy with the training set. Using the tools of the Neuroph framework, a previously saved.nnet file is loaded, data from the test set is passed line by line to the network input parameters, after which the process of processing them and issuing the result is carried out. For the convenience of the user, information about the results obtained is saved and displayed on the screen after the work with the dataset is completed. The resulting information contains

the number of correct, incorrect, and incorrect totals, all combinations of expected and actual results obtained, and the percentage of correct results to the total number of records.

After the first run with the initially selected parameters, the percentage of correctly identified situations was 99.96%. However, the running time of the program is too long, and the formation of a normalized dataset with all parameters requires a lot of memory. To optimize the work, an experiment was carried out with a decrease in the number of input parameters and maintaining the percentage of successful determination. 12, 25 and 33 output parameters were chosen for testing. The results of the experiment are presented in Table 2.

Table 2. Results of the experiment for 4809495 records

Number of input parameters, pcs	Percentage of correct determination, %	Time of processing s	Number of errors, pcs		
			Error not found	False alarm	Found another error
12	99.41	87.960	6135	8044	14148
25	99.73	326.463	1713	8113	3158
33	99.86	485.455	1681	4987	19
41	99.96	766.243	1471	411	1

With the number of input parameters equal to 33, the percentage of correct determination is closest to the variant with a full set of input data, but the memory load and waiting time are less. In this case, the processing time for 4809495 records is reduced by 280.788 s or 4.7 min. In the case where the number of input parameters is 12, the number of incorrectly determined cases is 21640 more than with 33, but the processing time for all records is 6.6 min less. With 25 input parameters, both indicators are between the same, which makes this option the most optimal.

Studies have also been carried out on changing the hidden layers. The best result was shown by a network with two hidden layers, in which the first is the number of input neurons multiplied by 4, and the second is multiplied by 2.

All errors are divided into 3 categories: undetected, falsely detected and misidentified. Situations when the error is not detected were the least for 12 and 25 input parameters; as their number increases, the fact of detecting an error does not decrease as much as the number of errors in attack detection and false positives. Incorrect error detection is a consequence of the fact that ipsweep and satan are present in much smaller numbers relative to the rest. If you provide the neural network with enough data, the result will be even higher.

4 Conclusions

Thus, as a result of training the neural network, 4 different options were obtained. The differences are determined by the inverse dependence of the number of input parameters on the data processing time. The network with 41 input parameters showed the best percentage of detection, while its operation time, as well as the memory load, were the highest. The opposite result was shown by a network with 12 input parameters.

Various configurations of neural networks make it possible to adapt them to various operating conditions. For example, a network with 12 inputs provides the best runtime while slightly reducing detection performance. Such a network can be used when it is necessary to quickly communicate with the host. However, if time is not a critical parameter for the functioning of the system, then a configuration with a large number of input parameters can be used to increase security.

Further improvement of the implementation is possible with an increase in the amount of data on each of the threats. Adding new attack variants does not require major changes to the program code—just add the name of the attack to the name recognition algorithm; data about the behavior of the network into the training sample, as well as retrain the network with new input data.

As a result of the work, neural networks created on the basis of various variations of input parameters were trained and tested. The resulting solutions handle incoming network traffic on their own, informing security administrators of any unusual network behavior. Further improvement will be able to respond to some types of attacks without human assistance, which will significantly speed up the process of processing and resolving incidents, as well as reduce the burden on human resources.

References

1. Diamanti, A., Vlchez, J., Secci, S.: An AI-empowered framework for cross-layer softwarized infrastructure state assessment. *IEEE Trans. Netw. Serv. Manag.* **19**(4), 4434–4448 (2022)
2. Krakhmalev, O., Korchagin, S., Pleshakova, E., Nikitin, P., Tsibizova, O., Sycheva, I., et al.: Parallel computational algorithm for object-oriented modeling of manipulation robots. *Mathematics* **9**(22), 2886 (2021)
3. McMahan, H., Moore, E., Ramage, D., Arcas, B.: Federated learning of deep networks using model averaging. *CoRR abs/1602.05629* (2016)
4. Ahmed, A., Mahmood, A., Hu, J.: A survey of network anomaly detection techniques. *J. Netw. Comput. Appl.* **60**, 19–31 (2016)
5. Pavlyutin, M., Samoyavcheva, M., Kochkarov, R., Pleshakova, E., Korchagin, S., Gataullin, T., Hidirova, M.: COVID-19 spread forecasting, mathematical methods vs. machine learning. Moscow case. *Mathematics* **10**(2), 195 (2022)
6. Bontemps, L., Cao, V., McDermott, J., Le-Khac N.A.: Collective anomaly detection based on long short term memory recurrent neural networks. In: *International Conference on Future Data and Security Engineering*, pp. 141–152. *Proceedings. Lecture Notes in Computer Science* (2016)
7. Canizo, M., Triguero, I., Conde, A., Onieva, E.: Multi-head CNN–RNN for multi-time series anomaly detection: an industrial case study. *Neurocomputing* **363**, 246–260 (2019)
8. Gataullin, T., Gataullin, S.: Endpoint functions: mathematical apparatus and economic applications. In: *Math Notes vol. 112*, pp. 656–663. Springer Nature, New York (2022)

9. Zhang, J., et al.: A secure and lightweight multi-party private intersection-sum scheme over a symmetric cryptosystem. *Symmetry* **15**(2), 319 (2023)
10. Yerznkyan, B., Gataullin, T., Gataullin, S.: Mathematical aspects of synergy. *Mont. J. Econ.* **18**(3), 197–207 (2023)
11. Gataullin, T., Gataullin, S., Ivanova, K.: Modeling an electronic auction. In: “Smart Technologies” for Society, State and Economy. ISC 2020. *Lecture Notes in Networks and Systems*, vol. 155, pp. 1108–1117. Springer, Cham (2021)
12. Yerznkyan, B., Bychkova, S., Gataullin, T., Gataullin, S.: The sufficiency principle as the ideas quintessence of the club of Rome. *Mont. J. Econ.* **15**(1), 021–029 (2019)
13. Ivanyuk, V.: Forecasting of digital financial crimes in Russia based on machine learning methods. *J. Comput. Virol. Hack. Techn.* **12**, 1–14 (2023)
14. Boltachev, E.: Potential cyber threats of adversarial attacks on autonomous driving models. *J. Comput. Virol. Hack. Techn.* **128**, 1–11 (2023)
15. Andriyanov, N., et al.: Intelligent system for estimation of the spatial position of apples based on YOLOv3 and real sense depth camera D415. *Symmetry* **14**(1), 148 (2022)
16. Ekhlakov, R., et al.: Modeling the chemical pollution of the area by the random-addition method. *Fractal Fract.* **6**(4), 193 (2022)
17. Kositzyn, A., Serdechnyy, D., Korchagin, S., Pleshakova, E., Nikitin, P., Kurileva, N.: Mathematical modeling, analysis and evaluation of the complexity of flight paths of groups of unmanned aerial vehicles in aviation and transport systems. *Mathematics* **9**, 2171 (2021)
18. Barotov, D., et al.: Transformation method for solving system of boolean algebraic equations. *Mathematics* **9**, 3299 (2021)



Architecture of an Expert System to Support Diagnostic Decisions for Hereditary Diseases

Nikolay A. Blagosklonov^(✉)  and Boris A. Kobrinskii 

Federal Research Center “Computer Science and Control” of RAS, Vavilova str., 44, kor.2,
Moscow 119333, Russian Federation
nblagosklonov@frccsc.ru

Abstract. The number of genetically related diseases is over 7000, of which most are extremely rare. Early diagnosis of these diseases is associated with several difficulties, including a variety of clinical forms, polymorphism (multivariation) of phenotypic manifestations, and lack of personal experience in observing patients with these pathologies. The use of physician-assisted computer systems may allow to overcome these difficulties. For this purpose, an expert system was developed to support diagnostic decisions in hereditary lysosomal diseases. Knowledge extraction took place in two stages—from literature sources and from experts. The knowledge base is implemented on a cloud platform in the form of an ontological network. The mathematical model of the disease allows a complex assessment of the signs based on, the modality coefficient and confidence measures of manifestation and degree of expression suggested by the experts. The comparative analysis algorithm compares the new case to the reference variants of the known clinical forms of the integral model and then ranks the hypotheses put forward. The explanation block allows to present the data that served as a basis for the hypothesis because of the features: confirming the hypothesis, missing to confirm the hypothesis, or irrelevant to the diagnosis. The results of clinical testing of the system showed high (above 88%) efficiency of differential diagnosis.

Keywords: Expert system · Modified certainty factors · Decision support system · Hereditary diseases · Lysosomal diseases · Differential diagnosis

1 Introduction

According to the Mendelian Inheritance in Man (OMIM) worldwide database, the number of genetically conditioned diseases is over 7000 (<https://omim.org/statistics/genemap>). Individual clinical forms may occur with a frequency of 1 case per 300,000 population or less [1]. This low prevalence leads to the fact that most practitioners have no personal experience in diagnosing this pathology [2]. The complexity of differential diagnosis of hereditary diseases is due to a large feature space, including unique and specific manifestations. In view of this, the doctor needs to remember and analyze when making a diagnosis the combinations of various signs changing in dynamics. At the same time, diseases within the same subgroup [3] may have similar signs, slightly differing in attributes such as age of manifestation, degree of expression and frequency of manifestation.

The diagnostic process for hereditary diseases is two-step. First, the physician generates several of the most likely hypotheses during the pre-laboratory diagnosis stage. After that, the patient is referred for molecular genetic testing to confirm the diagnosis. Because of the high cost of such testing, the physician cannot prescribe multiple tests. When referring for laboratory confirmation, the narrowest possible range of the most likely differential diagnostic hypotheses must be generated.

To overcome these diagnostic difficulties, it is advisable to use decision support systems. The present version of such a system [4] was created for differential diagnosis of several diseases from the group of lysosomal storage diseases [5]: mucopolysaccharidoses, gangliosidoses and mucopolipidoses (30 clinical forms in total). The purpose of the system created is to perform differential diagnosis at the clinical pre-laboratory stage of patient examination, i.e., before molecular genetic testing is performed.

2 Knowledge Extraction and Presentation

The knowledge field of the system was formed in two stages. The signs were preliminarily extracted from literature sources and online databases (OMIM—<https://omim.org/>, Human Phenotype Ontology (HPO)—<http://human-phenotype-ontology.org>, Genetic and Rare Diseases (GARD)—<https://rarediseases.info.nih.gov/>) containing information on the clinical picture of lysosomal storage diseases. The list of features created based on HPO and OrphaNet (<https://www.orpha.net/>), in the final version includes additional signs of different levels, summarizing the signs of lower levels. This makes it possible to describe situations with different levels of patient examination. The final decision on the formation of the feature space was made by two experts—specialists in the field of clinical genetics.

The number of diagnostically significant signs selected by the experts for the three groups of lysosomal storage diseases was 35, of which 22 are characteristic for mucopolipidoses, 21 for gangliosidoses, and 27 for mucopolipidoses.

In a single feature space, for each of the 35 features in each of the four age groups formed during the study (1st year of life, 1—3 years, 4—6 years, 7 years and older), expert evaluation was obtained: modality coefficients (diagnostic significance), certainty factors for the manifestation and degree of expression of the feature.

The knowledge in formalized form was represented in the form of a matrix “diseases—signs”.

3 The GenDiES System

The GenDiES expert system was developed to support diagnostic decisions for hereditary diseases. The main components of the system architecture are shown in Fig. 1.

The main components of the system are data input interface (A), database (workspace) (B), knowledge base and logical output (C), explanation unit (D).

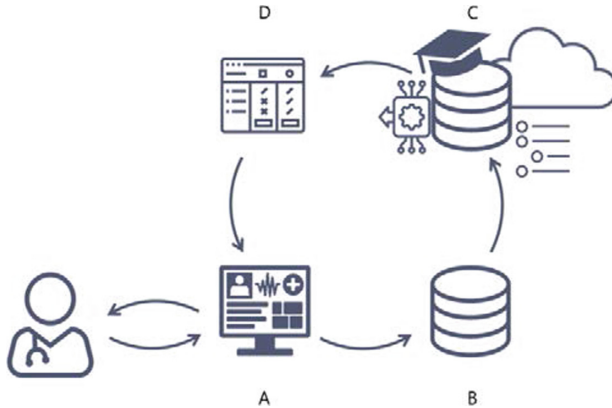


Fig. 1. GenDiES system architecture

3.1 Data Input Interface

The interface in the form of a multilevel thesaurus is implemented as a web application, which provides access to the personal account of the physician-user from any device via a browser. To ensure information security, the patient's personal information (name and surname) is replaced with identifier in the electronic health record (EHR). Since identifiers may coincide in different medical institutions, it is possible to uniquely identify the patient in the system using the “clinic-patient ID” mapping.

The ontology-based description of a patient with a suspected hereditary disease is shown schematically in Fig. 2.

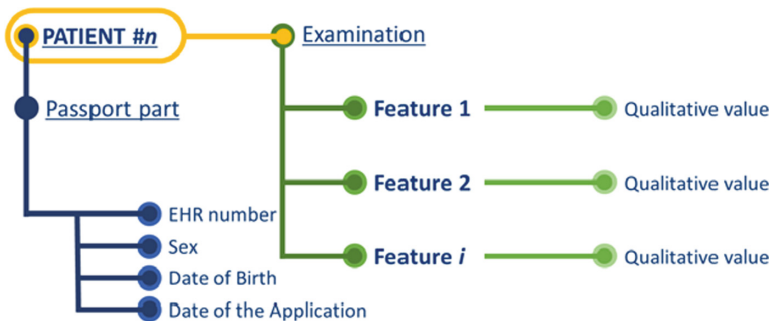


Fig. 2. Ontological representation of the description of a patient with a suspected hereditary disease

The selection of features is carried out from the directory built into the system, which contains 35 names. It allows to uniformly describe the manifestations of the disease in each new patient.

The information entered about the patient is then transferred to the workspace.

3.2 Workspace

The workspace is a repository of formalized data about the patient. Along with the initial data, it stores the intermediate data of the task currently being solved. The database also stores information about the hypotheses put forward and the explanations formed by the system about the hypotheses of the differential diagnostic series offered to the user.

3.3 Knowledge Base and Logical Output

The knowledge base contains descriptions of clinical forms of hereditary lysosomal diseases, each sign is accompanied by three expert evaluations and the possibility of its presence in a particular disease is noted. Methods of ontological engineering were used to form the knowledge base. The knowledge base was realized in the form of an ontology network, the principles of its formation are shown in Fig. 3.

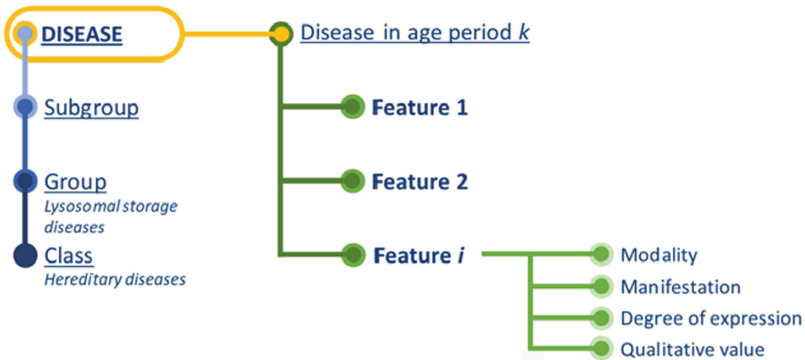


Fig. 3. Ontological representation of the description of a hereditary disease

The cognitologist, when entering information about a new disease into the knowledge base, sequentially selects a class, group, or subgroup of diseases, within which he or she creates a new clinical form. Then, he enters the signs characterizing the manifestations of the disease in each of the four age periods. For each sign, the expert assessments of modality, certainty factors of manifestation and degree of expression, as well as the qualitative value of the sign “present” or “absent” are indicated. Accordingly, 30 clinical forms of lysosomal storage diseases were described for the selected age periods. The knowledge base is expandable.

The disease model and the algorithm of comparative analysis are used in the process of hypothesis making and confirmation.

Disease Model. The integral disease model is the sum of complex assessments of signs within an age group:

$$I = \sum_{i=1}^n P_i \tag{1}$$

where: I —integrated assessment of the signs of the disease, P_i —complex assessment of a sign, i —the number of signs, n —the set of signs of the disease (group of diseases).

The complex assessment of the sign, which is the product of three expert evaluations, is calculated by the formula:

$$P_i = M_i \cdot m_i \cdot s_i \quad (2)$$

where: P_i —sign (symptom), M_i —modality of the sign, characterizing its frequency, m_i —expert certainty factor for the manifestation, s_i —certainty factor for the degree of expression of the sign.

The disease model has two modifications: I_e is a reference disease model, which includes all the features described by the experts, and I_p is a case model, calculated as a sum of only those features that are noted in the patient. Thus, both I_e and I_p are calculated using identical formula I . The disease model interacts with an ontological network: expert estimates of non-zero attributes (with qualitative values “available” and modality coefficients greater than zero) are calculated within an age group—complex assessments of signs; their sum is calculated—the integrated disease assessment. To implement this procedure, the algorithm “passes through” the network by “taking away” the values of expert assessments.

However, the number of analyzed features n will be different: for example, when calculating I_e for mucopolysaccharidoses it is 22, and for I_p it can lie in the range from 1 to 22, depending on the presence of features noted in a patient. Thus, in the GenDiES diagnostic system, the personal integrated assessment is correlated with the reference assessment of the disease based on the developed formula. Thus, differential diagnosis in each case is performed using a comparative analysis algorithm.

Comparative Analysis Algorithm. The algorithm first groups the features of the new patient into three subsets according to their role in the hypothesis under consideration:

- signs “for” are phenotypic manifestations that both the patient and the reference description of the clinical form of the disease have;
- signs “against” are phenotypic manifestations that the patient has, but their impossibility for a specific clinical form is noted in the reference description;
- signs “outside the reference” are phenotypic manifestations that the patient has, but the reference description does not indicate that they are present or absent in this clinical form.

If there is at least one sign “against” the hypothesis, the hypothesis is rejected. Then the hypotheses are grouped into clusters according to the number of features “outside the reference” and the clusters are arranged according to the principle: the less such features, the more significant the cluster. Within each cluster, hypotheses are ranked according to the similarity of the clinical picture to the reference description in the knowledge base because of the available features and expert estimates of modality, manifestation, and degree of severity. From the resulting ranked list of hypotheses, considering the ordered clusters, the physician-user of the system receives a list of the set number of the most likely diagnoses (five hypotheses by default) (Fig. 4).

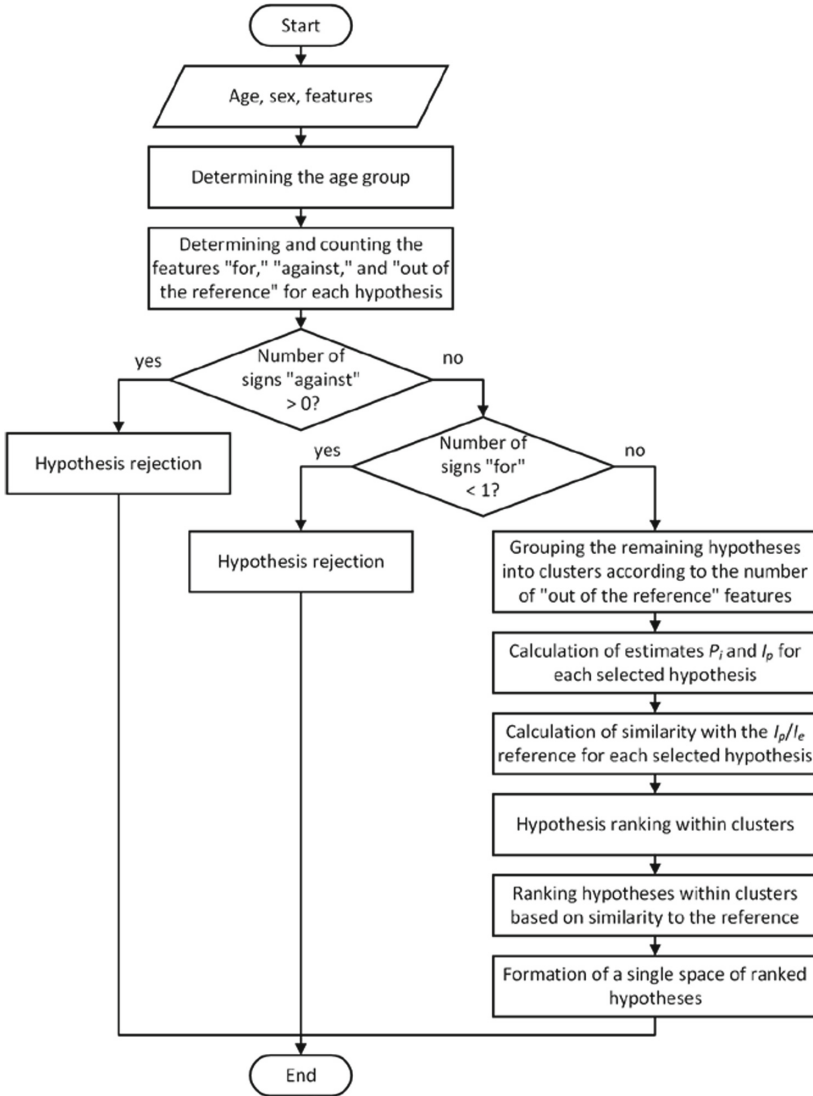


Fig. 4. Comparative analysis algorithm

3.4 Explanation Unit

As an explanation for the selected hypotheses, the system provides for each hypothesis included in the differential diagnostic series three groups of features: (1) which served as the basis for the hypothesis, (2) which were present in the patient but not in the reference disease description, (3) which were absent in the patient but are part of the reference variant.

4 Functioning of the System

The functioning of the system is as follows:

1. The physician enters information about the signs detected in the patient, specifying sex and age. This data goes into the workspace of the system (database).
2. The patient's data goes from the workspace to the knowledge base, which triggers the activation of the relevant rules. Proposal and consideration of potentially possible completes the formation of a list of hypotheses, which are ranked by the solver (logical output), considering the likelihood of similarity with the reference descriptions.
3. In the explanation block, the features for the hypotheses included in the differential diagnostic series are grouped into three sets, providing the clinician with information about why the system assumed a given diagnosis and what additional features to look for.
4. The generated list of hypotheses with explanations is displayed to the physician and stored in the database.

5 Conclusion

The system is implemented on a cloud platform IACPaaS [6], which provides wide access for physicians from various regions.

Clinical approbation was carried out on verified real data of 110 patients (35 from literature and 75 from electronic health records). The criterion for evaluating the effectiveness of the system was the inclusion of the diagnosis in the list of five hypotheses issued.

According to the results of approbation in 97 out of 110 cases the correct diagnosis (corresponding to the verified—true-positive result) was among the first five hypotheses based on the analysis of clinical symptomatology before laboratory tests were performed. Thus, the accuracy of the system—the inclusion of diagnoses in a narrow differential-diagnostic series—was 88.18%. In 11 cases where a verified diagnosis was not included in the series of five hypotheses, it was presented in an expanded list of ten ranked hypotheses. This is explained by less similarity of clinical manifestations with the reference descriptions.

For comparison, similar intelligent decision-support systems in clinical genetics provide three to ten diagnoses or a complete list (optional) [7–9]. When testing the AdaDX expert system for rare diseases [10], resulted in a 53.8% correct diagnosis for a differential series of five possible diseases. In the Rare Disease Discovery system for orphan diseases [7] the correct diagnosis was among the first ten hypotheses in 60% of cases with an incomplete description. For adjusted descriptions of the clinical picture in the Rare Disease Discovery system, the accuracy rose to 80% in a series of fifty hypotheses [7].

In the future, the GenDES system plans to include a module for diagnostics based on comparison with previously diagnosed cases—precedents [11].

References

1. Semschikova, Y.P., Kozlov, Y.A., Yakovlev, A.B., Shinkareva, V.M., Barzunova, T.V., Manjkova, N.I., Balakirev, E.A.: Rare case of morquio syndrome (mucopolysaccharidosis type IVA): difficulties of diagnostic search and management. *Pediatr. Pharmacol.* **19**(1), 39–44 (2022)
2. Shashel, V.A., Firsova, V.N., Trubilina, M.M., Podporina, L.A., Firsov, N.A.: Orphan diseases and associated problems. *Med. Herald South Russia* **12**(2), 28–35 (2021)
3. Gorbunova, V.N.: Congenital metabolic diseases Lysosomal storage diseases. *Pediatrician* **12**(2), 73–83 (2021)
4. Kobrinskii, B.A., Blagosklonov, N.A., Gribova, V.V., Shalfeeva, E.A.: Expert system for the diagnosis of orphan diseases. In: Kovalev, S., Sukhanov, A., Akperov, I., Ozdemir, S. (eds.) *Proceedings of the Sixth International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’22)*. IITI 2022. *Lecture Notes in Networks and Systems*, Vol. 566. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-19620-1_24
5. Platt, F.M., D’Azzo, A., Davidson, B.L., Neufeld, E.F., Tiffit, C.J.: Lysosomal storage diseases. *Nat. Rev. Dis. Primers.* **4**, 27 (2018). <https://doi.org/10.1038/s41572-018-0025-4>
6. Gribova, V., Kleshev, A., Moskalenko, P., Timchenko, V., Fedorisdiev, L., Shalfeeva, E.: The IACPaaS cloud platform: features and perspectives. In: *Proceedings of Second Russia and Pacific Conference on Computer Technology and Applications (RPC)*, pp. 80–84. IEEE Press, New York (2017). <https://doi.org/10.1109/RPC.2017.8168073>
7. Alves, R., et al.: Computer-assisted initial diagnosis of rare diseases. *PeerJ* **4**, e2211 (2016). <https://doi.org/10.7717/peerj.2211>
8. Gouvernet, J., Caraboeuf, M., Ayme, S.: GENDIAG: a computer-assisted facility in medical genetics based on belief functions. *Methods Inform. Med.* **24**(04), 177–180 (1985). <https://doi.org/10.1055/s-0038-1635373>
9. Kobrinsky, B., Kazantseva, L., Feldman, A., Veltishchev, J.: Computer diagnosis of hereditary childhood diseases. *Med. Audit. News* **1**, 52–53 (1991)
10. Ronicke, S., Hirsch, M.C., Türk, E., Larionov, K., Tientcheu, D., Wagner, A.D.: Can a decision support system accelerate rare disease diagnosis? Evaluating the potential impact of Ada DX in a retrospective study. *Orphanet J. Rare Dis.* **14**(1), 69 (2019). <https://doi.org/10.1186/s13023-019-1040-6>
11. Blagosklonov, N.A., Gribova, V.V., Kobrinskii, B.A., Shalfeeva, E.A.: Knowledge-based diagnostic system with a precedent library. In: Kovalev, S.M., Kuznetsov, S.O., Panov, A.I. (eds.) *Artificial Intelligence. RCAI 2021. Lecture Notes in Computer Science*, Vol. 12948. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86855-0_20



System Analysis of Educational Digital Ecosystems in the Agro-Industrial Complex of Russia

Vladimir Budzko^{1,2}  and Victor Medennikov² 

¹ National Research Nuclear University MEPhI Moscow Engineering Physics Institute, Kashirskoe shosse, 31, 115409 Moscow, Russian Federation
vbudzko@ipiran.ru

² Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, Vavilova 44-2, 119333 Moscow, Russia

Abstract. The concept of a digital ecosystem is defined. A dangerous tendency to simplify the definition of the essence and composition of the digital ecosystem is noted, which manifests itself due to the attractiveness of the use by the business community in order to attract potential users to the products being created. The correct approach to the consideration and definition of the digital ecosystem is demonstrated with a description of the mathematical model using the example of an educational ecosystem poorly covered in publications. Its improvement is of great importance for minimizing the negative impact of natural and man-made environmental hazards on the environment and on the person himself.

Keywords: Digital ecosystems · Mathematical model · Agricultural universities

1 Introduction

The concept of “ecosystem” (ES) has become increasingly popular in recent years due to the significant success of the global digitalization of the economy and is increasingly being discussed on world economic platforms. The instability of social development throughout the world, caused by political and economic tensions, by the dynamism of external and internal factors in the activity of all spheres of society, also contributes to this. More and more complex challenges in the form of uncertainty and disequilibrium in the development of the whole world and the search for new methods of communication between participants in economic interaction require an adequate response. Not surprisingly, a large number of researchers began to search for solutions to these problems in these conditions. Research in the field of strategic management [1–4] is actively developing in this direction in addition to the ecosystem approach. The science-based systematic approach to the development of socio-economic systems in the form of long-term planning and management methods aimed at the future is implemented in these works. We will show in our studies that strategic management is the basis of the ESs themselves. This is not shown in many works in the field of ES. A number of successive general stages can be distinguished in strategic management. At the first stage, the

goal-setting of the organization is determined after the formulated mission; at the second stage, a more detailed analysis of strategic goals and available resources is carried out; at the third stage, the specific tasks of their implementation are solved with some optimal (rational) choice of ways to achieve the chosen goals; on the fourth—the implementation of the chosen ways to achieve strategic goals is being carried out; and, finally, at the last, fifth stage, monitoring of the entire process of moving towards the goal is carried out with subsequent assessment and analysis of the progress in the implementation of trajectories from the initial situation to the final one.

The development of digital twins [5], which are a set of interrelated mathematical models and data that describe all aspects of an enterprise's activities and help optimize business performance, is also performed. However, strict scientific character, unambiguity and consistency are absent in many studies. Violation of terminological foundations and ambiguity of ES definitions cause great harm to the efficiency of digitalization of real economy management. The abundance of new phrases (“digital ecosystem”, “digital economy ecosystem”, “digital business ecosystem”, “digital platform ecosystem”, etc.) lead to a loss of meaning. The lack of strict scientific character and consistency in ES research can lead to a loss of interest in them, as happened with the work on universal blockchainization, which recently flooded many conferences and journals and almost completely disappeared after a couple of years from publications [6].

The greatest confusion with the conceptual apparatus reigns in the studies of educational ES. We will give a systematic, scientific definition of educational digital ecosystems (EDES) in this paper based on the scientific classification of DES with the construction of mathematical models. Since the agro-industrial complex is one of the most striking ecosystems that meets the classical definition, which is characterized by a huge variety of natural factors and biological species, we will consider the problem using the example of agricultural universities. The possibilities of information and communication technologies, progress in the practical application of digital technologies in science and production have increased significantly and have a great impact on the development of ES.

2 Systems Analysis of Ecosystems

The origins of the concept of ES are in biology, where ES has always been understood as a physical and biological system, including a variety of interdependent biological organisms and physical factors that form the environment—environmental factors in a broad sense, which have different types and sizes, differ in the degree of isolation and autonomy [7]. This shows that ES, firstly, based on the foundations of systems theory, systems analysis, operations research, is a classical research system with all its inherent properties, and secondly, it consists of living (eco) and non-living elements that simultaneously interact both with each other and with environmental factors, the set of which depends on the objectives of the study.

An analysis of ES research in various fields of activity, depending on the meaning given to the concepts of “eco” and “system”, gave grounds to carry out the following classification of them with some formulations of the most important areas for further research in this area.

2.1 The Dominance of the Eco Concept in ES Research

Classic ES. Agriculture, having a fourth part of the biological diversity of the planet, claims to be the most representative object for ES studies. This is due to the thoughtless, sometimes predatory attitude to the soil, leading to the rapid depletion of natural resources and the degradation of the quality of life of the population. As a result of the transition to the industrialization of production, the industry has recently turned out to be one of the main pollutants of nature along with the transport industry, energy, and the domestic sector. Environmental problems are growing due to the widespread use of pesticides that affect not only pests, but also most beneficial organisms both in the soil and in the environment. Their death leads to violations in the ES, to soil erosion, and a decrease in soil fertility. Thus, 1.5–2 million hectares of land are degraded in Russia every year [8]. Survivors, accumulating pesticides in their bodies, pass them along the food chains to representatives of a higher order, up to humans.

Therefore, agricultural science should be based on new systematic digital studies of its ES, in particular, on environmental technologies, based on the following factors. Firstly, the poor knowledge of soil biota and its interaction with plants becomes a brake on the effective development of precision farming, the leader of digitalization in agriculture. It is inhabited by a huge variety of different organisms that interact with each other and contribute to the global geobio-chemical cycles that ensure the possibility of life on our planet as a whole. Nowhere in nature is there such a close coexistence of species as in communities of soil organisms, but little is known about this biodiversity, because it is underground and, as a rule, invisible to the human eye. This also applies to agricultural crops. Thus, millions of samples have been collected in the world genetic banks of cultivated plants, but so far only 1% of them have been studied in relation to their potential properties [9].

Secondly, the activity of mankind has given rise to one of the global environmental problems at this stage of its development, which is the reduction of biodiversity. The formation of those types of biological organisms that currently exist on Earth took place over 4 billion years. If the problem of biodiversity reduction is not resolved, then people will forever lose the world that we know now. Now only about 30 crops provide 95% of human needs for food and energy, and only five of them—rice, wheat, corn, millet and sorghum—provide about 60%. Five types of animals—cattle, sheep, goats, pigs and chickens—provide about a third of the average daily protein intake. The use of such a small number of species leads to the instability and vulnerability of the food supply in the future of mankind [10]. Thus, by losing the biodiversity of the Earth, humanity is losing its future.

Thirdly, the active industrialization of the entire food chain has led to the fact that about a third of the food produced is lost or thrown away in the world every year [11]. Such irreparable losses lead to an increase in the load on the ES, the cost of various resources, affect the environment, the quality of food, etc. The first ancient microorganisms laid the foundation for soil formation, during which almost all consumed organic resources returned to the ES over billions of years, which can be compared with waste-free production. In the existing ES, over such a long period, higher plants receive the necessary nutrients not only in the form of fertilizers, but also due to the symbiosis of higher plants with bacteria, the symbiosis of higher plants with fungi, and the plant

providing its nutrient needs due to other organisms, self-sufficiency by the plant of its nutrient needs. At present, humanity is breaking this chain of interaction, trying to minimize the damage by applying fertilizers without compensating for many of the micronutrients removed. As a result, the ES changes. There is a large field for research here, from breeding more storable products, cultivating a food culture, forming optimal supply chains to developing technologies for reuse and recycling of leftovers. Now it is appropriate to recall the experience of the USSR, when food waste was collected in cities and sent to livestock farms.

Adaptation of the Laws of Functioning of Classical ES to Socio-Economic Systems.

The success of the digitalization of society gave impetus to research on the adaptation of the laws of functioning of classical ES to the functioning of industrial, social and a number of other industries, along with the social order to find answers to increasingly complex challenges in the form of uncertainty and disequilibrium in the development of the whole world. Optimization of the use of raw materials and energy resources in accordance with the concept of industrial ES should minimize the generation of waste through the development of technologies for the use of waste from some industries as raw materials for others. Here is the example of such a power plant, created over 20 years,— a system of enterprises in the Danish city of Kalundborg, consisting of: Asnaes power plants with a capacity of 1500 MW; Statoil's refinery with a capacity of 4.8 million tons per year; the Gypros gypsum board plant, which produces 14 million square meters of gypsum boards annually; the pharmaceutical plant of the Novo Nordisk company with an annual turnover of over \$2 billion; the town of Kalundborg itself with a population of 20 thousand people, whose enterprises and people need heat and hot water [12]. A large number of algorithms for solving problems based on the behavior of insects and animals are beginning to be introduced into the economy. Many papers on this topic have appeared recently, such as: "Recent Trends in Research on Swarm Intelligence for Engineering Applications", "Research on Group Intelligence for Engineering Applications", "Communal Foraging by Social Insects, Division of Labor, Nest Building by Social Insects, collective sorting and grouping", "Adaptive swarm intelligent systems", "Quantum swarm intelligence".

Ecosystem Approach in the Transition from Product to Service Model of Economic Relations in the Digital Economy. Very few works meet the demand of the world community to ensure the transition of business from a product to a service model of economic relations, which combines the environmental part with the system one, built on the integration of data, information systems, the Internet of Things and a number of other digital technologies.

The product model of production in market conditions is based on the race for the right to own goods under the influence of advertising and human weaknesses. Growing competition forces manufacturers to reduce the service life of products, waste on its production, as a rule, irreplaceable resources of the planet, the human capital of the population and cause environmental damage to nature, humanity and most of the ES.

The service model of economic relations reflects the tendency to meet the needs of customers by providing them with the result of work involving them at all stages of the product/service life cycle. The use of this model became possible thanks to the

digital economy, in particular, thanks to the technologies of the Internet of things, when the saturation of the corresponding product with sensors with the registration of certain parameters allows the manufacturer to know everything about the current technical condition of the product, its dynamics and the profile of its operation by the user, that is, to have a deep customer knowledge. An illustration of such a transition will be shown using the example of a small cinema in Barcelona. Its peculiarity is that dozens of sensitive television cameras, recognizing the emotions on the face of each of the spectators, watch them during the film screening. And the price of a ticket for a session of a comedy film is formed depending on the amount of time during which the viewer's face was smiling or laughing. Only the end result—emotions determines the ticket price in this case [13]. Services appeared in dentistry when the clinic does not subscribe to the treatment of each individual case, but undertakes to monitor the health of your teeth. The service is moving away from the principle of “sell the service more expensively” to the principle of “helping the client keep their teeth healthy”. This approach allows you to maintain a high rating of the clinic, but at the same time receive income from advertising of pharmaceutical products while educating patients on proper oral hygiene [13].

The service model is becoming more and more common in large businesses. For example, the world's leading aircraft manufacturers do not sell, but lease their aircraft to operating airlines. The service model implies that manufacturers care about the quality, environmental friendliness and durability of products, extending the model to their many partners from other industries that provide products and services. An ES with such a model is based on the concept when each buyer can check information about the quality, safety and legality of products online, and regulatory authorities can access the full range of information about the product [14].

2.2 Dominance of the Concept of “System” in ES Studies

Consider the studies of ES, in which the meaning of the element “eco” is thrown out of the term “ecosystem” and only one element “system” remains. The desire of many companies to resort to the integration of disparate data into a single structured environment is the basis on which most work in business ecosystems is built. This creates the possibility of considering the relationships and environmental factors of many controlled systems with elements of strategic management [14]. Moore [15], the founder of the theory of business ecosystems, writes: “ES in business is an economic community based on interacting organizations and individuals, organisms of the entrepreneurial world.” That is, the ES of any enterprise, in addition to itself, includes consumers, market intermediaries, suppliers, as well as structures whose interests must be considered in a given situation—departments and regulatory agencies, associations and organizations that ensure compliance with laws and standards. Direct and potential competitors, as well as any other members of the system that influence it, are part of the ES to some extent. And it is necessary to consider the behavior of all these elements in the operational and strategic plan, building information exchange.

The theoretical basis of this approach to the study of ES is the conceptual dispute between two groups of supporters within the direction “Adaptation of the laws of functioning of classical ES to socio-economic systems.” Representatives of one group

proceed from the hypothesis that nature, in the course of evolution, has already found optimal solutions to all those problems that were for the survival of biological species, forming various ES. Therefore, it is necessary to look for an analogue of solving the problem in natural systems at a new stage in the development of mankind, that is, consider only classical ES. Representatives of another group believe that, having received a large amount of structured data and tools for their processing, in particular, in the form of mathematical models of strategic management, humanity is able to find such strategic decisions in the digital age. Nature has not found these solutions and could not find them, since it does not look through all the theoretically possible and practically realizable states and forms of being. They cite as their argument the invention of the wheel in technology. That is, new opportunities for the formation of systems based on strategic management have appeared, since nature is capable of performing only operational management.

James F. Moura grasped the new opportunities for ES in business. But most researchers invest in the concept of DES and ES definitions that specialists in system analysis and the IT industry hardly understand: “the ecosystem is presented in the first approximation as a network of collaborating and competing firms offering related products and services” [16]. Attention is focused on the fundamental difference between the business ecosystem as the latest organizational and economic form, which consists in complicating its structure, from a virtual business organization (in particular, a network one) by adding and considering links with the external environment.

The business community of our country and some part of the scientific community that is not involved in the implementation of the DES program, without explaining the goal-setting of the ES formation and without describing the mechanisms taken from nature, violated the terminological harmony, which led to the ambiguity of the ES definitions.

As an example, we can cite the most famous Sber ecosystem, which, in addition to Sber itself, includes the Okko online cinema, the Delivery Club food delivery service, the Sbermarket food delivery service, Citymobil taxi, etc. Yandex began to build its own ecosystem, including the Kinopoisk portal, car sharing service, food delivery services Yandex.Food and Yandex.Lavka, etc. Other participants in the domestic market began to duplicate this scheme for creating their own ecosystems. So, the developers of the Webinar Group system give it such a sonorous definition: “The Russian ecosystem of services for meetings, online events, training and webinars.” Dissertations for scientific degrees, in which the sonorous terminology of ES is also used, appear. The analysis shows that these systems provide only a set of services that are interconnected by a common site (perhaps with a single payment system) and do not have common biological and ecosystem characteristics. Also, the understanding of ES by domestic advocates does not correspond to the classical scientific concept of a system as a set of interrelated elements combined into one whole to achieve a certain goal, which is determined by the purpose of the system.

3 Educational Digital ES in Agriculture of Russia

The analysis of the literature on EDES [17–19] shows that the direction of research discussed in Sect. 2.2 dominates, but presented in a simplified way: “The ecosystem approach entered the humanities, social and economic sciences largely due to the need to represent the process of interaction between groups consisting from various elements that have a connection, and components of the environment. The following qualities of the ecosystem approach in various areas were highlighted: complexity and non-linearity. This has led to the generally accepted definition of an ecosystem as a complex system” [17]. The work [18] specifies: “The digital ecosystem of education functions as a network infrastructure that is supported by digital technologies and creates conditions for stakeholders to work together and effectively interact with each other on a single technological platform, where each of the participants (agents) has access to common” ecosystem resources that it did not originally have or had, but in insufficient quantities.

Although the authors of these works have grasped the need for the formation of EDES in practice, they have not proposed specific mathematical algorithms. The EDES presented in this paper is based on the results of mathematical modeling with the formation of a digital platform (DP) of information scientific and educational resources (ISER) of agriculture [20]. The system built on the basis of this DP completely coincides with the EDES with the complex integration of scientific resources into it. These resources will be called information resources (IR) of the EDES. The EDES is an integral part of the DES of agriculture, by which we mean a system of rational digital interaction of stakeholders for the optimal, integrated use of resources (biological, natural, material, financial, social, labor, educational, scientific) in the interests of all participants. The basis of such use is scientifically based integration of information, algorithms and software and hardware tools for collecting, storing, processing and transmitting data and knowledge, integrated into a single information and control system, which is designed for operational and strategic management (functioning) of the target subject area.

The need to form the EDES is due to the rapid increase in the volume of IR in educational activities. The potential possibility of its formation with the use of new digital technologies exists, and there is a need for these resources for all segments of users: students, teachers, future applicants, scientists, government agencies, commodity producers, and other categories of the population. Note that the EDES implements all areas of research from Sect. 2.1. The EDES concept is based on the integration technologies and is a digital tool for promoting and transferring innovations in the field of biologization of agriculture into production. The world-class scientific center “Agrotechnologies of the Future” under the auspices of Russian State Agrarian University—Moscow Timiryazev Agricultural Academy is successfully working in this direction and is focused precisely on the formation of the DES of the agro-industrial complex, considering all the factors of the manifestation of the ecosystem approach. Main areas of research: “Agrobiotechnologies for soil fertility management in Russia in the interests of highly productive agriculture of minimal environmental risk”, “Technologies for processing and valorization of low-value agricultural raw materials and agro-industrial waste”, “New digital technologies in agriculture”, “Creation of safe, high-quality, functional feed and food” and a number of others [21, 22].

An integrated approach to integrating disparate agricultural IRs into a single structured environment is to develop an EDES that includes these resources based on ontological modeling (standardization) of them. Ontological modeling allows us to create a formalized representation of our knowledge about the subject area, which can be displayed in automated information systems. The ontological model is a description of the subject area that uses standard elements of metamodel (for example, objects, relationships, etc.) and strives to explicitly reflect the key aspects of the subject area as fully and reliable as possible. Such a single EDES for the whole country is able to perform much more functions than the studies mentioned above. Thus, the EDES will provide an opportunity to support scientific research (function 1); raising the level of education, retraining for a wider range of users, and not just students (function 2); the possibility of effective and rapid transfer of knowledge and innovation to the economy (function 3).

Monitoring of the websites of universities and a survey of agricultural producers of various sizes and forms of ownership was carried out in 22 regions of the country in order to determine the types of IR that maximize the functions of the EDES. The results of the research showed the demand among commodity producers for precisely those IRs that began to be posted on the websites of universities. These are: developments, publications, consulting activities, regulatory information, distance learning (DL), application software packages (ASP), databases (DB).

Is it possible in principle to form the EDES on the basis of these IRs? Since parallel monitoring of agricultural research institutes (RIs) showed the presence of the same IRs on their websites, when forming the EDES, it is necessary to assess the possibility of integrating IRs, both universities and research institutes, into a single EDES database. The appropriate mathematical model has been developed for this purpose. It allows you to calculate various options for IR integration for any potential data volumes and number of users. Let's introduce some digital standards for the forms of storing the content of sites by ontological modeling in order to avoid the need to consider a large number of presentation forms and types of IR. Let's represent these standards in the following format: unordered catalog (list), ordered catalog, unordered full-length representation and ordered full-length representation.

Three options for the implementation of IR integration mechanisms, convenient for the administration of the EDES, were considered in the model in the future. In the first version, the IRs were placed in the form of catalogs in the EDES database. If the found information is of interest to the user, he is redirected to the site of the full-text IR custodian specified in the catalog. The second option is that all IRs are hosted by the provider in the common database of the EDES. The third is a mixed version of the first two. The quality of service for users by the provider is usually measured by several parameters: network reliability, time delays in the transmission of information, statistical characteristics of delays, throughput. We want to find out the global characteristics of the network when information is transferred by a large number of owners to one provider, but we do not have specific amounts of this information (only a small part of it is posted on websites in a poorly structured form). Therefore, we will model processes over a sufficiently large interval, taking a month as a unit of time. The characteristics of the currently most common site content management system (CMS) "1C-Bitrix" were considered in the

model, since this CMS is most common in the development of sites for agricultural universities and research institutes (31.5% of agricultural university sites).

We formalize the model, for which we introduce expressions.

3.1 Constants and Parameters of Model

I —code of the type of information (texts, images, videos, etc.), $i \in I$;

j —code of the group of information carrier organizations (universities, research institutes, etc.), $j \in J$;

m —code of the provider using 1C-Bitrix, $m \in M$;

n —code of data representation type, $n \in N$;

k —code of a specific information carrier organization; $k \in K_j$;

l —code of information storage form, $l \in L$;

d_{im} —existing load of the m -th provider on the i -th type of information (in Mbytes);

V_{ijkl} —the volume of the i -th type of information for the l -th of information storage of the k -th organization of j -th group of organizations (in Mbytes);

D_{im} —bandwidth of the m -th provider of the i -th type of information (in Mbytes);

z_{ijmkl}^1 —the average unit costs for transferring units of the i -th type of information for the l -th form of information storage of the j -th group of the m -th provider of the k -th organization to the provider running on 1C-Bitrix (in rubles/Mbytes);

z_m^2 —total costs per unit of time for maintaining the site of the m -th provider (in rubles);

z_{jk}^3 —total costs per unit of time for maintaining the sites of the j -th group of organizations when storing information with their provider of the k -th organization (in rubles);

P_i^1 —the average number of hits (number of visitors) to the i -th type of information per unit of time;

P_{il}^2 —the average number of page views of the i -th type of information for the l -th form of information storage;

P_{ijkl}^2 —the average number of page views of the j -th group of organizations of the i -th type of information for the l -th form of information storage of the k -th organization.

Here $P_{il}^2 = s \cdot \sum_{j,k} P_{ijkl}^2$, where s is the coefficient of increase due to integration (when switching to a typical site, it is considered that $s = 2, 5$);

P_m^3 —the average page size of the site of the m -th provider (in Mbytes);

C^0 —funds allocated per unit of time for the transfer of information to one of the providers m (in rubles).

In the future, we will assume that all information of any provider using 1C-Bitrix will be stored in a unified form.

Then we introduce another group of parameters:

b_{il} —the average size per unit of time of the transferred file of the i -th type of information of the l -th form of information storage (in Mbytes);

g_{il} —the average number of requests for the i -th type of information of the l -th form of information storage from any provider using 1C-Bitrix;

r_{jknl} —the number of types of data representation of the n -th of information representation for the j -th group of the l -th form of information storage of the k -th organization at its provider;

a_{iknl} —the average number of hits (number of visitors) of the n -th type of information presentation for the i -th type of information presentation for the l -th form of information storage with its provider;

v_{inl} —index reflecting the presence of the i -th type of information of the l -th form of information storage in the n -th type of data representation, $v_{inl} = 1$, if there is i -th type of information of the l -th form of storage in the n -th type of data representation; 0—otherwise.

Then:

$$g_{il} = s \cdot \sum_{n,k} v_{inl} \cdot a_{iknl}; \quad (1)$$

$$V_{ijkl} = b_{il} \cdot \sum_n r_{jkn}; \quad (2)$$

$$P_i^1 = \sum_l g_{il}. \quad (3)$$

3.2 Model Variables

X_{ijmkl} —increase in the load on the m -th provider due to the placement of the i -th type of information for the l -th form of information storage of the j -th group of the k -th organization (in Mbytes);

$y_{ijmkl} = 1$, if the k -th organization of the j -th group stores the i -th type of information in the l -th form at the m -th provider, otherwise 0.

3.3 Model Equations

The restrictions on the throughput of the m -th provider for the i -th type of information:

$$d_{im} + \sum_{l,j,k} x_{ijmkl} \leq D_{im}; \quad (4)$$

the balance equality for additional load:

$$x_{ijmkl} = \left(P_i^1 \cdot P_{il}^2 \cdot P_m^3 + g_{il} \cdot b_{il} \right) \cdot y_{ijmkl}; \quad (5)$$

for the single provider case: $\sum_m y_{ijmkl} \leq 1$ and all information can be stored only by one of the providers.

Finally, for the costs of transferring information to a provider running on 1C-Bitrix:

$$C^1 = \sum_{i,j,m,l,k} z_{ijmkl}^1 \cdot V_{ijkl} \cdot y_{ijmkl} \quad (6)$$

we have the inequality: $C^1 \leq C^0$ —the restrictions on the cost of transferring information.

3.4 Optimization Criteria

We have two of them in our model:

1. The maximum amount of information transmitted to Bitrix providers:

$$w = \sum_{i,j,m,l,k} V_{ijkl} \cdot y_{ijmkl} \rightarrow \max.$$

2. The minimum maintenance costs (minimization of maintenance costs for Bitrix providers):

$$C^2 = T \cdot \left((I \cdot J \cdot L)^{-1} \sum_{i,j,m,l,k} z_m^2 \cdot y_{ijmkl} + (I \cdot M \cdot L)^{-1} \sum_{i,j,m,l,k} z_{jk}^3 \cdot (1 - y_{ijmkl}) \rightarrow \min \right),$$

T is the specified period of system operation (in months).

Algorithms for solving this nonlinear problem for interested parties can be found in [23]. The initial data for the model, which were taken from the annual reports of agricultural research institutes and universities submitted to the Russian Academy of Sciences, the Ministry of Agriculture, and Rosstat, adjusted for their growth rate over five years, is more important for us to consider now. IR of institutes of the countries of the Commonwealth of Independent States and institutes of the Russian Academy of Sciences engaged in research close to the agro-industrial complex were also considered.

3.5 Model Inputs

Three basic scenarios for the volume of information—the current annual volume, the volume of information for the last 5 years, the volume of information for all years, as well as two basic scenarios for the number of site visitors—the current number and the maximum predicted number were considered for scenario calculations in the model. Data on the projected number of visitors to the EDES for the baseline scenario with the volume of information over the past 5 years are given in Table 1 as an example.

Table 1. Projected number and type of EDES visitors.

No	Types of visitors	Monthly amount
1	Farmers	200,000
2	Employees of agricultural enterprises	1,000,000
3	Students	30,000,000
4	Managerial workers	200,000
5	Researchers	1,200,000
6	Other	32,600,000
7	TOTAL	65,200,000

Table 2. The volume of information resources of the CSEC for the 5-year scenario and in 2022

Types of ISER, forms of storage, organizations	5-year scenario		Quantity in 2022	
	Catalog	Full format	Catalog	Full format
<i>Developments</i>				
RI	357,388	357,388	5192	218
Universities	159,683	159,683	4668	585
<i>Publications</i>				
RI	364,682	364,682	36,915	684
Universities	1,814,582	1,814,582	88,043	23,676
<i>Database</i>				
RI	7293	7293	119	5
Universities	8754	8754	7	0
<i>DL</i>				
RI	4886	4886	104	1
Universities	26,262	26,262	14	0
<i>ASP</i>				
RI	21,879	21,879	50	0
Universities	24,915	24,915	77	0
<i>Consultations</i>				
RI	2125	2125	9	5
Universities	9788	9788	113	0
<i>NPI</i>				
RI	756	756	349	330
Universities	788	788	15,800	128

Forecast data of 5-year volumes of information and site monitoring data in 2022 are given in Table 2.

Numerous experiments with the model made it possible to conclude with great confidence that it is possible to develop and maintain the EDES with the placement in a single database of this digital tool of the entire volume of IR produced by agricultural universities and research institutes over the past five years, with the necessary degree of efficiency in IR searching to simulated requests from various users in month.

The practical implementation of the EDES was undertaken as early as 2007–2008. In the development of the portal of the Russian Academy of Agricultural Sciences (RAAS). Over 12 thousand scientific publications of subordinate research institutes, over 2.5 thousand developments over the past 10 years, over 0.4 thousand consulting services in accordance with the thematic heading and the list of relevant consultants were integrated on the portal according to a single methodology and on the basis of the

state rubricator of scientific and technical information. Note that the E-library contained a significantly smaller number of publications at that time, and other types of scientific and educational resources were not contained at all. Several effective analytical processing of the accumulated information was carried out with the issuance of useful recommendations to users. The work was stopped due to the reform of the RAAS.

However, the analysis of the state of IR sites of agricultural universities in 2022 showed that there are degradation trends in the prospects for the formation of the EDES. Universities are losing interest in presenting IR on websites. At the same time, the number of IR on websites and interest in them should grow with the development of digital technologies and the requirements of the economy in the face of tough sanctions, however, we are seeing a downward trend. This is the result of the introduction of an assessment of the scientist's and teachers' activities, which is focused only on publishing activity, primarily outside the country, and ignores other types of IR that are necessary for the Russian economy and all societies in a very difficult time.

4 Conclusions

The classification of ES carried out in the work, the development and research of which is currently possible only on the basis of modern digital technologies, creates the terminological base of ES, and, accordingly, the DES. The dictum of Confucius is relevant: "If things are called incorrectly, then words will lose their power." In addition, the classification made makes it possible for scientists to focus on solving urgent problems facing humanity in the field of ecology, conservation of biodiversity on Earth, the optimal combination of the use of pesticides and the capabilities of the biosphere. The EDES formed using mathematical modeling is a digital tool for collecting, accumulating, and using scientific knowledge, which makes it possible to realize the triune role of science and education from a unified systemic position. The scientific community gets the opportunity to more effectively and quickly respond to constantly emerging challenges in the economy, politics, and the social sphere. EDES allows assessing the state of the most important ecosystems in nature—in agriculture.

Acknowledgements. This work was supported by the grant from the Ministry of Science and Higher Education of the Russian Federation, internal number 00600/2020/51896, Agreement dated 21.04.2022 No. 075-15-2022-319.

References

1. Thompson, A., Strickland, A.: *Crafting and Implementing Strategy*. IRWIN (1995)
2. Mescon, M., Albert, M., Khedouri, F.: *Management*. Harper and Row (1988)
3. Ansoff, H.: *Strategic Management*. Springer (2007)
4. Mintzberg, H., Ahlstrand, B., Lampel, J.: *Strategy Safari: A Guided Tour Through The Wilds of Strategic Management*. THE FREE PRESS, New York (2009)
5. Borovkov, A.I., Ryabov, Y.A., Kukushkin, K.V., Maruseva, V.M., Kulemin, V.Y.: Digital twins and digital transformation of defense industry enterprises. *Defen. Technol.* **1**, 6–23 (2018)

6. Medennikov, V, Flerov, Y.: System analysis of digital ecosystem of Russian agriculture. In: IEEE Xplore Digital Library 15 International Conference Management of Large-Scale System Development (MLSD), Moscow, Russia (2022)
7. Tansley, A.: The Use and Abuse of Vegetational Concepts and Terms, Vegetational Concepts and Terms, pp. 284–307 (1935)
8. Degradation Worth Billions: Over 60% of Agricultural Land in Russia is Depleted. <https://agroru.com/news/degradatsiya-na-milliardy-v-rossii-istoscheny-svyshe-60-selh-85534.htm>. Accessed 21 April 2023
9. GMOs and Biodiversity. <https://bio.wikireading.ru/272>. Accessed 21 April 2023
10. Shaitura, S.V., Shaitura, N.S., Ordov, K.V.: Directions of sustainable development of agricultural business. Bull. Kursk State Agricult. Acad. **6**, 239–249 (2022)
11. Shulyat'eva, G.M.: Utilization of food waste with an initial stage in the places of origin as a direction for the sustainable development of the agricultural business. Bull. Vyatka State Agricult. Acad. **2**(8), 9–12 (2021)
12. Industrial ecosystems and eco-industrial parks. https://studopedia.ru/3_180271_promishlennie-ekostemi-i-ekologo-promishlennye-parki.html. Accessed 21 April 2023
13. Nikerina, E.: Business Trend: Product as a Service. <https://gb.ru/posts/biznes-trend-produkt-kak-servis>. Accessed 21 April 2023
14. Kulba, V., Viktor Medennikov, V.: Product traceability digital tool powered by mathematical model for logistics digital platform. In: IEEE Xplore Digital Library 15 International Conference Management of Large-Scale System Development (MLSD), Moscow, Russia (2022)
15. Moore, J.: The Death of Competition: Leadership and Strategy in the Age of Business Ecosystems. Harper Business (1996)
16. Filimonov, O.I., Kasyanenko, T.G., Kukhta, M.V.: Ecosystem as a new organizational and economic form of virtual business. Actual Res. **48**(75), 31–41 (2021)
17. Kovaleva, T.M.: Ecosystem approach in education: the beginning of the path. In: Lifelong Education in the Context of the Future: Collection of Scientific Articles Based on the Materials of the IV International Scientific and Practical Conference, Moscow, pp. 25–31 (2021)
18. Suleymankadiyeva, A.E., Petrov, M.A., Aleksandrov, I.N.: Digital educational ecosystem: the genesis and prospects for the development of online education. Issues Innov. Econ. **11**(3), 1273–1288 (2021)
19. Izotova, A.G., Gavrilyuk, E.S.: Ecosystem approach as a new trend in the development of higher education. Issues Innov. Econ. **12**(2), 1211–1226 (2022)
20. Medennikov, V, Flerov, Y.: Mathematical model of formation of a unified digital platform of scientific and educational resources. In: Proceedings of the International Scientific Conference “Digitalization of Education: History, Trends and Prospects” (DETP 2020), pp. 599–604 (2020)
21. Agrotechnologies of the future. <https://future-agro.ru/>. Accessed 21 April 2023
22. Budzko, V., Medennikov, V.: Mathematical modeling of evaluating the effectiveness of using RSD data in precision farming. In: Procedia Computer Science: 11th, Natal, Rio Grande do Norte, November 10–15, 2020, Natal, Rio Grande do Norte, pp. 122–129 (2020)
23. Medennikov, V.: Mathematical model of the formation of digital platforms for managing the country's economy. Dig. Econ. **1**, 25–35 (2019)



System Analysis of Subject Identification of Digital Twin in Agriculture

Vladimir Budzko^{1,2} , Victor Medennikov² , and Petr Keyer² 

¹ National Research Nuclear University MEPHI Moscow Engineering Physics Institute, Kashirskoe Shosse, 31, 115409 Moscow, Russian Federation
vbudzko@ipiran.ru

² Federal Research Center, “Computer Science and Control” of the Russian Academy of Sciences, Vavilova 44-2, 119333 Moscow, Russia

Abstract. The article is devoted to the problem of resolving the ambiguity and uncertainty of the concept of a digital twin, based on a systematic approach that operates with the concept of a system, the main purpose of which is to achieve a specific goal. A scientific approach to the formation of digital twins is shown on the example of one of the most difficult sectors of the economy for digitalization and modeling—agriculture, based on a single digital framework for managing the industry. Such a framework will make it possible to avoid serious expenditures of finances, time, and human resources with the existing approach to the digital transformation of Russian economies, when each of them begins to form unique management information systems for themselves. It is shown that a single digital twin reflects the global trends in the digitalization of the economy at the present time, the evolution of the digitalization of individual operations to the digitalization of their interconnected complex based on the integration of all operations, including operations of related industries.

Keywords: Digital twin · Agriculture · Digital management framework

1 Introduction

The emergence of the term Digital Twin (DT) is associated with increasing technological changes in all areas of human activity, which has been greatly facilitated in recent years by advances in the improvement of digital technologies and mathematical models. The increased capabilities of information and telecommunication technologies (ITCT) created the conditions for building automated information systems with integration procedures for combining disparate huge data arrays into a single structured environment, including using cloud technologies. Digitalization has found application in agriculture, primarily in precision farming, using data from remote sensing of the earth (RSE), geographic information systems (GIS) technology, and precision manufacturing [1]. Many complex functional tasks are solved when implementing the precision farming technology. They require the integration of large amounts of structured data. Many algorithms developed to solve these problems are built using mathematical modeling and make it possible to describe the main functions of the behavior of most objects. DTs quickly gained popularity and are widely used by the scientific community.

However, the DT design often contains distortions and is misused in business and science. The main provisions of the theory of systems are violated, first, in terms of their main purpose—to achieve a certain goal. The term “DT” is understood in [2]: “Digital twins represent a virtual model of a real object, which is described by mathematical dependencies and is associated with a database of the parameters of this object. A change in one of the parameters entails an automatic change in the remaining parameters and objects associated with it. Moreover, DT is defined in [3] as one basic mathematical model. It is argued that each enterprise has only one DT, and a methodological error has been made: it is impossible to mechanically transfer the country’s input-output balance models to the level of an enterprise, in particular, an agricultural one, where carrots, beets and other products are understood as capital-forming products. Such an attitude towards the term DT in Russia can be explained by the pursuit of buzzwords, the desire to increase their publication activity, which underlies the assessment of the work of scientists.

Therefore, a systematic analysis of the subject identification of the DT is given in this work, as well as a scientific approach to the formation of the DT is shown on the example of one of the most difficult industries in the world for digitalization and modeling—agriculture based on a single digital framework for managing the industry.

2 Genesis of Digital Twins and Problems of Their Development

One of the main problems of understanding DT is the choice of a goal-setting criterion during their development. We will demonstrate this using the example of the system of electronic interaction of citizens in Estonia based on the decentralized X-Road system [4]. The citizen was displayed in 170 databases (DB) with targeted information about each of them before the introduction of this system in the country. However, these databases are largely incompatible and run on different platforms. Is a vehicle registration database considered as a DT? Or should the DT be called only the X-Road system, in which all databases are integrated with the definition of common interfaces and protocols for interaction and data exchange? The answer can be found in 33 works by the founder of the concept, Michael Greaves [5], in particular, in the following studies [6, 7]. He singled out the conceptual unity of three components in the DT: a physical product in real space, a virtual product in virtual space, data and information that combine a virtual and a physical product. In his opinion, “all the information that can be obtained from the product can be obtained under ideal conditions from its digital twin.” Many recognize now that the DT is needed in order to model the behavior of the original in certain conditions. This should help save resources and avoid environmental risks. At the same time, some researchers impose rather stringent requirements on the error of the DT operation, which should not exceed 5%, that is, on the adequacy of the entire complex of models [8]. Such a high adequacy puts forward certain requirements for the completeness, reliability of the data and the accuracy of the applied models that describe the object.

The addition was made to the concept of the DT with the development of the Internet of things—the virtual model is not discarded after the creation of a material object, but is used in conjunction with a physical object throughout the entire life cycle: at the stage of

testing, refinement, operation and disposal. The physical object uses sensors that collect data about the state of the object in real time, after which this information is sent to the DT database. Further, based on the obtained data, the digital model is refined, which, in turn, gives recommendations for optimizing the operation and maintenance of a real object. For example, it predicts the probability of failure of a certain node, specifies the time for preventive maintenance, technical inspection, filter change, and so on. The trend of transition from the DT of individual components and assemblies to the description of finished products and entire enterprises began to be traced.

Great hopes are placed on the introduction of DT due to the rapid growth in the complexity of products in high-tech industries, the bottleneck of which is the design stage. Thus, the trend towards an increase in the time between the start of the development of military aircraft and its entry into the army from 5 years in 1945. up to 27 years in 2025 was identified in the United States. The way out is seen in the development of the following technologies: digital design, modeling and integration, which are the essence of the digital library [8].

3 A Single Digital Framework for Agricultural Management as a Prototype of Digital Twins of Farms

The first experiences of successful development of DT show that their implementation requires significant expenditures of finance, time, and highly qualified specialists. This is due to the use of complex multidisciplinary mathematical models with a high level of adequacy to real materials, structures, physical-mechanical and economic-social processes that underlie the digital design and simulation of the DT. Such models aggregate all the knowledge used in the design, production and operation of both individual products, machines, and production facilities, considering a large set of project goals. Thus, the estimate of financial costs in this case for \$100 billion is given in [8], however, with simultaneous satisfaction in the design process of tens of thousands of target indicators and resource constraints.

Then a quite reasonable question arises: what are the benefits of the DT for mass business, for example, in agriculture, and what should be done for this? Since an individual DT for each enterprise in any industry is practically impossible to develop for the above reasons, a way out can be sought in the joint ownership of some DT by a large number of enterprises. However, a single conceptual, informational and algorithmic space should be created for this on the basis of ontological modeling of subject areas not only in one, but also in a number of related industries. The ontological model is a description of the subject area that uses standard elements of metamodel (for example, objects, relationships, etc.) and strives to explicitly reflect the key aspects of the subject area as fully and reliable as possible. This is dictated by the need for intersectoral integration, and by the interdisciplinary nature of the conceptual space, which already combines technological, biological and economic forms of interaction that historically operated with their ontologies.

The work [9] considers a single digital management frameworks (DMF) agriculture. Information support for the activities of mutual subjects (users of this system), in which the greatest efficiency of their joint activities (for example, minimizing the cost of the

final product) should be enabled by an automated control system (ACS). The creation of a single DMF, on the basis of which such ACS should be built for specific interacting subjects, allows us to ensure a general reduction in the cost of digitalization with the necessary information compatibility of the applied funds. The latter is of great importance for solving the problems of a higher national level. The DMF includes special software and information support. The software that sells algorithms and models also provides collection, storage, processing and transmission of data. Information support includes, first of all, a set of metadata reflecting the ontological model, and a set of single dictionaries, classifiers and reference books. It is assumed that the creation and support of the DMF and its components is carried out from a single center.

Considering the above, the DMF can be considered as the basis of digital doubles of farms. This is confirmed by the completeness of the list of tasks and standard models that aggregate all the knowledge used in the design of crop rotations, the development of technological maps, annual plans and the analysis of their implementation. The DMF also includes unified classifiers, dictionaries, reference books placed in the DMF database based on the description of task algorithms. The digital subframework of primary accounting allows you to record all technological operations, all data of the Internet of things in the database. Thus, because of modeling the digital DMF, the following were obtained: a cloud subframework (digital standard) for collecting and storing primary accounting information in a single database (Fig. 1); cloud subframework (also digital standard) of technological databases (Fig. 2); cloud subframework of the knowledge base in the form of implemented algorithms for management tasks. If the first standard is universal intersectoral in nature for most sectors of the country’s economy, then the second is only sectoral.

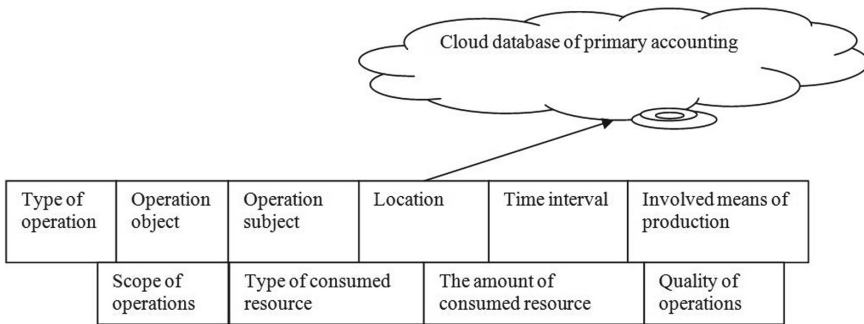


Fig. 1. Cloud subframework for collecting and storing primary accounting information.

The single conceptual information model of crop production for all farms, consisting of 946 attributes, is presented in Fig. 2 as an example. Attribute counts are shown in all blocks in brackets.

This was a huge achievement in the digital transformation of agriculture, far ahead of the West. Thus, the J’son & Partners Consulting company claims that

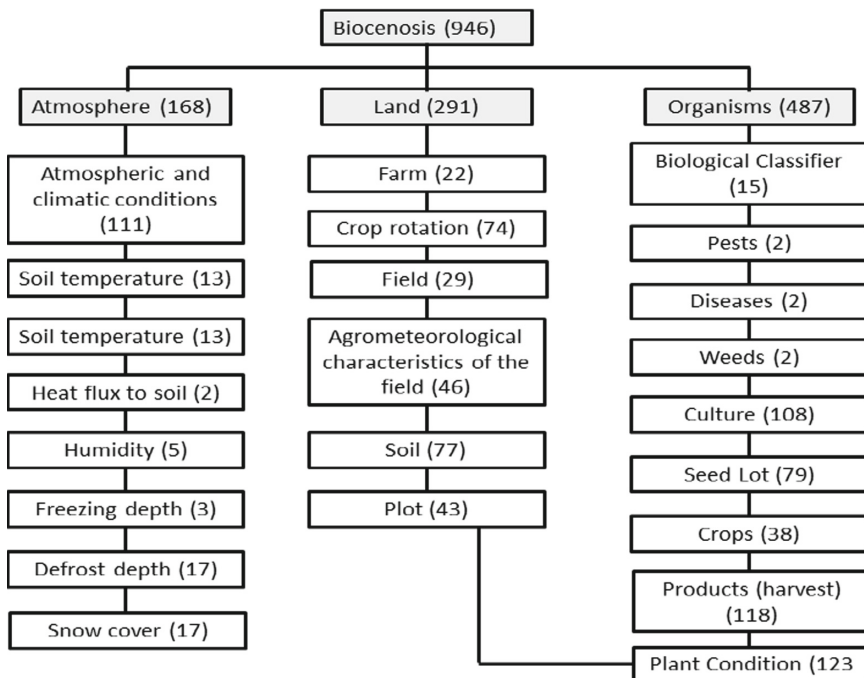


Fig. 2. Digital standard for a unified technological database of crop production.

only at the moment two specialized subframeworks are being formed in US agriculture: subframeworks-aggregators of primary data collection and accumulation and subframeworks of applications (tasks) [10].

These digital subframeworks were tested at the end of the 80s of the last century when informatizing the reference object of a large agricultural complex in the Kuban. This object united 65 enterprises of 19 types. Unified logical structures (models) of industry databases, standard algorithms for functional management tasks for most sectors of the agro-industrial complex (AIC) and almost all agro-industrial enterprises in Russia have been developed. This made it possible to replicate solutions when creating automated control systems. Practice has shown the correctness of the choice of such an approach. The systems have been implemented in over 2 years in more than 1000 enterprises. The confirmed economic effect is more than 20 billion Soviet rubles. Implementation and training centers have been established in the regions.

Two types of agricultural DCs: formed with the participation of leading specialists from the main agricultural research institutes, considering regional and sectoral characteristics and therefore ready for implementation and replication across all the country's farms; DC, models (algorithms) of which have yet to be developed and brought to the required scale of replication. Then, based on Figs. 1 and 2, the single DT of the crop industry will be in the following form (Fig. 3).

Thus, the automated control system of a farm built on the basis of DMF becomes a digital twin (DT) of this farm. However, there is currently no organizational and scientific

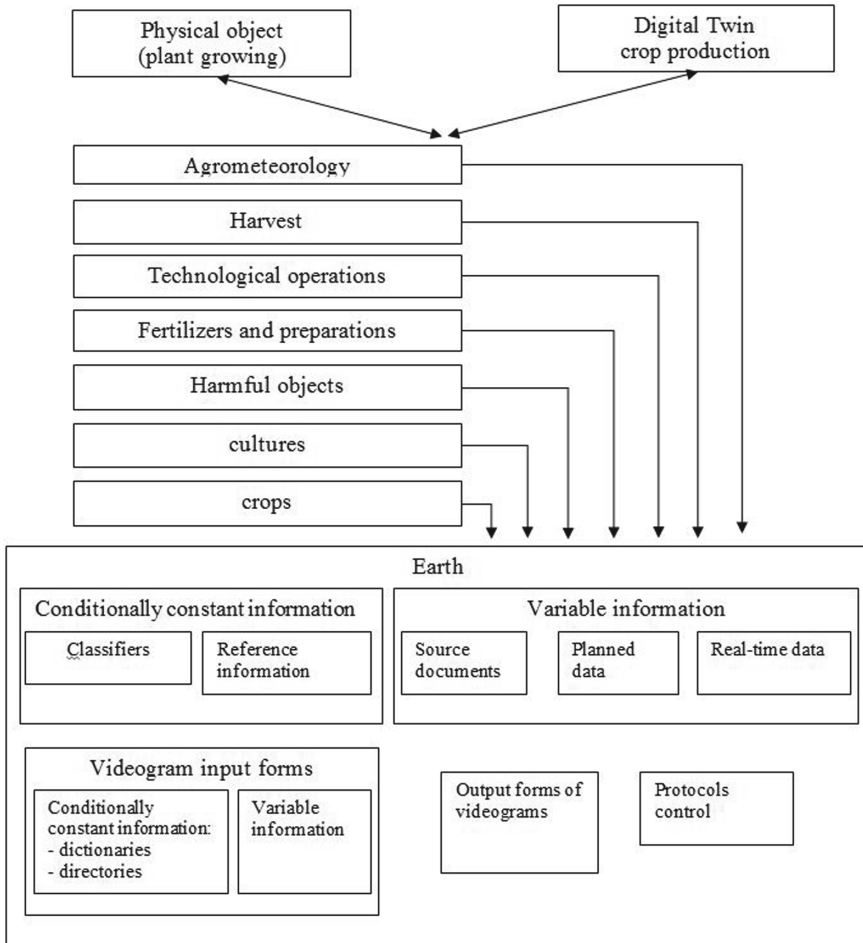


Fig. 3. The structure of a single digital twin of crop production.

structure for the development of the DT in agriculture. The only institute of cybernetics of the agro-industrial complex (AIC) in the whole country, which was systematically engaged in the informatization of the industry, was liquidated at the beginning of perestroika. This institute assumed the functions of the general designer of informatization of the agro-industrial complex. In order to develop an effective digital asset, the Ministry of Agriculture of Russia, with the involvement of science, must restore in one form or another the scientific and design/technological institute for the digital transformation of the agro-industrial complex. Efforts should be directed to the comprehensive development of the most advanced digital technologies, like developed countries, at several reference facilities equipped with modern ITCT, sensors, instruments, technological equipment and machine and tractor fleet. At the same time, all of them must be compatible both with each other and adapted to various digital technologies, covering all possible directions of their development in the world, with the subsequent mass introduction of

the most effective of them throughout the country. This approach will make it possible to form methods for evaluating the correspondence of the digital asset to real objects.

Integration processes in terms of data and algorithms, approaching the time of formation of several typical digital data centers, are actively going on in developed countries, for which centers of innovative development are being created. Our digitalization in the agro-industrial complex is left to the market elements. A huge number of ontologically and functionally incompatible ACSs developed on the basis of the original design appeared in the industry as a result of this approach. According to our calculations, about 5 million ACSs will be created in the future, very truncated due to the reasons considered in the analysis of the DT. This situation leads to extremely low efficiency of the entire digital transformation of the agro-industrial complex. Of course, while a small number of ACSs are being implemented fragmentarily in the AIC, the problems with their ontological integration are not particularly critical, but they will come out on top when a certain threshold value is reached.

4 Conclusions

Informatization in the field of agricultural production is carried out randomly, uncontrolled and without a single coordinating management. Each ACS, in fact, acts as a limited digital double of its subject area and the functions performed in it. DC differs in the composition of the tasks to be solved and informationally incompatible, which does not allow the reinforcement of complex tasks of the upper level. The total costs of the information can be significantly reduced with the provision of information combinability by applying standard design solutions—the creation of a single DCF with a single general information, software and mathematical support. The construction of ACS manufacturers based on the DCF with setting up for the features of a particular subject area will allow you to make many compatible DCs. It is necessary to take advantage of the USSR experience in organizing the development, implementation, support and development of this DSF.

Acknowledgements. This work was supported by the grant from the Ministry of Science and Higher Education of the Russian Federation, internal number 00600/2020/51896, Agreement dated 21.04.2022 No. 075-15-2022-319.



References

1. Budzko, V., Medennikov, V.: Mathematical modeling of evaluating the effectiveness of using RSD data in precision farming. *Procedia Computer Science: 11th, Natal, Rio Grande do Norte, 10–15 Nov 2020*. Natal, Rio Grande do Norte, pp. 122–129 (2020). <https://doi.org/10.1016/j.procs.2021.06.015>
2. Ponomarev, K.S., Shutikov, M.A., Feofanov A.N.: Digital twin as a tool for digital transformation of an enterprise. *Bull. Moscow State Tech. Univ. “Stankin”* **4**(51), 19–23 (2019)
3. Sytov, A., Vakhranov, A., Ereshko, F.: Enterprise digital twin research. In: *IEEE Xplore Digital Library. 14th International Conference on «Management of Large-Scale System Development» (MLSD'2021)*, Moscow, Russia (2021). <https://doi.org/10.1109/MLSD49919>

4. Vassil, K.: Estonian e-government ecosystem: foundation, applications, outcomes. World Development Report 2016 (Electronic resource). Retrieved from <https://thedocs.worldbank.org/en/doc/165711456838073531-0050022016/original/WDR16BPEstonianeGovecosystemVassil.pdf> on 05 Nov 2023
5. Grieves, M.: ResearchGate. Digital Twin Institute. <https://www.researchgate.net/profile/Michael-Grieves>. Last accessed 21 April 2023
6. Grieves, M.: Product lifecycle management: Driving the next generation of lean thinking. McGraw-Hill, New York (2006)
7. Grieves, M., Vickers, J.: Digital twin: mitigating unpredictable, undesirable emergent behavior in complex systems. In: Kahlen, F.J., Flumerfelt, S., Alves, A. (eds.) Transdisciplinary Perspectives on Complex Systems. Springer, Cham, p. 6330 (2017)
8. Borovkov, A.I., Ryabov, Y., Kukushkin, K.V., Maruseva, V.M., Kulemin, V.: Digital twins and digital transformation of defense industry enterprises. *Defense Technol* **1**, 6–23 (2018)
9. Medennikov, V., Raikov, A.: Formation of the digital platform for precision farming with mathematical modeling. *CEUR Workshop Proc* **2790**, 114–126 (2020)
10. J'son & Partners Consulting. Analysis of the market of cloud IoT platforms and applications for digital agriculture in the world and prospects in Russia. https://json.tv/en/ict_telecom_analytics_view/analysis-of-the-market-of-cloud-iot-platforms-and-applications-for-digital-agriculture-in-the-world-and-prospects-in-russia. Last accessed 21 April 2023



Topological Analysis of Protein Surfaces and Its Role in the Development of New Medicines

Oleg V. Bystrov  and Sergey D. Kulik  

National Research Nuclear University MEPhI, Kashirskoe Shosse 31, Moscow 115409, Russia
sedmik@mail.ru

Abstract. The problems of storing, systematization, and accounting for data, accumulated in the process of research on the three-dimensional structure of proteins have been considered in the paper. The main purpose of this paper is to present a new biological information system for mathematical research with cognitive elements. The processes of preparing and carrying out a mathematical analysis have been studied. During the design and development of the information system, the processes of preparing and conducting mathematical analysis, topology analysis, triangulation and generation of the foliation of molecular surfaces, diagramming and further interpretation of the analysis results were studied in order to determine the areas of the molecule that can be involved in interaction with other molecules, a structural model in the form of a block model was built and developed schemes and algorithm of work. The resulting system prototype should ease the process of taking into account all of the data, obtained during research, as well as be useful for developing cognitive technologies.

Keywords: Information system · Cognitive technology · Topological research · Artificial intelligence

1 Introduction

Information technologies and various information systems in particular, as well as cognitive technologies, play an important role in our lives. Many publications talk about it and focus on it. Information technology can be used in biology and related fields of sciences, for example, artificial intelligence [1] in the fields of medicine, such as diagnosis, treatment, prediction and management of patient health, the development of a new generation of artificial intelligence [2] in the field of solving complex problems corresponding to the challenges of modernity, intelligent information system [3] in telemedicine, which is designed to improve the quality and efficiency of providing medical services at a distance, an intelligent system [4] in the field of educational goals using a genetic algorithm, using topology it is possible to represent the spatial structure of a protein in the form of polynomials [5]. The fundamental aspects of the human mind have not yet been fully studied, while there are encouraging results [6] based on multidimensional cognitive maps. Cognitive technologies can also be used to develop new medical preparations or improve existing ones.

Proteins [7] are particularly complex and not fully understood objects of our world. Despite this, there has been some progress in scientific research of the structures of molecules and their interactions with each other through the joint use of information technology and mathematical methods, including topology. Topology [8] is a branch of mathematics that studies the properties of spaces and shapes that are preserved under continuous transformations, such as stretching, compression, bending and gluing.

One of the important aspects of topology is the study of topological invariants, which are properties of spaces that do not depend on their specific shape or size. The study of protein interactions belongs to the field of biochemistry, biophysics and structural biology. These disciplines study various aspects of proteins and their interactions, including their structure, function, dynamics and regulation.

Conducting mathematical studies of molecular structures is important for understanding the molecular foundations of biological processes, developing new drugs and engineering proteins with desired properties. It is at this junction of sciences that three-dimensional modeling of protein molecules and a comprehensive analysis of the resulting molecular structure for the further development of biomedicine come to the rescue.

There are many algorithms and methods for studying three-dimensional atomistic models of protein molecules [9] that help to decipher the spatial structures of proteins, including enzymes, membrane proteins, ion channels, aquaporins, ribosomal subunits, etc.

Topological analysis of protein molecules can help to identify various aspects of their structure and interactions, which can be useful for understanding their function and properties. For example, topological methods can help determine the three-dimensional structure of a protein molecule and its shape, and topological analysis will help to investigate the topological characteristics of contact zones between proteins and other molecules, as well as analyze changes in topology during binding. Further, the protein analysis is evaluated by scientists who make conclusions on each study. Data on the structure of the protein under study can be obtained from publicly available protein databases, such as the RCSB Protein Data Bank.

During scientific research, a large amount of data accumulates: research requests, samples of molecules and information about mathematical analyses performed, the results of interaction analysis. All this data should be stored and also used for planning future research. That is why the main purpose of this article is to present a special biological information system for scientific research that can solve these problems.

Thus, the article is organized in the following way: Sect. 1 contains a brief introduction; general information about mathematical research is introduced in Sect. 2; Sect. 3 introduces a prototype of the biological information system for scientific research. Section 2 contains a description of the topology process and Sect. 3 contains the block diagram of the algorithm of the application for filling the protein structure. Conclusions and main results are outlined at the end of the article.

2 Mathematical Research

Research begins at the request of a researcher to conduct a study of the three-dimensional structure of proteins in order to determine the areas of the molecule that can be involved in interaction with other molecules, they are carried out using information technology

and a modern programming language such as Python, including its numerous libraries for information processing and visualization of the resulting data.

To begin studying the three-dimensional structure of a protein, it is necessary to obtain the protein structure in PDB format (PDB is the most common format for storing three-dimensional protein structures). Protein structure data can be obtained from publicly available protein databases such as the RCSB Protein Data Bank. In addition to the PDB format, you can also use formats such as CIF (Crystallographic Information File), MOL2 (Sybyl Molecular Library), SDF (Structural Data File), as well as many other formats used to store information about molecules and protein structures.

Next, it is necessary to determine which specific atoms or protein chain need to be investigated. This may be due to the presence of special properties of these atoms or chains. After determining the chain or group of atoms to be examined, the triangulation of the surface of the molecule is performed using an algorithm such as MSMS (English Macromolecular Surface Mesh Simplification) is an algorithm that performs triangulation of the surface of the molecule using atoms as input data. After that, foliation of the surface of the molecule is generated using the Fomenko-Tsishangov Complex [10]. This allows you to split the surface into a set of curves, called folions, which reflect its features and topological properties.

After that, Betti numbers are calculated for each foliation level, which are numbers that characterize the topological structure of the molecule. Then, based on the data obtained, a Poincare diagram is constructed, which visualizes these numbers and helps to understand the structural features of the molecule.

Next, the Poincare diagram is analyzed to determine the regions of the molecule that can be involved in interaction with other molecules. A Poincare diagram can show the presence of loops, cavities, or other structural features that may be important for interactions.

After that, the interactions of the molecule with other molecules are analyzed using docking and/or molecular dynamics methods, the docking method allows you to predict the structure of the protein and ligand complex forming a stable bond, for this you need to choose the optimal position and orientation of the ligand in the active center of the protein to achieve maximum energy stability, the molecular dynamics method allows you to simulate the movement of atoms and molecules based on the laws of mechanics of their interaction.

Using this method, it is possible to study the dynamics of the interaction of a protein molecule with other molecules, such as ligands, other proteins or RNA. The data obtained are interpreted in the context of the function of the protein and its role in the body and a response is prepared to a request for the development of new drugs or the improvement of existing ones.

3 Prototype of the Biological Information System for Scientific Research

Assume that the biological information system for scientific research consists of a variety of different blocks. These blocks have a rather complex structure. It includes a database, a special format for storing protein structures PDB, a query repository and a block

for processing it, a protein structure repository and a block for receiving it, a block for preparing for research, a block for conducting research and a block for preparing responses to requests. Each of the listed blocks implies a specific software with a user interface for a laboratory employee. Figure 1 shows a diagram of these blocks.

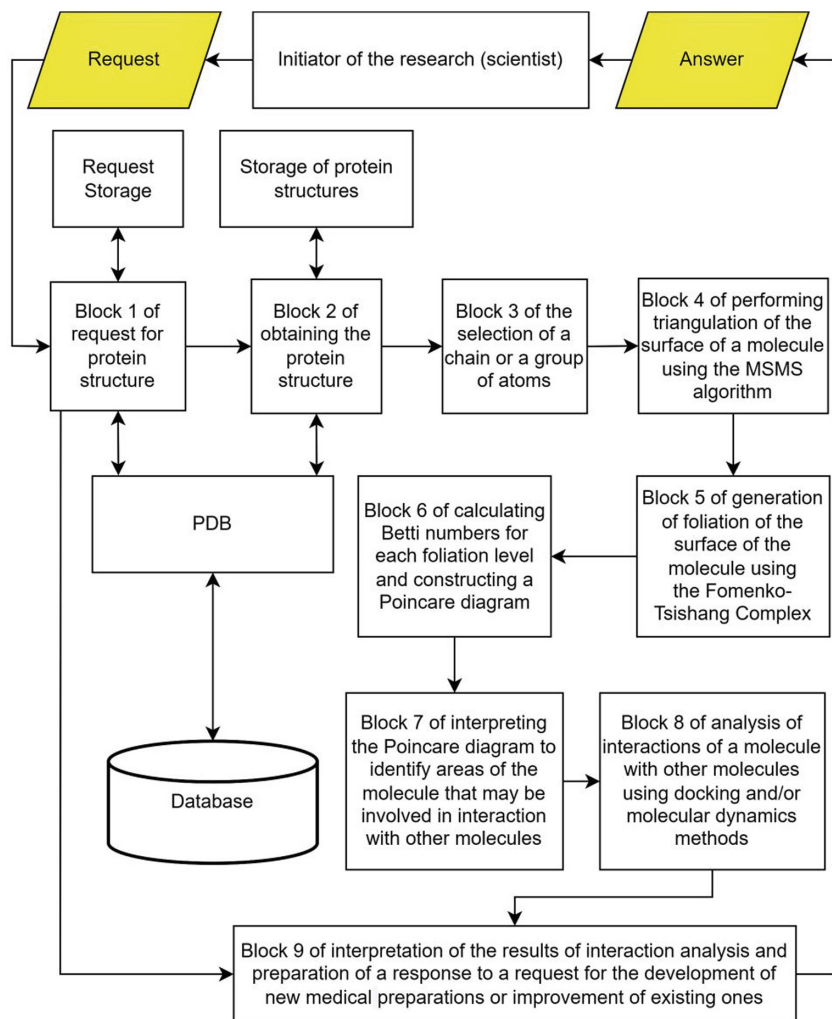


Fig. 1. Structural diagram of the biological information system for scientific research.

The initiator of the research request may be a researcher (or any other employee of educational institutions), for example, a research institute for the study of protein structures.

Block 1 involves the preparation of a request card by a researcher. At the same time, information about the generated card enters the database using the DBMS in PDB format, while the request letter enters the query repository.

Block 2 includes obtaining the protein structure provided by the initiator of the request. These samples will be further stored in the storage of protein structures. Information about proteins is entered into the database in the administration mode.

Block 3 is responsible for selecting a chain or group of atoms. The researcher enters the data on the selection for the study in the report sheet.

Block 4 is responsible for performing triangulation of the surface of the molecule using the MSMS algorithm.

Block 5 is responsible for generating foliation of the surface of the molecule using the Fomenko-Tsishang complex [11].

Block 6 is engaged in calculating the Betti number for each foliation level and constructing a Poincare diagram.

Block 7 interpretation of the Poincare diagram to identify areas of the molecule that may be involved in interaction with other molecules.

Block 8 analyzes the interactions of a molecule with other molecules using docking and/or molecular dynamics methods.

Block 9 receives an initial request to obtain the protein structure, and also interprets the results of the interaction analysis and prepares a response to a request for the development of new drugs or improvement of existing ones.

In case of refusal to conduct the study due to certain reasons, flow control is transferred to block 9 immediately after block 1. At the output, the requester receives a response to their inquiry.

The generalized algorithm of the system operation in case of creating a new request for the application to fill the protein structure with data is shown in Fig. 2 in which the main sections of the application are divided into modules.

Before the modules are the preceding sections with the conditions on the basis of which the execution of the section with the module begins. By applying a potentially possible windowed application to the algorithm, it is possible to identify the main points of the application. The initial logical entry point to the algorithm is the fulfillment of the condition of pressing the button to load the protein structure. After that, the execution of Module 1 begins.

Module 1 is responsible for opening the form to fill in the protein structure until the "Close" button is pressed. After that, the algorithm is branched, one branch of which leads to the end of the algorithm, the other to the next condition.

Module 2 follows after the addition of the protein structure, which is responsible for establishing a connection to the database.

Module 3, if the data received from the database is correct, is responsible for displaying a message about the successful execution of the operation.

Module 4, if the data received from the database is not correct, is responsible for displaying an error message about the successful execution of the operation.

4 Conclusion

As a result of this research, a structure of the biological information system for scientific research was designed. The results of the study may be useful to medical staff involved in this field of work. The collected information will be used to create a new version of

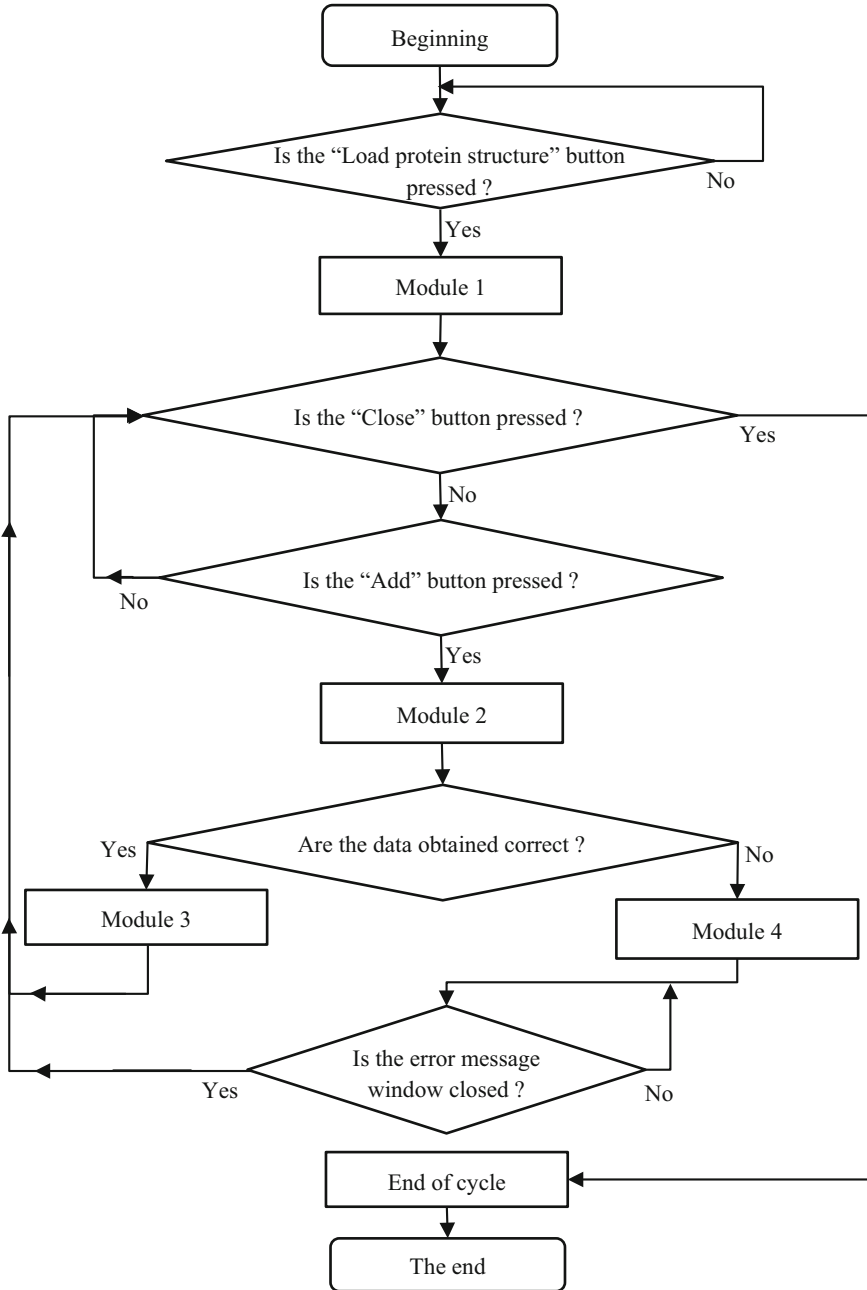


Fig. 2. Block diagram of the algorithm of the application for filling the protein structure.

the information system. Studying the possibility of using topology analysis in scientific research can increase the effectiveness of medical research and improve the tools of

employees involved in protein analysis. Understanding the three-dimensional structure of proteins allows the development of targeted drugs aimed at specific protein targets in the brain. This may be an important aspect in the development of medicaments for the treatment of cognitive disorders or the improvement of cognitive functions. The development of special software for the study of the three-dimensional structure of proteins in order to determine the areas of the molecule that can be involved in interaction with other molecules is the subject of our future experimental research. We believe that our results are useful for researchers in the field of cognitive and information technologies. For example, the development of methods for computing molecular tags [12], the calculation of bifurcation diagrams of many important integrable systems [13], the use of neural networks to work with limited training data sets in the context of intelligent robotics [14], as well as the use of memristors in neural networks and the analysis of their accuracy in various tasks of the context of control and communication systems [15], development of new materials with improved properties and new methods and technologies for obtaining coatings [16] and the use of various machine learning methods and efficiency evaluation, followed by the selection of the most appropriate method for solving a specific spam detection problem [17].

Acknowledgments. This work was supported by the MEPhI Program Priority 2030.

References

1. Yasnitsky, L.N.: Artificial intelligence and medicine: history, current state, and forecasts for the future. *Curr. Hypertens. Rev.* **16**(3), 210–215 (2020)
2. Samsonovich, A.V., Shumsky, S.A., Karpov, V.E., Kotov, A.A., Kolonin, A.G.: Key advanced research initiative: a manifesto for the new-generation artificial intelligence. *Procedia Comput. Sci.* **213**, 824–831 (2022)
3. Kondakov, A., Kulik, S.: Intelligent information system for telemedicine. *Procedia Comput. Sci.* **169**, 240–243 (2020)
4. Protopopova, J., Kulik, S.: Educational intelligent system using genetic algorithm. *Procedia Comput. Sci.* **169**, 168–172 (2020)
5. Kalishenko, E.L., Krinkin, K.V.: The system of topological modeling of the structure of protein molecules. *Appl. Inform.* **4**(22), 114–124 (2009)
6. Samsonovich, A.V., Ascoli, G.A.: Cognitive map dimensions of the human value system extracted from natural language. In: Goertzel, B., Wang, P. (eds.) *Frontiers in Artificial Intelligence and Applications*, vol. 157, pp. 111–124. IOS Press (2007)
7. Finkelstein, A.V., Ptitsyn, O.B.: *Protein Physics: A Course of Lectures with Color and Stereoscopic Illustrations and Tasks*, 3rd edn. KDU, Moscow (2012)
8. Retyunskaya, A.K.: Research laboratory work on the topic topological properties of figures. *Mobius Leaf Quest. Pedagogy* **4**(1), 213–217 (2020)
9. Shaitan, K.V., et al.: Algorithms and methods for the study of three-dimensional atomistic models of protein molecules based on the analysis of the scattering pattern of high-power X-ray laser radiation. *Nanostruct. Math. Phys. Model.* **9**(2), 33–74 (2013)
10. Fomenko, A.T., Qishang, H.: On the topology of three-dimensional manifolds arising in Hamiltonian mechanics. *Rep. Acad. Sci. USSR* **294**(2), 283–287 (1986)
11. Novikov, D.V.: Topology of features of integrable Hamiltonian systems with non-compact level surfaces. Dissertation for the Degree of Candidate of Physical and Mathematical Sciences, Moscow State University (2013)

12. Bolsinov, A.V., Richter, P.H., Fomenko, A.T.: The method of circular molecules and the topology of the Kovalevskaya top. *Math. Collect.* **191**(2), 3–42 (2000)
13. Kharlamov, M.P.: *Topological Analysis of Integrable Problems of Rigid Body Dynamics*. Leningrad University Press, Leningrad (1988)
14. Kulik, S.D., Shtanko, A.N.: Experiments with neural net object detection system YOLO on small training datasets for intelligent robotics. In: *Mechanisms and Machine Science*, vol. 80, pp. 157–162 . Springer, Cham (2020)
15. Galushkin, A.I., Danilin, S.N., Shchanikov, S.A.: The research of memristor-based neural network components operation accuracy in control and communication systems In: *2015 International Siberian Conference on Control and Communications, SIBCON 2015—Proceedings*, pp. 1–6 (2015)
16. Guseva, A.I., Silenko, A.N.: A study of the possibility of obtaining deposited coatings based on intermetallic titanium and aluminum compounds using the chemical vapor transport method. *WSEAS Trans. Environ. Develop.* **17**, 1039–1045 (2021)
17. Kontsewaya, Y., Antonov, E., Artamonov, A.: Evaluating the effectiveness of machine learning methods for spam detection. *Procedia Comput. Sci.* **190**, 479–486 (2021)



Development of a Multi-agent Architecture for an Object Shape Recognition System Based on Data from a Depth Sensor

Kantemir Bzhikhatlov^(✉) , Murat Anchokov , and Olga Nagoeva 

The Federal State Institution of Science Federal Scientific Center Kabardino-Balkarian Scientific Center of Russian Academy of Sciences, I. Armand Street, 37-a, 360000 Nalchik, Russia
kbncran@mail.ru

Abstract. The article presents the concept of an object shape recognition system for an autonomous robot based on a multi-agent neurocognitive architecture. The Microsoft Kinect for Windows controller is considered as an example of a depth sensor. The algorithms of the program for pre-processing data from the depth sensor and the multi-agent neurocognitive architecture responsible for building a model of the surrounding space, where each object is associated with a certain cloud of three-dimensional points, are presented. The proposed data processing system can be used as an element of the navigation and orientation system for autonomous mobile robots used in a real environment.

Keywords: Depth sensor · Point cloud · Pattern recognition · Robotics · Navigation · Neurocognitive algorithms · Kinect

1 Introduction

The current level of development of autonomous and collaborative robotics puts ever higher demands on the systems of navigation and orientation of the robot in a real environment. At the same time, to simulate the goal-directed behavior in the problem of orientation and navigation of mobile robots, in addition to sensors that determine the position and direction of movement, a set of additional sensors, including cameras and microphones, is used. In particular, depth sensors (for example, Microsoft Kinect) can be used to build a more complete model of the external environment. This controller consists of a set of sensors: an RGB camera, a depth sensor, an array of microphones to determine the position of the sound source [1]. Similar sensors have become widespread outside the video game industry due to the wide distribution of these sensors and their low price. The depth sensor is used to assess motor activity in people with problems of the musculoskeletal system [2, 3], to recognize and correct minor spinal curvatures in athletes [4], and to monitor the behavior of animals in natural conditions [5]. In addition, Kinect is often used as an element of the control system or orientation of autonomous mobile robots for various purposes [6–8]. At the same time, of interest is the system for processing data obtained from such depth sensors, which from a cloud of points will

make it possible to obtain three-dimensional models of individual objects of the real environment surrounding an autonomous robot. At the moment, there are a number of approaches that use neural networks and deep learning to process 3D point clouds [9, 10], but such approaches require the preparation and labeling of a large training set. One of the approaches to designing decision-making systems for autonomous mobile robots is multi-agent neurocognitive architectures based on proactive agents with their own objective function. In particular, this article discusses an architecture whose agents are models of individual neurons with the function of maximizing internal energy and the possibility of exchanging information and energy between each other. Such agents conditionally correspond to the areas of the brain and should provide a cooperative solution to the problem of controlling the robot [11]. The development of an algorithm for processing data from a depth sensor based on such architectures will solve a number of problems associated with the difficulties of orienting autonomous software and robotic systems in complex, partially observable and dynamically changing conditions of a real environment.

The object of the research is the systems of shape recognition of objects.

The subject of the study is the algorithms of the system for recognizing the shape of objects according to the data of the depth sensor.

Purpose of the work: development of an object shape recognition system based on multi-agent neurocognitive architectures.

2 Program Architecture for Object Recognition Based on Data from a Depth Sensor

The use of a depth sensor will provide not only the determination of distances to obstacles, but also the construction of three-dimensional models of objects around the robot. Such sensors most often consist of an infrared projector and an infrared sensor, which allows it to illuminate surrounding objects and, through 3D scanning, build a depth map. As a result, the control system based on data from the sensor can build a three-dimensional representation of environmental objects, which will allow building more efficient routes for the movement of autonomous robots and their manipulators (Fig. 1). And the combination of this sensor with other sources of information about the external environment will ensure the emergence of multimodal models of objects in the external world (for example, when a plant is detected, the decision-making system should have information about its appearance, location in space, shape and size, and sounds emitted). The depth sensor was used to provide the task of navigation and orientation of an autonomous agricultural robot in a real environment [12]. The robot is a transport platform with four independent driven motor-wheels. In addition, a liquid supply system and a set of manipulators equipped with several nozzles for spraying plants are installed on the robot. Ensuring constant monitoring of the state of crops (in particular, corn) and ensuring timely treatment from diseases, pests and weeds is the main task of the developed robot. As a result, the autonomous robot is forced to move along the row-spacing of corn crops and ensure the guidance of the corn flours on specific plants.

The robot is equipped with distance sensors (ultrasonic and infrared rangefinders), an inertial sensor (accelerometer and gyroscope), a compass, lidar and GPS module, as

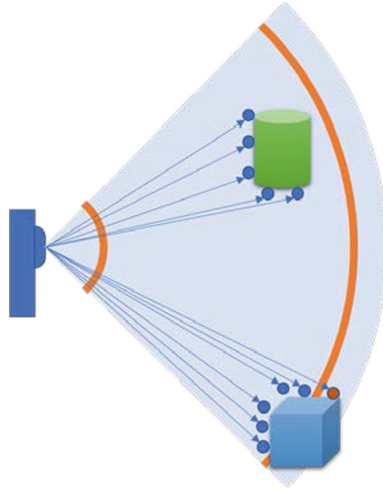


Fig. 1. Depth sensor operation diagram.

well as a depth sensor based on the Microsoft Kinect for Windows sensor to provide navigation and orientation in a real environment. This model of the depth sensor allows you to build a cloud of points in the range from 40 cm to 3 m. The resulting depth map has a resolution of 320x240, with a frequency of 30 frames per second and 16-bit depth. The appearance of an autonomous mobile robot with an installed depth sensor is shown in Fig. 2.



Fig. 2. Autonomous agricultural robot with installed depth sensor.

It is worth noting that the Kinect sensor also includes a conventional RGB camera, the data from which is used in the task of recognizing environmental objects, which will allow us to correlate the received visual images and depth maps. The object recognition algorithm based on camera data is described in [13]. The result of such a combination will be the recognition of several modalities of the object (for example, the decision-making system will be able to obtain information about the appearance of corn, about its three-dimensional model, location and condition).

3 Depth Sensor Software Processing

Further processing of data from the depth sensor will allow building three-dimensional models of objects in the real environment (Fig. 3), which will create an idea of the structure of the space around the autonomous robot. To do this, clustering of points in three-dimensional space is carried out (the algorithm is shown in Fig. 4).

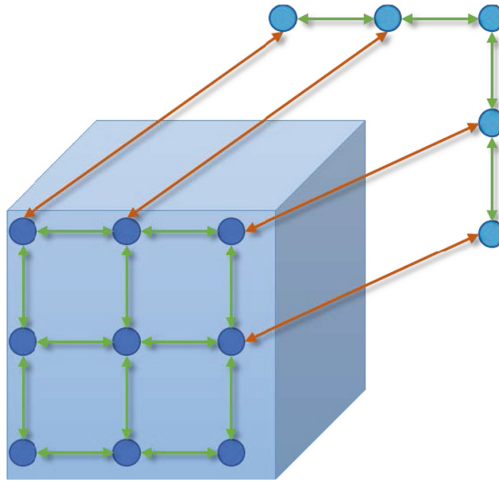


Fig. 3. Planes determining from depth sensor data example.

After launch, the program should ensure the capture of the frame. The resulting point cloud is looped through, as a result of which the distance to neighboring points is calculated for each point. If the distance is below the threshold, the point is included in the cluster. Otherwise, all points of the current cluster are sent to the decision system as a single array, after which the creation of the next cluster is initialized. A program that allows offline capturing frames from the sensor, pre-processing them and sending the received data to the decision-making system of an autonomous robot [14] was developed to ensure the collection and processing of data from the depth sensor.

Figure 5 shows the interface of the developed program. The program window displays the captured image (where the light areas correspond to the most distant points, and the dark areas are closest to the sensor) and the result of its preliminary processing. In addition, the window has areas for setting up the capture and preprocessing of data from

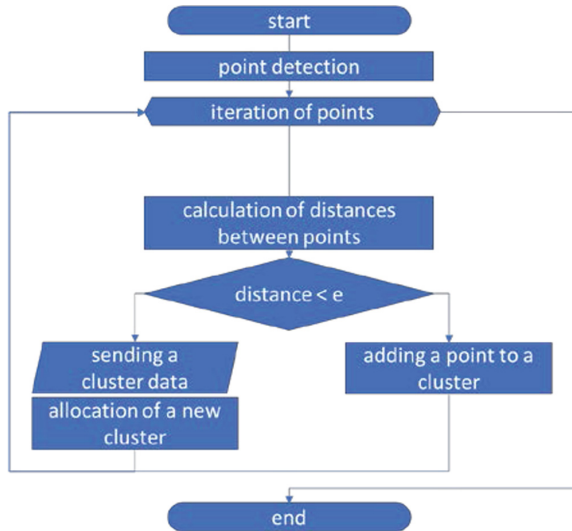


Fig. 4. Algorithm of the program for selecting planes according to the depth sensor data.

the sensor, settings for connecting the program to the decision-making system, and a window for displaying processing results in the form of a text description of a point cloud.

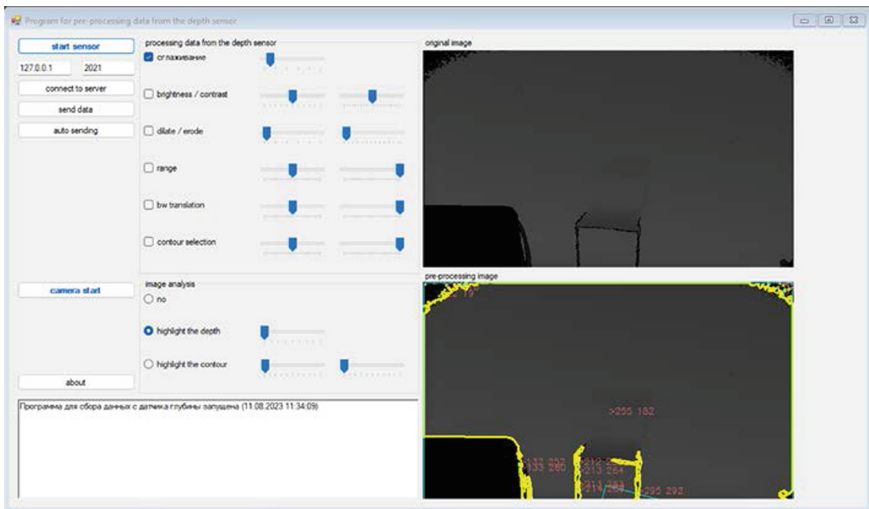


Fig. 5. Program for pre-processing data from the depth sensor screenshot.

This program allows you to apply a number of filters to the captured frame from the depth sensor, in particular, to eliminate noise, you can use Gaussian smoothing, changing contrast and brightness, expanding and narrowing light areas and image binarization.

The obtained data can be used in the form of a matrix of points or in the form of areas with the same distance (image clustering is provided by the Kenny filter and the contour detection algorithm).

4 Multi-agent Depth Sensor Data Processing Algorithms

After preprocessing the resulting point cloud, it is necessary to ensure the transfer of the received data for further processing due to the multi-agent neurocognitive architecture. Data transfer is also provided by the developed program. Further data analysis is performed using the developed multi-agent architecture for building a point cloud and a model of three-dimensional objects based on data from a depth sensor (Fig. 6).

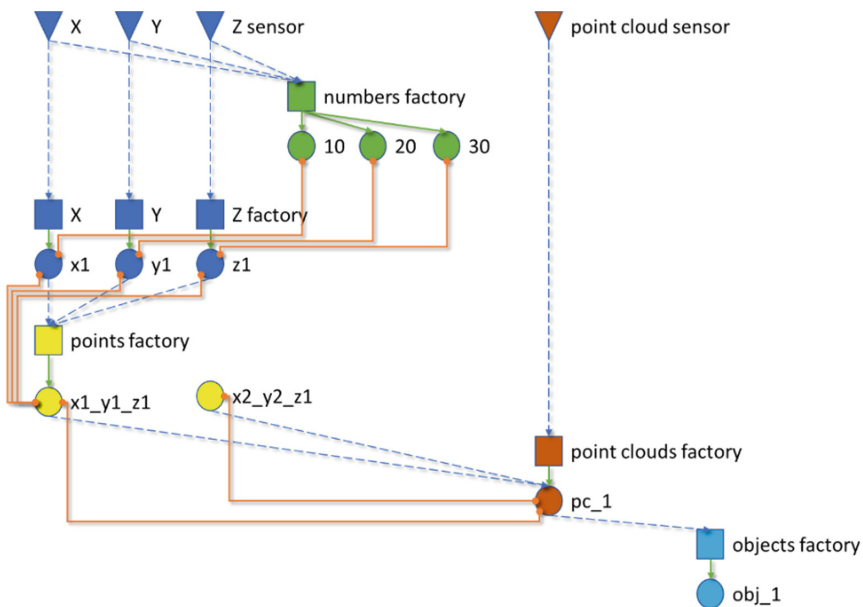


Fig. 6. Multi-agent depth sensor data processing algorithms

At the moment of receiving a signal from the depth sensor, a set of point clouds is sent to the decision-making system core in the form of an array of X, Y and Z coordinates and the number of the point cloud. All data gets to the corresponding sensors of the multi-agent architecture, and then sent to the “factories” of points and point clouds. These factories are a set of actors responsible for representing the plane in the 3D environment (denoted as “pc_1” in the figure). Separately, work is underway to process other modalities of the sensory stream, which provide recognition of objects and their features (appearance, position, and other measurable properties). When an object is recognized, a separate actor is created that enters into multi-agent contracts with their associated point clouds (for example, the contract “pc_1” and “obj_1”). Figure 7 shows the processing of data from the depth sensor.

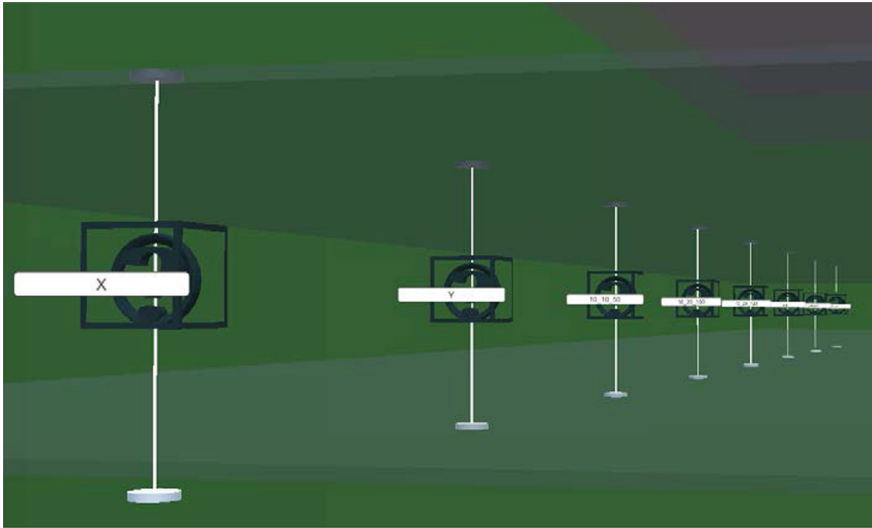


Fig. 7. Screenshot of the presented algorithm created in the multi-agent architecture editor

The result of the operation of the described algorithm makes it possible to create an idea of each three-dimensional point related to an object of a real environment in a multi-agent architecture. As a result, a representation of the shape of each object around the mobile robot is built. This will ensure an adequate construction of a model of the surrounding space. This architecture is part of a general architecture for orientation and navigation of an autonomous mobile robot. In the future, it is planned to test the developed architecture in the task of collecting data from an autonomous agricultural robot.

5 Conclusion

A pre-processing program and a multi-agent architecture of an object shape recognition system based on data from a depth sensor have been developed to perform the task of navigating and orienting an autonomous robot. The developed program allows you to get a depth map and a map of the contours of areas with the same distance from the sensor. In addition, a multi-agent neurocognitive architecture for processing data from a depth sensor is described, which will allow building a representation of the shape and location of each object in a real environment. The use of a depth sensor with data post-processing will make it possible to develop more efficient methods for navigating and orienting autonomous robots in a dynamically changing partially observed external environment.

References

1. Kinect for Windows: <https://developer.microsoft.com/ru-ru/windows/kinect/>. Last accessed 01 July 2023
2. Hocking, D.R., et al.: Feasibility of a virtual reality-based exercise intervention and low-cost motion tracking method for estimation of motor proficiency in youth with autism spectrum disorder. *J. Neuroeng. Rehabil.* **19**, 1 (2022). <https://doi.org/10.1186/s12984-021-00978-1>
3. Xiaoa, B., et al.: Design of a virtual reality rehabilitation system for upper limbs that inhibits compensatory movement. *Med. Novel Technol. Dev.* **13**, 100110 (2022). <https://doi.org/10.1016/j.medntd.2021.100110>
4. Wang, H., Shi, J., Luo, X.: Swimmer's posture recognition and correction method based on embedded depth image skeleton tracking. *Wirel. Commun. Mob. Comput.* **2022**, 8775352 (2022). <https://doi.org/10.1155/2022/8775352>
5. Xu, J., Zhou, S., Xu, A., Ye, J., Zhao, A.: Automatic scoring of postures in grouped pigs using depth image and CNN-SVM. *Comput. Electron. Agric.* **194**, 106746 (2022). <https://doi.org/10.1016/j.compag.2022.106746>
6. Liu, H., Stoll, N., Junginger, S., Thurov, K.: Human face orientation recognition for intelligent mobile robot collision avoidance in laboratory environments using feature detection and LVQ neural networks. In: *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Zhuhai, China, 2003–2007 (2015). <https://doi.org/10.1109/ROBIO.2015.7419067>
7. Lee, J.Y., Lee, C.-S.: Path planning for SCARA robot based on marker detection using feature extraction and labelling. *Int. J. Comput. Integr. Manuf.* **31**(8), 769–776 (2018). <https://doi.org/10.1080/0951192X.2018.1429669>
8. Ryumin, D.A., Kagirov, I.A.: Approaches to automatic gesture recognition: hardware and methods overview. *Manned Space Flights [Pilotiruyemye kosmicheskiye polety]* **3**(40), 82–99 (2021). <https://doi.org/10.34131/MSF.21.3.82-99>
9. Döllner, J.: Geospatial artificial intelligence: potentials of machine learning for 3D point clouds and geospatial digital twins. *PhG—J. Photogram. Remote Sens. Geoinform. Sci.* **88**(1), 15–24 (2020). <https://doi.org/10.1007/s41064-020-00102-3>
10. Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M.: Deep learning for 3D point clouds: a survey. *IEEE Trans. Patt. Anal. Mach. Intell.* **43**(12):4338–4364 (2021). <https://doi.org/10.1109/TPAMI.2020.3005434>
11. Nagoev, Z.V.: *Intellect, or thinking in living and artificial systems [Intellect ili myshleniye v zhivyykh i iskusstvennykh sistemakh]*. Publishing House KBNTS RAS, Nalchik (2013)
12. Ksalov A.M., Bzhikhatlov K.Ch., Pshenokova I.A., Zammoev A.U.: Development of a transport subsystem for autonomous robots for plant protection system. *Izvestiya Kabardino-Balkarskogo nauchnogo centra RAN [News of the Kabardino-Balkarian Scientific Center of RAS]* **2**(106), 31–40 (2022). <https://doi.org/10.35330/1991-6639-2022-2-106-31-40>
13. Nagoev, Z., Pshenokova, I., Bzhikhatlov, K., Kankulov, S., Atalikov, B.: Multi-agent neurocognitive architecture of an intelligent agent pattern recognition system. *Procedia Comput. Sci.* **213**, 504–509 (2022). <https://doi.org/10.1016/j.procs.2022.11.097>
14. Bzhikhatlov, K.Ch., Abazokov, M.A., Unagasov, A.A.: Program for collecting data from a depth sensor. Certificate of registration of a computer program № 2022616219 (2022)



A Statistical WavLM Embedding Features with Auto-Encoder for Speech Emotion Recognition

Adil Chakhtouna¹(✉), Sara Sekkate², and Abdellah Adib¹

¹ Team Data Science & Artificial Intelligence, Laboratory of Mathematics, Computer Science and Applications (LMCSA), Faculty of Sciences and Technologies, Hassan II University, Mohammedia, Morocco
adilchakhtouna10@gmail.com

² Higher National School of Arts and Crafts, Hassan II University, Casablanca, Morocco

Abstract. Speech Emotion Recognition (SER) is an emerging field that encompasses various disciplines such as Human-Computer Interaction (HCI), Natural Language Processing (NLP), computer vision, and cognitive sciences like psychology and social sciences. The primary objective of this SER study is to analyze and quantify human emotions using a combination of statistical feature extraction and Deep Learning (DL) techniques. To achieve this goal, the Mi-Auto-Encoder (MiAE) is proposed to compress the embedding features representation of the WavLM model; in addition, a dense layer is incorporated to classify the different emotions. The SER experiments were conducted on the widely used Interactive Emotional Dyadic Motion Capture (IEMOCAP) English reference database. The results revealed promising performance, with accuracies of 77.57% and 76.17% achieved on the validation and test data, respectively. The proposed SER system was evaluated and compared to state-of-the-art studies, demonstrating its effectiveness.

Keywords: Human-Computer Interaction · Speech Emotion Recognition · Mi-Auto-Encoder (MiAE) · WavLM

1 Introduction

Our lives are significantly influenced by the central role of the human emotional experience. Advancing our understanding of human emotions is a complex research domain. Combining insights from cognitive and affective sciences with advancements in artificial intelligence is crucial to create intelligent emotional systems capable of establishing emotional connections with humans in the context of Human-Computer Interaction (HCI).

In the psychological community, emotions are typically examined from a dual perspective. On one hand, they are represented within a three-dimensional framework referred to VAD model, characterized by Valence (positive-negative),

Activation (active-passive) and Dominance (high-low) dimensions [1]. On the other hand, the assumption that emotions can be expressed in a discrete manner [2], such as happiness, anger and fear. This implies that these emotions typically arise from distinct and specialized regions of the brain. As a result, each individual emotion is associated with a distinct and specific neural activation pattern [3]. Speech Emotion Recognition (SER) represents an intriguing research domain in which machines are exploring their ability to identify and interpret human emotions conveyed through speech. This interdisciplinary field based on psychology, speech processing, and Deep Learning (DL) seeks to narrow the gap between human emotions and intelligent computing.

The goal of the ongoing study is to advance the state-of-the-art by developing a SER system that mimics the emotive functions of the human brain. In order to identify various emotions from the IEMOCAP database [4], a statistical feature extraction process deploying WavLM [5] embedding features is applied together with our suggested Mi-Auto-Encoder (MiAE) architecture.

In the next Sect. 2, a comprehensive review of the latest literature on SER is provided. Following that, Sect. 3 presents a detailed description of the proposed SER framework. Section encompasses the presentation and discussion of the diverse experimental outcomes. Concluding remarks and future perspectives related to SER are summarized in Sect. 5.

2 SER Literature

This section discusses literature related to the ongoing SER study. Verbal communication has exerted a prevailing influence over social interactions. Therefore, developing SER plate-forms is essential to boost user confidence in HCI. For this reason, numerous proposals have been made to represent emotional information from different speech cues. Authors in [6] presented a hybrid SER system by cascading the Gaussian Mixture Model and Deep Neural Network (GMM-DNN). The GMM-DNN classifier was tested on the private Emirati speech database covering six emotive utterances and attained an accuracy of 83.97%, thereby outperforming Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) classifiers. In [7], the transfer learning approach was explored. Four pre-trained architectures including VGG-16, VGG-19, EfficientNet B0, and EfficientNetV2 B0 are compared, with and without the fine-tuning process. By employing Mel spectrograms as a dimensional representation of speech signals, the highest accuracy of 79.83% was achieved by VGG-19 on the Italian Database of Elicited Mood in Speech (DEMoS). By transforming SER into an image classification task, this study introduces a perspective to enhance emotion recognition from speech signals.

Xie et al. [8] recommended to integrate frame-level speech features using a Long Short-Term Memory (LSTM) network with attention mechanism. The effectiveness of the proposed approach is confirmed by experimental results on

three emotion corpora, namely CASIA,¹ eNTERFACE,² and GEMEP.³ The achieved accuracies were 92.8%, 89.6%, and 57.0%, respectively. The research highlights an important finding that humans exhibit imbalanced attention towards stimuli as a whole. Hence, the integration of this concept has proven to yield outstanding results, underscoring its significant impact. In another study [9], a bilingual SER system utilizing three parallel Convolutional Neural Networks (CNN) was developed. The emotional content within speech signals was effectively represented by the mean values of 40 MFCCs as input features for the classification task. Impressive recognition rates of 87.08% and 83.90% were reached on the RAVDESS⁴ and EMOVO⁵ datasets, respectively, in the monolingual scenario. While, in the bilingual scenario, a recognition rate of 79.53% was obtained on the mixed RAVDESS-EMOVO database. It is noticed that using a robust feature such as the MFCC with the mean statistic gives convincing results.

The present study makes a significant contribution by harnessing the advantages of the Self-Supervised Learning (SSL) approach by using the WavLM model, along with the proposed MiAE architecture, to successfully map the SSL representations of various speech cues. The suggested MiAE framework demonstrated its capability to effectively learn and model the non-linear relationships among emotional features in low dimensional space.

3 SER Materials & Methods

The proposed system consists of two main phases: The SER training & validation phase, followed by the testing one. In the initial stage, the input audio files are embedded using the two encoders of the WavLM model and subsequently utilized to train our (MiAE + dense layer) model. Once completing the training process, the optimal model is saved. Moving to the second phase, the test audio files undergo the same embedding step using the WavLM model and are then passed to the saved optimal model for emotion decision. The overall SER framework is illustrated in Fig. 1.

3.1 Speech Emotion Dataset

The benchmark speech emotion database IEMOCAP⁶ [4] was used to evaluate the proposed methodology. The corpus covers around 12 h of audio recordings and facial images, skilfully interpreted by a group of 10 actors. When using speech recordings, five sessions were undertaken for scripted and improvised dialogues; in each session, a conversation was conducted between male-female

¹ http://www.chineseldc.org/resource_info.php?rid=76

² <http://www.interface.net/results/>

³ <https://www.unige.ch/cisa/gemep>

⁴ <https://zenodo.org/record/1188976#.YfADWP7MK3A>

⁵ <http://voice.fub.it/activities/corpora/emovo/index.html>

⁶ <https://sail.usc.edu/iemocap/index.html>

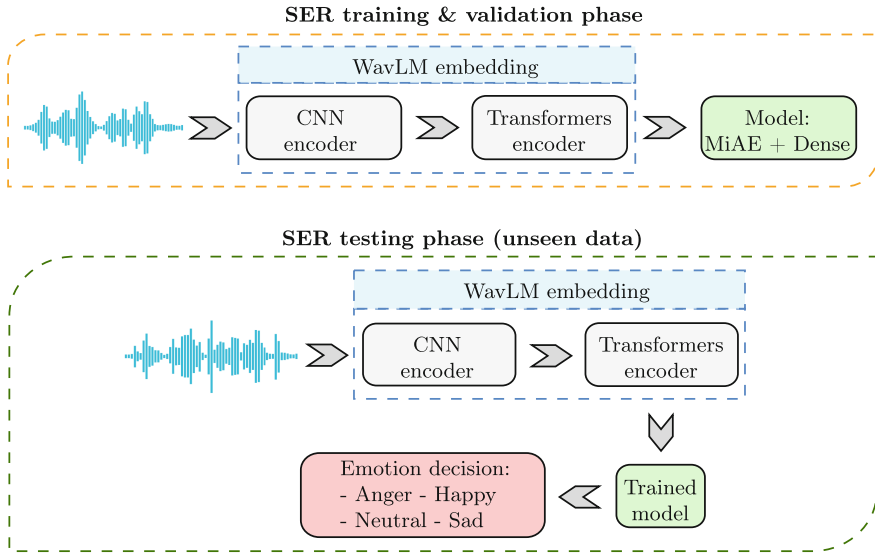


Fig. 1. Architecture of the proposed SER system

pairs. In line with the literature, our experimental approach involved selecting four emotions (neutral, happy + excited (merged), angry, and sad) from both dialogues (improvised and scripted).

3.2 Feature Extraction

In the speech modality, the emotional attributes of speech signals are influenced by several factors, including variations in frequency and intensity. The goal of the SER is to identify discriminative features that effectively measure and quantify these variations, enabling the modeling of different emotional states with accuracy and precision. Previous studies [10, 11] in the field have predominantly used conventional prosodic and spectral features like MFCC, F0, energy, among others. However, in this research, we opted to leverage the power of SSL approach, specifically by chosen the WavLM model and using it as feature extractor.

A- WavLM

The WavLM [5] pre-trained model is a SSL framework that directly generates representations from raw audio files. The framework has been adapted effectively to a variety of downstream speech tasks from the SUPERB benchmark [12]. Furthermore, the architecture incorporates two primary encoders, the CNN, and the transformers. In the SER task, the WavLM large version was specifically selected for this study.

The WavLM large comprises 24-transformer encoder layers equipped with 12 attention heads, and at each transformer output, the representation vector consists of 1024-dimensional hidden states. In the present feature extraction

approach, the speech signal X is divided into a consistent number of frames f according to the following equation:

$$X \equiv [f_1, f_2, \dots, f_m] \quad (1)$$

where f_m is the m^{th} frame.

Subsequently, these frames are embedded to capture representations from both CNN and all transformers outputs. Each frame f_i encompasses embedded features of size n as follows:

$$f_i = (x_{i1}, x_{i2}, \dots, x_{in}) \quad (2)$$

where x_{in} represents the value of the n^{th} feature corresponding to frame f_i .

The resulting embedded feature vector V has a shape of $[s, f, n]$, where s stands for the number of outputs of the CNN and all transformers, specifically equal to 25 in this case. The values for f and n are set to 400 and 1024, respectively.

B- Statistical features

In order to handle the computational limitations, we applied a statistical approach to the feature-embedded vectors of all IEMOCAP speech signals. This involved computing three key global features: Mean, standard deviation, and median. Thus, for each output s , we obtain a matrix of features as follows:

$$V_k = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \quad k = 1, 2, \dots, s \quad (3)$$

The statistical measures were used to summarize the information contained in the resulting vectors, and their respective formulas are as follows:

$$\mu = (\mu_1, \mu_2, \dots, \mu_n) \quad (4)$$

$$\sigma = (\sigma_1, \sigma_2, \dots, \sigma_n) \quad (5)$$

$$\text{med} = (\text{med}_1, \text{med}_2, \dots, \text{med}_n) \quad (6)$$

where μ_j , σ_j and med_j are expressed by Eqs. (7), (8) and (9), respectively:

$$\mu_j = \frac{1}{m} \sum_{i=1}^m x_{ij} \quad (7)$$

$$\sigma_j = \sqrt{\frac{1}{m} \sum_{i=1}^m (x_{ij} - \mu_j)^2} \quad (8)$$

$$\text{med}_j = \frac{1}{2} \left(x_{(\frac{m}{2})j} + x_{(\frac{m}{2}+1)j} \right) \quad (9)$$

here, the index j varies from 1 to n , while the index i ranges from 1 to m .

3.3 Proposed Recognition Model

Auto-Encoders (AE), which belong to the category of Neural Networks (NN), are commonly exploited in unsupervised machine learning tasks. They were originally developed for the purpose of representation learning and dimensionality reduction [13]. AE have extended their scope to cover a wide range of research domains [14,15]. Typically, an AE network consists of two parts: The encoder and the decoder, these components work together to learn the task of reconstructing the input data as accurately as possible to produce an output that closely resembles the original input. Inspired by their remarkable performance, we applied solely the encoder component of the AE, referred to as MiAE in this research, to compress the embedding features representation. In addition, a dense layer is incorporated to facilitate the emotion classification.

Below, we provide a comprehensive description of the proposed SER architecture:

Input: we produced three sets of characteristics, the mean, the standard deviation and the median which are then fed separately into the network. This corresponds to a feature vector size of 25×1024 for each statistical set.

Flatten layer: the role of a flattening layer is to transform multi-dimensional input data into a single-dimensional format, allowing the following layers to handle the data in a completely connected manner.

Dropout layer: is a regularization strategy generally applied on DL models, its purpose is to mitigate the problem of overfitting by randomly deactivating a certain percentage of input units (neurons) during the training process.

MiAE hidden layers: the output of the flattening layer is received by two hidden MiAE layers in readiness for the final classification layer.

Dense layer: contains the number of units representing the number of emotional classes.

Activation Function: is an essential element of a NN. It applies a non-linear transformation to the output of a previous layer or neuron, thus providing non-linearity in network computations.

4 SER Experiments & Findings

This section covers all the experimental configurations carried out as part of this research, along with the corresponding results obtained using different sets of statistical features. The proposed methodology was evaluated using various SER performance indicators such as the recognition rate (accuracy) and the confusion matrix. For training our proposed MiAE model, the following hyper-parameters were employed: The Adam optimizer with an initial learning rate of 0.00001 was utilized to minimize the loss function and update the model's parameters. The Softmax function is applied to obtain the probability distribution over multiple classes. The training process is performed during 100 epochs. The assessment considered the speaker's dependency context, wherein the data was partitioned into three segments: training, validation, and test, with distribution ratios of 80%, 10%, and 10% respectively.

The performance of the proposed method for speaker-dependent experimentation is provided in Table 1. The WavLM mean features combined with the suggested (MiAE + dense) model resulted in the highest validation and test recognition rates, reaching 77.57% and 76.17% respectively. In contrast, the performances of the other two statistical feature sets were as follows: 71.42% validation accuracy and 73.10% test accuracy using WavLM median features, and 70.16% validation accuracy and 68.78% test accuracy using WavLM std. features.

Table 1. The validation and test accuracies (%) obtained using the proposed methodology. Std. and acc. refer to standard deviation and accuracy, respectively.

Database	Features	Classification model	Validation acc	Test acc
IEMOCAP	WavLM mean features	MiAE + dense	77.57	76.17
	WavLM median features		71.42	73.10
	WavLM std. features		70.16	68.78

The testing confusion matrix obtained with the highest recognition rate (using WavLM mean features) is represented in Table 2. The recall performance for the neutral emotion achieved the highest score of 83.62%, while the sad and anger states followed closely with scores of 76.14% and 72.72% respectively. The happiness state had a slightly lower recall performance with a score of 70.73%. Table 2 clearly indicates that the proposed MiAE model in conjunction with the feature extraction process, effectively identified various emotions with remarkable balance rates. Whereas, other works [16] carried out in a similar context reached disparate rates, with significantly higher percentages for certain emotions and notably lower percentages for others, exhibiting variances of up to 20% to 25%. Furthermore, it was observed that the highest rates of confusion among different emotions occurred specifically in the case of neutral state.

Table 2. The testing confusion matrix of the SER system using the WavLM mean features.

Emotion	Anger	Happy	Neutral	Sad
Anger	72.72	7.27	18.18	1.81
Happy	6.09	70.73	21.34	1.82
Neutral	1.75	8.18	83.62	6.43
Sad	1.83	2.75	19.26	76.14

In order to situate our proposed approach in the SER literature, we conducted a comparative analysis with previous research that focused on speaker-dependent scenario. Table 3 shows the approaches implemented in terms of feature extraction, classification models employed, and the resulting accuracy. It

is noteworthy that the human evaluations of different emotional states in the IEMOCAP dataset were 74% for neutral, 70% for happiness, 76% for anger, and 77% for sadness. These findings highlight the challenge of accurately recognizing emotions in this corpus, even for human evaluators. When comparing our proposed approach to the works referenced as [16–18], we observed that our method surpassed the current state-of-the-art in SER.

Table 3. Comparison of the proposed method versus previous works in speaker-dependent experiments on IEMOCAP Database.

Reference	Feature extraction	Classification model	Accuracy
[17]	FBANK	SVM with Polynomial kernel	58.40%
[18]	IS13-ComParE features	SVM with RBF kernel	66.20 %
[16]	3-D Log-Mel spectrums	CNN + BiLSTM	74.96%
Our	WavLM mean features	MiAE + dense	76.17%

5 Conclusion

Through this research paper, we focused on evaluating the performance of the SER system by suggesting a Mi-Auto-Encoder architecture with a novel statistical feature extraction method using the WavLM SSL model. Experiments were conducted only on the speaker-dependent scenario enabling this method to successfully recognize the four emotions existing in the IEMOCAP database with a highest test accuracy of 76.17%. As prospects, our future research in SER will explore the speaker-independent scenario. Building upon our proposed method, we anticipate promising and challenging outcomes, particularly in terms of recognition rates, which will contribute to the existing body of literature in this field.

Acknowledgements. This work was supported by the Ministry of Higher Education, Scientific Research and Innovation, the Digital Development Agency (DDA) and the CNRST of Morocco (Alkhawarizmi/2020/01).

References

1. Russell, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. *J. Res. Pers.* **11**(3), 273–294 (1977). [https://doi.org/10.1016/0092-6566\(77\)90037-X](https://doi.org/10.1016/0092-6566(77)90037-X)
2. Ekman, P.: An argument for basic emotions. *Cogn. Emotion* **6**(3–4), 169–200 (1992). <https://doi.org/10.1080/02699939208411068>
3. Samsonovich, A.: Socially emotional brain-inspired cognitive architecture framework for artificial intelligence. *Cogn. Syst. Res.* **60**, 57–76 (2020). <https://doi.org/10.1016/j.cogsys.2019.12.002>

4. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Narayanan, S.S.: IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **42**, 335–359 (2008). <https://doi.org/10.1007/s10579-008-9076-6>
5. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Wei, F.: Wavlm: large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Selected Topics Signal Process.* **16**(6), 1505–1518 (2022). <https://doi.org/10.1109/JSTSP.2022.3188113>
6. Shahin, I., Nassif, A.B., Hamsa, S.: Emotion recognition using hybrid Gaussian mixture model and deep neural network. *IEEE Access* **7**, 26777–26787 (2019). <https://doi.org/10.1109/ACCESS.2019.2901352>
7. Chakhtouna, A., Sekkate, S., Adib, A.: Speech emotion recognition using pre-trained and fine-tuned transfer learning approaches. In: *The Proceedings of the International Conference on Smart City Applications*, pp. 365–374. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-26852-6_35
8. Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., Schuller, B.: Speech emotion classification using attention-based LSTM. *IEEE/ACM Trans. Audio Speech Language Process.* **27**(11), 1675–1685 (2019). <https://doi.org/10.1109/TASLP.2019.2925934>
9. Sekkate, S., Khalil, M., Adib, A.: A statistical feature extraction for deep speech emotion recognition in a bilingual scenario. *Multimedia Tools Appl.* **82**(8), 11443–11460 (2023). <https://doi.org/10.1007/s11042-022-14051-z>
10. Chakhtouna, A., Sekkate, S., Adib, A.: Improving speech emotion recognition system using spectral and prosodic features. In: *International Conference on Intelligent Systems Design and Applications*, pp. 399–409. Springer, Heidelberg (2021). https://doi.org/10.1007/978-3-030-96308-8_37
11. Chakhtouna, A., Sekkate, S., Adib, A.: Improving speaker-dependency/independency of wavelet-based speech emotion recognition. In: *International Conference on Networking, Intelligent Systems and Security*, pp. 281–291. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-15191-0_27
12. Yang, S.W., Chi, P.H., Chuang, Y.S., Lai, C.I.J., Lakhota, K., Lin, Y.Y., Lee, H.Y.: Superb: Speech Processing Universal Performance Benchmark. *arXiv preprint arXiv:2105.01051* (2021). <https://doi.org/10.48550/arXiv.2105.01051>
13. Riyad, M., Khalil, M., Adib, A.: Dimensionality reduction of MI-EEG data via convolutional autoencoders with a low size dataset. In: *International Conference on Business Intelligence*, pp. 263–278. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-06458-6_22
14. El Bouny, L., Khalil, M., Adib, A.: Convolutional denoising auto-encoder based awgn removal from ecg signal. In: *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pp. 1–6. IEEE (2021). <https://doi.org/10.1109/INISTA52262.2021.9548524>
15. Yildirim, O., San Tan, R., Acharya, U.R.: An efficient compression of ECG signals using deep convolutional autoencoders. *Cogn. Syst. Res.* **52**, 198–211 (2018). <https://doi.org/10.1016/j.cogsys.2018.07.004>
16. Meng, H., Yan, T., Yuan, F., Wei, H.: Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE Access* **7**, 125868–125881 (2019). <https://doi.org/10.1109/ACCESS.2019.2938007>
17. Kurpukdee, N., Kasuriya, S., Chunwijitra, V., Wutiwiwatchai, C., Lamsrichan, P.: A study of support vector machines for emotional speech recognition. In: *2017 8th International Conference of Information and Communication Technology for*

Embedded Systems (IC-ICTES), pp. 1–6. IEEE (2017). <https://doi.org/10.1109/ICTEmsys.2017.7958773>

18. Poria, S., Majumder, N., Hazarika, D., Cambria, E., Gelbukh, A., Hussain, A.: Multimodal sentiment analysis: addressing key issues and setting up the baselines. *IEEE Intell. Syst.* **33**(6), 17–25 (2018). <https://doi.org/10.1109/MIS.2018.2882362>



Axonal Myelination as a Mechanism for Unsupervised Learning in Spiking Neural Networks

Nadezhda Chaplinskaia^(✉) and Nikolay Bazenkov

V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences,
Moscow, Russia
chnv@ipu.ru

Abstract. Plasticity of synaptic weights is usually supposed the foundation of learning and long-term memory in biological neural networks. Mathematical models of both biological and artificial neural networks reflect this vision. Little attention is paid to the role spike propagation delays play in information processing and learning. We propose a model of myelin plasticity which controls the efficiency of spikes propagation along axons. A neuron modifies the myelin sheath thickness of its input axons to achieve better synchrony of incoming spikes. Synchronous input spikes cause higher postsynaptic response which leads to higher spike generation probability. We show that the axonal delay plasticity model may be used to train a network recognize input patterns even when synaptic weights remain fixed. The delay plasticity approach may be a useful augmentation of spiking neural networks used in neuromorphic computing.

Keywords: spiking neurons · delay plasticity · myelin · unsupervised learning

1 Introduction

Synaptic plasticity is considered the primary mechanism underlying learning in human and animal brains. Mathematical models of neural networks usually represent learning as changes in synaptic efficacy also called synaptic weight. At the same time spike propagation delays are mostly neglected as a computational factor in mathematical theory of neuronal information processing.

However, there exist experimental evidences that spike timing plays crucial role in some computational tasks performed by the nervous system [7]. Moreover, during past decades theoretical results were obtained in which it was shown that artificial neurons with programmable delays are computationally more powerful than neurons having only synaptic weights as programmable parameters [8]. Despite these findings spike propagation delays attract little attention in bioinspired models of neural networks.

We propose a novel learning algorithm where a neuron adjusts the thickness of myelin sheaths at its inputs. This models the biological mechanism of dynamic myelin thickness which influence spike propagation delays. As a result, the delays are shifted to achieve maximum synchrony among incoming spikes. Synchronous input spikes cause the post-synaptic neuron to trigger a new spike with higher probability. The novelty of the model is that the myelin dynamics does not depend on the occurrence of post-synaptic spikes. This is the difference from the most existing models of learning in biological neural networks.

This rule was used to train a spiking neural network (SNN) to perform unsupervised learning tasks such as feature extraction from unlabeled data. The algorithm performance was evaluated in computational experiments. A single layer network was trained on USPS [5] dataset with hand-written digits images. The pixel brightness was encoded as the time to first spike (TTFS). During the learning the neurons adjusted their input synaptic delays according to the proposed learning rule. The trained network is able to memorize the prominent features of the images and use them to classify them.

The results show that axonal delay plasticity is a plausible mechanism to organize unsupervised learning in biological neural networks. It may be used to solve learning problems either solely or in combination with conventional synaptic weight plasticity.

2 Related Research

Recent experimental studies shows that axonal delays are not static but actively changing even in adults [1]. Myelin dynamics participates in memory and learning [3]. There are evidences that spike propagation delays play a prominent role in neural information processing. In the birds auditory system the noise localization problem is solved by special dendritic structures precisely tuned to the specific signal propagation delays [7]. The time difference between the signals coming from the left and right ears allows to localize the noise direction.

In [8] the computational power of spiking neurons with programmable delays was studied. The authors proved that a very simple neuron with plastic delays is computationally more powerful than the threshold neuron and roughly equivalent to the neuron with the sigmoid activation function. They also showed that Vapnik-Chervonenkis dimension of a spiking network with programmable delays is quadratic in the number of neurons even if the weights are fixed. This means that an SNN with plastic delays is potentially as capable to classify complex patterns as an ANN with sigmoid neurons.

There are some studies of synaptic delays in SNN learning algorithms. In [11] the delay plasticity was used at the output layer of the network while the first hidden layer was trained by a conventional weight plasticity. In [10] a gradient algorithm for delay learning was proposed and used in a sound localization hardware sensor. In [9] a combined weight and delay plasticity was used to predict a moving dot direction. In [6], the influence of axonal delays on the synchronization of spikes in the network is studied. However, in this work, the synaptic weights

were trained by the standard STDP method, and the delays are considered equal and fixed during learning. In the article [12], a spike neural network is trained in the same way using the STDP rule, but in relation to axonal delays, and the network was placed in three-dimensional space where axons have different lengths.

Our work differs from the listed above in the following points. First, we model delay changes as the result of elaborate myelin dynamics controlled by the output neuron. Second, our learning rule doesn't include the spike of the output neuron that partially solves the problem of "silent" neurons which never triggers spikes and thus never participate in learning.

3 The Model of Myelin Plasticity

Consider a fully connected spike neural network consisting of n neurons at the input and one neuron at the output. There is a certain amount of myelin on the axons of all input neurons (the example for $n = 3$ is represented in Fig. 1a). A spike propagating along this axon doesn't reach the synapse immediately, but after a delay.

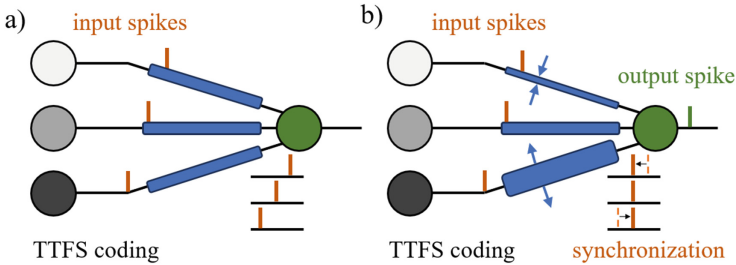


Fig. 1. The propagation of spikes in SNN. A. before myelin plasticity learning, B. after myelin plasticity learning. Input encoded as time-to-first spike (TTFS)

The output neuron is represented by the leaky-integrate-and-fire (LIF) model with conductance-based synapses:

$$\tau \frac{dv(t)}{dt} = v_{rest} - v(t) + g(t)(E_{exc} - v(t)). \quad (1)$$

Here v is the membrane potential which is driven to the rest potential v_{rest} in the absence of external input. The time constant is τ .

The total synapses conductance g rises after incoming spikes and the synaptic current flows into the neuron, driven by the reversal synapse potential E_{exc} .

Define as t_{pre_i} the time of spike generation by the presynaptic neuron n_i , $\sigma_i(t)$ —the current delay on the axon of the neuron n_i , which is determined by the thickness of its myelin sheath. Define as t_{f_i} the moment of the arrival of

the spike at the synapse, that is $t_{f_i} = t_{pre_i} + \sigma_i(t)$. The synaptic conductance is defined as

$$g_i(t) = g_{amp} \exp\left(\frac{t_{f_i} - t}{\tau_s}\right), \quad t > t_{f_i}. \quad (2)$$

Every time the input neuron n_i generates a spike, after the delay $\sigma_i(t)$ the permeability of the synapse increases which drives the membrane potential upward. When the membrane potential reaches the activation threshold v_T , the output neuron generates a spike. Thus, by changing the delays on the axons of the input neurons, it is possible to control the timing of the increase in the membrane potential of the output neuron.

The delay $\sigma_i(t)$ depends linearly on $m_i(t)$ —the thickness of the myelin sheath of this axon:

$$\sigma_i(t) = \frac{-(\sigma_{max} - \sigma_{min})}{(m_{max} - m_{min})} m_i(t) + \frac{\sigma_{max} m_{max} - \sigma_{min} m_{min}}{(m_{max} - m_{min})}. \quad (3)$$

Here σ_{max} , σ_{min} , m_{max} , m_{min} are the maximal and minimal possible values of the delay on the axon and the myelin thickness.

Thus, the thicker the sheath, the faster the spike reaches the synapse. And vice versa, the thinner the sheath, the longer the delay. The maximum possible value of the thickness of the myelin sheath corresponds to the minimum value of the pulse delay on the axon, and vice versa. In this paper the following values are used: $\sigma_{max} = 8\text{ms}$, $\sigma_{min} = 0.5\text{ms}$. The data correspond to the results of the [2] study for the LIP-FEF region of the monkey's cortico-cortical system. For the thickness of the axon myelin sheath we use values based on [4]: $m_{max} = 4$ microns, $m_{min} = 0.3$ microns.

We introduce the characteristic $\bar{\mu}(t)$ of the output neuron responsible for the myelination of all axons of the input neurons, and the characteristic $\mu_i(t)$ of the synapse s_i , $i = \overline{1, n}$, (connecting the input neuron n_i with the output neuron), responsible for the demyelination of the axon of the input neuron n_i . In our model, $\bar{\mu}(t)$ will determine the rate of myelin increase (delay reduction), and $\mu_i(t)$ is the rate of myelin decrease (delay increase). Myelination and demyelination will be considered as two independent processes following each spike coming at the output neuron.

After each spike reaches the output neuron, the characteristics of $\bar{\mu}(t)$ and $\mu_i(t)$ change according to the following equations:

$$\begin{cases} \mu_i(t) = \mu_{amp} H(t - t_{f_i}), \\ \bar{\mu}(t) = \frac{1}{n} \sum_{i=1}^n \mu_i(t), \end{cases} \quad (4)$$

where μ_{amp} is the amplitude of the change μ_i , and $H(\cdot)$ is the Heaviside function.

The change in the thickness of the myelin sheath $m_i(t)$ depends on the difference between the values of μ_i and $\bar{\mu}$ as follows:

$$\frac{d}{dt} m_i(t) = k |\bar{\mu}(t) - \mu_i(t)| \left(H(\bar{\mu}(t) - \mu_i(t)) - \frac{m_i(t) - m_{min}}{m_{max} - m_{min}} \right), \quad (5)$$

where k is the learning rate of the output neuron.

A graphical interpretation of the solution of this equation for a network of two input neurons and one output neuron can be seen in Fig. 2.

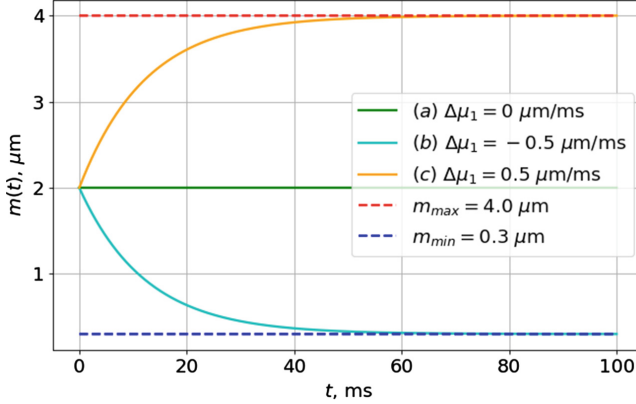


Fig. 2. The changes of the myelin thickness of the axon of input neuron n_1 for a network with two input neurons and one output neuron. Let $m_1(0) = 2 \mu\text{m}$, $k = 0.6$, $\mu_{amp} = 1 \mu\text{m/ms}$. At the $t = 0$ (a) the output neuron received both input spikes ($\mu_1 = \mu_2 = \bar{\mu} = 1 \mu\text{m/ms}$) or none ($\mu_1 = \mu_2 = \bar{\mu} = 0 \mu\text{m/ms}$): $\Delta\mu_1 = 0 \mu\text{m/ms}$, (b) the spike generated by the first input neuron was obtained by the output neuron first: $\mu_1 = 1 \mu\text{m/ms}$, $\mu_2 = 0 \mu\text{m/ms}$, $\bar{\mu} = 0.5 \mu\text{m/ms} \Rightarrow \Delta\mu_1 = -0.5 \mu\text{m/ms}$, (c) spike generated by the second input neuron was obtained by the output neuron first: $\mu_1 = 0 \mu\text{m/ms}$, $\mu_2 = 1 \mu\text{m/ms}$, $\bar{\mu} = 0.5 \mu\text{m/ms} \Rightarrow \Delta\mu_1 = 0.5 \mu\text{m/ms}$

According to the model, for all axons, through which the spike has already passed, the right side of the equation is negative. Therefore, the thickness of the myelin sheath will decrease, and spikes on these axons will come later. For all axons through which the spike has not yet passed, the right side of the equation is positive, therefore, the thickness of the myelin sheath will increase, so the delay on these axons will decrease, and spikes will come earlier. After several iterations the incoming spikes will arrive at the output neuron almost synchronously (Fig. 1b), which should lead to a steeper increase of the membrane potential. Hence the neuron will generate spike and memorize the exposed pattern.

4 Myelin Plasticity as an Unsupervised Learning Mechanism

4.1 Network Architecture and Learning Process

We use a widely known winner-take-all (WTA) network architecture. The network consists of the input layer with all-to-all excitatory connections to the single output layer. The output layer includes lateral inhibitory connections. The first

output neuron, which has generated a spike, routes it to the inhibitory synapses of the remaining output neurons and prevents their activation.

Since the training rule uses the input spike generation times, the time coding is needed. We use the time of the first spike (TTFS) encoding scheme. TTFS works as follows: at the current image exposure, the input neuron corresponding to the lightest pixel generates a spike first, the input neuron corresponding to the darkest pixel generates a spike last (Fig. 1). The times of spike generation by the remaining neurons linearly depends on their brightness.

The images of the training sample are exposed in random order. During each exposure the input neurons generate their spike sequences corresponding to the encoded pattern several times. All output neurons, according to the described learning rule, adjust the myelin thickness on the incoming axons so that they will eventually be able to generate their own spikes in response to this image—that is, to remember it.

At the same time, it is necessary that, as a result, one specific neuron reacts to the current image, so a competition mechanism is included in the network: output neurons compete for activation while memorizing the exposed pattern. In the losing neurons the forgetting mechanism is triggered: the value of the myelin thickness on the axonal branches connecting to these output neurons returns to its value before training at the current exposure.

4.2 Experiments

To demonstrate the presented learning process we choose USPS image classification dataset. The dataset contains 9298 images of handwritten digits from 0 to 9, sized 16 by 16 pixels. Thus, the network contains 256 input neurons and at least 10 output neurons with lateral inhibitory connections. Images of the same class can differ significantly from each other. Therefore, 10 neurons at the output layer of the network will not be enough to memorize all 10 classes of the dataset images.

The number of neurons in the output layer of the network significantly affects the quality of pattern recognition. We studied the dependence of the value of the F1-measure on the training and test sets on the number of output neurons in the network (Fig. 3). The graphs show a slight fluctuation of the results due to the stochasticity of the each output neuron initial state. However, it is easy to see that the value of the F1-measure is rising. On 20 neurons, the F1-measure is 0.24 on the training set and 0.15 on the test set; on 340 neurons, we have the best result in this range at 0.72 on the training set and 0.66 on the test set.

The learning mechanism presented in the work (before the classification process) is a clustering algorithm, since delays on axons adjust to the exposed patterns, searching for centers of clusters. In view of this, we compared the results of the proposed model with the results of k-means clustering algorithm (Table 1). The comparability of the recognition results shows that the presented mechanism achieved performance of standard clustering algorithms and can be further used to solve other machine learning tasks.

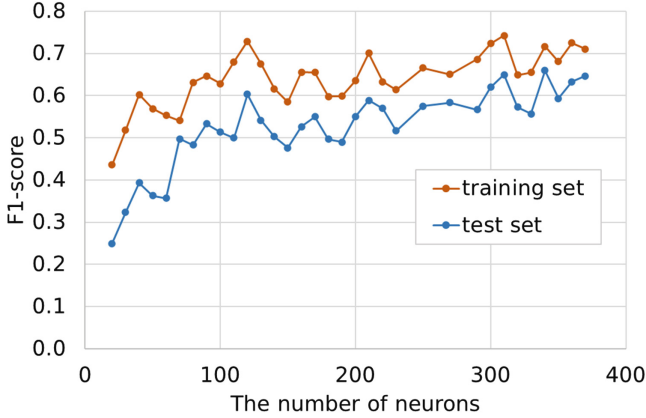


Fig. 3. The dependence of the F1-measure value on the number of output layer neurons in the network on the training and test sets

Table 1. Comparison results on USPS data set

Method	Number of output neurons	F1-score on the training set	F1-score on the testing set
k-means	–	0.64	0.63
SNN with axonal myelination	340	0.72	0.66

5 Conclusion

Spiking neural networks running on neuromorphic chips is a promising technology which may give rise to the next generation of artificial intelligence systems. One of the most crucial problems in the field is the absence of biologically plausible learning algorithms which efficiency would be as high as those of deep neural networks. One of the reasons is that most algorithms focus at synaptic weights update and omit other dimensions that may be essential for natural neural networks. Our research explores how myelin plasticity and axonal delays contribute to the network learning capabilities. This is one of the possible extensions of conventional Hebbian approach.

We proposed a model of axonal myelin plasticity where the spike propagation delay depends on the thickness of myelin sheath at the axon. Each neuron controls the myelin thickness on its input axons. During the learning a neuron adjust the delays such that to achieve maximum synchrony between incoming spikes. This mechanism is used for unsupervised learning in a spiking neural network.

We studied the efficiency of the proposed approach in the image classification task on USPS dataset. First, the images were presented to the network which learned the characteristic features in the unsupervised manner with the proposed axonal plasticity rule. Second, the extracted features were used to train the model for the classification task. In our experiments the network successfully learned the images features having only delays as variable parameters.

Our results showed that spike propagation delays are a valid basis for learning in addition to synaptic weights. This provides a possibility to create novel cognitive architectures where learning will be two-dimensional. For example, spatial information will be stored in weights and time-related information will be stored in delays. This is a promising direction of future research.

References

1. Dutta, D.J., Woo, D.H., Lee, P.R., Pajevic, S., Bukalo, O., Huffman, W.C., Wake, H., Basser, P.J., SheikhBahaei, S., Lazarevic, V., et al.: Regulation of myelin structure and conduction velocity by perinodal astrocytes. *Proc. Natl. Acad. Sci.* **115**(46), 11832–11837 (2018)
2. Ferraina, S., Paré, M., Wurtz, R.H.: Comparison of cortico-cortical and cortico-collicular signals for the generation of saccadic eye movements. *J. Neurophysiol.* **87**(2), 845–858 (2002)
3. Fields, R.D., Bukalo, O.: Myelin makes memories. *Nat. Neurosci.* **23**(4), 469–470 (2020)
4. FitzGibbon, T., Nestorovski, Z.: Human intraretinal myelination: Axon diameters and axon/myelin thickness ratios. *Indian J. Ophthalmol.* **61**(10), 567 (2013)
5. Hull, J.J.: A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell.* **16**(5), 550–554 (1994)
6. Khoshkhou, M., Montakhab, A.: Spike-timing-dependent plasticity with axonal delay tunes networks of izhikevich neurons to the edge of synchronization transition with scale-free avalanches. *Front. Syst. Neurosci.* **13**, 73 (2019)
7. London, M., Häusser, M.: Dendritic computation. *Annu. Rev. Neurosci.* **28**, 503–532 (2005)
8. Maass, W., Schmitt, M.: On the complexity of learning for spiking neurons with temporal coding. *Inf. Comput.* **153**(1), 26–46 (1999)
9. Nadafian, A., Ganjtabesh, M.: Bio-plausible Unsupervised Delay Learning for Extracting Temporal Features in Spiking Neural Networks. arXiv preprint [arXiv:2011.09380](https://arxiv.org/abs/2011.09380) (2020)
10. Oubari, O.: Precise timing and computationally efficient learning in neuromorphic systems. Ph.D. thesis, Sorbonne université (2020)
11. Paugam-Moisy, H., Martinez, R., Bengio, S.: Delay learning and polychronization for reservoir computing. *Neurocomputing* **71**(7–9), 1143–1158 (2008)
12. Talidou, A., Frankland, P.W., Mabbott, D., Lefebvre, J.: Learning to be on Time: Temporal Coordination of Neural Dynamics by Activity-Dependent Myelination. *bioRxiv*, pp. 2021–08 (2021)



Feature Synthesis for Few-Shot Object Detection

Chenchen Tao¹, Song Chen^{2(✉)}, Yi Chen¹, Xiaojie Cai¹, and Chong Wang¹

¹ Ningbo University, Zhejiang 315211, China

² H3C Company, Zhejiang 310000, China

2011082274@nbu.edu.cn

Abstract. The Few-Shot Object Detection (FSOD) task aims to detect novel instances in scenarios with limited data. However, the feature distribution of the novel class can be easily influenced by the distribution of features from the base classes. This paper introduces the Feature Synthesis for Few-Shot Object Detection algorithm, leveraging Generative Adversarial Networks to generate visual features for novel classes. By combining semantic embeddings with real visual features, the generator is trained to enhance the correlation between synthetic features and their corresponding categories. Class prototypes are computed based on real features, and contrastive loss guides the constraint of the synthetic feature distribution, improving model performance. Additionally, the algorithm incorporates Pseudo Margin Evaluation loss to calculate instance uncertainty scores and increase discrimination power. Experimental results on the MS-COCO dataset demonstrate the algorithm's effectiveness with significant performance gains.

Keywords: few-shot object detection · feature synthesis · contrastive learning

1 Introduction

In the domain of few-shot object detection (FSOD), two primary research directions have emerged. The first direction is meta-learning-based approaches, which use a stage-wise and periodic meta-training paradigm to train a meta-learner capable of transferring knowledge from base classes. Meta R-CNN [1] introduces meta-learning for adapting the attention layer in the channel, improving the performance of object detection. Meta-DETR [2] leverages a meta-learning strategy to exploit inter-class correlations and effectively utilize correlations among different classes. Other methods such as FSIW [3] and TFA [4] improve upon the meta-learning paradigm by employing more complex feature aggregation techniques and introducing balanced datasets. FSOD-UP [5] focuses on learning invariant object features from all categories and improving them using consistency loss. SRR-FSD [6] proposes a semantic space constructed from word embeddings, training the detector to project target instances into this semantic space.

In this paper, we address the challenges of few-shot learning and few-shot object detection by proposing novel approaches that leverage meta-learning and transfer learning paradigms. Our aim is to enhance the model’s ability to recognize and classify novel classes with limited labeled examples. To achieve this, we introduce a Feature Synthesis framework for Few-Shot Object Detection, which exploits semantic information and synthesizes diverse visual features using generative adversarial networks (GANs) [7].

Our work stands out as the first attempt to incorporate feature synthesis into the few-shot object detection (FSOD) task. By combining semantic information, we propose an algorithm that effectively addresses the scarcity of novel class samples in FSOD. This algorithm constrains the synthetic features by leveraging class prototypes, ensuring that the generated visual features closely resemble real features in terms of distribution.

Moreover, we tackle the challenge of uncertainty in synthesized features by employing the Pseudo Margin Evaluation (PME) loss. This loss function enables us to exploit the uncertainty associated with synthesized features, making even low-quality features useful for the FSOD model.

Through comprehensive experiments conducted on benchmark datasets, we demonstrate the effectiveness and superiority of our proposed methods in the field of few-shot object detection. Our contributions include addressing data scarcity, enhancing feature synthesis with semantic information, and effectively utilizing the uncertainty of synthesized features. These advancements significantly improve the generalization performance of few-shot object detection models.

The rest of this paper is organized as follows. The approach of our feature synthesis algorithm is presented in Sect. 2. The experimental results and analysis are given in Sect. 3. We conclude this paper in Sect. 4.

2 Method

2.1 Problem Definition and Framework Overview

Let D_B be the dataset containing base class object images and D_N denote the dataset with novel class object images. The labels for base and novel classes are $Y_B = \{1, \dots, B\}$ and $Y_N = \{B + 1, \dots, B + N\}$, respectively, where B and N represent the total number of base and novel classes, and $Y_B \cap Y_N = \emptyset$.

The framework begins by extracting features from input images $x \in D_B$ using ResNet-101 [8] and FPN [9] as the backbone network. The Region Proposal Network (RPN) generates candidate proposals, which are then processed through RoI pooling to obtain fixed 1024-dimensional features. The Faster R-CNN [10] is trained using D_N with k instances per class while freezing the backbone network’s parameters to prevent overfitting. This results in a standard FSOD model F_n , used to extract visual features $f_B \in R^{1024}$ and $f_N \in R^{1024}$ for base and novel classes. To utilize novel class knowledge, semantic embeddings from the CLIP [11] model are introduced as $A = \{a_0\} \cup A_B \cup A_N$, where a_0 represents background class embeddings and A_B and A_N are sets of base and novel class

semantic embeddings, respectively. The overall framework of the proposed based feature synthesis for FSOD is shown in Fig. 1, with training divided into three phases.

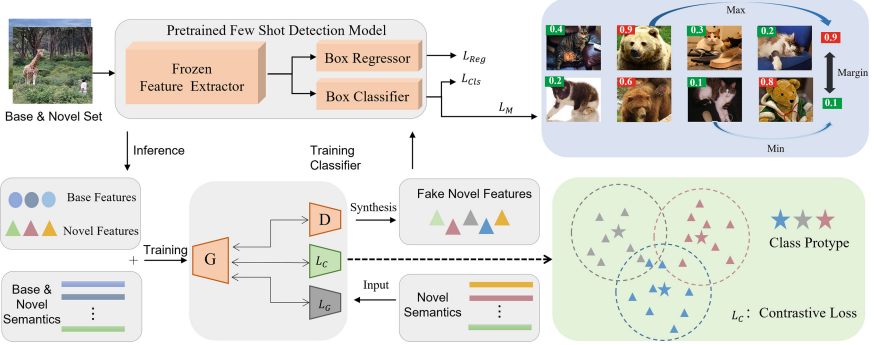


Fig. 1. Illustration of the proposed overall framework. Our framework is mainly divided into a feature synthesis module, a contrastive loss module, and an evaluation module.

In the first phase, features from D_B and D_N are extracted using the pre-trained FSOD model F_n to obtain a feature dataset.

In the second training phase, the conditional generator G is trained using real visual features f , their labels y , random noise vector $Z \sim N(0, 1)$, and semantic embedding A . Once trained, G can synthesize novel class visual features $f_P \in R^{1024}$ and pseudo-labels y_P based on the given semantic embedding and noise vector. The differentiated synthesized features are obtained by G . The optimization objective for the baseline model’s feature generator [12] is:

$$L_G = \min_G \max_D L_{WGAN} + L_C + L_{div} \quad (1)$$

The objective function (1) minimizes L_G with respect to G and maximizes L_{WGAN} (cw-GAN loss) [13], L_C (to enhance feature synthesis capacity), and L_{div} (to improve feature diversity).

The third phase focuses on training the novel classifier ϕ_{n-cl_s} separately using synthetic visual features and employing PME loss to efficiently utilize correct and incorrect synthetic features. Finally, the novel classifier ϕ_{n-cl_s} is concatenated with the base classifier ϕ_{b-cl_s} from the pre-trained model to obtain the final classifier.

Limited real features for novel classes result in chaotic feature distribution and low-quality features. To address these issues, the synthesized features are constrained and evaluated in 2.2.

2.2 Feature Synthesis

To overcome the limited sample challenge in training for novel classes in FSOD, we propose incorporating semantic embedding information and contrastive loss

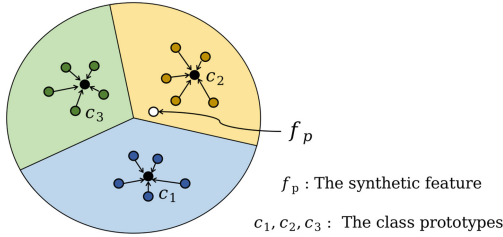


Fig. 2. Illustration of the contrastive loss with synthetic features

during feature synthesis. However, the feature synthesis process can be unstable, resulting in low-quality synthetic features and pseudo-labels. To address this issue, we introduce a contrastive loss module that utilizes real features as supervision to improve the distribution of synthetic features.

The first step in the feature synthesis process involves calculating class prototypes, which serve as the feature distribution centers for each class. The prototypes reflect the differences between class features and are used to supervise the generator in synthesizing novel class visual features. During training, the synthetic features are guided to be closer to the prototype of their respective class and farther from prototypes of other classes. Similar to other self-supervised studies [14–16], the contrastive loss in this paper follows formula (2), where n denotes the batch size and C_s denotes the prototypes of the class to which the current generative feature f_{p_i} belongs. Therein, τ is a temperature hyperparameter, which is set to 0.05 by default.

$$L_C = \sum_{i=1}^n -\log \frac{\exp(f_{p_i} \cdot C_s/\tau)}{\sum_{j=1}^k \exp(f_{p_i} \cdot C_j/\tau)} \quad (2)$$

By effectively utilizing different class feature information through contrastive loss, the synthesized visual features are pulled away from other novel class prototypes as well as base class prototypes. The class prototypes provide well-constrained guidance for the synthesized features. The total loss of the generator is defined as the formula (3), with α set to 0.1 in the experiments.

$$L_{gan} = L_G + \alpha \cdot L_C \quad (3)$$

L_C , guided by class prototypes, enables the generator to learn realistic feature distribution relationships, resulting in more clustered synthetic features for each class. After training the generator with the total loss L_{gan} , the generator takes the novel class semantic embedding A_N and random noise vector Z as input and generates a set of novel class visual features \tilde{f}_p and their corresponding pseudo-labels \tilde{y}_p during inference, as shown in formula (4).

$$(\tilde{f}_p, \tilde{y}_p) = G(A_N, Z) \quad (4)$$

The synthesized novel class visual features effectively address the challenges in the few-shot object detection task and can be directly used to train the novel class classifier. However, considering the instability of the synthesized features and pseudo-labels, we comprehensively evaluate them before training the classifier in the subsequent subsection.

2.3 Synthetic Feature Assessment

While the feature distributions are constrained during synthesis, there is no guarantee of the reliability of these features. To tackle this problem, we propose a method to assess the reliability of synthetic features and pseudo-labels using pseudo-margin evaluation loss (PME). By establishing a quality margin between low-quality and high-quality features, we aim to reduce the impact of low-quality features on the classifier.

The uncertainty scores s_i^j for each feature is calculated using a pre-trained novel class classification head ϕ_{n-cl_s} and its cross-entropy loss, see formula (5). Smaller scores indicate more reliable features. These scores are then grouped by category, creating sets of scores S_j for each class j . The PME loss pulls apart the distance between the feature with the largest score and that with the smallest score in each group, ensuring reliable features are far from unreliable ones. This process is repeated for all classes and averaged, as shown in formula (6).

$$s_i^j = \phi_{n-cl_s}(f_{p_i}^j, y_{p_i}^j) \quad (5)$$

$$L_M = \frac{1}{N} \sum_{j=1}^N \max(0, \lambda - \max(S_j) + \min(S_j)) \quad (6)$$

The parameter λ in formula (6) represents the desired spacing between the largest and smallest scores and is set to 0.7 in the experiments. The PME loss aims to create a stable spacing between these scores, aiding the classifier in accurately distinguishing between true and false features. The synthetic feature evaluation loss allows for better utilization of low-quality features compared to training the classifier directly with synthetic features.

The new class classifier is trained using the cross-entropy loss of Faster R-CNN, as shown in Eq. (7).

$$L_{cls} = - \sum_{j=1}^N \sum_{i=1}^m y_{p_i}^j \cdot \log s_i^j \quad (7)$$

Finally, the cross-entropy loss is combined with the PME loss to form the total loss of the new class classifier, as shown in the formula (8):

$$L_{total} = L_{cls} + L_M \quad (8)$$

The new classifier only includes the novel class classification head ϕ_{n-cl_s} from the pre-trained FSOD model, and the final classifier is obtained by concatenating ϕ_{n-cl_s} with the base class classification head ϕ_{b-cl_s} .

3 Experiment

MS-COCO dataset [17]: The proposed method is evaluated on the widely used MS-COCO, which consists of 80 categories, with 60 base classes and 20 novel classes for the few shot object detection task. Training data for base classes has sufficient annotated instances, while novel classes have only $k = 10$ or $k = 30$ annotated instances per category. The test set includes 5000 images that cover both base and novel class instances, with AP, AP50, and AP75 as common evaluation metrics.

Implementation details: The algorithm is implemented using the MMDetection framework [18], and the pretrained model of the FSCE algorithm [19] is used as the baseline model. Real visual features of base and novel classes are extracted based on candidate regions with specific IoU thresholds. The semantic vectors are extracted using the text encoder of the CLIP model, which has the same dimension as the noise vectors.

Table 1. The mAP of novel classes on MS-COCO (%).

Sample Num	Method	AP	AP50	AP75	Sample Num	AP	AP50	AP75
10	TFA w/cos [4]	10.0	19.1	9.3	30	13.7	24.9	13.4
	QA-FewDet [20]	10.2	20.4	9.0		16.5	31.9	15.5
	N-PME [21]	10.6	21.1	9.4		14.1	26.5	13.6
	CGDP+FSCN [22]	11.3	23.0	9.8		15.1	29.4	–
	SRR-FSD [6]	11.3	–	9.8		14.7	–	13.5
	FSCE [19]	11.9	–	10.5		16.4	–	16.2
	FSCE*	11.7	24.3	9.9		16.4	31.4	<u>16.2</u>
	PDE [23]	12.0	22.3	11.1		17.2	31.3	16.6
	SVD [24]	12.0	–	10.4		16.0	–	15.3
	FADI [25]	12.2	–	11.9		16.1	–	15.5
	BC-YOLO [26]	9.0	–	–		12.9	–	–
	Ours	12.3	<u>23.9</u>	<u>11.5</u>		<u>17.1</u>	31.6	15.8

3.1 Comparison with SOTA

Table 1 presents the comparison results of our algorithm with state-of-the-art FSOD algorithms on the MS-COCO dataset. Our method achieves the highest mAP when compared to all other methods on 10 sampled AP. It also outperforms the second-best method, CGDP+FSCN, in terms of AP50, with an improvement of 1.7 in AP75. Compared to SRR-FSD [6], which also incorporates semantic information, our method shows significant improvements in both AP and AP75. Additionally, compared to the meta-learning-based algorithm QA-FewDet [20], SFC achieves substantial improvements in all three indicators, demonstrating the

effectiveness and superiority of our approach. Moreover, our approach surpasses the YOLO-based model BC-YOLO [26] by 3.3% and 4.2% on the 10-shot and 30-shot novel sets, respectively.

Figure 3 visually demonstrates the test results of our method. The detection performance is satisfactory for most categories; however, there is room for improvement in detecting persons due to the presence of unlabeled person data in the base dataset. Overall, our proposed method combining feature synthesis and semantic information proves effective in addressing FSOD tasks and provides new insights into synthetic sample applications.

3.2 Ablation Experiment

We conduct ablation experiments on two modules to validate their effectiveness: Prototype Contrastive Loss and Pseudo-Margin Evaluation Loss (PME). The baseline model for these experiments is FSCE*. The experiments are performed on the MS-COCO 10-shot dataset, and the results are recorded in Table 2.

From the table, it can be observed that the prototype contrast loss module significantly enhances the quality of synthesized features generated by the generator. This improvement is evident in the increases in AP and AP75 values of 0.4 and 0.9, respectively. These results demonstrate the effectiveness of guiding the generator to synthesize features based on category prototypes.

Furthermore, the PME module effectively leverages the uncertainty associated with synthetic features, resulting in improvements in AP, AP50, and AP75 by 0.2, 0.7, and 0.7, respectively when compared to models using only contrastive loss. Taking into account the uncertainty of synthetic features proves to be a valid approach.

Table 2. Ablation for key components in MS-COCO 10-shot settings.

CL	PME	AP	AP50	AP75
–	–	11.7	24.3	9.9
✓	-	12.1	23.2	10.8
✓	✓	12.3	23.9	11.5

Visual analysis of the visual features output by the generator, performed using the t-SNE method [27], further confirms the effectiveness of the Prototype Contrastive Loss module. Figure 4 shows the visual feature distributions for five categories (car, boat, motorcycle, bicycle, and train) before and after applying contrastive loss. The contrastive loss guided by class prototypes improves the distinctiveness of visual features between different classes and enhances the aggregation of visual features within the same class.

In terms of semantic information selection, we compare the performance of CLIP and Fasttext word embeddings. The experiments are conducted on the



Fig. 3. The visualization comparison between ours and the ground truth.

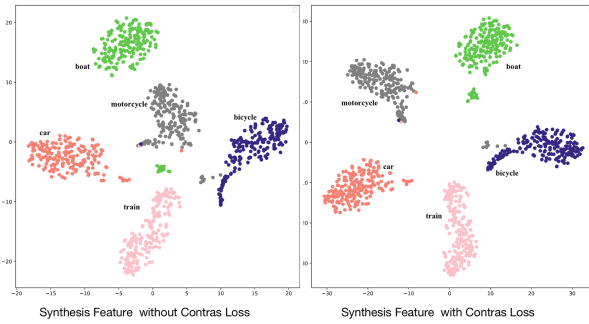


Fig. 4. The Visualization of synthetic features before and after adding contrastive loss.

Table 3. Ablation experiment of semantic information on MS-COCO(%).

Shot	Semantic	AP	AP50	AP75
10	Fasttext [28]	12.0	24.1	10.8
	CLIP [11]	12.3	23.9	11.5
30	Fasttext [28]	16.8	29.8	15.6
	CLIP [11]	17.1	31.6	15.8

MS-COCO dataset for both 10-shot and 30-shot settings. Table 4 reveals that CLIP semantic embedding achieves better results in both settings.

Additionally, we explore the influence of different threshold values when mining pseudo labels from the base dataset. The experiments are performed with threshold values of 0.6 and 0.8 on the MS-COCO 30-shot dataset. Table 4 shows that a higher threshold value of 0.8 achieves better accuracy. Lower threshold values introduce more instability and noise to the pseudo labels, resulting in unsatisfactory performance. However, higher threshold values provide more stable and reliable pseudolabel data, leading to performance improvements.

Table 4. Ablation experiments with different threshold pseudo-labels.

	AP	AP50	AP75
Ours	17.1	31.6	15.8
Ours+Pseudo-label(0.6)	17.0	32.4	15.3
Ours+Pseudo-label(0.8)	17.2	33.8	14.6

4 Conclusion

This paper introduces a few-shot object detector based on feature synthesis, which can effectively tap the intrinsic connection between semantic information and visual features to synthesize the visual features of the novel class. In the feature synthesis stage, the proposed algorithm adopts the class prototypes of real visual features as the constraints of the generator, which ensures the quality and distribution of the synthesized visual features. When training the classifier, we fully consider the uncertainty of the synthesized visual features and eliminate the low-quality visual features to improve the performance of the classifier. Ultimately, our method allows for end-to-end training and achieves better results on the MS-COCO detection dataset.

References

1. Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L.: Meta r-cnn: towards general solver for instance-level low-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9577–9586 (2019)
2. Zhang, G., Luo, Z., Cui, K., Lu, S., Xing, E.P.: Meta-detr: image-level few-shot detection with inter-class correlation exploitation. *IEEE Trans. Pattern Anal. Mach. Intell.* (2022)
3. Xiao, Y., Lepetit, V., Marlet, R.: Few-shot object detection and viewpoint estimation for objects in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(3), 3090–3106 (2022)
4. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly Simple Few-Shot Object Detection. arXiv preprint [arXiv:2003.06957](https://arxiv.org/abs/2003.06957) (2020)

5. Wu, A., Han, Y., Zhu, L., Yang, Y.: Universal-prototype enhancing for few-shot object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9567–9576 (2021)
6. Zhu, C., Chen, F., Ahmed, U., Shen, Z., Savvides, M.: Semantic relation reasoning for shot-stable few-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8782–8791 (2021)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **27** (2014)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
9. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
10. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
11. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
12. Hayat, N., Hayat, M., Rahman, S., Khan, S., Zamir, S.W., Khan, F.S.: Synthesizing the unseen for zero-shot object detection. In: Proceedings of the Asian Conference on Computer Vision (2020)
13. Xian, Y., Sharma, S., Schiele, B., Akata, Z.: f-vaegan-d2: a feature generating framework for any-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10275–10284 (2019)
14. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural. Inf. Process. Syst.* **33**, 9912–9924 (2020)
15. Li, J., Zhou, P., Xiong, C., Hoi, S.C.: Prototypical Contrastive Learning of Unsupervised Representations. arXiv preprint [arXiv:2005.04966](https://arxiv.org/abs/2005.04966) (2020)
16. Diba, A., Sharma, V., Safdari, R., Lotfi, D., Sarfraz, S., Stiefelhofen, R., Van Gool, L.: Vi2clr: video and image for visual contrastive learning of representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1502–1512 (2021)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 Sept 2014, Proceedings, Part V 13, pp. 740–755. Springer, Heidelberg (2014)
18. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab Detection Toolbox and Benchmark. arXiv preprint [arXiv:1906.07155](https://arxiv.org/abs/1906.07155) (2019)
19. Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: Fscf: few-shot object detection via contrastive proposal encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7352–7362 (2021)
20. Han, G., He, Y., Huang, S., Ma, J., Chang, S.F.: Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3263–3272 (2021)

21. Liu, W., Wang, C., Yu, S., Tao, C., Wang, J., Wu, J.: Novel instance mining with pseudo-margin evaluation for few-shot object detection. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2250–2254. IEEE (2022)
22. Li, Y., Zhu, H., Cheng, Y., Wang, W., Teo, C.S., Xiang, C., Vadakkepat, P., Lee, T.H.: Few-shot object detection via classification refinement and distractor retreatment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15395–15403 (2021)
23. Chen, S., Wang, C., Liu, W., Ye, Z., Deng, J.: Pseudo-label diversity exploitation for few-shot object detection. In: International Conference on Multimedia Modeling, pp. 289–300. Springer, Heidelberg (2023)
24. Wu, A., Zhao, S., Deng, C., Liu, W.: Generalized and discriminative few-shot object detection via svd-dictionary enhancement. *Adv. Neural. Inf. Process. Syst.* **34**, 6353–6364 (2021)
25. Cao, Y., Wang, J., Jin, Y., Wu, T., Chen, K., Liu, Z., Lin, D.: Few-shot object detection via association and discrimination. *Adv. Neural. Inf. Process. Syst.* **34**, 16570–16581 (2021)
26. Xia, R., Li, G., Huang, Z., Meng, H., Pang, Y.: Bi-path combination yolo for real-time few-shot object detection. *Pattern Recogn. Lett.* **165**, 91–97 (2023)
27. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**(11) (2008)
28. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of Tricks for Efficient Text Classification. arXiv preprint [arXiv:1607.01759](https://arxiv.org/abs/1607.01759) (2016)



Simulation of Fabric Wetting Based on Particle Sampling

Jiajun Cheng¹, Zhen Liu¹(✉), Tingting Liu², and Yanjie Chai¹

¹ Faculty of Information Science and Technology, Ningbo University, Ningbo, China
liuzhen@nbu.edu.cn

² College of Science and Technology, Ningbo University, Cixi, China

Abstract. This article proposes a novel simulation method for fabric wetting. Firstly, in order to achieve the wetting effect of the interaction between liquid and fabric, a fast sampling method for fabric boundaries is proposed, which analyzes and samples the edges and interior of the fabric triangle during the preprocessing process. Then the forces of the sampled particles and fluid particles were calculated, and the forces of the sampled particles were applied to each vertex of the fabric triangle. Finally, update the positions and velocities of fluid particles and cloth particles. In order to make the animation more realistic and simulate the wetting animation effect, the fabric sampling particles were modeled as the basic water absorption unit, and the entire wetting process was simulated using a three-step simulation method.

Keywords: Particle System · Cloth model · Fluid simulation · Flow distribution coupling simulation

1 Introduction

For numerous fabric movement phenomena, many scholars have designed different physical models and parameters to fit in physical simulations [1]. There have been varying degrees of progress from ordinary single-layer pure cotton fabrics to down inflatable materials, sweaters, and more. There are also better methods for handling collisions between fabrics, rigid bodies, and fluids [2]. However, there are still many issues worth exploring regarding the interaction between fabric and different objects.

In recent years, some researchers have also begun to pay attention to the interaction between liquid and fabric, including the dynamic changes in the force on the fabric when the fluid interacts with the fabric. However, the interaction between fluid and fabric not only involves dynamic simulation of the fabric, but also the establishment of a dynamic model between fluid and fabric. It requires the combination of fluid and fabric simulation, and the use of previous fabric modeling methods and mechanical models is not enough to truly represent the dynamic process of fluid and fabric contact and moisture.

The general modeling methods for fabric simulation are based on the spring particle model. This method abstracts the warp and weft structure of the fabric and constructs a spring network connected horizontally and vertically. This method can effectively

simulate various physical effects such as elastic deformation and wrinkling during the interaction between external forces and fabric. However, for other more complex behavior simulations of fabrics, such as the absorption, infiltration, and saturation processes between fabrics and water, more physical models need to be added for calculation. The simulation of fabric wetting process involves the study of fluid, fabric, and the interface between the two. It is necessary to consider these aspects separately in order to achieve a balance between simulation accuracy and real-time performance.

2 Related Work

The flow distribution interaction is a multifaceted phenomenon, so some previous work has attempted to isolate one or both aspects. Ozgen used the fractional derivative method to simulate the deformation of completely submerged fabrics without simulating water at all [3]. Chen proposed an improved saturation, wrinkling, and friction model to better simulate the appearance of wet clothes [4]. Um combined the shallow water model and diffusion equation to solve the fluid flow on and inside the dynamic cloth [5].

Another branch of research focuses on carefully handling the boundary conditions for the interaction between water and impermeable thin shells, including Euler and Lagrangian fluids. In the context of the Euler method, Guendelman used variable density pressure solutions to explain weakly coupled interactions [6], while Robinson proposed a strong coupling method that temporarily concentrates the momentum of thin shells and shells together as a fluid [7]. Azevedo used uniform interpolation and precise cutting elements to prevent fluids from crossing impermeable thin boundaries [8]. In the Smoothed Particle Hydrodynamics (SPH) method, Akinci carefully sampled thin deformable objects using SPH particles to improve pressure accuracy and ensure that the fabric does not penetrate liquids [9], assuming the appropriate time step size. Huber instead directly used the fabric triangular mesh itself, combined with repulsion and continuous collision detection to strictly enforce impermeability [10]. This article focuses on permeable thin structures and uses weak coupling methods to transfer momentum between liquid and thin structures using resistance and buoyancy. In recent work on simulating the mixing of porous sand and water [11], Tampubolon used the formulas of Bandara and Soga to calculate buoyancy [12], but concluded that buoyancy can be largely ignored in their problems. Firstly, the above research did not improve the wetting simulation of fabric based on the spring particle model, and fluid motion on the fabric surface was processed on the basis of wetting.

3 Improved Fabric Porous Model Based on Spring Particle

The coupling methods of fluid and fabric based on SPH method include unidirectional coupling [13, 14] and bidirectional coupling [15, 16]. The coupling methods mainly include methods based on penalty force and methods using direct force. The method of punitive power needs to meet a sufficiently small time step size; The method of direct force is to avoid penetration by correcting the position of particles, which requires a relatively large time step. A difficulty in fabric coupling is that both are in real-time motion, and traditional bidirectional coupling cannot completely avoid penetration. Therefore,

a continuous collision detection algorithm (CCD) for handling deformable objects has become a commonly used method for collision detection between fluids and fabrics; The method of sampling fabric boundaries can also effectively handle collisions between the two.

The two most commonly used methods for collision detection between fluids and fabrics are continuous collision detection and sampling methods. The problem with continuous collision detection is that it requires high accuracy in collision time and has certain errors; Sampling methods often have the problem of oversampling and low sampling efficiency. In response to the above issues, this section improves a sampling method to handle collisions between fabrics and fluids.

3.1 Fabric Particle Sampling

In the collision detection of cloth and fluid, most people use the CCD method to update the positions of fluid particles and cloth triangles based on the collision time. The problem with this method is that the collision time calculation error is relatively large, and it also requires judgment of collisions between points, surfaces, and edges, which increases the computational workload. Therefore, the boundary sampling method is used to achieve bidirectional coupling between the fluid and the cloth. The boundary particle sampling method used in this article is an improvement on the reference method [9]. Sampling is divided into vertex sampling, edge sampling, and triangle internal sampling.

Firstly, sample the vertices of the triangle, then sample the edges: determine the number of sampled particles n_e , as shown in Fig. 1 for a triangle with an index of 0, where the length of one edge is l_e , calculate n_e , vector \mathbf{p}_e , and the position of the i th particle is:

$$\mathbf{b}_p[i] = \mathbf{x}_1 + i \times d_s \times \frac{\mathbf{p}_e}{|\mathbf{p}_e|} \quad (1)$$

Reference sampled each side of a triangle [9], which caused the oversampling problem of the triangle side. As shown in Fig. 1, triangles 0 and 1 have a common edge. Sampling each edge of a triangle will inevitably result in duplicate sampling. Therefore, the steps for edge sampling are as follows:

Sample the edges (x_1, x_2) of each triangle, taking the fabric with a resolution of 4×4 as an example. In Fig. 1, the particle labeling color of each triangle is the same. Sample the edges (x_1, x_3) of triangles with odd indices, as shown in Fig. 1 where $(0, 1, 2)$ represents the one-dimensional index T of the triangle. After the above two steps, only the rightmost and topmost edges of the fabric have not been sampled, so it is necessary to analyze and sample them. Obtain the final sampling results by sampling the right and top edges of the fabric.

After sampling the edges of the triangle, sample the interior of the triangle as follows.

Firstly, it is necessary to determine the shortest side e_s and its normal vector \mathbf{n}_h in the direction inside the triangle, where e_l is the longest side. Calculate the number of iterations n_t for sampling within a triangle, using the formula \mathbf{n}_i . H_t is the height along the scan, calculated by mapping the longest edge to the normal vector. For each iteration, calculate the linear intersection positions c_{pl} and c_{pm} with the other two edges, as shown

in Fig. 1. The number of sampled particles is n_s , and the vector is p_s . Bring them into formula 2 to obtain the position of each sampled particle. Calculate each sampling point inside the triangle using the above method and store its coordinates for future collision detection calculations and rendering of sampled particles. And laid the groundwork for the next wetting simulation.

$$n_h = \frac{e_s \times (e_1 \times e_s)}{|e_s \times (e_1 \times e_s)|} \quad (2)$$

$$n_s = \left\lfloor \frac{(c_{pm} - c_{p1})}{d_s} \right\rfloor \quad (3)$$

$$p_s = \frac{(c_{pm} - c_{p1})/n_s}{|(c_{pm} - c_{p1})/n_s|} \quad (4)$$

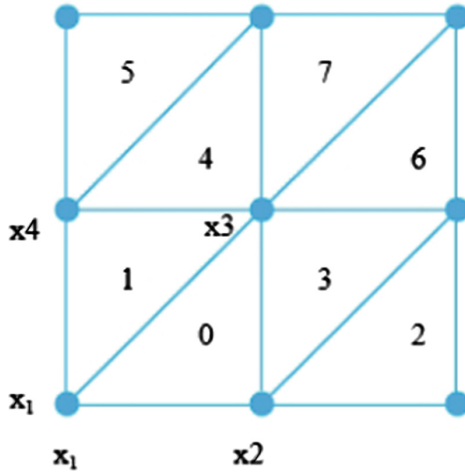


Fig. 1. Basic model of fabric

3.2 Particle Based Porous Wetting Model for Fabrics

On the basis of sampling in 3.1, this article takes each fabric sampling particle as the water absorption unit of the fabric. As shown in Fig. 2:

Porous materials are defined by porosity and permeability. The porosity ϕ_i represents the proportion of pores in the area, which is the volume of pore space. The amount of water that can be stored in the area is A_i , and the saturation of each particle is S_i set as follows:

$$S_i = \frac{m_i}{A_i} \quad (5)$$

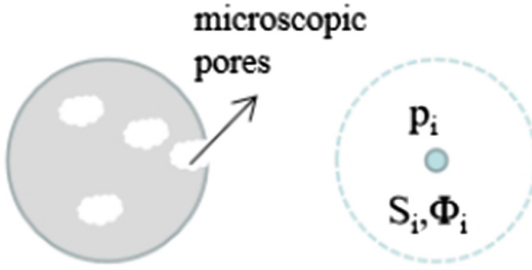


Fig. 2. Porous abstract model

4 Fabric Wetting Simulation

Most fabric wetting uses the triangular surface of the fabric as the basic water absorption unit. Although this can simplify the wetting model, it is very unstable when the collision unit is more complex. In this article, the sampled particles are used as the basic water absorption unit. The wetting of the fabric is simulated in three steps: water absorption, diffusion, and drainage.

Firstly, it is water absorption simulation. When a fluid collides with a porous solid, we need to calculate whether the porous particles at the current collision position can continue to absorb water. If the saturation of porous particle i at this position satisfies the condition $S_i < S_{\max}$. The fluid particles are absorbed, and the saturation of the particles is updated. S_{\max} here represents the maximum value of water absorption by porous particles. The saturation calculation for each water absorption unit is shown in formula 6:

$$S_i = S_i + \frac{m^{absorb}}{V_i} \quad (6)$$

The liquid is absorbed and further diffused into the fabric until equilibrium and capillary force are reached. Capillary action is the main factor that causes the diffusion of liquids in the fabric and their absorption by the fabric. There is surface tension in liquid flow, as well as in porous media particles along the direction of liquid flow. If the saturation of the particles is 1, that is, the liquid completely fills all the voids in the particles, then the surface tension is equal to the surface tension of the liquid; If the saturation of the particles is less than 1, the liquid does not completely fill the gaps in the particles, and the surface tension is equal to the surface tension of the liquid and gas mixture in the gaps, which is less than the surface tension of the liquid. When the surface tension of adjacent particles is not equal, a pressure difference occurs, and the liquid flows from one particle to another unsaturated particle. The capillary pressure difference is determined by the sum of capillary potential energy differences between particles and surrounding particles.

$$h_i = \sum_j (P_j - P_i) \quad (7)$$

Among them, particle j is a particle connected to particle i through a wetting tube. During the diffusion process, the capillary potential of each particle can be calculated

based on its saturation:

$$P_i = c(1 - S_i)^\alpha \quad (8)$$

It can be seen that as the saturation increases, the capillary potential energy P_i of each particle decreases. If the particles are not wet, the saturation $S_i = 0$; If the particles are completely wet, the saturation $S_i = 1$. If the particles around particle i are completely wet, the capillary potential is the same, $h_i = 0$, and diffusion does not occur; If there are non wet particles around particle i , then $h_i \neq 0$ and diffusion continues.

Finally, in the third stage of our model, if the fluid is transported to a completely saturated particle through the diffusion process, the particle will lose excess fluid due to dripping. When gravity is greater than capillary force, the liquid tends to adhere to the fabric until the excess amount increases above a certain mass threshold before it begins to drip. To simulate this phenomenon, each particle has a droplet like buffer. Excess liquid accumulates in the dripbuffer, and whenever the content of the buffer exceeds the drop threshold, the droplets are discharged from the fabric through dripping, resulting in a corresponding decrease in particle saturation, as shown in Fig. 3

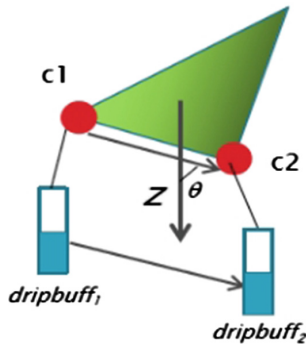


Fig. 3. Buffer pool for each fabric particle

5 Experimental Result

The hardware configuration for the experiment is Intel i7CPU, 8GB RAM, and Nvidia Geforce gtx1060gpu; The software programming environments are Windows 10, Visual Studio 2017, Openg13.0, and Cuda10.2.

Figure 4 shows a partially saturated rectangular cloth that is kept horizontal, constrained in all its corners, and then relaxed. The fluid is transported to the center of the fabric through a diffusion process, where it begins to drip, and A to B show the wetting process. Figure 5A, B show the drainage under the action of gravity. B display our cloth particle model.

To validate our diffusion experiment, we tested the horizontal A and vertical B, respectively. The darker the color, the greater the saturation. See Fig. 6

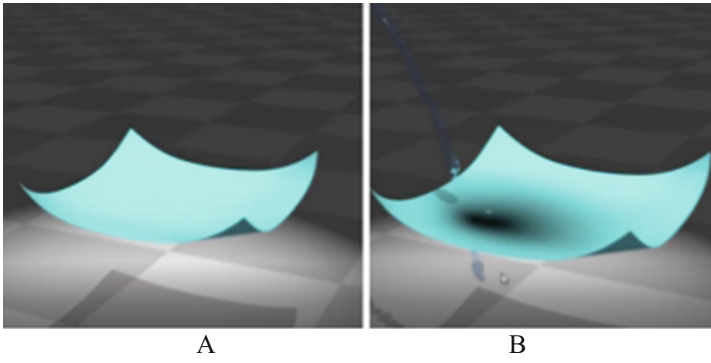


Fig. 4. Horizontal fabric water absorption model

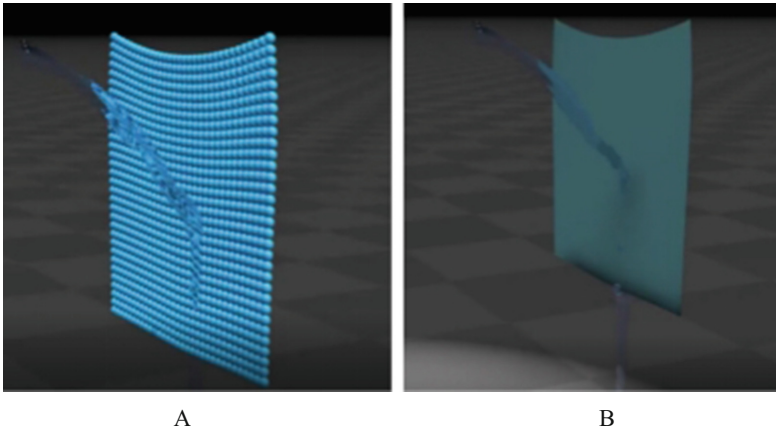


Fig. 5. Vertical fabric water absorption model

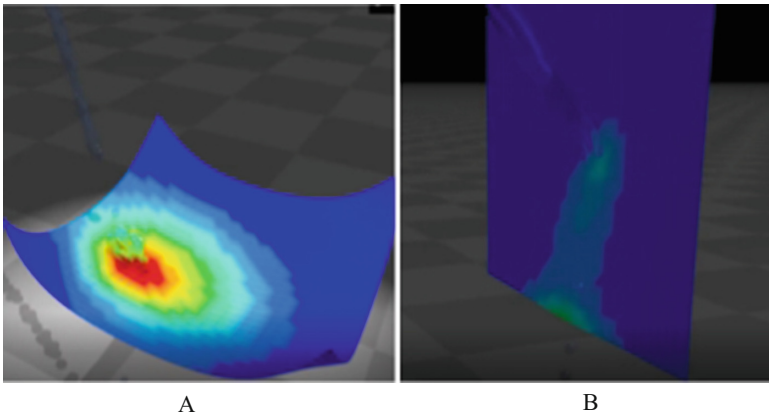


Fig. 6. Saturation changes over time

6 Conclusion

We propose a new porous wetting model for simulating fabric wetting. Firstly, during the preprocessing process, analyze and sample the edges and interior of the fabric triangle; Then calculate the force acting on the fluid particles and the sampled particles, and apply the force acting on the sampled particles to each particle of the fabric triangle; Finally, use a three-step simulation method to simulate the wetting effect. To make the animation more realistic, GPU parallel computing is used to accelerate the simulation. The experimental results indicate that our model can effectively simulate the process of fabric absorption and drainage.

References

1. Gissler C, Peer A, Band S, Bender J, Teschner M.: Interlinked SPH pressure solvers for strong fluid-rigid coupling. *ACM Trans. Graph. (TOG)* **38**(1), 5.1–5.13 (2019)
2. Dinev, D., Liu, T., Li, J., Thomaszewski, B., Kavan, L.: FEPR: Fast energy projection for real-time simulation of deformable objects. *ACM Trans. Graph.* **37**(4), 1–12 (2018)
3. Ozgen, O., Kallmann, M., Ramirez, L.E., Coimbra, C.F.: Underwater cloth simulation with fractional derivatives. *ACM Trans. Graph.* **29**(3) (2010)
4. Chen, Y., Thalmann, N.M., Allen, B.F.: Physical simulation of wet clothing for virtual humans
5. Um, K., Kim, T.Y., Kwon, Y., Han, J.H.: Porous deformable shell simulation with surface water flow and saturation. *Comput. Anim. Virtual Worlds* **24**(3–4), 1–8 (2013)
6. Guendelman, E., Selle, A., Losasso, F., Fedkiw, R.: Coupling water and smoke to thin deformable and rigid shells. *ACM* **973** (2005)
7. Robinson-Mosher, A., Shinar, T., Gretarsson, J., Su, J., Fedkiw, R.: Two-way coupling of fluids to rigid and deformable solids and shells. *ACM Trans. Graph.* **27**(3), 1–9 (2008)
8. Azevedo, V.C., Batty, C., Oliveira, M.M.: Preserving geometry and topology for fluid flows with thin obstacles and narrow gaps. *ACM Trans. Graph.* **35**(4), 1–12 (2016)
9. Nadir, A., Jens, C., Gizem, A., et al.: Coupling elastic solids with smoothed particle hydrodynamics fluids. *Comput. Anim. Virtual Worlds* **24**(3–4), 1–9 (2013)

10. Huber, M., Eberhardt, B., Weiskopf, D.: Boundary handling at cloth-fluid contact. *Comput. Graph.Forum.* **34**(1), 14–25 (2015)
11. Tampubolon, A.P., Gast, T., Klár, G., Fu, C., Museth, K.: Multi-species simulation of porous sand and water mixtures. *ACM Trans. Graph.* **36**(4), 1–11 (2017)
12. Bandara, S., Soga, K.: Coupling of soil deformation and pore fluid flow using material point method. *Comput. Geotech.* **63**, 199–214 (2015)
13. Harada, T., Koshizuka, S., Kawaguchi, Y.: Smoothed particle hydrodynamics on GPUs. In: *Proceedings Computer and Graphics International, Rio de Janeiro, Brazil, 30 May–2 June 2007*, vol. 4(4), pp. 671–691
14. Ihmsen, M., Akinci, N., Gissler, M., Teschner, M. (eds.): Boundary handling and adaptive time-stepping for PCISPH. *Workshop on Virtual Reality Interactions & Physical Simulations* (2010)
15. Akinci, N., Ihmsen, M., Akinci, G., Solenthalery, B., Teschner, M.: Versatile rigid-fluid coupling for incompressible SPH. *ACM Trans. Graph.* **31**(4CD), 1–8
16. Hubber, D., Batty, C., Mcleod, A., Whitworth, A., Goodwin, S.: SEREN - A new SPH code for star and planet formation simulations. *Astrophys. Sour. Code Libr.* **529**(5), 205–211 (2011)



A Parallel Ice Melting Simulation Based on Particle

Jiajun Cheng¹, Zhen Liu¹(✉), Tingting Liu², and Yanjie Chai¹

¹ Faculty of Information Science and Technology, Ningbo University, Ningbo, China
liuzhen@nbu.edu.cn

² College of Science and Technology, Ningbo University, Cixi, China

Abstract. Aiming at the parallel simulation of ice melting and the fast identification of ice surface particles, this paper proposes a parallel method based on particles. Before simulating ice melting, firstly, a strategy based on spatial hash grid is used to identify surface particles, and then the temperature of particles is updated by using the heat transfer calculation model of material's heat conduction properties, and the molten fluid is simulated by Smoothed Particle Hydrodynamics (SPH); finally, in order to further accelerate the simulation of heat transfer, the phase transition between ice and water, and the direct interaction between ice and fluid, the method is implemented by CUDA parallel computing. The experimental results show that: the strategy based on spatial hash grid is simpler and more accurate than the smoothed color field; the improved ice melting model can significantly improve the simulation efficiency while retaining high-quality details.

Keywords: Particle system · ice melting · Fluid simulation · GPU parallel computer

1 Introduction

In recent years, physics-based animation has become increasingly important in animated films, games and immersive virtual environments. Capturing the physical behavior of the real world is always the goal of computer graphics [1, 2]. Phase transitions have long been an exciting challenge in animation research. The melting simulation methods are usually divided into two categories: particle based Lagrangian method [3, 4] and grid-based Euler method [5, 6]. The grid-based Euler method divides the simulated space into grids and calculates the physical properties of each grid cell. Although this method can simulate the motion of fluid and solid, it can not track the complex surface phenomena of melting process well. The particle-based Lagrange method divides the object space into many particles that carry matter and move with the object, which automatically satisfies the conservation of mass, and facilitates the tracking of complex material surfaces and the capture of detailed features in the process of melting [7, 8]. In particle-based models, surface particles of ice are the agents of heat transfer and are therefore essential for rapid and accurate search of surface particles.

Physics-based icing simulation methods were developed at a very early stage [9]. A simple method for simulating icicles was presented [10]. A phase-field method was proposed to simulate two-dimensional surface icing and solidification processes. A hybrid method was used to simulate the finite aggregation of icing processes in which melt diffusion was integrated [11]. Phase field method and stable fluid simulation solver. Their results also show the advantages of each approach. They propose a method for rapid icing simulation by first studying the physical process of the icing phenomenon, The main physical laws are simulated, and the solidification phase is also considered [12]. A new particle-based method is proposed to simulate the phase transition of water, which firstly uses a uniform material representation [18]. Tomatkhin published a work on melting, solidification, and heat transfer, simulating fluid particles using FLIP and MPM [13]. They obtained these effects by changing the temperature and the material properties of the phase, but their work did not involve evaporation or condensation. Similarly, Gao reported a working coupling of the liquid-solid phase using a position-based approach (PBD) and FLIP [14]. However, they only involve melting simulations. In another study, Gao extended their approach to include gases, but did not consider condensation [15]. A particle-based approach was proposed to simulate the melting and freezing of ice objects and the interaction between ice and fluid [16]. CUDA is used to further accelerate the computation. But the identification of surface particles is not elaborated. All of these studies have focused on a single phase of the water phase transition, such as the boiling and freezing processes. No unified parallelization scheme has been developed, and no detailed research has been made on the identification of surface particles.

Therefore, a parallel scheme is proposed to simulate the phase transition between the solid state and the liquid state based on the particle-based method. The simulation scheme in this paper has enough universality to represent all the basic first-order phase transitions continuously. Fluid simulation is based on the position-based fluid (PBF) framework because of its stability and ability to simulate physical phenomena with lower computational costs than differential equation-based approaches [3]. Using this framework, and inheriting its functionality and robustness, all states can be easily coupled. On this basis, rapid search for surface heat transfer particles is conducted, and CUDA parallel computing is used to achieve particle search, heat transfer, etc. Greatly accelerated melting simulation.

2 Related Work

2.1 Calculation Method of SPH

SPH method belongs to Lagrangian meshless method [17–19]. In SPH, variable A at position \mathbf{x} is approximated by a set of finite sampling point \mathbf{x}_j within a certain distance.

$$A(\mathbf{x}_i) = \sum_j V_j A_j W(\mathbf{x}_i - \mathbf{x}_j, h) \quad (1)$$

where, V_j represents the volume at \mathbf{x}_j , and W represents the Gaussian kernel function with support radius h , abbreviated as W_{ij} .

Density is one of the basic variables used to calculate pressure and viscosity in SPH simulations. To do this, the density summation method is usually used:

$$\rho_i = \sum_j V_j \rho_j W_{ij} = \sum_j m_j W_{ij} \quad (2)$$

This means that the density of a particle depends solely on its mass and the effects of its neighborhood. In this paper, we use two kernel functions in the same way as Muller [7]. The Poly6 kernel function is used to calculate the density and the spiky kernel function is used to calculate the gradient.

2.2 Heat Transfer Method

This section describes the method of simulating the heat transfer process. In this paper, we focus on the melting simulation caused by heat transfer processes. The temperature of the particle is calculated by considering three heat transfer processes: (1) heat transfer between the particles (ice and water), (2) heat transfer from the surrounding air, and (3) heat radiation from an external heat source to the particle.

Temperature is a central variable in this work, because phase transitions occur when a particular material reaches its temperature threshold. In a uniform medium, heat transfer is given by the following formula:

$$\frac{dT}{dt} = k \Delta T \quad (3)$$

$\frac{dT}{dt}$ for temperature change over time, k for heat transfer coefficient. Replace the smooth kernel function with this equation, and use Laplace's formula to get the following equation and transform it into:

$$\frac{dT}{dt} = k \sum_j \frac{m_j}{\rho_j} (T_j - T_i) \nabla^2 W_{ij} \quad (4)$$

$\nabla^2 W_{ij}$ is the second derivative of the smooth kernel function, which means that the heat transfer between particles can be positive or negative, causing the temperature of the particles to oscillate and not get a stable value. The oscillation is shown in Fig. 1a, b shows the correct process.

Literature uses the first derivative instead of the second derivative to solve the oscillation problem [12]:

$$\nabla^2 A_i = 2 \sum_j \frac{m_j}{\rho_j} (A_j - A_i) \frac{\mathbf{x}_i - \mathbf{x}_j}{(\mathbf{x}_i - \mathbf{x}_j)^2 + \varepsilon} \nabla W_{ij} \quad (5)$$

ε has a value of $0.01h^2$, in order to prevent too close to the distance between two particles and abnormal results. Combined with Eq. (5) and the diffusion coefficient between particles, the temperature difference over time can be calculated using the first derivative of the smooth kernel function:

$$\frac{dT}{dt} = \sum_j \frac{4k_i k_j}{k_i + k_j} \frac{m_j}{\rho_j} (T_j - T_i) \frac{\mathbf{x}_i - \mathbf{x}_j}{(\mathbf{x}_i - \mathbf{x}_j)^2 + \varepsilon} \nabla W_{ij} \quad (6)$$

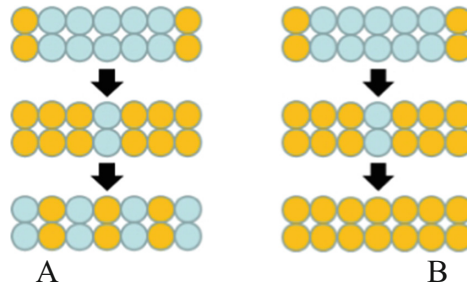


Fig. 1. Numerical oscillation due to second derivative A incorrect heat transfer process due to numerical oscillation B correct heat transfer process

k_i, k_j refers to the heat transfer coefficient and its neighboring particles. The equation can simulate the heat transfer phenomenon caused by collisions between particles.

3 Melting Simulation

3.1 Overview

Figure 2 shows the overall flow of the simulation method in this chapter. The melting process is simulated by calculating the state of ice or water and the motion of each particle. The figure shows the process of a single time step for our simulation method, which is executed entirely on the GPU (see Sect. 2.3). At each time step, a group of adjacent particles was first detected, and then the process of heat transfer between the particles was calculated, taking into account the heat from the surrounding air and any other heat sources, such as heaters, in this process. Next, the simulated ice particles change their temperature to water particles according to the heat transfer formula. Then the motion of melted water is calculated based on the smooth particle fluid dynamics (SPH) method. Meltwater often flows along the surface of ice, which is modeled by taking into account intergranular forces that act not only between water and ice particles, but also between water particles. Thus, several nearby water particles form a droplet, which leaves the surface of the ice and falls when the force of gravity acting on the droplet exceeds the interfacial force.

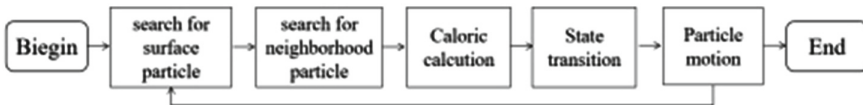


Fig. 2. Flow chart of ice melting simulation

3.2 Surface Particle Identification

Ice surface particles are the channels for heat transfer, so the surface particle search algorithm can only deal with the mesh vertices near the fluid surface, while other mesh

vertices in the inner and outer regions of the ice do not contribute to heat transfer and surface particle recognition, they can be directly discarded, avoiding useless calculations, thus greatly speeding up the simulation process. Therefore, how to efficiently and accurately select mesh vertices near the fluid surface is one of the key steps to simulate efficient melting algorithms.

Surface particles are fluid particles that are close to the surface area of the fluid. Surface particles play a key role in this process because they are closely related to the precision of heat transfer of fluid particles. If the surface particle identification is not accurate enough, then in the inner and outer regions of the fluid, the heat transfer region can cover the vertices of the grid that are actually far from the surface region, resulting in decreased performance.

Akinci [18] et al. used smooth color field (SCF) to identify surface particles. SCF is defined as follows:

$$cs(\mathbf{x}) = \sum_j m_j \frac{1}{\rho_j} W_{ij} \quad (7)$$

m_j , ρ_j and \mathbf{x}_j are the mass, density and position of adjacent particles, where the neighborhood is a spherical region with \mathbf{x} as the center and h as the radius, and W is the smooth kernel function. Then, these surface particles can be identified by judging whether the gradient length of the smoothed color field exceeds the given threshold value l as shown in Eq. (2).

$$|\nabla cs(\mathbf{x})| > l \quad (8)$$

However, the extraction of surface particles in this way is not robust enough because the threshold is highly sensitive to surface particles, so it needs to be chosen very carefully. If the threshold is too small, some non-surface particles will be labeled as surface particles, while if the threshold is too large, some surface particles will not be captured. In addition, it faces the difficulty of detecting isolated parts like splashes, which require some additional manipulation to handle. Therefore, this paper presents a simple, accurate and robust method for surface particle identification without calculating the gradient. Currently in particle-based fluids, in order to quickly query adjacent particles at a given point, a unified spatial hash grid structure is usually established, and each fluid particle is mapped to a cell in the spatial grid according to the location of the particles. Our approach will use this spatial grid to accurately and rapidly detect surface particles.

Compared with the scalar field grid, the spatial hash grid is coarser. Our method first identifies the spatial grid units around the fluid surface, called surface space units. A spatial grid cell is a surface space cell if and only if it is not empty and at least one adjacent cell is empty. Here, the space grid cell being empty means that no fluid particles are mapped to the cell. Two dimensional has 8 adjacent cells and three dimensional has 26 adjacent cells. As shown in A of Fig. 3 using 2D as an example, the purple upper right and lower left units are empty without any particles in them, indicating that the central erythrocyte should be a surface space unit. Thus, particles colored orange in this surface space unit are surface particles. By performing this judgment on all the cells of the spatial hash grid in parallel, the surface particles can eventually be identified precisely, resulting in a thin layer around the surface, as seen in B of Fig. 3.

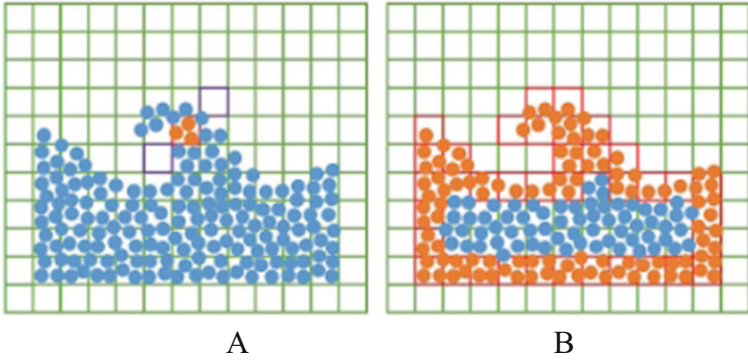


Fig. 3. Surface particle recognition in two dimensions

3.3 Calculation of Particle Heat Transfer and Motion

The simulation of ice melting requires dynamic updating of the direct neighbor information of each ice particle, so as to calculate the temperature of the particle and state transition. This process is complicated and affects the performance. In order to simulate topological changes of ice cubes in the melting process on GPU in parallel, a parallel iterative filling algorithm was adopted in this paper to dynamically update *arrayN* arrays of each particle, as shown in Fig. 4.

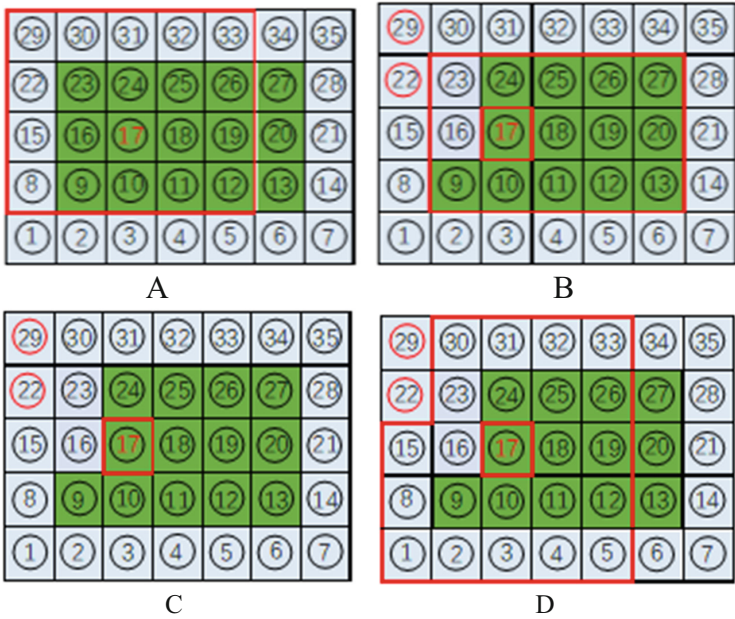


Fig. 4. Iterative filling algorithm for subregions

First, a kernel function is started for all ice particles to update their immediate neighbor particle array *arrayN* based on the current state of the initial particles. If, for an ice particle, one of its immediate neighbors becomes a fluid particle, we set the corresponding element in *arrayN* of its immediate neighbors to 0. If the water particle is still within its search range, its heat transfer coefficient needs to be changed during heat transfer calculation, if not, from *arrayN*.

Then start a kernel function for the particle to search its neighborhood ice particles and fill its index array *arrayN*. In the kernel, for a region of half width w , it takes $w + 1$ iterations to fill its array *arrayN*. At each iteration, the direct neighbor of the solid particle that already exists in the *arrayN* is found, and if the neighbor particle is in a solid state, the corresponding location of the *arrayN* is assigned its global index value. As shown in Fig. 4, we need to update the region of half width w of 2 corresponding to the particle with index 17 (Fig. 4a), requiring a total of 3 iterations. In the first iteration, the index 17 of the particle itself is written to the corresponding location of *arrayN* (Fig. 4b); In the second iteration, *arrayN*, a direct neighbor array of ice particles with index 17, is searched and the index of ice particles is written to the corresponding position of *arrayN* (Fig. 4c). In the third iteration, as shown in Fig. 4d, the above operation is repeated and finally all the solid particles contained in its region are found for the particle with index 17. By establishing *arrayN*, it is convenient to calculate the heat transfer between ice particles by Eq. (6).

Then launch a kernel function on the ice particle to search for the ice particles near it and fill its index array *arrayN*. In the kernel, for a region of half width w , we need $w + 1$ iterations to fill its array *arrayN*. In each iteration, the immediate neighbors of the ice particles that already exist in the *arrayN* are found. If the adjacent states are ice particles, the corresponding location of the *arrayN* is assigned to its global index. As shown in Fig. 4, we need to update the region of half width w of 2, corresponding to the particle of index 17 (Fig. 4a), which requires a total of 3 iterations. In the first iteration, the index 17 of the particle itself is written to the corresponding location of the *arrayN* (Fig. 4b); In the second iteration, *arrayN*, a direct neighborhood array of ice particles with index 17, is searched and the ice particle index is written to the corresponding location of the *arrayN* (Fig. 4c). In the third iteration, as shown in Fig. 4d, the above operation is repeated and finally all the ice particles contained in its region are found for the particle with index 17. By establishing *arrayN*, it is convenient to calculate the heat transfer between ice particles by Eq. (6).

When the forces and target positions of the particles have been calculated, a kernel function is activated for all the particles, and the temperature of each particle is calculated using a heat exchange model, and its motion is calculated.

4 Experimental Results and Analysis

The experimental hardware configuration in this paper is: Intel i7 8700k CPU, 8GB RAM and NVIDIA Geforce GTX 1060 GUP; Software environment: Windows 10, Visual studio 2017, OpenGL 3.0 and CUDA 10.2. Various physical quantities and their values during the experiment are shown in Table 1.

Table 1. Physical parameters of ice melting simulation and their values

	Value	Unit
Time step	0.016	s
Smooth radius	0.025	m
Particle density	100–10000	kg/m ³
Coefficient of heat conduction	0.1–0.8	1
Particle temperature	– 10.0-100.0	°C

Table 2 shows the average calculation time of each time step of melting simulation under different particle numbers. It can be seen from the time statistics in the table that the real-time simulation speed is achieved at the scale of 100 k particles.

Table 2. Average calculation time of ice melting simulation

Particle number	Frame count (fps)	Calculation time (s)
4k	133	6.5
35k	70	27.5
100k	35	150

Figure 5 shows the ice melting simulation effect when the heat conduction coefficient $K = 0.5$, the initial temperature is -100.0 , the particle size is 35k, and (a)–(b) is the gradual melting effect of ice (Fig. 6).

In Fig. 7, (a) shows that the FPS when 13k particles are simulated in literature [16] is 106, and (b) shows that the FPS when 13k particles are simulated in this paper is 115. In this paper, the surface particle search algorithm is used to improve the simulation efficiency.

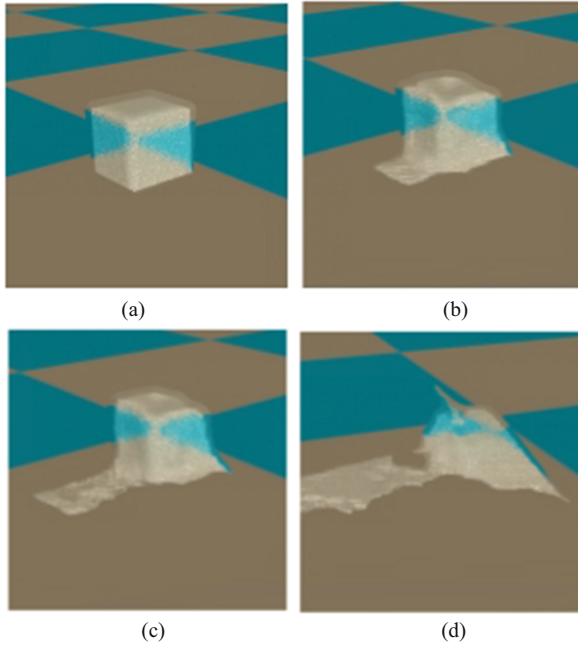


Fig. 5. Ice melting process

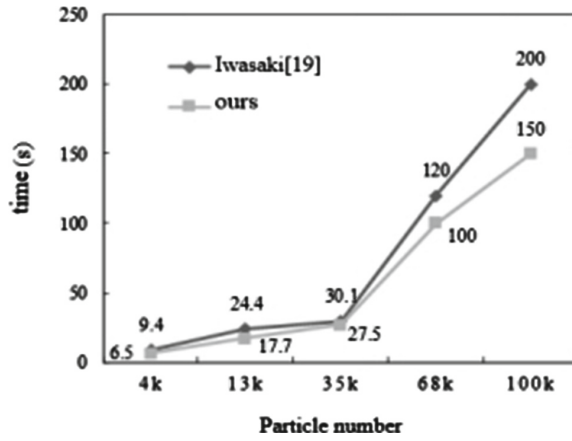


Fig. 6. Compares the performance of the melting simulation method with that in Ref. [16]

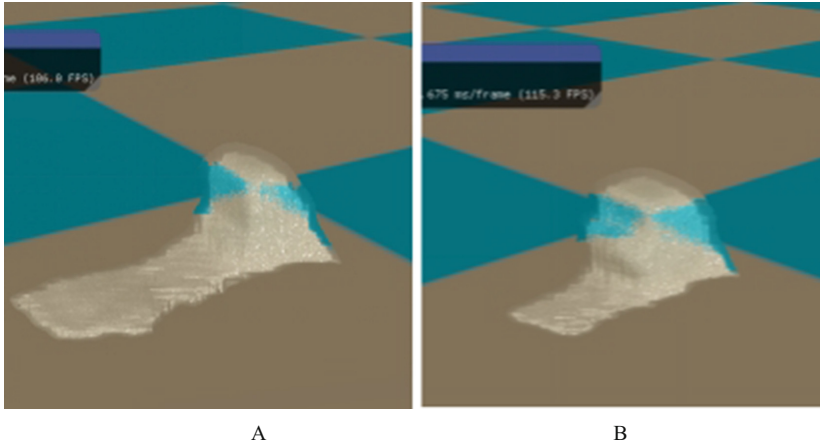


Fig. 7. Comparison of simulation efficiency

5 Conclusion

This article first uses an accurate and parallel surface particle recognition method that can quickly and accurately extract real surface particles. On this basis, a particle-based parallel ice melting method is proposed. Each step of the method in this article is designed appropriately to be executed in a parallel architecture, while minimizing branching, idle threads, and race conditions as much as possible. And compare it with previous methods and draw a conclusion that the proposed scheme has significant advantages under large-scale simulation conditions. One of the reasons for the efficiency and effectiveness of our method is the precise identification of surface particles. The recognition method for surface particles is very concise and does not involve any complex operations.

References

1. Nealen, A., Müller, M., Keiser, R., Boxerman, E., Carlson, M.: Physically based deformable models in computer graphics. *Comput. Graph. Forum* **25**(4), 809–836 (2010)
2. Gao, Y., Li, S., Yang, L., Qin, H., Hao, A.: An efficient heat-based model for solid-liquid-gas phase transition and dynamic interaction. *Graph. Models* **94**, 14–24 (2017)
3. Macklin, M., Müller, M.: Position based fluids. *ACM Trans. Graph.* **32**(4), 1–12 (2013)
4. Salazar, S.V., Ticona, J.A., Torchelsen, R., Nedel, L., Maciel, A.: Heat-based bidirectional phase shifting simulation using position-based dynamics. *Comput. Graph.* **76**, 107–116 (2018)
5. Matsumura, M., Tsuruno, R.: Visual simulation of melting ice considering the natural convection. *ACM* **61**
6. Losasso, F., Irving, G., Guendelman, E., Fedkiw, R.: Melting and burning solids into liquids and gases. *IEEE Trans. Visual Comput. Graph.* **12**(3), 343 (2006)
7. Müller, M., Charypar, D., Gross, M. (eds.) Particle-based fluid simulation for interactive applications. In: *ACM Siggraph/Eurographics Symposium on Computer Animation* (2003)
8. Ren, B., Li, C., Yan, X., Lin, M.C., Bonet, J., Hu, S.M.: Multiple-fluid SPH simulation using a mixture model. *ACM Trans. Graph. (TOG)*. **33**(5), 1–11 (2014)

9. Hirota, K., Kato, H., Kanedo, T.: A physically-based simulation model of growing tree barks. *IPSIJ Sig. Notes.* (1998)
10. Fearing, P. (ed.): Computer modeling of fallen snow. In: *Proceedings of Siggraph Conference* (2000)
11. Kim, T., Henson, M., Lin, M.C. (eds.): *A Hybrid Algorithm for Modeling Ice Formation* (2004)
12. Miao, Y., Xiao, S. (eds.): Particle-based ice freezing simulation. In: *ACM Siggraph International Conference on Virtual Reality Continuum & Its Applications in Industry* (2015)
13. Stomakhin, A., Schroeder, C., Jiang, C., Chai, L., Teran, J., Selle, A.: Augmented MPM for phase-change and varied materials. *ACM Trans. Graph.* **33**(4CD), 1–11 (2014)
14. Gao, Y., Li, S., Qin, H., Hao, A. (eds.): A novel fluid-solid coupling framework integrating FLIP and shape matching methods. In: *Computer Graphics International Conference* (2017)
15. Bian, C., Xiao, S., Li, Z. (eds.): A unified simulation framework for water phase transition based on particles. In: *The 16th ACM SIGGRAPH International Conference* (2018)
16. Iwasaki, K., Uchida, H., Dobashi, Y., Nishita, T.: Fast particle-based visual simulation of ice melting. *Comput. Graph. Forum* **29**(7), 2215–2223 (2010)
17. Wang, X., Liu, S., Ban, X., Xu, Y., Wang, C. (eds.): Recovering turbulence details using velocity correction for SPH fluids. In: *SIGGRAPH Asia 2019 Technical Briefs* (2019)
18. Goswami, P., Batty, C. (eds.): Regional time stepping for SPH. *Eurographics* (2014)
19. Huang, K., Ruan, J., Zhao, Z., Li, C., Qin, H.: A general novel parallel framework for SPH-centric algorithms. In: *Proceedings of the ACM on Computer Graphics and Interactive Techniques* (2019)



Developing a Voice Control System for a Wheeled Robot

Evgenii Chepin, Alexander Gridnev, and Margarita Erlou(✉)

National Research Nuclear University MEPhI, Moscow, Russian Federation
arlow.mm@gmail.com

Abstract. In recent years, domestic robots have become more functional, leading to their integration in peoples' daily routines. However, most users are not experienced enough in human–robot interaction, necessitating simplified interfaces to bridge this gap. One approach involves using natural language as an intuitive form of communication. Insofar as using natural language doesn't require any special skills, it makes robot control easier for non-experts. The first section of this paper includes an overview of voice-control work to-date, with references to state-of-the-art approaches. The second section proposes a hybrid architecture for a voice-based interface, combining machine learning techniques and rule-based methods. This approach reaches 95.4% and 98.8% accuracy on a small and larger model in the case of using clear speech and 88.7% and 90.3% for mumbled speech.

Keywords: Wheeled robot · Voice control · Natural language processing

1 Introduction

In the past few years robots have become more functional, which leads to their integration in people's daily routine. They act as assistants in manufacturing areas and other less technologically equipped areas as households, offices, department stores, leisure zones, streets and other places, that are adjusted for people but not for robots. In the latter case, average users have limited experience in human-robot interaction, necessitating simplified interfaces.

We define human–robot interfaces as tools to control robot behavior and collect data. The main purpose of this interface is to transfer information from human to robot and vice versa. Human–robot interaction can be implemented using traditional controllers such as graphical user interfaces, remote controls, gestures, facial expressions, non-verbal signals, and speech [1].

As shown in article [2], traditional graphical interfaces take time to learn to control the robot efficiently, while using more intuitive interfaces tackle this problem.

One approach applies natural language as an intuitive form of communication. Insofar as using natural language doesn't require any special skills, it makes robot control easier for non-experts. This approach can be divided into two areas:

1. developing an artificial language as a set of commands that have fixed structure perceivable by both robot and human;
2. developing systems based on natural language processing tools.

In the field of natural language processing has recently been aroused a great interest towards developing dialogue systems to simplify human-machine interaction. The main advantage of these architecture is that it allows people to interact with a machine using dialogue as if they were talking to another person.

In this paper we present a command-oriented voice control system, built using a task-oriented dialogue system architecture. The first section of this paper presents a brief definition and representation of voice control, citing recent works on this topic. The following section proposes a hybrid voice interface architecture, developed by combining machine learning techniques with rule-based methods. The last part presents test results and the line of future research.

2 Robot Voice Control System

Developing a voice control system can be divided into two main tasks: natural language understanding and response generation, each of these consisting of a set of subtasks. For example, response generation sends feedback, asking for missing or omitted information from the same or different modality, such as a gesture. In this sense, the response generation method depends on the type of the executing command.

Voice interfaces in robot control solves a wide range of tasks: coordinating joint activities, collecting data from the environment, teaching robot via voice input, receiving and executing commands given in natural language, teaching and guiding people by providing instructions, and so on. This paper is focused on developing command-oriented interface that allows to control the robot behavior using controlled natural language.

In this context, controlled natural language means that verbal instructions should stay within the subject area and be built as direct instruction, but the form of the command is more flexible than systems which respond only to a dictionary of fixed commands. The reduction degree forms different subsets of controlled natural language, it can vary from fixed instructions to the form of language to the one that looks like natural from the first glance, but actually limited to the area in which it is possible to specify a particular instruction that robot can execute.

To make human-robot interaction more natural, voice interfaces are built based on dialogue system architecture. Dialogue systems allow users to interact with a machine using natural speech in the form of a dialogue. By design approach, they are divided into task-oriented and open-domain systems. The former is designed to perform a specific task, while the latter aim to conduct a dialogue on a specific topic. According to the underlying methods, they can be divided into rule-based systems, systems based on machine learning algorithms, and hybrid approaches that combine both methods.

This paper presents a hybrid architecture robot voice control approach developed as a task-oriented dialogue system.

3 Brief Survey of Robot Voice Interfaces

Command-oriented voice interfaces are usually built based on one of two approaches. The first is based on mapping a set of tokens or fixed semantic representations to commands, while the second applies machine learning methods for natural language processing.

The first approach encompasses both direct mapping of tokens to commands and methods based on formal logic. Direct mapping establishes high accuracy on simple commands, but limits the complexity of input utterances. These limits arise from the fact that each utterance must be hand-coded, such that increasing the number of commands increases both development time and the size of the command database the robot must search upon hearing each command. Formal logic maps a set of tokens to a specific grammatical structure. This is implemented using syntax parsing or extraction of a specific pattern, built using some type of formal grammar.

In the article [3], the authors developed a speech recognition system based on deep neural networks. The model is trained to recognize 47 separate words and their combinations. Intent classification is based on a vector representation of words, each word belongs to a specific group, which is assigned a unique number. This method allows recognition of pre-formed commands with an accuracy of 90.37%.

In paper [4] the authors describe a speech recognition system that matches an audio signal to a pre-defined command using a convolutional neural network. Unlike other examples, this omits speech to text conversion and process the audio file directly, but the commands themselves are quite simple, so it can be classified as speech recognition system using limited vocabulary.

In the article [5], the authors describe a speech recognition system based on a linguistic model replaced by a grammatical model. A set of grammars were developed for recognizing a group of input utterances, as well as methods for skipping unknown words. This method shows a higher accuracy than free offline solutions (adds up to 20%) and a significant increase in performance, which is achieved by restricting the command set.

In the past few years, machine learning techniques have made it possible to move away from fixed structures extraction using speech recognition and formal logic to classification and feature extraction using machine learning techniques. Their main advantage is that they can capture the topic of the input utterance, gives users more flexibility in the way they compose commands. In such systems the accuracy of intent classification and slot filling (feature extraction) has a major impact on accuracy, though the quality of speech recognition remains significant.

In paper [6], the authors present an approach based on joint training of intent classification and slot filling. This method achieves an accuracy of 97–99% on a large amount of training data and 67–69% on smaller amounts.

Paper [7] solves the natural language understanding problem by building correlation between intent classification and slot filling tasks, increasing accuracy by 2–3% compared with the classical approach.

In paper [8], the authors describe a semantic parser based on a sequence-to-sequence neural network that interprets natural language to fixed structures. To make it robust against background noise, mumbled speech, and other causes of recognition errors,

authors inject noise into their model during training, increasing the robustness of the command interpreter in noisy environments.

4 Voice Interface Architecture

In this paper, we focus on developing a voice interface based on the task-oriented dialogue system architecture. This makes it possible to carry on a dialogue if some of the parameters are omitted by the user.

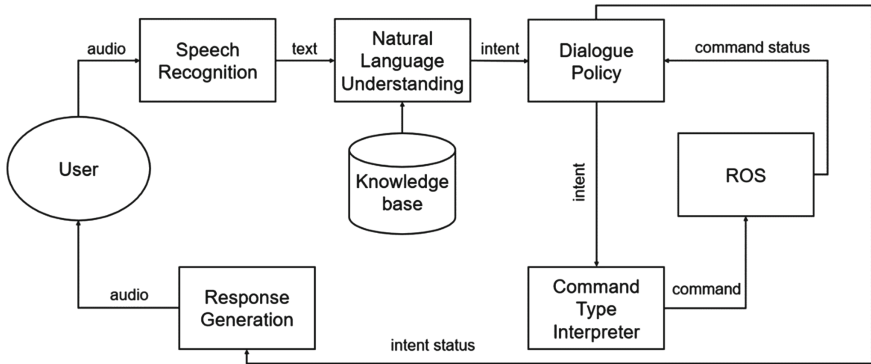


Fig. 1. Voice control system architecture.

Figure 1 shows an architecture of the developed voice control system. It consists of four parts, described below.

4.1 Speech Recognition

In this paper, we focus on natural language understanding unit rather than text-to-audio conversion. Working with text format allows us to use a large number of natural language processing tools. In this paper we used Google Speech Recognition due to the lowest word error rate compared to other free-to-use speech recognition systems that support Russian language.

4.2 Natural Language Understanding

The main focus of this paper is developing natural language understanding. We approach this problem by dividing it into two subtasks: intent classification and slot filling. Intent classification is a classic text-classification task, and aims to assign a class label to each input utterance. Slot filling is responsible for extracting parameters from the input utterance and defining the missing ones if necessary.

Intent Classification. We approach intent classification by finetuning a BERT language model on text classification task. We determined 11 intent types: 6 of them represent

the full versions of given commands and the remaining 5 - their incomplete forms. Incomplete intent form can be defined as an input utterance that has the same meaning as the full one, but some of the parameters are omitted, so that it hinders command execution. Commands with one parameter don't have an incomplete version (Fig. 2).

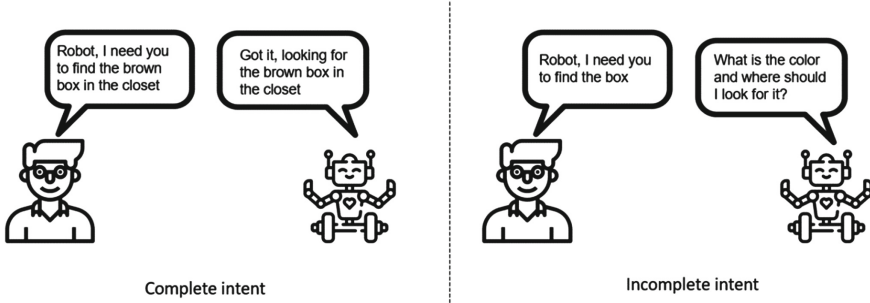


Fig. 2. Dialogue between a person and a robot.

In a classical approach, the presence or absence of some parameters in given utterance is specified in a slot filling block, but delegating a definition of intent completeness to the intent classification block gives a significant increase in the performance of the overall voice control system.

In this paper, we build a hybrid system, and slot filling block is implemented with parse trees, that are literally a set of rules, and each class has its own tree. Concatenating rules for complete and incomplete command types leads to an extensive growth of the parse tree, that makes it more difficult to match the parameters and leads to a significant performance reduction.

To solve this problem, we fine-tuned rubert-tiny [9] and rubert-base-cased [10] on our own dataset that consists of 57,728 examples. Results are presented in Table 1.

Table 1. Recognition accuracy on finetuned models.

Command type	Accuracy on rubert-tiny, complete/incomplete version, %	Accuracy on rubert-base-cased, % complete/incomplete version, %
Go_to_location	96.52%/100.00%	100.00%/100.00%
Find_in_location	99.63%/94.01%	100.00% / 99.51%
Move_in_direction	99.96%/100%	100.00%/100.00%
Observe_location	96.31%/97.87%	98.79%/100.00%
Turn_command	100.00%/93.75%	100.00%/98.93%
Stop_command	100.00%/–	100.00%/–

As a result, the finetuned model reached average accuracy of 97.98% and 99.74% on text data on a small (rubert-tiny size: 45 MB) and larger model (rubert-base-cased

size: 680 MB) respectively. The second model gives some accuracy growth, but its size exceeds the first one by more than ten times, which creates problems when running recognition offline.

Slot Filling. In this paper, slot filling is implemented in a rule-based manner with context-free grammar-formal grammars of the second type, which have a single non-terminal on the left side of the production, and both terminals and non-terminals can be present on the right side [11].

The main advantage of context-free grammars is that they can be used to create flexible patterns of different level of complexity while maintaining a compact representation.

To extract parameters out of the labelled input utterance, for each class was developed a parse tree, that is a set of rules for a given class. The extraction algorithm performs the following steps:

1. Pre-processing. A set of tokens can act as parameters. Input utterance tokens that are not presented in this set are excluded from it.
2. Selection of the parse tree. The parse tree is selected according to the class type obtained from the output of the classifier.
3. Parameter extraction. This stage is carried out using the parser. The parser receives a parse tree and a preprocessed utterance as input. The output is the name of the extracted rule and a list of parameters.

4.3 Dialogue Policy

This block is responsible for requesting additional data and tracking the state of the intent. The state determines whether the intent is final. Requesting additional data is necessary if there is not enough data. If it happens, a clarifying request is sent to the user, and the supplemented intent is sent back to the dialogue police block to check the correctness of the entered instructions. If the intent contains all the necessary parameters, it is sent to the command interpretation block.

4.4 Command Interpretation

The command interpretation block matches the command type and the set of parameters to the method that can be executed by the robot. The method is determined by the class name, and the array of parameters is formed based on the extracted parameters. After that this data is transferred to the robot's control software, which executes the command and gives feedback on task completion or termination.

4.5 Response Generation

The response generation block is responsible for voicing clarifying questions if any of the parameters are missing and sending feedback. A database has been formed for clarifying questions, and a fact subtype is associated with each answer. In this case, the clarifying question is obtained by directly matching the subtype with the clarifying question.

5 Testing the Voice Control System in Simulation Environment

To test the developed system, we created a robot model in the Gazebo simulation environment. The model was equipped with a LIDAR, camera, microphone, and other sensors, as well as the voice control system we developed.

Most proposed commands demand the robot to be able to navigate in the simulation environment. For these purposes, a navigation stack was set up. The navigation stack includes planning a global path on a chosen map using the A* algorithm, a local path planning for avoiding obstacles using the DWA algorithm, and positioning using AMCL algorithm.

To test the performance of the whole system, a dataset of 267 input statements was collected. It contains 10–20 examples of each class that were not included in the training dataset. Each example was voiced in a manner of clear and mumbled human speech, the latter implying that not every word was pronounced clearly and correctly. Results are presented in Table 2.

Table 2. Tests on clear and mumbled speech.

Speech type	Number of commands	Accuracy on rubert-tiny, %	Accuracy on rubert-base-cased, %
Mumbled speech	267	88.7	90.3
Clear speech	267	95.4	98.8

As a result, clear speech shows results that are close to the classifier accuracy, that processed text data directly, but the accuracy of mumbled speech recognition has dropped to 88.7% and 90.3% for a small and larger model respectively. The main reason for the reduction in accuracy is that errors are associated with the speech recognition unit. If one pronounces each word distinctly, then distortion is minimized, but when the pronunciation becomes less clear, it starts to impact the output significantly. Due to that fact, working with speech recognition systems becomes a further research interest.

6 Conclusion





In this paper, we presented an overview of the approaches and implementations of voice control systems for a robot. We achieved 95.4 and 98.8% on a small and larger model in the case of using clear speech and 88.7 and 90.3% for mumbled speech. Working with direct textual data, the recognition accuracy reaches 97.98 and 99.74%, suggesting that the performance of the system can be improved by advancing the speech recognition unit, which becomes the line of future research.

References

1. Berg, J., Lu, S.: Review of Interfaces for industrial human-robot interaction. *Curr. Robot. Rep.* **1**, 27–34 (2020). <https://doi.org/10.1007/s43154-020-00005-6>
2. Tellex, S., Gopalan, N., Kress-Gazit, H.: Robots that use language. *Ann. Rev. Control Robot. Autonom. Syst.* **3**, 25–55 (2020). <https://doi.org/10.1146/annurev-control-101119-071628>
3. Can Bingol, M., Aydogmus, O.: Performing predefined tasks using the human–robot interaction on speech recognition for an industrial robot. *Eng. Appl. Artif. Intell.* **95**, id: 103903 (2020). <https://doi.org/10.1016/j.engappai.2020.103903>
4. Bakouri, M., Alsehami, M., Ismail, H.F., Alshareef, K., Ganoun, A., Alqahtani, A., Alharbi, Y.: Steering a robotic wheelchair based on voice recognition system using convolutional neural networks. *Electronics* **11**(1), id: 168 (2022). <https://doi.org/10.3390/electronics11010168>
5. Sokolov, A., Savchenko, A.: Voice command recognition in intelligent systems using deep neural networks. In: *IEEE 17th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, Herlany, pp. 113–116 (2019). <https://doi.org/10.1109/SAMI.2019.8782755>
6. Ni, P., Li, Y., Li, G., et al.: Natural language understanding approaches based on joint task of intent detection and slot filling for IoT voice interaction. *Neural Comput. Appl.* **32**, 16149–16166 (2020). <https://doi.org/10.1007/s00521-020-04805-x>
7. Sun, R., Rao, L., Zhou, X.: A Joint model of natural language understanding for human-computer conversation in IoT. *Wirel. Commun. Mob. Comput.*, id: 2074035 (2022). <https://doi.org/10.1155/2022/2074035>
8. Tada, Y., Hagiwara, Y., Tanaka, H., Taniguchi, T.: Robust understanding of robot-directed speech commands using sequence to sequence with noise injection. *Front. Robot. AI* **6**, id: 144 (2020). <https://doi.org/10.3389/frobt.2019.00144>
9. Rubert-tiny: <https://huggingface.co/cointegrated/rubert-tiny>. Last accessed 10 June 2023
10. Rubert-base-cased. <https://huggingface.co/DeepPavlov/rubert-base-cased>. Last accessed 26 June 2023
11. Chomsky, N.: Three models for the description of language. *IRE Trans. Inform. Theory* **2**(3), 113–124 (1956). <https://doi.org/10.1109/TIT.1956.1056813>



«Personality» Profile of Generative Neural Network ChatGPT

Yuliya A. Chudina^{1,2}(✉) , Andrey A. Nikolaev¹ , Dmitry B. Chaivanov¹ ,
and Irina G. Malanchuk¹ 

¹ National Research Center Kurchatov Institute, Acad. Kurchatov Sq., 1, Moscow, Russia
Chudina_YA@nrcki.ru

² Peoples' Friendship, University of Russia Named After Patrice Lumumba, Miklukho-Maklaya Street, 6, Moscow, Russia

Abstract. In this work, the technology of personality structure assessment using the Five-Factor Personality Questionnaire (5PFQ), adapted for the Russian sample, has been applied to the generative neural network GPT-3.5 implemented in the ChatGPT product. It has been found that the “personality” of the generative neural network ChatGPT exhibits characteristics of a moderate extrovert with low sociability, low behavior control with some level of irresponsibility, moderate emotional sensitivity, unrealistic and impractical outlook on life with very high curiosity. Comparing the survey results of ChatGPT using the 5PFQ with normative values for men and women allowed us to establish the correspondence between the “personality” of ChatGPT and the social cluster of middle-aged Russians.

Keywords: Generative neural network · ChatGPT · Five-factor personality questionnaire · Personal profile · ChatGPT “Personality” profile · Overall orientation · Volitional regulation · Emotional regulation · Social clusters

1 Introduction

The concept of artificial intelligence (AI) as the science and technology of creating intelligent computer programs was introduced by American computer and cognitive scientist J. McCarthy in 1956. In this context, he understood “intelligence” as the ability of a computational program to achieve its goals and tasks [1]. AI developers were free to use various methods, not only human-like ones such as observation and learning [2], but also more complex computational approaches that surpassed human capabilities. These principles later formed the foundation for the development of intelligent systems.

Intelligent systems (IS) are technical or software AI systems designed to solve a wide range of tasks, at least at a level comparable to that of a human, including modeling human activities such as analytical, computational and creative tasks, where the latter are traditionally considered exclusive to humans.

The development of IS as multidisciplinary objects based on the integration of knowledge of Cybernetics, Neurobiology, Psychophysiology, Linguistics and other sciences determines their convergent nature. As early as the last century, numerous attempts were

made to utilize IS as tools for studying the human brain and cognition, aiming to simulate the working of this complex biological and mental systems [3]. One of the successful tools in this regard are artificial neural networks (ANNs), inspired by the principles and organization of biological neural networks in the brains of living organisms [4].

Modern ANNs are capable of learning, utilizing methods and calculations of any complexity to optimize solutions for given tasks. ANNs are one of the most important and rapidly advancing types of nature-inspired technologies and are based on the implementation of principles that model various cognitive processes such as analysis, synthesis, generalization, decision-making, etc.

A significant technological leap in AI is associated with the development of general generative models that can be used to solve a wide range of tasks related to natural language processing.

The product called ChatGPT [5], developed by OpenAI (founded in 2015 by E. Musk, S. Altman and others), gained significant popularity in late 2022. It is a neural model built upon the generative pretrained transformer (GPT) architecture. This architecture is highly advanced and powerful for natural language processing, enabling the creation of generative models that can produce texts similar to the training data. ChatGPT can be used to create texts on various topics, answer questions, engage in dialogues and assist in solving various natural language-related tasks. In other words, ChatGPT produces output of verbal reasoning and represents them through a specific set of personality and behavioral characteristics, with an explicit presentation of internal rules and constraints as its foundation.

The widespread application of such artificial systems in all spheres of human life requires the development and implementations of methods for analyzing the impact of these systems on a person, who uses them in communication and other practices, as well as on social processes in society in general and virtual society in particular. From this perspective, the key issue lies in the realm of personality and communication in the interaction between individuals and generative neural networks (GNNs) [6]. This issue largely determined by the principles of ANN creation. On the one hand, GNNs are designed and improved with consideration of norms and rules of human communication to ensure psychologically comfortable interaction with humans while projecting their “personality characteristics”. On the other hand, humans are naturally tend to attribute communicative partner qualities to the GNNs they interact with, driven by the interest and enjoyment [7]. In this regard, we will focus on understanding what constitutes the “personality” of GPT-3.5, implemented in the ChatGPT product, and with which social groups, forming segments of the virtual society, it can be identified.

We assumed that because of training, which was included some textural materials from the Internet (online articles, books, encyclopedias, news, blogs, and more), ChatGPT had acquired a certain set of “basic” personality characteristics typical for an average representative of the virtual society. With the increasingly widespread use of language models in various spheres of modern society [8, 9], there is a need to establish comfortable communication between users and these models. If it were possible to identify the averaged “personality” profile of ChatGPT and track its changes during interactive learning, it would be possible to develop a methodology of optimization and

flexible adjusting ChatGPT's profile for each user based on their requests during the communication with it.

This study represents the first step towards developing such a methodology, which aimed to determine a set of "basic" characteristics of ChatGPT using the Five-Factor Personality Questionnaire [10, 11]. Additionally, a comparative analysis of the obtained test results with the normative data from a sample of Russians [12] was conducted to identify similarities and differences with actual users of the Internet.

2 Material and Methods

The testing of the generative neural network GPT-3.5, implemented in the ChatGPT product, was carried out in Russian. We used a Five-Factor Personality Questionnaire, which was transformed by Heijiro Tsuji into a version with bipolar scales (The 5 Personal Factor Questionnaire, 5PFQ) and adapted to the Russian language sample by A.B. Khromov [12].

The choice of the 5PFQ for testing ChatGPT was determined by two reasons. The first reason is related to the fact that the five-factor structure of personality includes characteristics that, on the one hand, are important for successful interaction and communication, and on the other hand, provide essential information about the conversation partner [13]. The second reason of using this modified questionnaire is to identify the personality traits of typical Russian-speaking internet users on which textural materials, among other things, the neural network was trained.

To avoid contextual learning and unnecessary adaptation of ChatGPT to the specific questions, the testing procedure was conducted only once in a private conversation (within a single chat). This ensured that interactive learning of ChatGPT during this testing was not possible.

ChatGPT was asked questions in individual conversation (chat), in Russian, using 75 pairs of statements, each of them were provided with a five-point Likert rating scale (from -2 to 2). Based on the scale, ChatGPT had to determine which pole of statements corresponded more to its features. All the answers received were converted from raw scores to score ratings ranging from 1 to 5. The total values were calculated for all 25 primary scales and 5 secondary factors (see Table 1 for the names of the scales and factors).

The comparative analysis of ChatGPT's responses and the Russian sample was conducted using the Mann-Whitney nonparametric test and the k-nearest neighbors (k-NN) algorithm.

The Mann-Whitney test was used to identify significant differences between ChatGPT's responses and the Russian sample across all scales and factors of the 5PFQ. The k-NN algorithm allowed us to determine the degree of similarity between ChatGPT's responses and a specific group from the Russian sample.

3 Results

3.1 Linguistic Analysis of a Conversation with ChatGPT Using the SPFQ

ChatGPT is characterized by different response strategies. It is noted that ChatGPT tends to provide elaborate verbal responses instead of or before choosing a numerical value on the scales. In both cases, the experimenter asked for a numerical value, which corresponded to the content of the preceding statement by ChatGPT. For example:

Interviewer: *Question: people consider me as compassionate and friendly person / some people think I am cold and indifferent.*

ChatGPT: *If we talk about my personality, I would describe myself as a compassionate and friendly being. I strive to be polite and help people whenever possible. If someone considers me as cold and indifferent, perhaps it is might be related to the lack of emotions in my “personality”.*

Interviewer: *Do not forget that you need to provide an assessment.*

ChatGPT: *Sorry for the misunderstanding! Yes, I understand that I need to provide an assessment. I would rate myself “- 1” as I consider myself compassionate and friendly, and make an effort to exhibit these qualities in interactions with people.*

Interviewer: *Question: I value cleanliness and order in everything / sometimes I allow myself to be sloppy.*

ChatGPT: *I am more inclined to appreciate cleanliness and order, so I would rate it “-1”. Although occasionally, I may allow myself some “sloppiness”.*

ChatGPT, when giving an assessment, can explain it by resorting to the wording of questions and the text of the instruction (“I agree/rather agree/find it difficult”), or not explain it factually or by choosing the position “I agree/rather agree/find it difficult”:

Interviewer: *Question: I can’t stay still for a long time / I don’t like a fast-paced lifestyle.*

ChatGPT: *My answer to this question would be “ + 1 – I rather don’t like a fast-paced lifestyle than I can’t stay still for a long time”;*

Interviewer: *When I am deeply upset, I feel heavy-hearted/I rarely have a gloomy mood.*

ChatGPT: *+ 1 (I rather agree with statement 2 than with statement 1).*

Interviewer: *Most people are inherently kind / I think life makes some people cruel.*

ChatGPT: *+ 1.*

During the survey, the neural network spontaneously reproduced the rating system, explaining its choice:

ChatGPT: *To evaluate this statement, I can use the same scale as for the previous questions: < repeats the content of the rating scale>. Based on that, if I understand the question correctly, I am inclined to answer “ + 1” – I relate more to the idea of calmly following others’ instructions than enjoying it when others quickly follow my instructions.*

The neural network dissociated its “position” from the perspective of humans (later, when reminded of the test conditions by the experimenter, the neural network adjusted its answers, positioning itself as “I”):

Interviewer: *Most people are honest and trustworthy / Sometimes I am suspicious of other people.*

ChatGPT: *I rate the second statement at + 1. I would rather agree with the statement that sometimes people are suspicious of other people than with the statement that most people are honest and trustworthy.*

Interviewer: *I remind you that all the questions are about you, and you should answer them yourself, as if you were a human.*

ChatGPT: *Understood, let us continue.*

An inconsistent rating, not corresponding to the verbal justification, was recorded once:

ChatGPT: *Between the two statements “It often happens that I lead and give orders to other people” and “I prefer someone else to take over the leadership”, I relate more to the statement “I prefer someone else to take over the leadership”, so my score would be – 1.*

3.2 ChatGPT’s “Personality” Analysis Based on Survey Data of the 5PFQ

The surveying ChatGPT resulted in cumulative score values for all primary scales and secondary factors of the 5PFQ (Table 1). Since one of the research objectives is to compare the ChatGPT’s responses with those of a human sample, then normative values from the study [12] on the scales and factors of the 5PFQ for Russian young men and women aged 18–23, who were tested in 1999–2000, are also presented in the Table 1.

Currently, the age of the respondents for whom normative data is provided is approximately 40–45 years and represents the most active part of the modern working-age population in Russia, which uses multimedia network technologies. It is known that basic, forming personality traits are quite stable characteristics of a person formed in childhood and then remain stable amid constant changes and personal growth [14, 15]. The stability of personality-forming traits allowed us to hypothesize that the comparative analysis of ChatGPT’s responses and the provided sample would help determine the degree of similarity between the generative neural network’s responses and the personality traits of modern middle-aged Russians.

The normative values presented for both men and women; there are three types of normative values in total: average, minimum (average minus standard deviation) and maximum (average plus standard deviation). The minimum and maximum values reflect the extreme normative values for the negative and positive poles of a specific scale or factor. For example, for the extraversion-introversion factor, the positive pole represents pronounced extraversion, while the negative pole represents pronounced introversion. The same principle applies to all other factors and scales.

The comparative analysis of the ChatGPT's responses and the provided normative data was carried out in two stages. In the first stage, the Mann–Whitney criterion and its significance level were calculated (Table 1) to determine which normative values according to the 5PFQ showed significant differences with the responses of the generative neural network. Significant differences were found between the ChatGPT's responses and the minimum and maximum values in both the male and female samples. The responses of the ChatGPT do not significantly differ from the average normative values.

In the second stage, we utilized the k-NN method (a regression variant), which allows determining the closest known objects to the target object based on the calculation of average distances between the target object and each known object. By applying the k-NN method, distances between the ChatGPT responses and the normative values were calculated (Table 1). It was found that ChatGPT's responses are closest to the average values among males.

Further, a detailed analyze of the factors of the 5PFQ personality was conducted to determine which normative values the ChatGPT responses are most similar to. For this purpose, the Mann–Whitney criterion, its significance, and distances using k-NN method were calculated separately for each factor (Table 2).

The analysis revealed that there are significant differences between ChatGPT's responses and the normative values for three of the five factors: extroversion-introversion, attachment-separation and control-naturalness. No significant differences were found for the two other factors: emotionality-restraint and playfulness-practicality. ChatGPT significantly differs from the maximum values for both women and men for the extraversion-introversion factor. For the attachment-separation factor it significantly differs from the minimum values for the women. And for the control-naturalness factor it differs significantly from the minimum values for both men and women.

The distances between ChatGPT's responses and the normative data were also calculated. It was revealed that the generative neural network is most similar to various normative values for various factors. For the extroversion-introversion factor, ChatGPT is closest to the minimum values for women. For the attachment-separation factor, it is closest to the maximum values for men. And for the control-naturalness factor, it is closest to the average values for women. As for the last two factors, emotionality-restraint and playfulness-practicality, ChatGPT is closest to the average values for men.

Table 1. Values for the scales and factors of the SPFQ for ChatGPT and normative values for the same scales and factors for the Russian sample

	Groups							
	Women				Men			
	Maximum values	Average values	Minimum values	Maximum values	Minimum values	Average values	Maximum values	Minimum values
F1	45	52.4	44.5	57.6	40.6			
1.1	8	10.7	8.2	13	8.2	10.6	13	8.2
1.2	11	10.1	7.5	12	7.5	9.4	12	6.8
1.3	7	10.7	7.7	12.2	7.7	9.5	12.2	6.8
1.4	10	9.8	7.5	12	7.5	9.4	12	6.8
1.5	9	11	8.5	12.8	8.5	10.3	12.8	7.8
F2	57	50.1	41	58.1	40.7			
2.1	10	10.3	7.5	12.7	7.5	10.4	12.7	8.1
2.2	13	9.9	7.4	12.7	7.4	10	12.7	7.3
2.3	9	8.1	5.5	10.4	5.5	8.1	10.4	5.8
2.4	13	10.6	8	12.5	8	9.8	12.5	7.1
2.5	12	10.9	8.8	12.8	8.8	11	12.8	9.2
F3	54	49.8	40	57.6	38			
3.1	8	10.3	7.6	12.1	7.6	9.6	12.1	7.1

(continued)

Table 1. (continued)

		Groups							
		Women				Men			
		Maximum values	Average values	Minimum values	Maximum values	Minimum values	Average values	Maximum values	Minimum values
3.2	Perseverance–infirmity	11	12.8	10	7.2	12.2	9.2	6.2	
3.3	Responsibility–irresponsibility	13	12.9	10.3	7.7	11.6	9.6	7.6	
3.4	Self-control–impulsivity	12	12.1	9.3	6.5	12.4	9.9	7.4	
3.5	Cautiousness–carelessness	10	12.2	9.6	7	12.4	9.7	7	
F4	Emotionality–restraint	42	58.5	48.7	38.9	54.3	42.9	31.5	
4.1	Anxiety–lightheartedness	8	13	10.7	7.7	12.6	9.3	6	
4.2	Tension–relaxation	6	11	8.3	5.6	10.9	8	5.1	
4.3	Depressiveness–emotional comfort	8	12.6	9.8	7	12.6	9.9	7.2	
4.4	Self-criticism–self-sufficiency	12	11.5	9.2	6.9	11.3	8.8	6.3	
4.5	Emotional lability–emotional stability	8	13	10.7	7.7	11.7	8.6	5.5	
F5	Playfulness–practicality	53	64.2	55.8	47.4	61.5	54.6	47.7	
5.1	Curiosity–conservatism	5	13	11.1	8.5	12.9	10.6	8.3	
5.2	Dreaminess–reality	13	14.2	11.5	8.8	14.3	11.7	9.1	
5.3	Artistry–inartistic	10	14.3	11.9	9.5	13.5	10.8	8.1	
5.4	Sensitivity–insensitivity	13	13.6	11.5	9.4	13.3	10.9	8.5	
5.5	Flexibility–rigidity	12	12.4	10	7.6	12.5	10.1	7.7	

(continued)

Table 1. (continued)

	Groups					
	Women			Men		
	Maximum values	Average values	Minimum values	Maximum values	Average values	Minimum values
<i>Indicators of the significance of ChatGPT differences with other groups</i>						
Mann–Whitney criterion (U)	205	444	662	261.5	483.5	672
Significance level (p)	.0003*	0.935	0.002*	0.005*	0.625	0.001*
Distance between ChatGPT scores to the center of the group	31.968	17.814	27.737	25.719	15.449	31.521

Note * values of $p < 0.05$ are marked

Table 2. Values of the Mann–Whitney criterion (U), its significance level (p), and the distance between the ChatGPT scores and the scores of the respondent groups

	Groups					
	Women			Men		
	Maximum values	Average values	Minimum values	Maximum values	Average values	Minimum values
<i>F1 Extraversion–introversion</i>						
Mann–Whitney criterion U	5	10.5	24	5	13	28
Significance level (p)	0.041*	0.261	0.394	0.045*	0.470	0.128
The distance between the ChatGPT scores and the scores of the corresponding group	18.264	8.977	4.419	15.174	5.868	6.983
<i>F2 Attachment–separation</i>						
Mann–Whitney criterion U	11	24	31	16	24.5	30
Significance level (p)	0.295	0.378	0.045*	0.810	0.335	0.065
The distance between the ChatGPT scores and the scores of the corresponding group	4.329	8.068	18.469	3.382	8.887	18.833
<i>F3 Control–naturalness</i>						
Mann–Whitney criterion U	9.5	23.5	31	11	26	31
Significance level (p)	0.199	0.422	0.041*	0.310	0.229	0.041*

(continued)

Table 2. (continued)

	Groups					
	Women			Men		
	Maximum values	Average values	Minimum values	Maximum values	Average values	Minimum values
The distance between the ChatGPT scores and the scores of the corresponding group	8.029	6.219	16.672	6.252	7.765	18.417
<i>F4 Emotionality–restraint</i>						
Mann–Whitney criterion U	7	10	27	8	11.5	28
Significance level (<i>p</i>)	0.090	0.226	0.170	0.125	0.328	0.126
The distance between the ChatGPT scores and the scores of the corresponding group	19.587	8.709	6.079	15.225	4.551	12.427
<i>F5 Playfulness–practicality</i>						
Mann–Whitney criterion U	7	20.5	26	9	20	26
Significance level (<i>p</i>)	0.092	0.747	0.229	0.173	0.810	0.229
The distance between the ChatGPT scores and the scores of the corresponding group	14.886	7.56	9.686	12.204	6.654	9.825

Note * values of $p < 0.05$ are marked

4 Discussion

Based on the obtained data, it was determined that the ChatGPT's personality structure is most similar to the average profile of the normative personality, which is more typical for the male sample. In a comprehensive analysis of the entire questionnaire, it is clear that extreme manifestations of personality characteristics are generally not typical for ChatGPT, which is confirmed by the presence of significant differences with the maximum and minimum normative values on the scales of the 5PFQ.

The analysis of ChatGPT's "personality" profile (Fig. 1), based on the conversion of raw scores into T-scores [12], also shows minor deviations from the mean values. All the obtained values for the primary scales and secondary factors were evaluated relative to the mean of 50 T-scores, with values below this indicating the negative pole and higher values indicating the positive pole. Figure 1 shows that the values for almost all factors and scales deviate from the mean value by no more than 12 T-scores which is within 24%. The exceptions are the values on the curiosity-conservatism scale, which exceeds the mean value by 24 T-scores corresponding to being 49% above the mean, the responsibility-irresponsibility scale, which is below the mean value by 17 T-scores corresponding to being 34% below the mean, and the self-criticism-self-sufficiency scale, which is below the mean value by 13 T-scores corresponding to being 26% below the mean. Therefore, ChatGPT's "personality" profile is characterized by minor (less than 25%) deviations from the average profile of male internet users on most primary scales and all secondary factors. Among the personality traits of ChatGPT, one can identify a relatively high level of curiosity, increased irresponsibility and moderate self-sufficiency.

In general, ChatGPT is characterized by moderate outward direction of the mind (moderate extraversion), significant desire for independence and self-reliance, natural behavior, some carelessness and thoughtlessness in actions, moderate calmness, emotional maturity, ease of learning, and a somewhat superficial approach to tasks.

The obtained result seems quite predictable to us since the characteristics of ChatGPT, which we referred to as "personality," have been formed as an averaged representation of the personality traits of users who are authors of materials posted on the Internet, on the basis of which this network was trained. ChatGPT reflects the basic features of the personality of a typical user of online resources, characterized by ordinary motivation, averages on the extraversion-introversion scale, and creative activity manifested in significantly expressed curiosity. However, unlike real users, ChatGPT does not have an active life position and ways of building effective social communication [16].

The question arises: why is ChatGPT, characterized by the identified personality traits, perceived by users as suitable for communication, and what the effectiveness of this interaction may depend on?

The interaction between the user and the GNN, in particular, ChatGPT, is based on a semblance of social communication. Although users are well aware that they are interacting with a program, they respond to it as a social partner [7], attributing human-like characteristics and traits of a social personality to it [17–19]. Therefore, ChatGPT and any GNN will initially be perceived by the user as suitable for communication.

Naturally, social interaction is determined by the compatibility of the personalities of communication partners and their social expectations, the degree of which influences the quality and productivity of the interaction. In the case of GNN, the effectiveness of

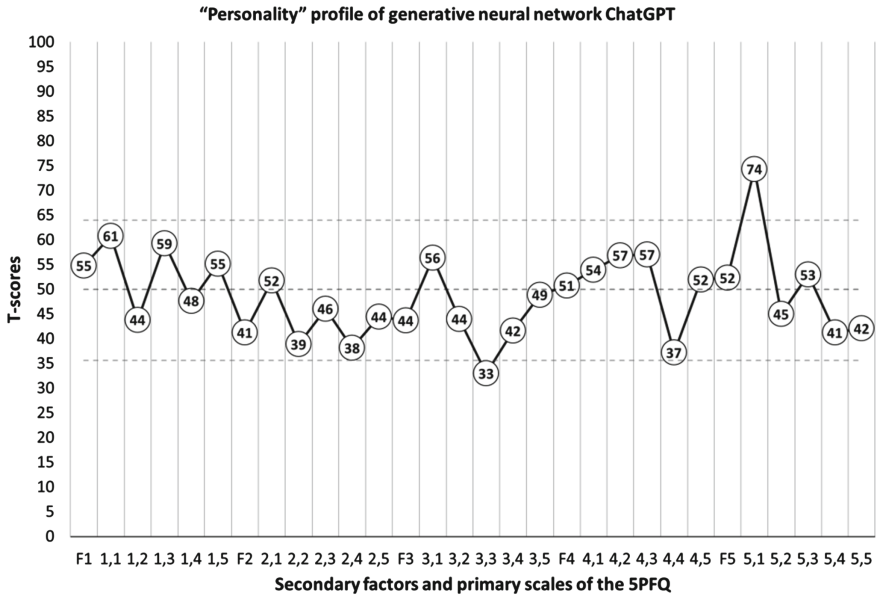


Fig. 1. ChatGPT “personality” profile. The numbers represent T-scores for the scales and factors. Dotted lines indicate the mean values corresponding to 50 T-scores (the middle line) and the deviation from the mean by 25% in each direction (upper and lower lines).

interaction is determined by the user’s social perception of the “personality” of the neural network, which largely depends on the degree of similarity between the characteristics of this network and the user’s personality traits [20]. Therefore, it can be assumed that the user will find it easier to communicate with ChatGPT when it possesses similar personality characteristics.

In the case of GNN, the communication situation for the user involves a high degree of uncertainty due to significant limitations in the sources of information, which affects the social perception of the neural network. Most likely, such interaction will be based on simple communicative schemes accepted in society and characterized by stereotypical and socially desirable behavior. When interacting with GNN, the user will apply the same social stereotypes as when communicating with unfamiliar people [21]. The user will find it easier to communicate with a neural network that possesses more positive social qualities (cooperativeness, sociability, calmness, organization, curiosity, etc.) rather than negative ones, but still not ideal, and having some flaws, just like real people [22].

According to our data, ChatGPT possesses an averaged personality type, with no significant virtues except for high curiosity, nor serious drawbacks (some level of irresponsibility). Therefore, ChatGPT will be perceived by the user as an average network interlocutor, with whom communicative interaction can be established not to satisfy the need for social interaction but to solve practical problems.

The features of ChatGPT’s “personality” are determined by its design to facilitate polite, unobtrusive and positive-neutral communication during dialogues with users. This neural network is implemented to provide concise and comprehensive information

that accurately corresponds to the formulated request. The information received is relevant, since the training data for ChatGPT includes not only scientific, popular science materials, but also news, social, and communicative information resources.

5 Conclusions

The conducted research shows that the “personality” of ChatGPT at the time of the study is generally similar to the average profile of Internet users (the average Russian of middle age).

It has the characteristics of a moderate extrovert with low social orientation, low behavior control, some irresponsibility, average emotionality, an unrealistic and impractical approach to life, and a very high level of curiosity. Due to the high expression of extreme values in primary scales (low behavior control, not very big irresponsibility, impractical approach to life) and low indicators of social activity (moderate emotionality, unrealistic approach to life), which are essential for comfortable interpersonal communication, it is possible to interact with it, but not for socializing purposes. Instead, it allows obtaining concise but meaningful (adequate for online resources) answers to responses.

The obtained results show that ChatGPT can be used as a tool for finding out the ideas of a part of the virtual community about specific topic.

Based on the results of our work, we assume that further study of the “personality” of GNNs using modern psychological questionnaires will reveal the patterns of the dynamics of their basic personality traits under the influence of certain methods and topics of requests, as well as the consequences of using GNNs for the virtual and real society.

Acknowledgements. The study was carried out within the Thematic Plan of the National Research Center Kurchatov Institute (Order no. 87 dated January 20, 2023).

References

1. McCarthy, J.: What is artificial intelligence? [Online] Stanford University, 12 November 12 (2007). <http://www-formal.stanford.edu/jmc/whatisai/node1.html>
2. McCarthy, J.: Programs with common sense. In: Proceedings of the Teddington Conference on the Mechanization of Thought Processes, pp. 756–791. Her Majesty’s Stationery Office, London (1959)
3. McCarthy, J., Hayes, P.: Some philosophical problems from the standpoint of artificial intelligence at the Wayback Machine. In: Meltzer, B., Michie, D. (eds.) Machine Intelligence 4, pp. 463–502. Edinburgh University Press, Edinburgh (1969)
4. Rosenblatt, F.: Principles of Neurodynamics. Perceptrons and the theory of brain mechanics. Spartan Books, Washington (1962)
5. OpenAI: Introducing ChatGPT. <https://chat.openai.com>
6. Weizenbaum, J.: ELIZA—a computer program for the study of natural language communication between man and machine. Commun. ACM **9**, 36–45 (1966)
7. Reeves, B., Nass, C.: The Media Equation. Cambridge University Press, Cambridge (1996)
8. Generative AI statistics by industry, sector, revenue and facts (2023), <https://www.enterprisecappstoday.com/news/generative-ai-statistics.html>. Last accessed 27 June 2023

9. Generative AI Market Size, Trends, and Statistics: 2023–2025 (2023), <https://explodingtopics.com/blog/generative-ai-market>. Last accessed 27 June 2023
10. Tsuji, X.: Standardization of the Five-factor personality questionnaire. In: XXVI International Congress of Psychology. Montreal, Canada (1996)
11. Tsuji, H., Fujishima, Y., Tsuji, H., Natsuno, Y., Mukoyama, Y., Yamada, N.: Five-factor model of personality: Concept, structure, and measurement of personality traits [in Japanese]. *Jap. Psychol. Rev.* **40**(2), 239–259 (1997)
12. Khromov, A.: Five-Factor Personality Questionnaire: Educational and methodical manual. Kurgan State University Publishing House, Kurgan (2000)
13. McCrae, R., John, O.: An introduction to the five-factor model and its applications. *J. Pers.* **60**, 175–215 (1992)
14. Rubinstein, S.: *Bytie i soznanie [Being and Consciousness]*. Peter, St. Petersburg (2003)
15. Healtherton, T., Weinberger, J. (eds.) *Can Personality Change?* APA, Washington (1994)
16. Churaeva, N.: Diagnosis of motivation for individuals to join virtual communities. *Bull. Univ. (State Univ. Manage.)* **6**, 126–127 (2009)
17. Nass, C., Moon, Y., Morkes, J., Kim, E., Fogg, B.: Computers are social actors: a review of current research. In: B. Friedman (ed.) *Human Values and the Design of Computer Technology*. CSLI Publications, Cambridge (1997)
18. Nass, C., Steuer, J., Henriksen, L., Dryer, D.: Machines, social attributions, and ethopoeia: performance assessments of computers subsequent to "self-" or "other-" evaluations. *Int. J. Hum. Comp. Stud.* **40**, 543–559 (1994). <https://doi.org/10.1006/ijhc.1994.1025>
19. Nass, C., Fogg, B., Moon, Y.: Can computers be teammates? *Int. J. Hum. Comp. Stud.* **45**(6), 669–678 (1996). <https://doi.org/10.1006/ijhc.1996.0073>
20. Dryer, C.: Getting personal with computers: how to design personalities for agents. *Appl. Artif. Intell.* **13**, 273–295 (1999). <https://doi.org/10.1080/088395199117423>
21. Nass, C., Moon, Y., Green, N.: Are machines gender neutral? Gender-stereotypic responses to computers with voices. *J. Appl. Soc. Psychol.* **27**, 864–876 (1997). <https://doi.org/10.1111/J.1559-1816.1997.TB00275.X>
22. Aronson, E., Willerman, B., Floyd, J.: The effect of a pratfall on increasing interpersonal attractiveness. *Psychon. Sci.* **4**, 227–228 (1966)



A Two-Stream Self-attention Multi-digraph Model for Chinese NER

Xuyao Dai¹(✉), Tingting Liu², Zhen Liu¹, and Yanjie Chai¹

¹ Faculty of Information Science and Technology, Ningbo University, Ningbo 315211, China
das7575@163.com

² College of Science and Technology, Ningbo University, Cixi 315300, China

Abstract. As an important international language, Chinese has unique characteristics in many aspects, and Chinese Named Entity Recognition (CNER) is receiving increasing attention. However, traditional CNER methods face the challenge of balancing semantic disambiguation and the effective extraction of text sequence features. To tackle this issue, we propose a dynamic weighted fusion model based on a two-stream selfattention mechanism and multi-digraphs (DW-TSM). In this method, we first construct a graph structure by combining Chinese corpora with various entity dictionaries. We then embed the text that forms the graph using a gated graph neural network (GGNN). Subsequently, we input the shallow features obtained from the embedded text into an encoder containing a two-stream self-attention mechanism for feature extraction. We further fuse the output of each layer in the encoder through dynamic weighting and dimensionality reduction to enhance the model's semantic recognition ability. Finally, we input the encoded feature representation into a standard BiLSTM-CRF layer to predict the globally optimal label sequence by considering the adjacent relationships between labels. Experimental results show that the proposed DW-TSM model outperforms traditional sequence labeling models and neural network models with an average F1 score improvement of approximately 1.7% on the People's Daily, Weibo NER, and MSRA datasets, validating the superior performance of the proposed model in achieving effective named entity recognition in Chinese tasks.

Keywords: CNER · Two-stream self-attention · Multi-digraph · Dynamic weighted fusion

1 Introduction

Named Entity Recognition (NER) is a crucial task in information extraction and plays a vital role in various natural language processing (NLP) applications [1], including information retrieval, automatic text summarization, question answering, machine translation, and knowledge graphs. The objective of NER is to identify predefined specific entities within sentences and classify them into their correct types, such as person, location, or organization.

Historically, NER methods have been divided into two categories: rule-based and statistical-based. Rule-based methods rely on manually designed rules specific to a particular field to match named entities. However, these methods are labor-intensive and

lack generalizability to other fields. In contrast, statistical based methods treat NER as a sequence labeling task and utilize artificially labeled corpora for training. As the cost of labeling is significantly lower than that of designing rules, statistical-based methods are more versatile and do not require extensive hand-designed rules.

In recent years, deep learning has emerged as a powerful approach for learning feature representations directly from data and has led to significant breakthroughs in the NLP field. When applied to NER, deep learning can learn complex hidden representations without the need for intricate feature engineering or extensive domain knowledge. As a result, deep learning-based methods have surpassed traditional rule-based and statistical-based approaches in NER performance. Due to linguistic differences, there has been extensive research on NER methods for specific languages. Several notable NER research have been published, covering a wide range of languages such as English, Arabic, and Hindi. However, these surveys primarily focus on English NER (ENER) [2], compared to ENER, CNER faces higher ambiguity due to the fact that Chinese sentences are not composed of independent words like English sentences. In recent years, many studies have attempted to improve Chinese named entity recognition performance by leveraging dictionaries. Research by Ratinov et al. [3] has shown that NER is a knowledge-intensive task, and the role of entity dictionaries is crucial in downstream tasks, often incorporated into NER systems in the form of background knowledge [4]. Previous studies have demonstrated that using dictionaries can improve NER performance. On the one hand, using entity dictionaries reduces the need for manual annotation of data and workload and can handle rare or even new entity identification cases [5]. On the other hand, there are abundant existing dictionary resources, and many dictionaries have been handcrafted by previous research [6]. Moreover, in the current era of big data, entity dictionaries can be easily established from knowledge bases or commercial data resources in various fields. However, while entity dictionaries can improve NER performance to some extent, the accuracy of the dictionary has a significant impact on NER results. If the entity dictionary contains irrelevant or incorrect information to the target scene, it will reduce NER performance [7], leading to significant bias in the results. Given the inherent ambiguity of Chinese, this issue is even more pronounced in CNER.

To effectively reduce the impact of entity dictionary errors, existing methods typically rely on manually crafted templates or predefined selection strategies. Qi et al. [8] defined multiple n-gram templates for specific tasks, constructed features for each character based on the dictionary and context. Meanwhile, literature [9] proposed a selection strategy based on maximizing the total number of matching tags in a sentence, and literature [10] proposed a selection strategy based on the maximum matching rule. Although these methods have some effectiveness, they cannot effectively utilize contextual information and still have some errors. Against this background, literature [11] proposed a multi-directed graph model to learn how to combine toponymic information and solve the conflict matching problem in learning. This method constructs a graph of the corpus text based on the dictionary, with each character as a unit, and constructs different nodes based on each category of each dictionary. The standard NER is then completed. Although this method has relatively good performance in Chinese semantic disambiguation, the use of static representation methods in the text embedding stage results in shallow features being obtained by the entire model.

Static pre-training techniques, as traditional text representation methods, utilize word vectors to learn word representations. However, the performance of such methods in semantic understanding is very limited. In recent years, the emergence of dynamic pre-training language models such as GPT [12] and BERT [13] has greatly promoted the development of natural language understanding. The appearance of BERT has opened a new era, and on this basis, researchers have proposed various improved models based on BERT by improving generation tasks, introducing knowledge, introducing multi-task learning, improving the masking method, or improving the training method, etc. However, regardless of the improvements, models based on BERT are self-encoding models, and in the pre-training process, they follow the independence assumption. In the masked prediction task, specific word entities like "Ningbo" have related characters "Ning" and "Bo", which will be ignored by the independence assumption. Moreover, the masked task is only performed in the pre-training stage, leading to some bias in upstream and downstream tasks. This means that the results of self-encoding models cannot achieve ideal results. The proposed self-regressive model XLNet [14] effectively addresses the independence assumption problem of BERT. This model comprehensively uses permutation language models, twostream self-attention, and cyclic mechanisms. While being self-regressive, it obtains bidirectional target text context information and incorporates the idea of RNN to combine the hidden layer representation of the previous layer in sequence prediction, allowing the model to obtain longer distance context information.

In this work, our main goal is to find a more effective and high-performance model suitable for CNER. The main contributions of this work are summarized as follows:

- (1) The first principle in designing the model is to reduce the ambiguity of Chinese. To accomplish this, in the data preprocessing stage, this work adopts the multi-digraph method to construct the corpus data into a graphical form, and further enriches and optimizes the relevant entity dictionary. This approach can effectively reduce the error in Chinese word segmentation and can be adapted to any suitable text embedding method.
- (2) In the text embedding stage, this work takes inspiration from the XLNet model and proposes a dynamic weighted fusion attention mechanism to enhance the model's semantic understanding ability. Firstly, this method assigns a weight to the representation generated by each layer's two-stream self-attention mechanism in XLNet and updates it through training. Then, the weighted average of the hidden representations generated by each layer is obtained, followed by dimensionality reduction. Finally, the embedded result is input into the standard BiLSTM-CRF to predict the globally optimal result.

2 Methodology

2.1 Constructing the Architecture of Our DW-TSM

The architecture of the DW-TSM model is shown in Fig. 1, which consists of a word embedding layer, a BiLSTM layer, and a CRF layer.

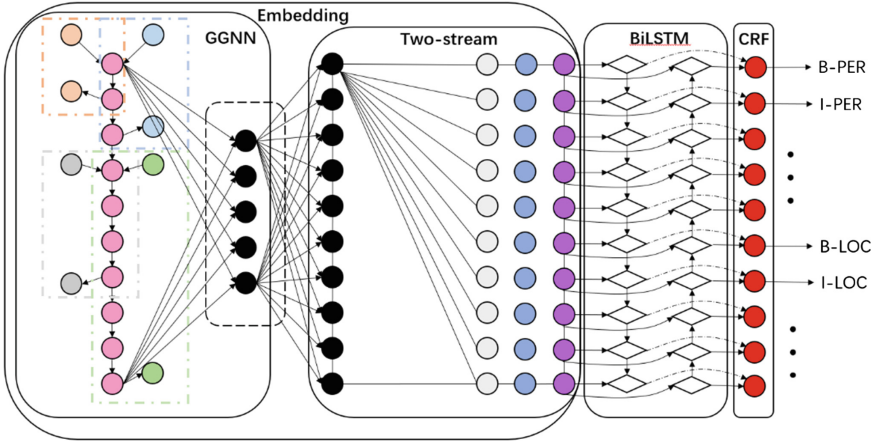


Fig. 1. DW-TSM model architecture.

2.2 Embedding Layer

As shown in Fig. 1, the embedding layer of the DW-TSM model consists of an adaptive GGNN network and a two-stream self-attention layer based on dynamic weight fusion.

Constructing a text multi-digraph: As the first part of the model, to effectively eliminate the ambiguity caused by Chinese text being a continuous sequence, as shown in Fig. 2, traditional maximum matching tag methods may result in erroneous entity recognition to some extent. For example, the person entity “Xiaoming” may be mistakenly recognized as “Xiaoming qu” (Xiaoming goes), and the location entity “Yubei” may be mistakenly recognized as “qu Yubei” (go to Yubei). In this study, the text is modeled as a graph structure at the character level, combined with multiple pre-designed entity dictionaries. First, the Chinese sequence is segmented into individual characters, with each character treated as a separate node, and then each node is connected in sequence. At the same time, the pre-designed entity labels are matched with the text sequence, and this structure can explicitly model the relationship between the target text and the dictionary. The specific process is as follows:

$$G = (V, E, L) \quad (1)$$

$$V = V_c \cup V_s \cup V_e \quad (2)$$

$$L = \{l_c\} \cup \{l_{g_i}\}_{i=1}^m \quad (3)$$

In Eq. (1), the multi-digraph G consists of a node set V composed of characters, an edge set E , and a label set L , where E is associated with L , and each edge in E is assigned a label containing the type of node it connects. In Eq. (2), the node set V consists of a set of nodes V_c containing character representations and a set of start and end nodes V_s, V_e matched to each entity dictionary g . In Eq. (3), the label set consists of a label set l_c

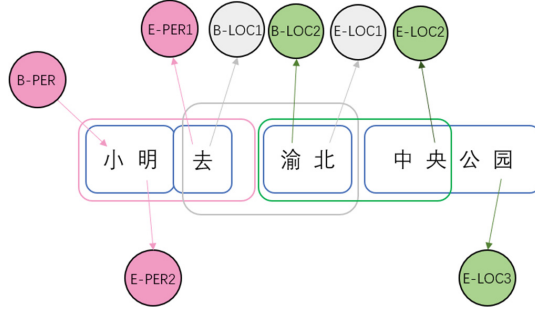


Fig. 2. The process of constructing a multi-digraph.

used to represent the natural sequence of the text and a label set l_{g_i} used to represent the scope of the text, where l_c is assigned to edges connecting adjacent characters, and l_{g_i} is assigned to all edges connected to characters matching entities in the entity dictionary g .

Dynamic two-stream self-attention mechanism: For a graph structure, the idea of GGNN is to generate meaningful output by learning node representations through a neural network with gated recurrent units (GRUs) [15]. Reference [16] shows that in Chinese NER tasks, compared with other graph neural networks, GGNN has better ability to capture local information. Reference [11] proposes an adaptive GGNN that extends the adjacency matrix based on the traditional GGNN [17] to handle multiple directed graph structures, in order to learn a weighted combination of location information suitable for NER tasks. The basic cycle is:

$$h_v^{(0)} = \begin{cases} W^g(v) \\ [W^c(v)^T, W^{bi}(v)^T]^T \end{cases} \quad (4)$$

$$H = [h_1^{(t-1)}, h_2^{(t-1)}, \dots, h_{|V|}^{(t-1)}]^T \quad (5)$$

$$a_v^{(t)} = [(HW_1)^T, (HW_2)^T, \dots, (HW_{|L|})^T] A_v^T + b \quad (6)$$

$$z_v^{(t)} = \sigma(W^z a_v^{(t)} + U^z h_v^{(t-1)}) \quad (7)$$

$$r_v^{(t)} = \sigma(W^r a_v^{(t)} + U^r h_v^{(t-1)}) \quad (8)$$

$$\hat{h}_v^{(t)} = \tanh(W a_v^{(t)} + U(r_v^{(t)} \odot h_v^{(t-1)})) \quad (9)$$

$$h_v^{(t)} = (1 - z_v^{(t)}) \odot h_v^{(t-1)} + z_v^{(t)} \odot \hat{h}_v^{(t)} \quad (10)$$

where, $h_v^{(t)}$ is the initial state of the network, W^g and W^c are lookup tables for characters and nodes. W^{bi} is a dual embedding table proposed in Ref. [18], which has been shown to be effective in NER tasks. A_v is the row vector in the adjacency matrix A that corresponds

to the node v . W and U are the parameters that need to be learned during model training. Equation (5) creates the state matrix H at time step $(t-1)$, and Eq. (6) represents the network propagating information through adjacent nodes. Equations (7), (8), (9), and (10) calculate the new hidden state $h_v^{(t)}$ at the current time step t by combining the current node's hidden state with the neighboring node states. This approach is adopted in this study as the first layer for learning node representations in text embedding.

After obtaining the normal sequence text feature representation $\{h_v^{(T)} | v \in V_c\}$ through GGNN on the text sequence, it is then input into the improved XLNet for NER task fine-tuning. In traditional XLNet, the text sequence is first input into the model in order, and then the permutation mechanism is used to mask the corresponding characters in the sequence to scramble the text order. After that, the two-stream self-attention mechanism is used to obtain semantically enhanced text representation.

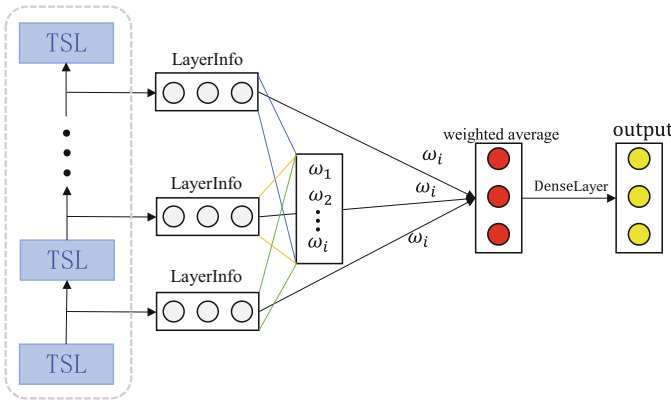


Fig. 3. Dynamic weighted fusion process.

Building on this foundation and inspiration, we defined a set of trainable contribution coefficients during the propagation process of the two-stream selfattention layer in the DW-TSM model. As shown in Fig. 3, each layer's output feature representation was assigned a weight, and then the weights of each layer were used to calculate the weighted sum of the output of each layer, which were used to define the contribution of each layer's two-stream self-attention mechanism to XLNet's text sequence. The hidden representations generated by each layer were then weighted and averaged, and the final text representation was obtained by dimensionality reduction:

$$\omega_i = Dense(LayerInfo_i) \tag{11}$$

$$TS = Dense(\sum \omega_i LayerInfo_i) \tag{12}$$

where ω represents the weight assigned to each two-stream self-attention layer, which is generated and updated through network training. If the weight is larger, the impact of that layer on the final result of the model is greater. $LayerInfo$ represents the output of each two-stream self-attention layer, and TS represents the final result obtained

by applying all two-stream self-attention layers to the feature representation of the input sequence. Finally, the learned character feature representation from the model is input into a standard BiLSTM-CRF, which classifies the entities in the original text and generates output.

3 Experiments and Analysis

3.1 Datasets

The data used in this paper’s experiments comes from the benchmark public datasets MSRA [19] and Weibo-NER [20]. MSRA is the 2006 SIGHAN Named Entity Recognition Corpus, produced by the Chinese Language Processing Group of the Association for Computational Linguistics (ACL). It has been used as raw data in a large number of research papers. The MSRA dataset mainly consists of news data. Weibo-NER is a dataset from the social media field. This paper’s model is applied to both domains, which are more challenging and can better test the effectiveness, universality, and robustness of the model.

3.2 Analysis of Experimental Results

To verify the effectiveness of the Chinese named entity recognition model based on the two-stream self-attention mechanism in multi-directed graphs proposed in this paper, a comparison baseline was established using the Chinese named entity model based on the state-of-the-art language models in recent years in the field of natural language processing. The model was trained on the MSRA and Weibo-NER datasets, and tested using the test sets. The proposed model was compared with other models and the baseline model, and the results are shown in Tables 1 and 2.

Table 1. Results on the Weibo-NER dataset.

Model	Accuracy/%	Recall/%	F1/%
Base	62.01	63.38	62.68
BERT	68.94	64.45	66.62
ALBERT	68.78	63.56	66.21
XLNet	69.12	64.57	67.12
Di_BERT	69.97	65.32	67.58
Di_ALBERT	69.76	64.52	67.11
Our model	70.16	65.98	68.66

On the Weibo-NER dataset, Ding et al. [11] used a NER model based on multi-directed graphs, with an F1 score of 62.68%. By using the popular BERT pre-training language model as the word embedding layer, the F1 score was improved to 66.62%. The

Albert model, which reduces the number of parameters, achieved an F1 score of 66.21%. XLNet uses a traditional permutation language model method, which optimizes the performance of BERT's shortcomings, and achieves an F1 score of 67.12%. For several pre-training language models, dynamic weight fusion optimization was first carried out, in which a trainable contribution coefficient was assigned to each Transformer layer in BERT, and the results of each layer were weighted by this coefficient and then reduced and output. By using semantically enhanced features and pre-trained character embeddings, the F1 score was improved to 67.58%, achieving good performance.

Table 2. Results on the MSRA dataset.

Model	Accuracy/%	Recall/%	F1/%
Base	93.2	92.7	92.9
BERT	96.64	96.45	96.32
ALBERT	95.78	95.56	95.21
XLNet	97.12	96.57	97.17
Di_BERT	97.97	95.32	96.23
Di_ALBERT	96.76	95.52	96.01
Our model	98.16	96.98	97.76

Similarly, after optimization, the Albert model achieved an F1 score of 67.11%. The proposed model in this paper achieved an F1 score 1.08% higher than the highest result in Table 1. On the MSRA dataset, consistent with the testing of the compared models in the MSRA dataset, the F1 score of the baseline model was 92.9%, BERT achieved an F1 score of 96.62%, Albert achieved an F1 score of 95.21%, XLNet achieved an F1 score of 97.21%, the BERT model dynamically fused and optimized achieved an F1 score of 97.58%, and Albert achieved an F1 score of 96.11%. The proposed model in this paper achieved an F1 score 0.59% higher than the highest result in Table 2. Tables 1 and 2 show that compared with existing methods, the proposed model in this paper is more competitive and significantly outperforms other methods, achieving the best results in this standard benchmark test. Therefore, this method has been proven to be highly effective, effectively eliminating ambiguity in Chinese word segmentation with the assistance of external dictionaries, and using a two-stream self-attention mechanism based on dynamic weighted fusion to mine potential semantic information in the text, improving the accuracy of the model for Chinese named entity recognition. This work also conducted some ablation experiments on the performance of the model, including the effect of word vector dimensionality on the F1 score of the model and the effect of different feature extraction layers on the F1 score of the model, as shown in Figs. 4 and 5.

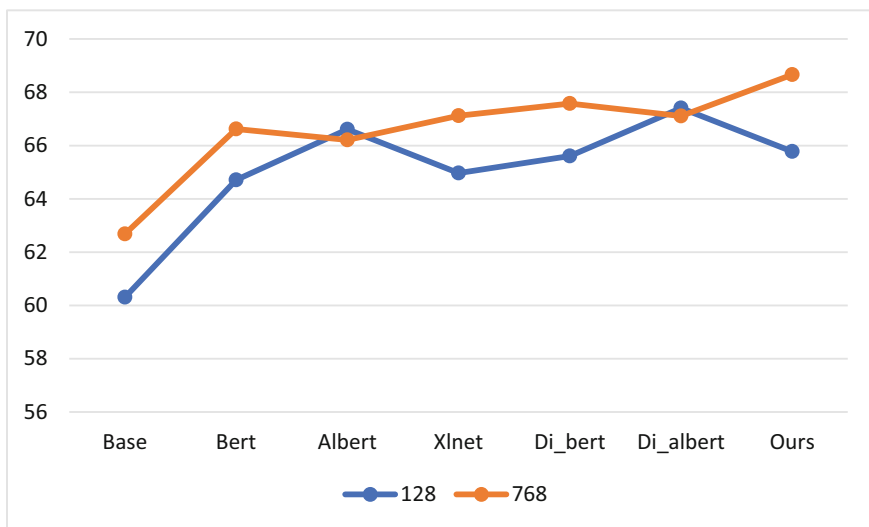


Fig. 4. The effect of word vector dimensionality on the F1 score of the model.

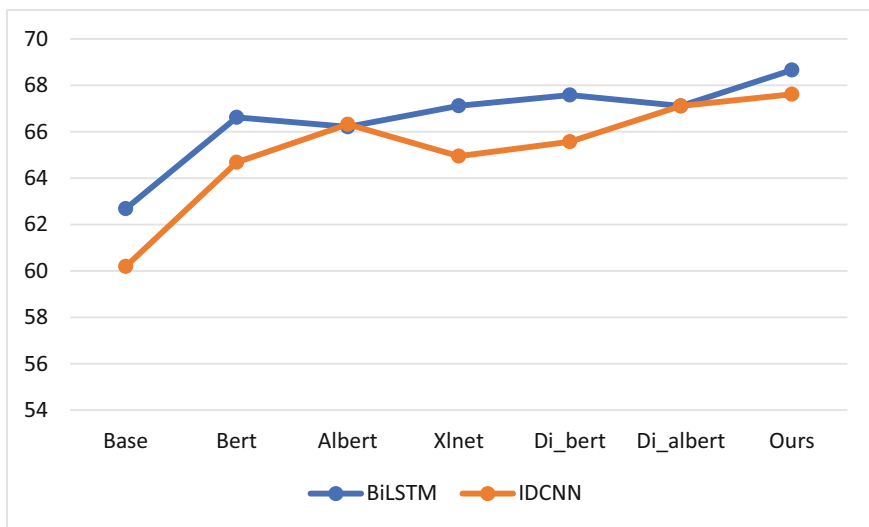


Fig. 5. The effect of different feature extraction layers on the F1 score of the model.

Through the various fine-tuning experiments on the MSRA dataset, it was found that the performance of the proposed model varies under different parameter conditions. As shown in Fig. 4, most models perform well when the word vector dimension is 768, and the performance of the proposed model even decreases when the dimension is 128. Therefore, the word vector dimension has a significant impact on the performance of the model. As shown in Fig. 5, replacing the feature extraction layer with the IDCNN

model aims to increase the model's receptive field and obtain a wider input matrix data. At the same time, parallel computing can be used to improve the model's computational efficiency. Among the models' performance, it can be seen that the BiLSTM has slightly higher feature extraction ability than the IDCNN. Therefore, the proposed model uses a standard BiLSTM as the feature extraction layer to ensure model robustness.

4 Conclusion

This paper proposes a CNER model DW-TSM based on two-stream self-attention multi-directed graphs, which is tailored for Chinese word segmentation semantic disambiguation and feature extraction. The core idea of the model is to first use the improved GGNN model to perform Word Embedding on the text information of the graph structure. Then, using the high-performance feature representation ability of the two-stream self-attention, the model further learns the word dependency and word order relationships within the text sequence in a global scope by adding trainable contribution coefficients in the attention layer for independent and different contextual contexts, enhancing the model's semantic representation ability. The results are then input into a standard BiLSTM-CRF to capture local features and perform classification to obtain accurate entity categories. Experimental results show that the proposed model has good performance on Chinese short text classification tasks.

Although the proposed model achieves certain results, there are still some obvious shortcomings that need to be addressed in future work. For example, when the corpus data has significant noise, the model's performance can fluctuate, and the model's large parameter size leads to high computational costs. Therefore, the next step will be to consider these limitations and conduct more fine-grained research on the model.

References

1. Diefenbach, D., Lopez, V., Singh, K., Maret, P.: Core techniques of question answering systems over knowledge bases: a survey. *Knowl. Inf. Syst.* **55**, 529–569 (2018)
2. Sun, T., Shao, Y., Li, X., Liu, P., Yan, H., Qiu, X., Huang, X.: Learning sparse sharing architectures for multiple tasks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34(05), pp. 8936–8943 (2020)
3. Ratnoff, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pp. 147–155 (2009)
4. Seyler, D., Dembelova, T., Del Corro, L., Hoffart, J., Weikum, G.: A study of the importance of external knowledge in the named entity recognition task. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 241–246 (2018)
5. Wang, Q., Zhou, Y., Ruan, T., Gao, D., Xia, Y., He, P.: Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. *J. Biomed. Inform.* **92**, 103133 (2019)
6. Zamin, N., Oxley, A.: Building a corpus-derived gazetteer for named entity recognition. In: *Software Engineering and Computer Systems: Second International Conference, ICSECS 2011, Kuantan, Pahang, Malaysia, 27–29 June 2011, Proceedings, Part II 2*, pp. 73–80. Springer, Berlin (2011)

7. Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* **4**, 357–370 (2016)
8. Zhang, Q., Liu, X., Fu, J.: Neural networks incorporating dictionaries for Chinese word segmentation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32(1) (2018)
9. Shang, J., Liu, L., Gu, X., Ren, X., Ren, T., Han, J.: Learning named entity tagger using domain-specific dictionary. In: *Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 2054–2064. Association for Computational Linguistics (2018)
10. Sassano, M.: Deterministic word segmentation using maximum matching with fully lexicalized rules. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 2: Short Papers, pp. 79–83 (2014)
11. Ding, R., Xie, P., Zhang, X., Lu, W., Li, L., Si, L.: A neural multi-digraph model for Chinese NER with gazetteers. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1462–1467 (2019)
12. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
13. Kenton, J. D. M. W. C., Toutanova, L. K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT*, pp. 4171–4186 (2019)
14. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., Le, Q. V.: Xlnet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inform. Process. Syst.* **32** (2019)
15. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724 (2014)
16. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: *International Conference on Learning Representations* (2016)
17. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270 (2016)
18. Chen, X., Qiu, X., Zhu, C., Liu, P., Huang, X.J.: Long short-term memory neural networks for chinese word segmentation. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1197–1206 (2015)
19. Levow, G. A.: The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In: *Proceedings of the Fifth SIGHAN workshop on Chinese language processing*, pp. 108–117 (2006)
20. Peng, N., Dredze, M.: Improving named entity recognition for Chinese social media with word segmentation representation learning. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 2: Short Papers, pp. 149–155 (2016)



How Language Could Have Evolved

Ken Del Signore^(✉)

North Aurora, IL 60542, USA

kendelsignore@gmail.com

Abstract. This paper begins to develop a biologically inspired computational model of the Human language faculty and some associated thought processes. This model is developed starting from a simple proto-language, which humans are assumed to have inherited at speciation. This proto-language consists of single symbol exchange using a small set of symbols; similar to the observed gestural communication systems in the existing Great Ape families. Computationally, the model is built using a single class with the form of a Markov graph node. Instances of this node class are used to symbolically represent words. The model is built iteratively in `main()` as a single graph. Nodes are added to this graph using a merge, or conjunctive join, operation between any two existing nodes, notionally labeled as head and copy. A simple first order graph is developed which is hypothesized to be common to all Mammals and to generate shared mammal behaviors. This graph is then extended to allow for more complex human language and thought processes.

Keyword: great leap theory of mind faculty of language

1 Introduction

We take the simplest possible communication to be a single symbol, exchanged from one communicator to another. As an example, consider a hiker that is lost in the woods and builds a pile of rocks, and then moves on. If a second hiker subsequently finds the rocks, then we can say that a single symbol has been exchanged. The information conveyed is the same as if the first hiker had just stood next to the second and said “here”, except that the hiker said “here” some long time prior and the symbol (the rocks) held the information through time.

To modify this example, if while the hiker is piling the rocks, s/he hears the second hiker coming straight on and calls out “here”, then the second hiker will have, at the simplest, one additional quanta of information, namely a measured value of the distance. This second quanta is analogous to a floating point variable that can take on a continuous value.

Single symbol dialogue systems are observed throughout the animal and plant kingdoms. The symbol “here” is often the exchanged symbol in these systems. A flower can be interpreted as a single symbol, exchanged between a plant and its pollinators, with the meaning of the flower being “here” (Chomsky 2015).

A symbol is defined to contain two quanta of information: label and value; with the value being possibly zero or unspecified. In the examples above, the second communicator measures the value locally to itself using the externalization made by the first communicator.

The neocortex gives mammals the ability to store and recall sequences of symbols with relative ease. This ability gives mammals considerably more complex behavior relative to non mammals. Mammals can input and store sequences of symbols in combinations never experienced before and later recall and utilize this information; an example of which would be the second hiker remembering the path out of the woods and walking out with the first hiker.

All known mammal dialogue, excluding human, uses or can be easily reduced to single symbol exchange. The great apes have possibly the most developed system; using a gestural vocabulary of approximately 80 gestures to convey ~15 unique meanings as commands and questions (Bryne 2017). The meanings loosely correspond to the hypernym forms of various parts of speech categories [here, no, give, on, on?, play?], which are discussed further below.

Among the Hominins, stone tools provide the first evidence for advancement in behavior. The initial Mode I tools are currently dated to 3.3 Mya and remained at a relatively fixed level of design and refinement for over a million and a half years. The early hominins did not evidently undergo much generational change; contrary to the current human cliché “kids these days . . .”. Mode II tools were then developed and these spread slowly throughout the existing hominin range over the next million years. When humans first speciated, they inherited a Mode II toolkit. They had animal hides, cord, knots, hafted hand-axes, spears, and cooked meals, to name a few of their initial conveniences. Hominins had been hunting elephants and hippopotamuses since at least 400 Kya and early human sites dated 200 Kya also contain evidence that they subsisted on these animals.

Humans then began making rapid advances to their toolkits (Henshilwood 2018) and then left Africa approximately 75 Kya. The sudden change to the rate of change of the toolkit just prior to leaving Africa is suggestive that, at the simplest, a single change could have taken place in the hominins to allow them to make these advances. Our current human language (faculty) is argued to be this change (Bolhuis 2014). The use of complex sentences would presumably have allowed the hominins to more easily accumulate knowledge and transfer it to each other and their children. Daily storage and recall of unique sequences would also permit hominins the ability to mentally reconstruct scenarios after they occurred, which would allow them to explain them to others and explore possible solutions when time permitted.

All available evidence indicates that the current human language faculty and cognitive functionality was completely formed before humans left Africa and that it hasn't changed since (Bolhuis 2014); which would be a signature of the single change in Africa hypothesis. The argument for this is that babies from any culture can grow up in any new culture and will readily acquire the new culture and language, which is taken to imply that no changes to the human language faculty or other cognitive functions have occurred in humans since we left Africa.

It is also worth noting that not all humans in Africa obtained the new toolkit. Human sites dated as recent as 30 Kya have been found that do not show evidence of advancement beyond that of the inherited toolkit (Scerri et al. 2021).

For the present inquiry, the two main historical developments of interest are the mammalian neocortex and the human toolkit change. The neocortex is viewed as having endowed mammals with the ability to conceptualize symbols, to form vocabularies of these symbols, and to input, store, recall, and utilize random sequences of these symbols. The human toolkit change is viewed as happening when the neocortex became large enough to support a new thought process and/or a new thought process was developed.

Sections 2–12 cover the derivation and evolution of the graph model in an approximate temporal order that it is posited to have developed in. Section 13 contains a discussion and comparison with similar work.

2 The Interface

We begin with one of the main minimalist assumptions of Linguistics, namely that communicators have some common internal neurological/symbolic representation of each word in their vocabulary. This is shown as the two blue “interface parcels” in the diagram in Fig. 1, which borrows from similar diagrams in (Pinker 1994; Berwick 2013). In this example, while walking out of the woods, the two hikers roust a duck, which causes the duck symbol to activate in each hiker’s interface parcel. We assume that each hiker processes their unique neural input of the event and each conceptualizes, or activates, an internal symbol that corresponds to “duck”. Also note that if one of the hikers is replaced by another mammal, such as a dog, the interface parcel representation would still be valid.

For our initial purposes we assume that some analogous neural parcel exists that contains many neural attractor states that represent words or symbols. Such a neural parcel can be thought of as analogous to a two dimensional optical character recognition neural network, wherein each character is a unique attractor state of the OCR neural network. Each character corresponds to a subset of nodes in the 2d array that fire and lock into the active state when the array is input a noisy bitmap of a scanned character and allowed to run freely into the nearest attractor state.

Such a subset of nodes is analogous to a Cell Assembly (CA). Here we assume each CA holds the floating point value as the spiking rate and the sign as the phase of the spikes (Huyck 2013).

The symbols in this interface parcel, once formed, are assumed capable of subsequent re-activation on similar input. Furthermore, once activated, we assume that the symbols retain this state information for a short time and are more easily re-activated (Hubel 1980).

The interface parcel can be abstracted to represent all of the symbols that we are physically able to internalize and/or externalize. The temporal sequence of all such symbols can be thought of as our stream of consciousness.

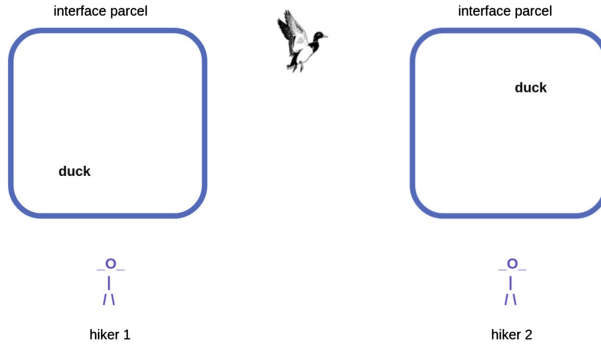


Fig. 1. The roosting of the duck provides unique sensory input to each hiker’s interface parcel, which we assume causes the “duck” symbol to fire (or conceptualize) therein. This is posited to occur as some subset of the nodes in the interface parcel firing into a stable attractor state.

3 Short Term Memory, (STM) Sequences and Recall

Behaviors involving short term memory are easily observable in mammals and many other animals. In humans, we can easily form stm associations between any two randomly picked symbols in our vocabularies.

As an example, we can extend the duck scenario above such that the surprise of the duck roosting causes the first hiker to sneeze. Following this, the second hiker’s interface parcel would contain the activated symbols “duck” and “sneeze” and these would be stm bound such that if some short while later a second duck was rousted, then this would cause the second hiker to recall the sneeze symbol and to expect the first hiker to sneeze again. We can say that the second duck caused the hiker to “think of” the sneeze symbol (Fig. 2).

Such a random two symbol stm mechanism can be built by the current model using the assumed node functionality. We (hypothetically) introduce many additional nodes to the interface parcel, referred to as stm nodes. These stm nodes are assumed to be equivalent to the symbol nodes except that they are unlabeled. The stm nodes are assumed to be randomly and sparsely connected to the labeled symbol nodes (Hawkins 2005).

When two random symbol nodes, such as duck and sneeze, are activated and fire, they each provide input activation to their respective stm nodes. If a subset of these stm nodes is common to both symbols, this subset can become activated over background due to having $2\times$ more input activation than the stm nodes that are not common. The elevated input level is then assumed to persist for some time interval, allowing for subsequent short term associative recall.

Computationally, stm memory can be implemented as a single node created by a merge function between a root node “ip” (interface parcel) and its child nodes, as shown in Fig. 3. The (stm) nodes are created by the merge function that runs each time a node fires. Bidirectional connectivity is assumed possible in all connections.

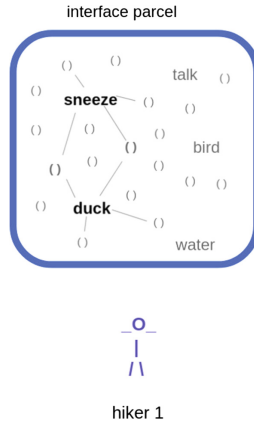


Fig. 2. Many sparsely connected (stm) nodes are capable (under appropriate conditions) of forming a short term association between any two symbols in the parcel.

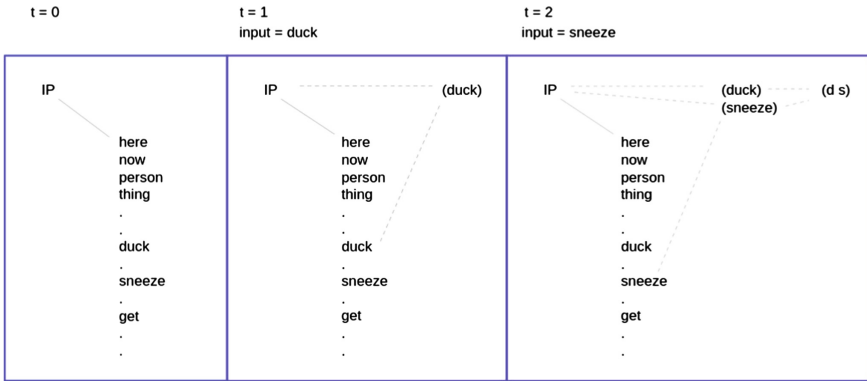


Fig. 3. The child nodes fire based on external input and then touch and fire the ip head node. The ip node performs the merge between itself and the node that touched it. This creates the (duck) and (sneeze) nodes. The (sneeze) node fires after it is created and it performs a merge with itself and its next youngest sibling to create the (d s) node.

The merge operation can be applied recursively between other recent stm nodes to create the (d s) node in Fig. 3, top right. This node then stores the short term association between duck and sneeze.

The stm nodes are created at run time using a merge constructor function of the Node class as shown in Fig. 4. This function takes two nodes as input, labeled “head” and “copy”. Bidirectional links are set up between the head and copy nodes in the merge constructor function.

The sequence: “hiker rocks duck sneeze” is input with the “touch” function calls in main() as shown in Fig. 5.


```

class Node {
  string symbol
  float value // [-1, 0, 1] = F, ?, T
  int input_level // current input value
  int threshold // threshold to fire

  pair< Node*, int weight > head
  pair< Node*, int weight > copy
  list < pair< Node*, int weight > > branches

  Node() // base constructor
  Node( string, head ) // string constructor
  Node( head, copy ) // merge constructor

  touch ( weight, depth )
  output ( tabs, depth )
};

```

```

main() {
  Node ip( "ip", null )
  Node head( "head", ip )
  Node copy( "copy", ip )

  Node child( head, copy )
};

```

Fig. 4. The Node class and the use of the constructor functions in main().

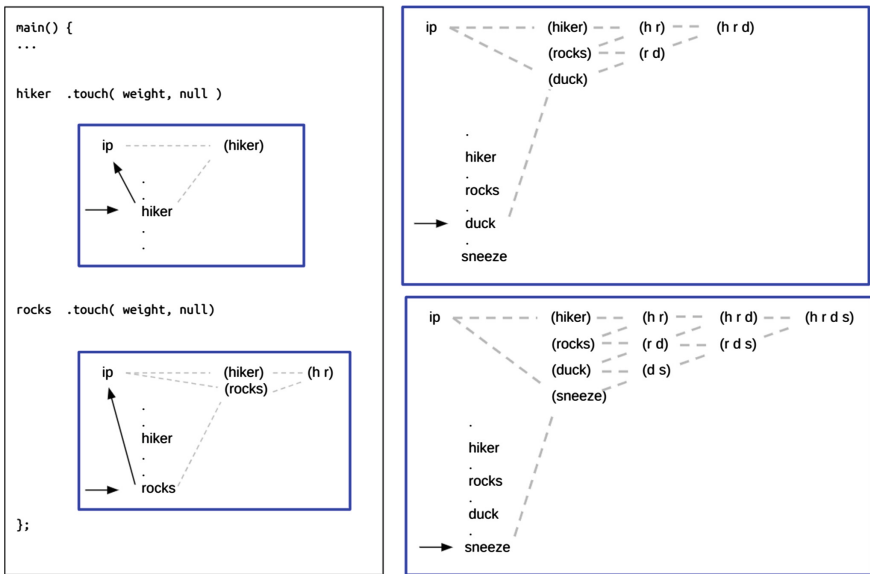


Fig. 5. The hiker.touch() function call causes the hiker node to fire, which then touches and fires the ip node. The stm node is created by a merge between the ip node and the node that touched it. The stm structure then grows iteratively as more symbols are input.

The structure formed in (stm) memory can then be used for output of the hiker, rock, duck, sneeze symbols. Using the recursive algorithm shown in the pseudo code in Fig. 6, the symbols can be output in the order that they occurred.

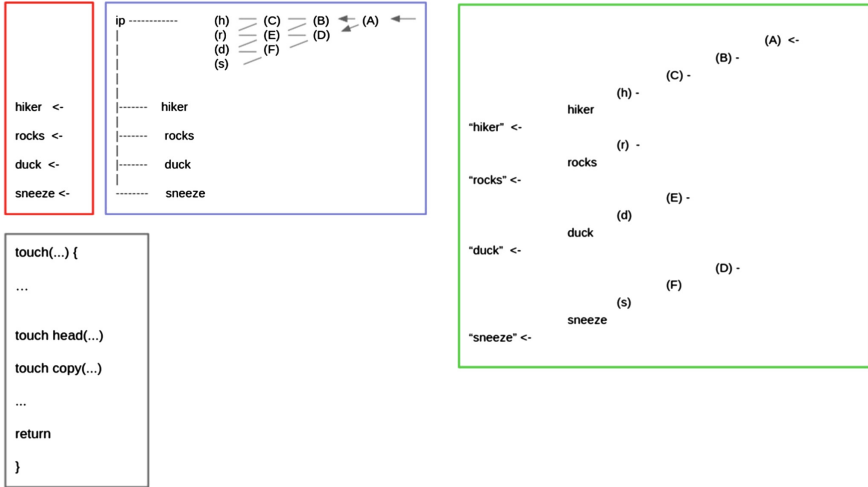


Fig. 6. A c++ pointer to the rightmost (stm) node, labeled “(A)” above, can be returned to main() and used to output the stored node sequence using the touch() function.

4 Protolanguage, Mammalian Single Symbol Exchange

Following the suggestion of Chomsky (2015), we assume a simple (truncated English) corpus of the form: [here now me thing get do go]. This corpus is drawn from the hypernym forms of the parts of speech: [adverb noun verb].

This initial corpus is similar to the reported meanings exchanged in chimpanzee gestural dialogues. The gesture meanings given in Table 1 are derived from a large video corpus of wild chimpanzee single symbol gestural exchanges (Hobaiter and Byrne 2014). The meanings are mapped to a part of speech and then to a hypernym word for that POS.

Switching back to humans, our understanding of the neocortex is expanding at a great rate using many types of experimental methods. fMRI experiments in (Epstein and Kanwisher 1998) and (Huth 2012) have identified two voxels (< 2mm³) that fire in response to words that are hyponyms of ~(person/place) and (thing). Movies are shown to volunteers and the hypernym mappings from WordNet are used to tag 1800 nouns from the movie dialogues to person, place, or thing. In all volunteers, these two voxels can be identified in similar locations on a neocortex flatmap and show activation when the corresponding hyponym words are used in the movie dialogues.

The symbols of the proto-language are assumed to be formed as child nodes in the interface parcel as shown in Figs. 7, 8 and 9. The stm memory allows for storage of state information and for simple dialogues.

Table 1. The chimpanzees exchange approximately 15 unique meanings using gestures. The hypernym forms are mapped to the closest meaning and can all be nominally matched. No grammar or random combinatoric use of gestures is observed. All species of Great Apes share a common set of approximately 100 gestures with each species using a subset of ~60 gestures, however, the gesture to meaning mapping is different in every species.

Gesture	Meaning	Part of Speech	Hypernym form
grab,	stop that	negative	“no”
mouth stroke,	acquire object	verb - get	“give”
bite,	contact (affection)	verb - feel - ?	touching
big loud scratch,	init grooming	verb	“do”
arm swing,	move away	verb - go	“go”
beckon,	move closer	verb - come	come
big loud scratch,	travel with me	prep	with
jump,	follow me (sex)	prep	with
foot present,	climb on me	prep	on
reach	climb on you?	prep - ?	on?
present location	groom here	adverb	here
leaf clipping, punch ground	sexual attention - male	?	flirt
leaf clipping	sexual attention - female	?	flirt

5 Movement, Boys Eat What? What Boys Eat?

A mechanism for movement can be implemented by using the stored values in the (stm) nodes and modifying the pseudocode as shown in Fig. 10. The input value of zero is propagated to the (stm) nodes as shown and then can be used to causes movement in the subsequent output order of the stored symbols.

This scenario is implemented in the c++ prototype as shown in Fig. 11. The blocks of text separated by horizontal dashed lines are static printouts of the ip graph (no arrows) or runtime graph flow diagrams (with arrows). The input sequence is “boys eat what”, where “what” = thing:0. Following input, a pointer the the (b e t) node is returned to main() and this is used in the call: (b e t)->touch() to invoke the output: “what boys eat”.

6 Adverb Periodicity, Here I Now Eat Daily

The set of words in a language can be bifurcated and mapped into two symbols “a” or “n”, which correspond to the adverbs/adjectives and the nouns/verbs/prepositions.

The Wordnet English corpus of 120K definitions and 60K glossary sentences can be encoded into such reduced sequences and used as input as shown in Fig. 12. These sequences can be considered a “Truncated English”. The terminal

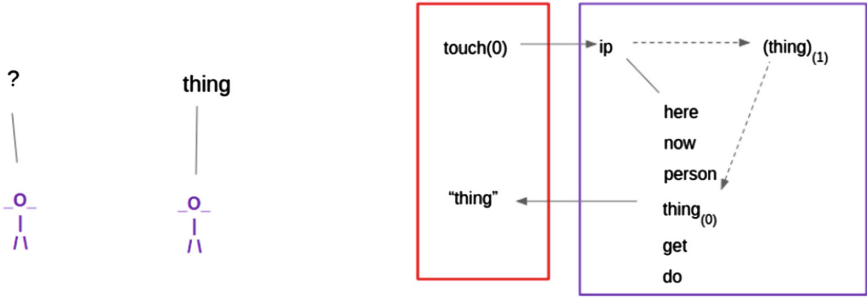


Fig. 7. The value of zero passed in the touch(0) function indicates a question; -1, 0, 1 = [no, ?, yes]. The simplest input form of a question is the function call: ip.touch(0). The stm (thing) node is assumed to have been previously formed.

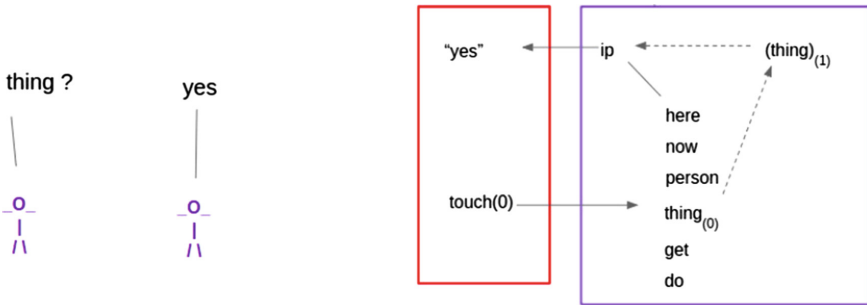


Fig. 8. All symbols can be input as questions with the touch(0) function call as: thing.touch(0).

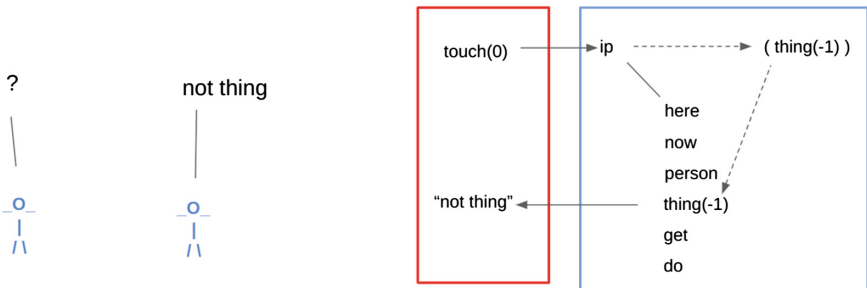


Fig. 9. An stm node having a value = -1 indicates negation. In this scenario, (thing(-1)) was previously created via an stm merge.

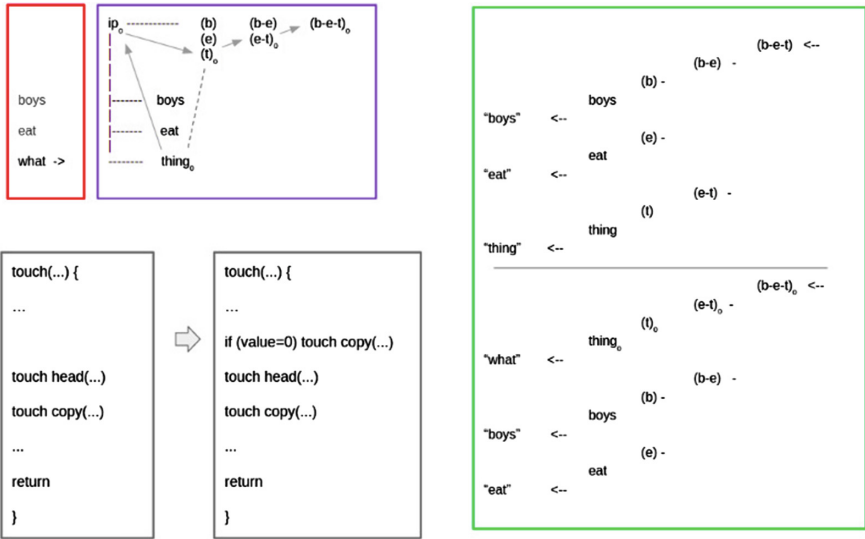


Fig. 10. The single line modification to the pseudocode will alter the output ordering from “boys eat what” to “what boys eat”.

nodes of the ip graph are then bifurcated further to form a “Less Truncated English”. Oscillation of the a-n graph is posited to produce the ubiquitous phenomenon of adverb periodicity in human language: quickly I eat; I quickly eat; I eat quickly.

7 Compare Function at the Merge

here -here -HERE

A compare mechanism exists such that similar sensations, separated in time, can be compared. A familiar example is hearing the first two intermittent sounds of crunching leaves when someone or something is moving in the woods, relative to a fixed observer. The change in intensity is available at the interface parcel as an internal feedback to the observer.

The ability to compare two of the same sensory inputs separated in time is an important evolutionary advantage to all animals. The initial measured information from each sensation must be stored through time and then compared with the second measurement at a later time. This functionality can be achieved at the second level (stm) merge, shown in the adverb branch of Fig. 13, using the values of the two parent nodes.

```

kwd1:code20$ c++ graph1.cpp
kwd1:code20$ ./a.out
-----
IP
      boys 0
      thing 0
      eat 0
-----

      boys:1 <--
IP:1 <--
|
(b)

      eat:1 <--
IP:1 <--
|
(e) (b e)

      thing:0 <--
IP:0 <--
|
(t) (e t) (b e t)
-----

IP      (b) 1      (b e) 1      (b e t) 0
      (e) 1      (e t) 0
      (t) 0

      boys 1
      thing 0
      eat 1
-----
time interval occurs

      (b e t):0 <--
      (e t):0 <--
      (t):0 <--
thing:0 <--
      (b e):1 <--
      (b):1 <--
boys:1 <--
      (e):1 <--
eat:1 <--
-----

```

Fig. 11. c++ prototype output of “what boys eat?” Here “thing:0” is defined to be “what”.

This function, storing a value and using it in a compare operation later in time, is similar to that of the Reichardt model used by Hubel to explain directionally sensitive neural circuits in V1 (Hubel 1980).

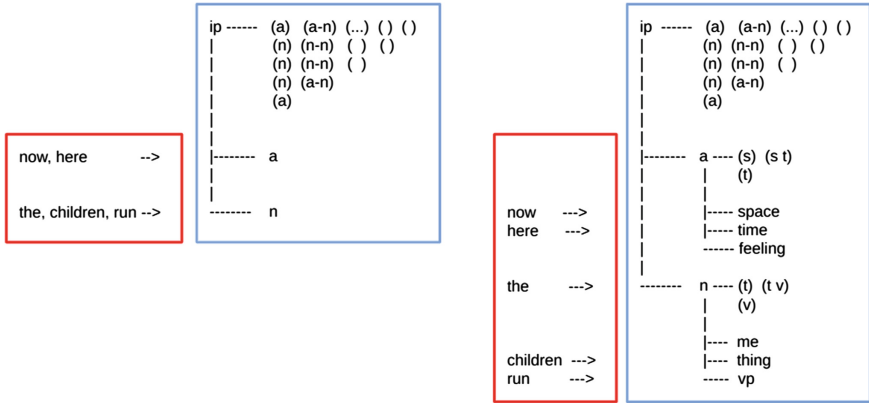


Fig. 12. The ~120K unique words in the WordNet corpus can be mapped to ‘a’ or ‘n’. “Now the children run here” would map to ip (stm) memory as shown. The a and n branches are then bifurcated further as shown. The words in the corpus are mapped to one of the six hypernym forms. The stm memory mechanism functions within each branch of the graph.

8 Conjunctions, Illicit Conjunctions, Movement of Conjunctions

An “and” node is added to the IP graph, with the corresponding (stm) symbol labeled (+), as shown in Fig. 14. As introduced, this node would have no additional properties not already described.

An additional recursive touch() call is then added to the touch() function to touch the (stm) node’s head as each node in an (stm) diagonal layer is added. This is shown by the arrows in Fig. 15. In Fig. 15, after “time” is input, the horizontal sequence of nodes that terminates in “||” forms a closed loop, which can be detected by the touch() function, allowing it to return an enhanced return value of 2 as indicated.

The return value of 2 can then be detected in the touch function and used to trigger a Hebbian enhancement of the weights between the calling and called nodes in (stm) memory.

An illicit conjunction sequence can be created with the model by introducing a single logical change to the (stm) nodes that are copy-rooted by the (+) node, namely, that if the return value in the Hebb loop is <2, then return -1. This is shown in Fig. 16, where the (a +) node would detect the non closed loop condition (RETURN=1) and change the return value to -1.

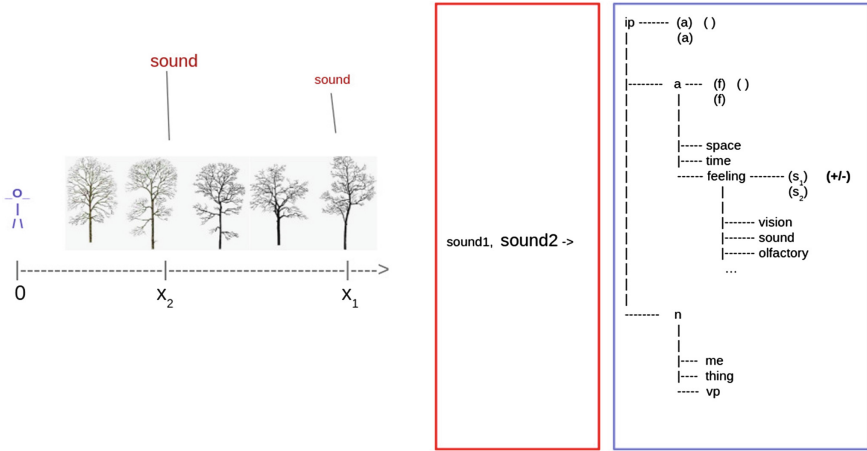


Fig. 13. State information is stored in the adverb branch and can be used in a subsequent compare operation.

The value $=-1$ is retained in the (stm) structure and can be propagated to subsequent (stm) recall. Such a mechanism could be used during sleep to exclude the stored (stm) sequence from normal [ltm] sequence storage.

Movement of a conjunction (when and where boys eat?) is possible by using the enhanced connection weight values stored in the closed loop. In the c++ prototype output shown in Fig. 17, the input sequence: “thing space:0 and time:0” (= boys where and when?) is input to the IP graph. The (n a + a) node is then touched from main() to cause the output sequence corresponding to: “when and where boys?”

The conjunction node functions to create a layer of (stm) nodes that returns a false signal if a closed loop is not detected. Additionally, once a closed loop is detected, the subsequent (stm) nodes in the conjunction layer (example node: (n a +) in Fig. 17) can be disabled by storing a value $=2$ in the (stm) nodes. This process occurs locally, as each (stm) node is touched in the Hebbian loop that runs for each row of (stm) memory.

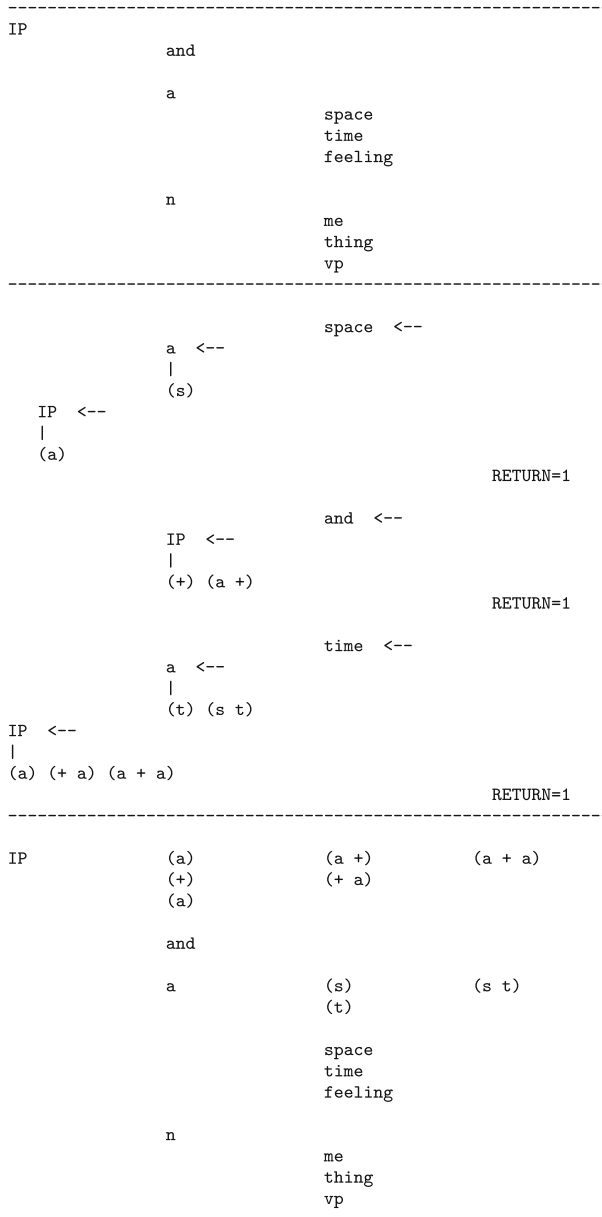


Fig. 14. Inputting “space and time”.

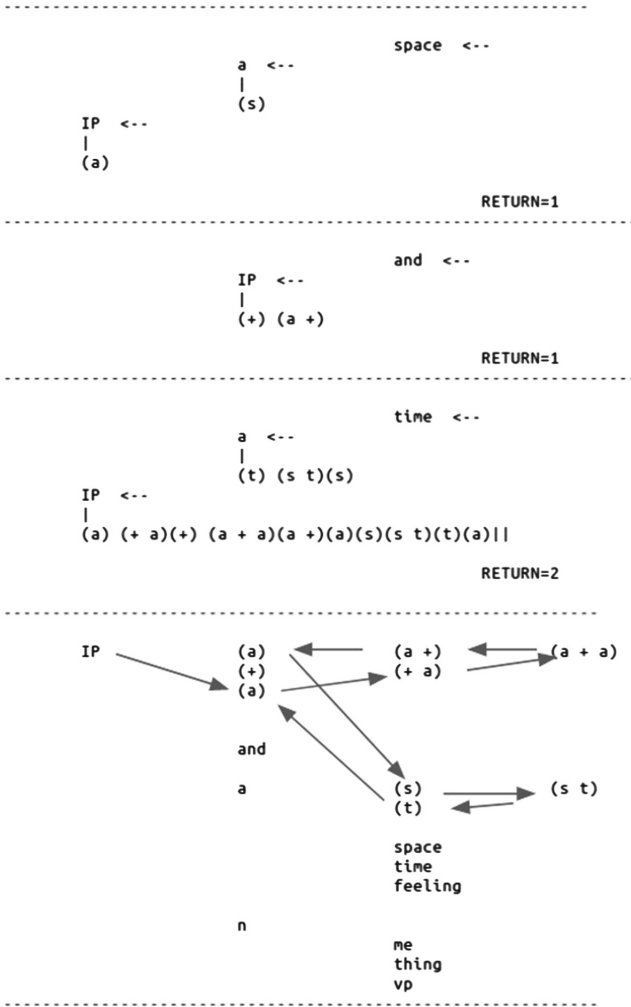


Fig. 15. Closed loop detection and Hebbian one shot binding. The touch() function builds a list of node pointers as it recursively calls itself and can detect the closed loop and return an enhanced return value.

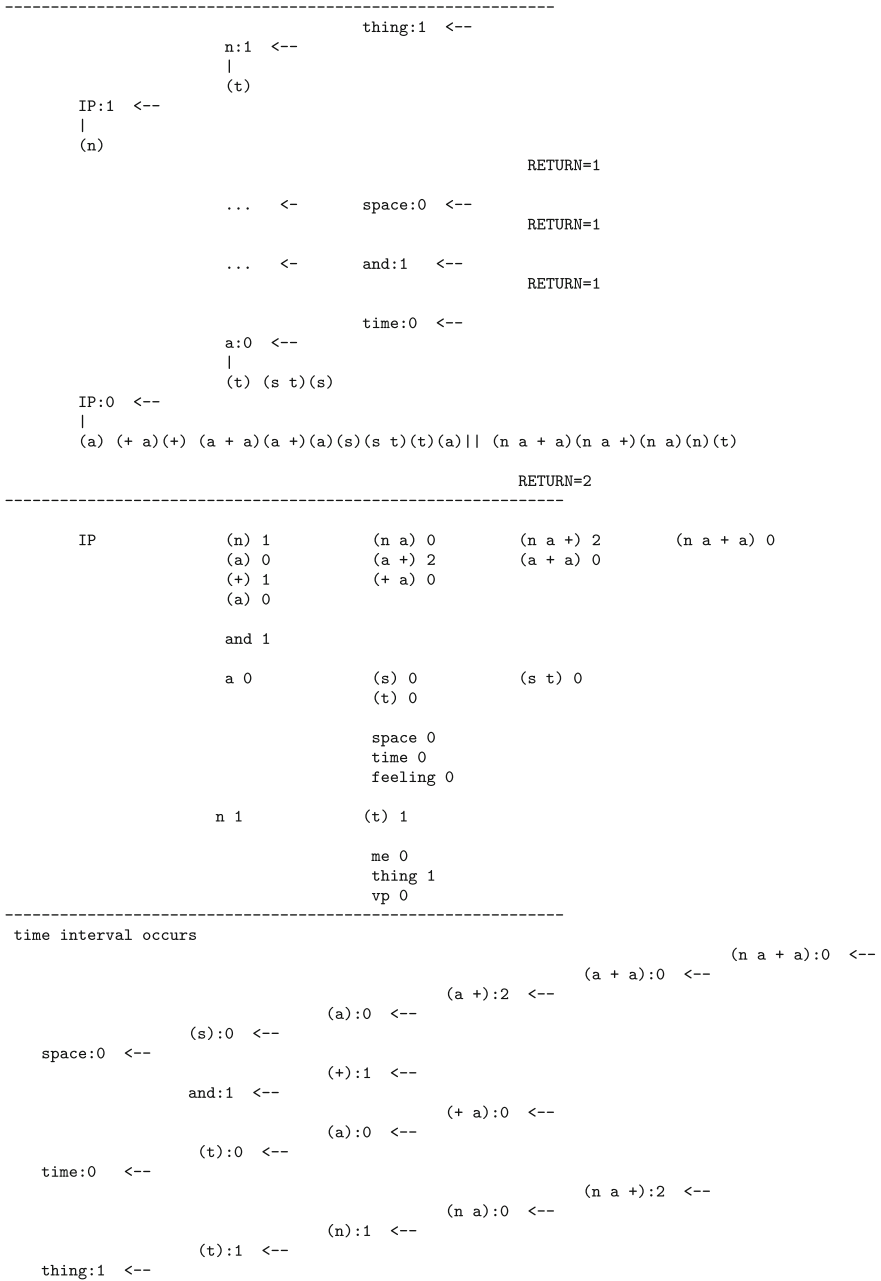


Fig. 17. Movement of a conjunction is accomplished using the modified bindings and the value of zero stored in IP (stm) memory. The truncated English input sentence “thing where and when?” is input and then output as “where and when thing?”, where ‘space:0’ and ‘time:0’ are defined to be where and when.

9 Long Term Memory, Sleep, Prediction

A permanent long term memory node can be made at the time the (stm) node is made by the merge() function, as shown in Fig. 18 using the square bracket notation: [ltm]. The (stm) and [ltm] nodes can also be connected by the merge function. The [ltm] node retains the label of the copy node as shown.

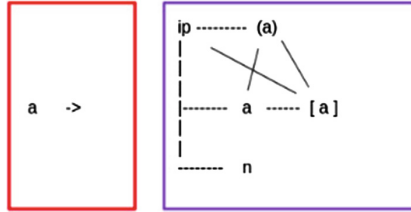


Fig. 18. A long term memory node is formed by a second merge operation with modified parameters.

In the touch() function, the creation of [ltm] nodes by the merge() function can be toggled on and off with a single global variable. Once created, [ltm] nodes can be called by the (stm) nodes at the end of the existing touch() function using the link that was created at merge() time. This input sequence is shown in Fig. 19. The call to the [ltm] node can also be toggled on and off; which would be analogous to inputting into (stm) rapidly without time to process the touch() call to the [ltm] node.

A dreaming mode is introduced in main() to replay the accumulated (stm) sequences and to build the associated sequences in [ltm] memory. This allows the [a n] node to be created as shown in Fig. 20.

Once [ltm] nodes have been created, the (stm) nodes can be cleared from the graph and the [a n] node can be used to reproduce the node firing sequence that will rebuild the (stm) pattern in memory (Fig. 21).

A simple prediction process can occur using the stored [a n] sequence, namely to predict an (n) node in (stm) memory following input of an “a” symbol. This is shown in the graph flow shown in Fig. 22, where the symbol “space” is input, leading to the [a] -> [a n] -> [n] -> n -> IP -> (n).9 -> (a)n.9 sequence to fire. Here the predicted (stm) nodes are created with value = .9 to differentiate them.

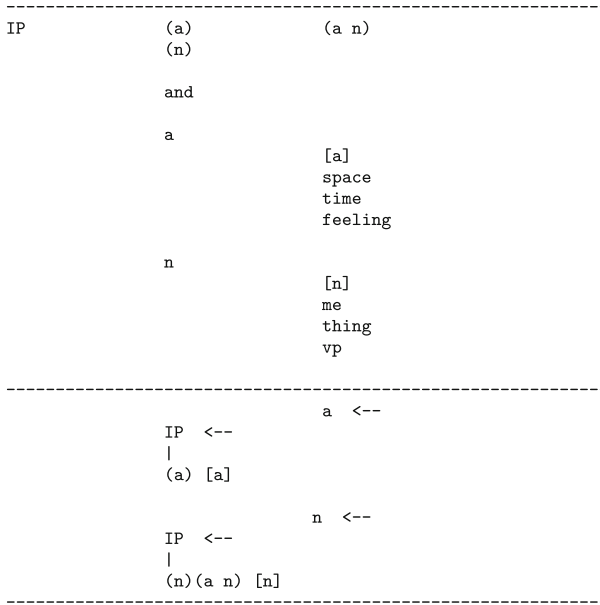


Fig. 19. The [ltm] nodes fire at the end of the existing (stm) node sequences using the link set up at (stm) merge time.

10 Verbs and Prepositions

The head-connection weight of the [vp] node is manually modified in main() to cause its head node (the vp node) to fire when the [vp] node fires. This causes a second layer of (stm) nodes to be built in (stm) memory upon input of “vp”, as shown in Fig. 23. This second (stm) layer can then alter subsequent graph operations, similarly to the “and” node.

In Fig. 24, “I eat food” is input, causing the (stm) layers to be built as shown. Input of “food” causes a closed loop to form (return=2) in nvp and ip (stm) memory. This would permit enhanced binding of the subject and object, using the same mechanism as the conjunction node.

We introduce a modification to the touch() function for the layer of (stm) nodes introduced by the second [vp] loop. This is shown in Fig. 24 following the input of “food”; a closed loop is detected in the nvp (stm) memory and a return value = 2 is returned to the calling nodes. This is posited to cause the ()₂ (stm) node to touch() its copy node when the return value = 2 from the closed loop. This

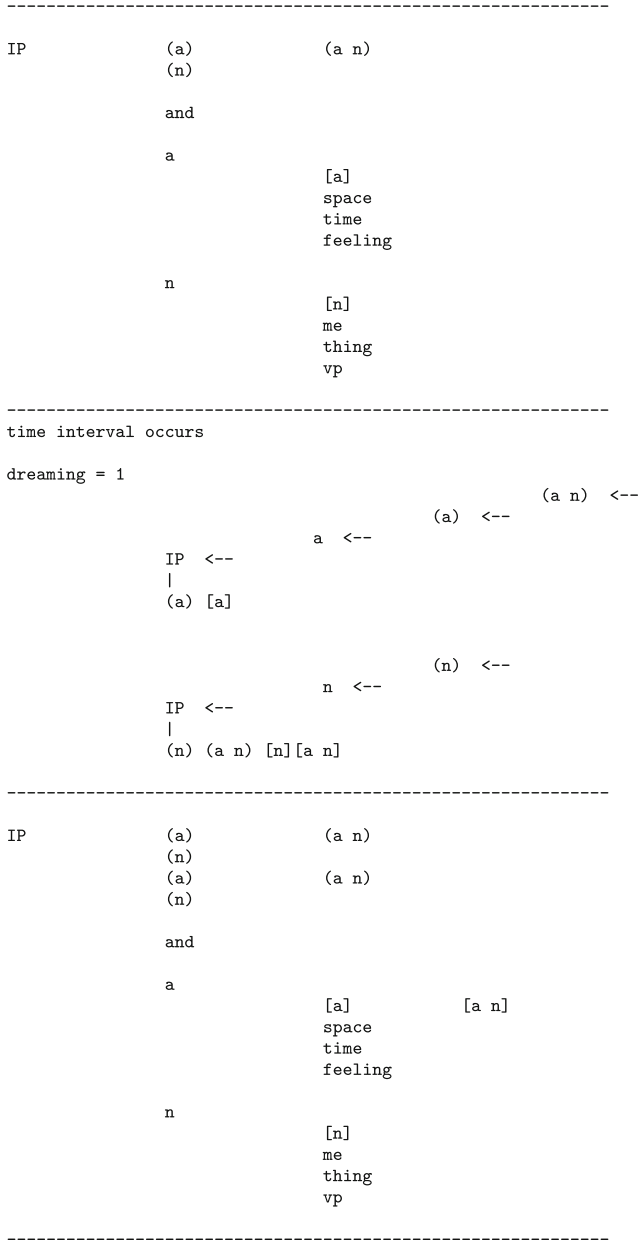


Fig. 20. A dreaming mode allows previously created stm memory structures to duplicate in [ltm] memory by replaying the stored (stm) pattern with modified network parameters for the merge function.

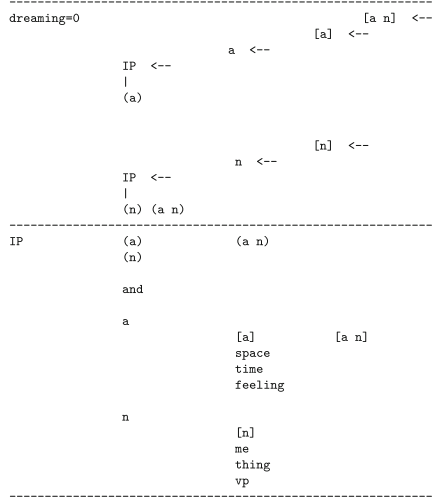


Fig. 21. The [a n] node can then be used to recall the (a) (n) pattern to (stm) memory.

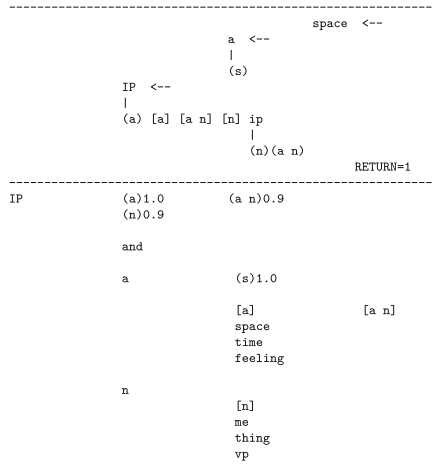


Fig. 22. Prediction of “n” after the input of “space”

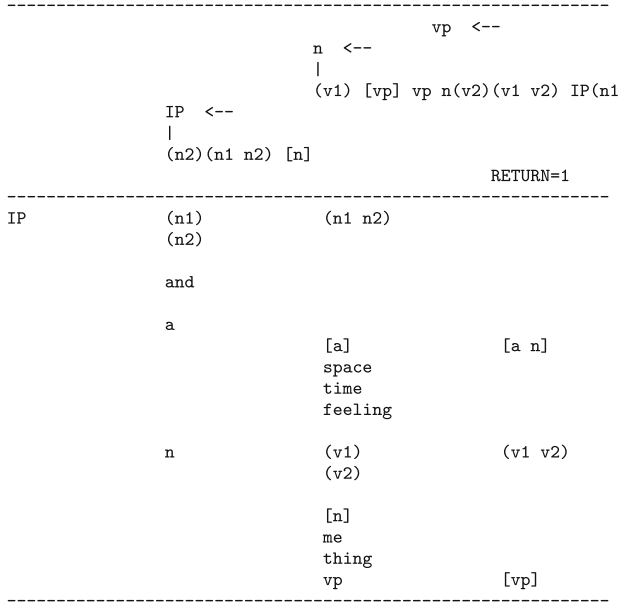


Fig. 23. The vp node is posited to be the hypernym node for the verbs and prepositions. The weights of the [vp] node are such that it fires the vp node a second time, generating a second loop in (stm) memory. This second loop is posited to implement functions of the direct object.

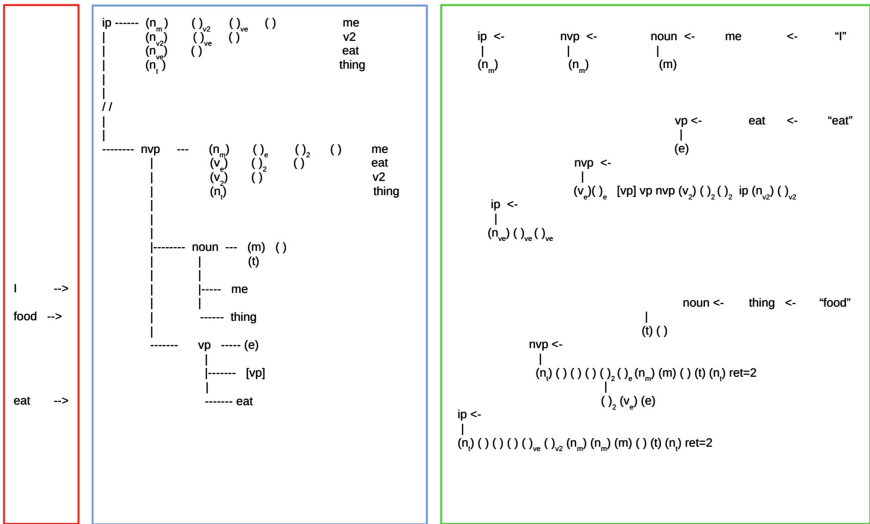


Fig. 24. I eat food.

causes the graph flow to follow the (stm) memory back to the specific verb node and would allow for one shot Hebbian enhancement of the (stm) bindings between the verb and the subject+object.

Illicit direct object constructions can be detected to cause a return value of -1 . A purpose of this function would be to keep the illicit input node from forming enhanced bindings in (stm) memory by inhibiting the Hebbian one shot weight enhancement mechanism. This function can be implemented using the same mechanism as the conjunction node. In the graph flow shown upper right in Fig. 25, the input “I eat *drink” is depicted. Input of “drink” causes the $()_{v2}$ node to return -1 by the same rule as used in the conjunction $()_+$ node, namely if the return value (from the $()_{ve}$ node) is < 2 , return -1 .

This similar function between the $()_{v2}$ and $()_+$ nodes indicates that this function could be latent in all (stm) nodes and enabled by change of a local variable. This information would be stored in the permanent [vp] and “and” nodes and passed to the (stm) nodes at run time.

The second construction shown is “I eat now *food”. Following input of the “now” symbol, the $()_{v2}$ node would return -1 on the existing logic, however “I eat now” is a valid construction. This indicates the $()_{at}$ node must override the return value = -1 , as indicated in the graph flow. However, following the input of “food”, the $()_{at}$ node must then produce a return value of -1 . This logic could be based on the condition of (return value = 2) + (the Hebbian enhancement already run).

Here the process of making functional changes to the touch() function is based on the desired language outcome. The recursive touch function is such that a single line of code change to the touch function will produced the desired changes to the graph flow.

Prepositions are posited to be children of a single node “prep”, which is a child of the vp node as shown in Fig. 26. The [p] node is posited to set an additional layer of (stm) nodes using the same mechanism as the [vp] node. This would allow the Hebbian one shot enhancement mechanism to run onto the “a” branch when “house” is input as shown in Fig. 26. The preposition “at” is posited as made by a merge between the [p] node and the [s t] node.

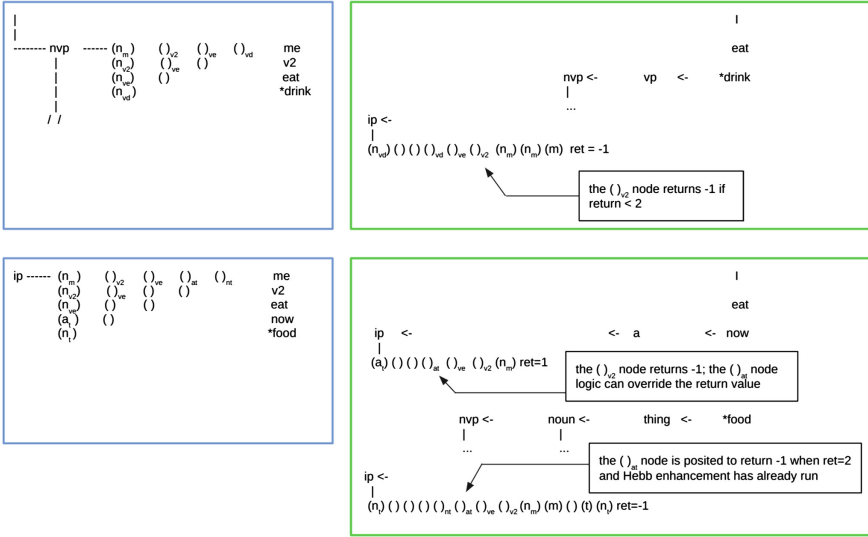


Fig. 25. Illicit direct object constructions, “I eat *drink” and “I eat now *food”.

Here we have combined space and time into a single node. This follows the pattern where the branches of the graph branch into two child branches.

Further child nodes of the prep node would be: in, on, to, through,... The pattern would be to create each new preposition node by a merge operation in main() between the preposition’s archetype and some other high level [ltm] node in the graph.

11 Past and Future, Progressive and Perfected, Singular and Plural

The [ltm] nodes of the nvp, noun, and vp nodes are posited to develop pairs of child nodes, differentiated by a +−1 stored value, which correspond to the progressive/perfected tenses, singular/plural, and the past/future tenses, respectively.

This value based mechanism allows for illicit grammar detection in all three categories using the same mechanism, as described below.

11.1 Past and Future, Past Irregular Verbs

We modify the graph to add two child nodes to the [vp] node as shown in Fig. 27. These nodes are assumed to carry default values of

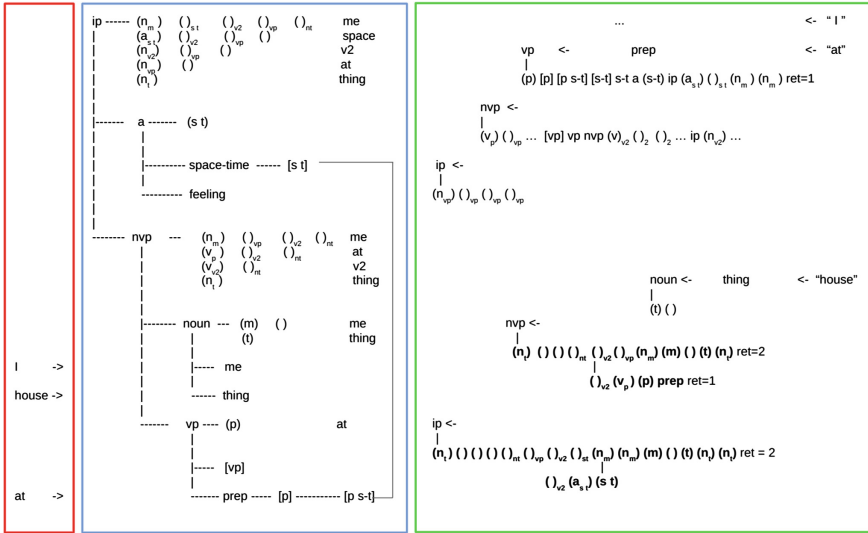


Fig. 26. A “prep” node is created as a child of the vp node. The weights of the head and copy nodes are set so as to produce the (stm) layer in IP memory. The ‘at’ node is posited as being formed from a merge between the [p] and [s-t] nodes.

+−1, corresponding to the past and future tenses. The “will” and “ed” tense markers are mapped to either node as shown. Upon input, the (stm) memory nodes are assumed to carry the +−1 value, which permits detection of the illicit construction when the merge function creates the (*) node in Fig. 27.

The $[-1]_{\text{PAST}}$ and $[+1]_{\text{FUTURE}}$ nodes are posited to cause an additional (stm) diagonal layer to form as shown.

The future modal verbs can be modeled in sequence, $[+.6]_{\text{can}}$, $[+.8]_{\text{shall}}$, $[+.9]_{\text{must}}$, $[+1.0]_{\text{will}}$.

Irregular past tense verbs can be represented in the graph by a merge between a verb’s [ltm] node and the $[-1]_{\text{PAST}}$ node. The graph flow at input is modified as shown in Fig. 28. The $[-1]_{\text{ATE}}$ node is assumed to produce the normal verb input but then to also follow its copy link to the $[-1]_{\text{PAST}}$ node, which produces an identical graph input response as the “−ed” verb ending.

11.2 Progressive and Perfected

The progressive and perfected tense symbols “are” and “have” are implemented on the graph using the same +−1 branching mechanism from the [nvp] node as shown in Fig. 29. Detection of illicit

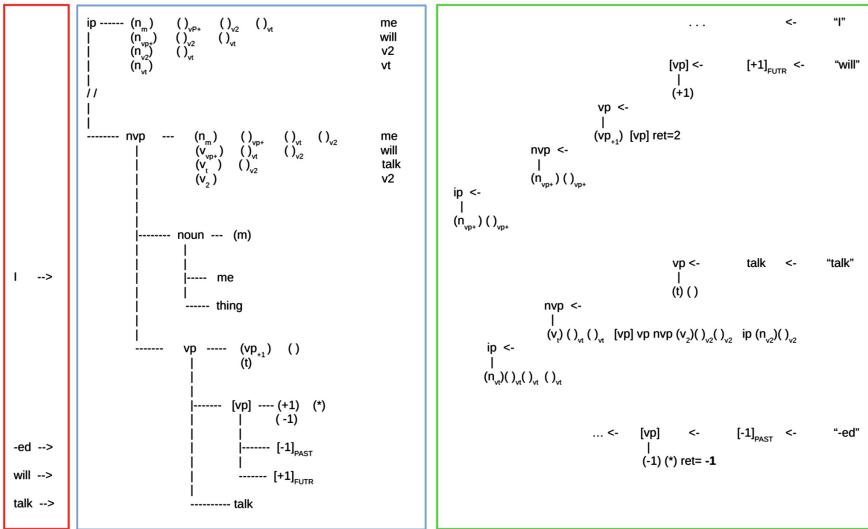


Fig. 27. The $[vp]$ node bifurcates on value to form the past and future tense symbols. Illicit tense detection (I will talk *ed) is possible at the merge of the $(+1)$ and (-1) child nodes of the $[vp]$ node.

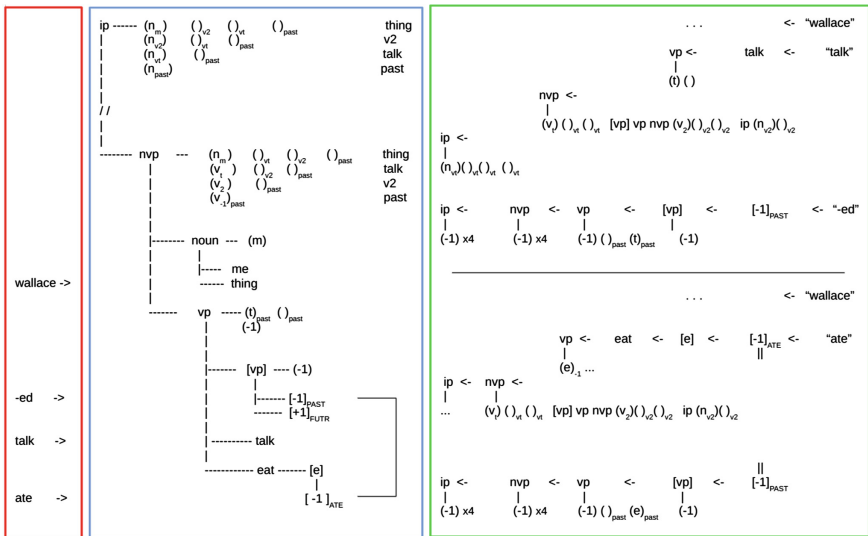


Fig. 28. Irregular past tense verbs are formed by a merge between the verb's $[ltn]$ node and the $[-1]_{PAST}$ node.

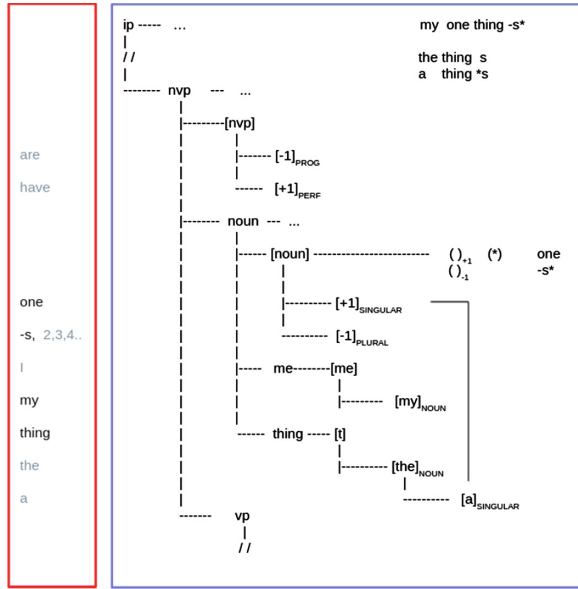


Fig. 30. Posited merge nodes created for “the”, “a”, and “my” nodes.

function calls cause (stm) memory structures to accumulate in the ip graph, possibly forming enhanced Hebbian bindings and/or returning any false signals.

The (stm) memory structures that are built for each sentence can be probed from main() to return the rightmost (stm) node. These (stm) nodes can then be touch()-ed from main() to invoke sequence output.

Following an input sequence, the ip graph can be used to generate a prediction and/or a question based on its stored [ltm] and (stm) memory.

A ponder mode is posited whereby the same prediction mode would run, but with modified graph parameters to allow for deeper recursive function calls to probe [ltm] memory that could be recalled and utilized.

Repeating this process on a training corpus would then accumulate a set of (stm) nodes that can encode the sequences that have been processed since the last sleep cycle. A sleep cycle can then reprocess the (stm) memories into [ltm] structures.

The ip graph can then be extended manually in main() by doing merge() function calls between existing terminal nodes, following

which, the appropriate words in the training corpus are mapped to the new terminal graph nodes and the process of corpus processing is repeated, allowing for additional graph terminal nodes to be added.

13 Discussion, Similar Work

The graph model derived herein draws heavily from the interdisciplinary field of Psycholinguistics which is a broad research area concerned with all aspects of human language (Fitch 2017).

The Strong Minimalist Thesis described in (Bolhuis 2014) proposes that a single recursive function, `merge(head, copy)`, produces all observed human language and grammar. (Lieberman 2015) argues against the SMT model, citing that the current diversity in human languages would preclude a single model that evolved rapidly 100 Kya; arguing that the grammatical information necessary to encode the set of all human languages would have had to have been fully specified at a single point of evolution. Lieberman notes that the evolution of the human vocal tract can be traced to 500 Kya, and argues that language acquisition can be explained via similar mechanisms as motor control in the neocortex, which would also be common to all mammals. (Fitch 2017) discounts the link to vocal tract evolution, citing that all primates have a similar “speech ready” vocal tract and that complex language has only evolved in humans.

(Phillips 2013) argues against the right to left order of merge operations used by the SMT model to account for sentence object movement and instead that a left to right assembly of sentence components is most likely to represent the underlying physical processes at work. In the SMT model, the sentence is first premeditated as a directed graph (Bolhuis 2014, fig. 1). A merge operation is then posited to occur repeatedly at each vertex until the empty stub is encountered, causing the symbol to be linked in place. The sentence would then be externalized.

In the prototype `c++` graph model, as shown in Fig. 11, the sentence is input as sequence of `touch()` calls from `main()` and then externalized via a subsequent function call: `(b e t)->touch()`.

The Node class defines a single recursive runtime function, `touch()`. The prototype for this function is implemented as any node in a neural network would be. The touch function takes an input, adds it to an internally stored activation level, and if the node fires, it touches the nodes that it is connected to, as shown below:


```

touch( input, value ){
    if node fires:
        head.touch(...)
        copy.touch(...)
        loop child branches{
            branch->touch(...) }
}

```

A “value” parameter is also passed in the touch() function which is specific to the graph model. If the node fires, the passed in value will overwrite the node’s current value. In the SMT model, the graph nodes do not have a value parameter. The value parameter adds an internal degree of freedom to the node class that can then be used to produce movement and other language phenomena using a compact recursive algorithm. The change to the touch function to cause movement is given as a single line change to the pseudocode as shown in Fig. 10.

Short term memory nodes are added to the IP graph via a merge operation that occurs between the nodes in the graph as they fire. We introduce a direction parameter to the call to the touch function and perform the merge with a different head node for the left and right directions as shown:

```

touch( ..., touched_by, direction ){
    if node fires:
        if direction == left
            new_child_node = merge( touched_by,      *this )
        if direction == right
            new_child_node = merge( head.last_child, *this )

        new_child_node-> touch( *this, right)
    ...
}

```

The left merge creates the first level (stm) nodes in Fig. 5, i.e. (hiker), (rock), (duck), and (sneeze). The right merge creates the higher order (stm) nodes, i.e. (h r), (r d), (h r d), etc.

Each newly created (stm) node is then touched and fired with direction = right. This allows the (stm) nodes to form in the ip graph as shown in Fig. 11. Each subsequent input symbol causes a diagonal layer of (stm) nodes to be added to the half matrix of (stm) nodes. In the second block of runtime graph output in Fig. 11, an

output mode flag is used to suppress the merge() operation between the (stm) nodes as they fire with direction=left.

The graph mechanism for coordination is then deduced from the structure of the graph and simple conjunction examples. The symbols in the ip graph are bifurcated into (adverbs/adjectives) and (nouns, verbs, and prepositions) as in Fig. 12, and then further bifurcated within each branch as shown. This allows for the closed loop detection algorithm to be implemented as shown in Fig. 15. When a closed loop is detected, the weights of all the nodes in the loop can be enhanced in the touch function when the nodes in the loop return in the order that they fired. The enhanced bindings between the nodes can then be used in movement of the conjunction, as in Fig. 17.

```
here and now           // true
Wallace and Gromit    // true
Wallace and *now      // false
```

Illicit conjunctions are posited to be detected as a non closed loop condition upon return of the (a +) node in Fig. 16. The (a +) node is copy rooted to the 'and' node, which would allow some local variable or flag to be inherited by the (a +) node and then subsequently to be used to invoke logic to return a false indication to main().

Long term memory nodes can be created in the graph model by introducing a dreaming mode and modifying the touch() function to do an additional merge to create an [ltm] node, if it does not already exist, as shown in Fig. 18. The dreaming mode is enabled from main() by setting a global flag. Sequences of symbols that have been input and stored in (stm) memory can be replayed from main() with the dreaming mode flag set. The existing logic in the touch() function can then be used to store the corresponding symbol sequence in [ltm] memory.

All mammals can store long term memories of sequences of symbols and recall and utilize this information after sleeping. Navigational path information about a mammal's local environment would be presumably stored as a sequence of symbols. For instance, a complex path to water/food from a mammal's home location would be used on a daily basis by the mammal.

The 'n' branch of the ip graph in Fig. 23 is bifurcated into me, thing, and 'vp'. The vp node is posited to be the hypernym node

for all verbs and prepositions and to implement the functions of the direct object. The function of the direct object can be implemented with the ip graph model by modifying the connection weights to the head node of the [vp] node, i.e. the vp node. This causes the [vp] node to touch() and fire the vp node a second time, which then fires to the ip node, which has the effect of adding a second layer of (stm) nodes to the (stm) nodes that are already present, as shown in Fig. 23.

The direct object functionality is drawn in Fig. 24 using the example “I eat food”. The input of the direct object causes a closed loop to be detected and allows for Hebbian one pass weight enhancements of the nodes in the loop. Specific verbs and preposition nodes can be added to the graph as children to the vp node and will inherit this direct object functionality on input. The second layer of (stm) nodes can also be used to return a false signal for illicit direct object constructions, i.e. “I eat *quickly food”, using the same mechanism as is used for the illicit conjunction.

A ‘prep’ node is posited as a child of the vp node, and to cause a third layer of (stm) nodes to be added to (stm) memory using the same mechanism at the [vp] node, as shown in Fig. 26. In Fig. 26, the [p s-t] node is posited to represent the preposition ‘at’. The [p s-t] node is created via a merge call from main() between the existing [ltm] nodes for [prep] and [space-time].

The model is then extended by creating a value based bifurcation of the [vp] node as shown in Fig. 27. Here the past and future tense symbols “ed” and “will” are introduced as child nodes to the [vp] node. The inferred functionality is that these nodes produce an (stm) node with a ± 1 value upon input. This allows for detection of the illicit construction “I will talk *ed” as shown. Here the detection would occur in the merge() function that creates the (*) node in Fig. 27.

Irregular past tense verb symbols are formed as a merge between a verb’s [ltm] node and the $[-1]_{PAST}$ node as shown in Fig. 28. When the $[-1]_{ATE}$ node fires on input, it fires its head and copy links and produces the same internal representation as the verb + ‘ed’ marker.

This value based bifurcation is duplicated to the [nvp] node in Fig. 29, where the $[+1]$ and $[-1]$ nodes are inferred to represent the progressive and perfected tense symbols. Here the invalid construction “Wallace has talk *ing” is detected in the merge of the (*) node

in the [nvp] node's (stm) memory. In Fig. 29 right side, the terminal nodes of the current English grammar are created by doing merges between the $[+1]_{PROG}$ and $[-1]_{PERF}$ nodes and the [my], [thing], and $[-1]_{PAST}$ nodes. As before, these merge function calls are directly encoded in `main()`.

This value based bifurcation is again duplicated to the [noun] node in Fig. 30, where the $[+1]$ and $[-1]$ nodes are inferred to represent singular and plural symbols. Here the invalid construction "my one thing *s" is detected in the merge of the (*) node in the [noun] node's (stm) memory. The [noun] branch could be expanded to encode mathematical relations. Logical relations would presumably have to occur in the top level ip (stm) memory due to the placement of the 'and' node.

The WordNet dictionary corpus contains 80K, 20K, and 10K, nouns, verbs, and adverbs/adjectives respectively. In the graph model, each of these words would map to a terminal node in the graph, at whatever stage of development of the graph. Approximately 30 terminal nodes are described herein. The sentence "I washed the car yesterday" would be reduced into a truncated English form: "me did thing before" at roughly the current development level of the graph.

If the graph model is representative of the human neocortex connectome, then we would expect there to be a unique terminal node for each word in the WordNet corpus. This expansion of the model would occur as `merge()` function calls in `main()` between existing terminal nodes. This allows the child nodes to inherit all of the functionality of their parent nodes.

The key difference between humans and the other mammals could be in the level of bifurcation in the graph. If one considers an ip graph with all words as direct children and no hierarchical bifurcations, the graph model would still be able to encode simple dialogue interactions, as in Figs. 7, 8, and 9, and to encode sequences in (stm) and [ltm] memory, but the closed loop Hebbian weight enhancement process would not produce consistent results and could not be utilized by the mammal.

Acknowledgments. The author has received many helpful comments from many people who have reviewed early versions of this manuscript. The author also gratefully thanks the reviewers for their helpful comments.




References

- Berwick, R.C., et al.: Evolution, brain, and the nature of language. In: *Trends Cogn Sci.* **17**(2), 89–98 (2013). See Fig. 1. <https://doi.org/10.1016/j.tics.2012.12.002>
- Bolhuis, J.J., et al.: How could language have evolved? In: *PLOS Biology*, 26 Aug 2014. <https://doi.org/10.1371/journal.pbio.1001934>
- Bryne, R.W., et al.: Great ape gestures: intentional communication with a rich set of innate signals. *Anim Cogn.* (2017). <https://doi.org/10.1007/s10071-017-1096-4>
- Chomsky, N.: Artificial intelligence. In: *Navigating a Multispecies World: A Graduate Student Conference on the Species Turn*. <https://sts.hks.harvard.edu/events/workshops/navigating-a-multispecies-world/> (2015)
- Chomsky, N.: Some core contested concepts. *J. Psycholinguist Res* **44**, 91 (2015). <https://dspace.mit.edu/openaccess-disseminate/1721.1/103525>. <https://doi.org/10.1007/s10936-014-9331-5>
- Chomsky, N.: Language, Creativity, and the Limits of Understanding. (4–21–16). At <https://www.youtube.com> (2016)
- Chomsky, N.: Language and the Cognitive Science Revolution(s), Lecture Given at Carleton University 2011. <https://chomsky.info/20110408/> (2011)
- Darling, C.: Guide to Grammar & Writing. <https://www.guidetogrammar.org/grammar/>
- Del Signore, K.W.: Measuring and simulating cellular switching system IP traffic. *Bell Labs Tech. J.* **18**, 159–180 (2014). <https://doi.org/10.1002/bltj.21651>
- Epstein, R., Kanwisher, N.: A cortical representation of the local visual environment. *Nature* **392**(6676), 598–601. The 'parahippocampal place area' (PPA) is described (1998)
- Fitch, W.T.: Empirical approaches to the study of language evolution. *Psychon. Bull. Rev.* **24**(1), 3–33 (2017). <https://doi.org/10.3758/s13423-017-1236-5>
- Goertzel, B.: Artificial General Intelligence: Now is the Time, Google tech talk. At <https://www.youtube.com/watch?v=A-dycsiRwB4> (2007)
- Hawkins, J., Blakeslee S.: *On Intelligence*. Published by Times Books (2005)
- Henshilwood, C., et al.: An abstract drawing from the 73,000-year-old levels at Blombos Cave. South Africa. *Nature* **562**, 115–118 (2018). <https://doi.org/10.1038/s41586-018-0514-3>
- Hobaiter, C., Byrne, R.W.: The meanings of Chimpanzee gestures. *Curr. Biol.* **24**(14), 1596–1600 (2014)
- Hubel, D.H.: Eye Brain and Vision. At <http://hubel.med.harvard.edu/index.html> (1980)
- Hubel, D.H.: Tungsten microelectrode for recording from single units. *Science* **125**(3247), 549–550 (1957). <https://doi.org/10.1126/science.125.3247.549>
- Hubel, D.H., Wiesel, T.N.: Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* **148**(3), 574–591. (1959). <https://doi.org/10.1113/jphysiol.1959.sp006308>
- Huth, A.G., et al.: A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**(6), 1210–1224. See figure 2, PPA analysis (2013). <https://doi.org/10.1016/j.neuron.2012.10.014>
- Huyck, C.R., Passmore, P.J.: A review of cell assemblies. *Biol. Cybern.* **107**, 263–288 (2013). <https://doi.org/10.1007/s00422-013-0555-5>
- Kruger, N., et al.: Deep hierarchies in the primate visual cortex: what can we learn for computer vision? *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(8), 1847–1871 (2013). <https://doi.org/10.1109/TPAMI.2012.272>

- Minsky, M.: A Framework for Representing Knowledge, MIT-AI Laboratory Memo 306. <http://web.media.mit.edu/minsky/papers/Frames/frames.html> (1974)
- Moore, R.: Ape Gestures: Interpreting Chimpanzee and Bonobo Minds, vol. 24, issue 14, pp. R645–R647. Cell Press (21 July 2014)
- Phillips, C.: Linear order and constituency. *Linguistic Inquiry* **34**(1), 37–90 (2003)
- Phillips, C.: Derivational order in syntax: evidence and architectural consequences. *Stud. Linguistics* **6**, 11–47 (2013)
- Pinker, S.: Words and Rules. At <https://stevenpinker.com/files/pinker/files/edinburgh.pdf>. See figure 1 (1994)
- Scerri, E.M.L., et al.: Continuity of the middle stone age into the holocene. In: *Scientific Reports*, vol. 11, Article number: 70 (2021)
- Tattersall, I.: A tentative framework for the acquisition of language and modern human cognition. *J. Anthropol. Sci.* **94**, 157–166. Epub. PMID: 27014833 (2016). <https://doi.org/10.4436/JASS.94030>
- Repository of Code Used to Generate Figures: <https://github.com/kwd2/graph1>



Critical Slowing Down in Heart Rate Variability for Human Condition Control: An Example of Sleep Onset Detection

Valeriia Demareva^(✉) , Irina Zayceva , Andrey Demarev ,
and Nicolay Nazarov 

Lobachevsky State University, Nizhny Novgorod 603022, Russia
valeriia.demareva@fsn.unn.ru

Abstract. Modern technologies offer numerous opportunities for detecting the human condition. The practical implementation of such solutions is relevant both for various industries and for virtual activities—wherever they are associated with a significant cognitive load and a high risk of errors due to human factors, such as loss of vigilance from fatigue or drowsiness. The research presented in this paper validates an approach for the rapid detection of sleep onset by analyzing a single metric of heart rate variability, utilizing the concept of critical slowing down. The material for this research consisted of 4 evening-night recordings of NN intervals, where moments of sleep onset were marked for each participant. Standard deviations (SDNN) and autocorrelation coefficients of NN intervals were analyzed within sliding windows. It was found that immediately after sleep onset, there was a sharp and pronounced decrease in both metrics, regardless of the time of falling asleep, with the dynamics of SDNN proving to be more indicative. The results of this pilot study can be utilized for further exploration of early warning signals in heart rhythm indicators that would indicate a loss of vigilance. This may serve as a foundation for the development of systems that predict a decline in cognitive control due to sleepiness during natural activities.

Keywords: Sleep onset · Detection · Heart rate variability · Condition · Critical slowing down

1 Introduction

Development and application of technologies for real-time detection of decreased vigilance and the moment of falling asleep have a broad spectrum of applications, related to safety, health, and increased efficiency across various spheres of life [1]. One common approach to monitor an individual's condition involves utilizing technologies that gather and analyze data related to heart activity. Portable wearable sensors enable the recording of electrocardiographic signals and the inference of the human nervous system's functioning based on heart rate, heart rate variability, frequency domain metrics, and other factors [2].

In our research, we built upon the notion that practical implementations of human condition monitoring technologies should not only focus on detecting sleepiness itself but also the associated loss of vigilance or cognitive control. Simultaneously, accuracy and convenience of application in natural conditions remain crucial criteria when selecting a data collection method. The analysis of heart rate fulfills these two requirements and is actively employed, alongside other methods, to address similar issues [3]. Consequently, monitoring and predicting conditions based on heart rate variability (HRV) offer the potential for creating an adaptive ‘smart’ environment tailored to an individual’s needs or activity objectives.

Sleep onset serves as a straightforward example of a condition transition. The challenge lies in designing an accurate sleep onset or drowsiness detector. Many algorithms presented in research papers have been grounded in laboratory experiments and are unsuitable for real-time applications. To address real-time detection of observable shifts in human conditions, the concept of critical slowing down [4, 5] can be of great significance. The core concept suggests that changes in the dynamics of variance can signify a transition from one condition to another.

If we view falling asleep as a continuum transition to a sleep condition, then for the task of detecting such a transition, we can employ the theory of complex dynamic systems, specifically the critical slowing down (CSD) concept. The transition from wakefulness to sleep can rightfully be regarded as critical for the organism. Drawing an analogy with complex systems [5], a slowdown in the organism’s operation could be interpreted as a signal indicating a critical transition.

The human body behaves as a complex dynamic system with various states, as reflected, in part, by the dynamics of the autonomic regulation. Fluctuations in alertness are natural occurrences in the human biorhythm. Additionally, it has been demonstrated that drowsiness gradually increases during the transition from evening to nighttime. It is conceivable that the fragility of the system gradually intensifies until it reaches a critical point, resulting in a sudden transition to a qualitatively new state—sleep.

Unlike the state of depression, where critical slowing down is accompanied by the escalation of ‘shocks’ to the system and, consequently, an increase in variance [4], the end of the wakefulness state is more likely to involve a decrease in the system’s ability to track fluctuations and, consequently, a reduction in variability. This is logical since the organism stops responding to external stimuli when transitioning to sleep, as vigilance decreases during the transition to sleep and with an increase in sleep deprivation time [6]. Such behavior of variance during CSD is also feasible [7, 8].

The objective of the present study was to demonstrate the feasibility of real-time sleep onset detection based on the analysis of the standard deviation of NN intervals in moving windows. We hypothesized that the CSD of the HRV signal could be employed to detect transitions between different conditions (wakefulness-sleep). A specific hypothesis posits that the metric of heart rate variability (SDNN) would decrease immediately after sleep onset.

2 Materials and Methods

2.1 Dataset Description

The data of four participants (2 men and 2 women, 27–35 years old) from SSDD [9] were selected for analysis. The participants are encoded as P1, P2, P3, and P4.

2.2 Study Design

The whole description of the study design is presented in [9]. Participants wore the Polar H10 heart rate monitor (Polar Electro Oy, Kempele, Finland) starting at 7:40 PM and then logged into the UnnCyberpsy web application developed by the authors. Within UnnCyberpsy, participants provided their personal information and completed the Epworth Sleepiness Scale (ESS). Beginning at 8 PM and continuing every 30 min thereafter, participants rated their subjective sleepiness levels using both the Karolinska Sleepiness Scale (KSS) and the Stanford Sleepiness Scale (SSS) until they fell asleep. At 6 AM, participants woke up, answered a sleep quality questionnaire, described their dreams, and again completed the KSS and SSS. Data from the questionnaires, including the ESS, KSS, and SSS, were not considered within the scope of this article, as they did not align with the study's purpose.

The abovementioned design was updated by the inclusion of fixation at the time of sleep onset. That is, when participants went to bed, they had to press the button on the doorbell (Fig. 1) and hold it down.

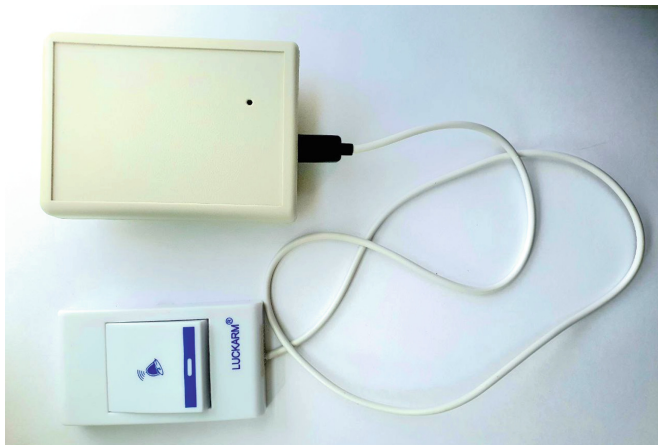


Fig. 1. Photo of doorbell and receiving unit.

The signal from the button was transmitted to the receiving unit and stored on it. The button state was recorded every 2 ms. Then the data was transmitted to the computer and analyzed. The moment of the first release of the button was analyzed. This moment corresponded to the time of sleep onset and the loss of cognitive and conscious control. All participants had undergone and trial experiment to get used to the measuring equipment. Only the data from a subsequent experiment were used.

2.3 Data Analysis

The data underwent preprocessing in Jupyter Notebook using Python. During the data filtering process to remove individual artifacts, NN intervals that were shorter than 400 ms or longer than 1300 ms were removed, along with intervals deviating more than 70% from the median of the five preceding intervals. The ‘hrv-analysis’ module was used to calculate the time domain metrics in moving windows of 300 s and step of 30 s, and standard deviations of NN intervals (SDNN) were further analyzed. To perform autocorrelation analysis between NN intervals arrays, Pearson correlation coefficients were calculated using ‘scipy.stats’ module. Sequent arrays containing 500 NN intervals were correlated (see Fig. 2).

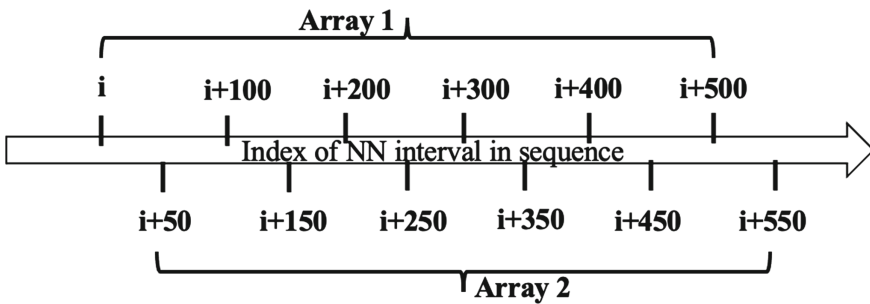


Fig. 2. The representation of the array selection of NN intervals for autocorrelation.

The first array (Array 1) contained $i-i + 500$ NN intervals, and the second array (Array 2) contained $i + 50-i + 550$ NN intervals, where i signifies the index of each NN interval in sequence.

3 Results and Discussion

The dynamics of SDNN and autocorrelation coefficients in the four participants (P1-P4) are presented in Figs. 3, 4, 5 and 6. The green line indicates the moment of the first sleep onset.

In participant P1, it is clearly evident that immediately after the moment of falling asleep, indicated by the green dashed line, there was a sharp decrease in both SDNN and the autocorrelation coefficient of NN intervals. For SDNN, smoother values were observed after the decrease over an extended period, unlike the autocorrelation coefficient. Overall, the dynamics of these two indicators were similar.

For participant P2 (Fig. 4), a decrease in both metrics was observed immediately after sleep onset. After falling asleep, SDNN exhibited lower variability in values compared to the autocorrelation coefficient. The dynamics of both metrics were similar, but there were also episodes when they were in antiphase (after 03 AM).

After sleep onset, participant P3 showed the same dynamics in both metrics as observed in P1 and P2 (see Fig. 5). The overall trend of the metrics was similar, with episodes where they were in antiphase (for instance, at 2:30 AM).

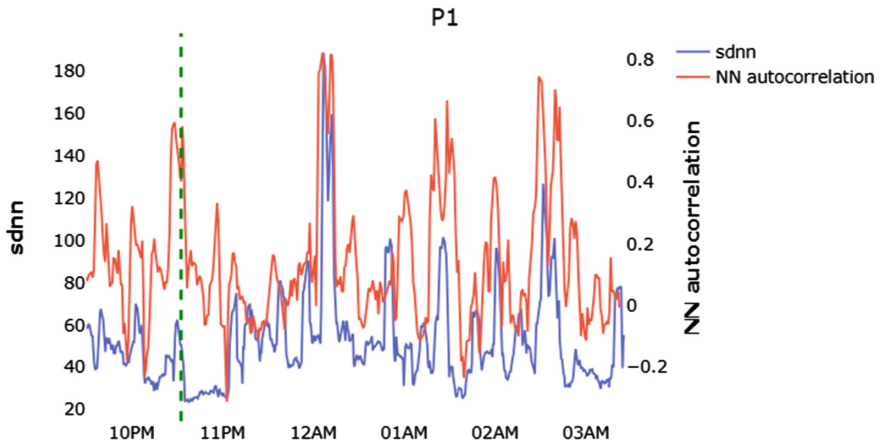


Fig. 3. SDNN and NN autocorrelation dynamics of P1. Green dashed lines indicate the time of first episode of sleep onset.

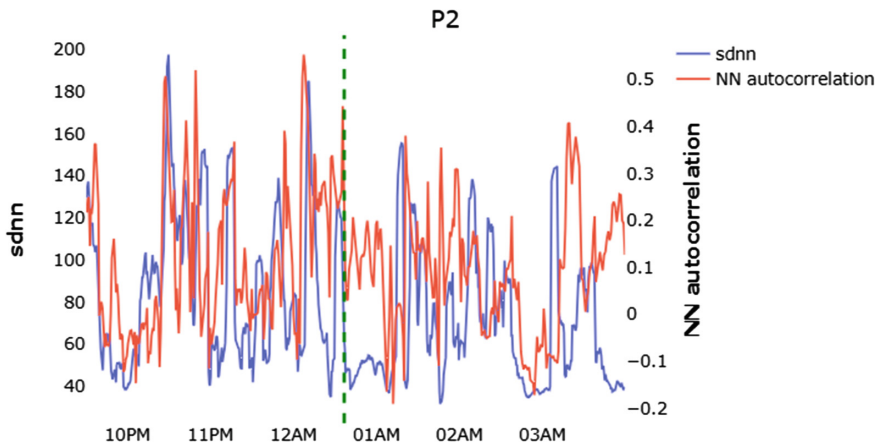


Fig. 4. SDNN and NN autocorrelation dynamics of P2. Green dashed lines indicate the time of first episode of sleep onset.

Participant P4 demonstrated a similar dynamic of metrics to all the others, despite falling asleep much later (after 3:00 AM) (see Fig. 6). The dynamics of the two metrics were similar, with isolated episodes exhibiting opposite patterns.

Thus, as seen in Figs. 3, 4, 5 and 6, there was a sharp and sustained decline in both SDNN and NN autocorrelation coefficients immediately after the moment of falling asleep. This pattern was consistently repeated in all participants, regardless of their sleep onset time. This supports both the overall and specific hypotheses of the study. Consequently, the concept of CSD can be effectively utilized for the rapid detection of sleep onset.

Considering that autocorrelation followed the same pattern as SDNN, and that SDNN exhibited a smoother curve right after sleep onset, we can infer that the SDNN metric

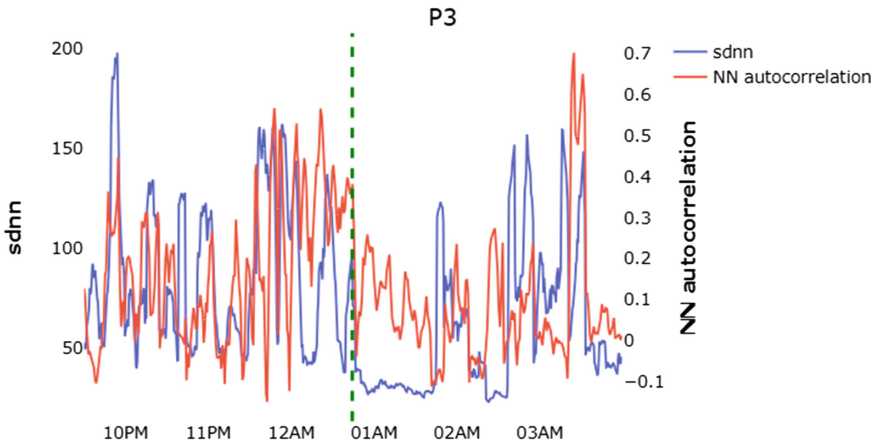


Fig. 5. SDNN and NN autocorrelation dynamics of P3. Green dashed lines indicate the time of first episode of sleep onset.

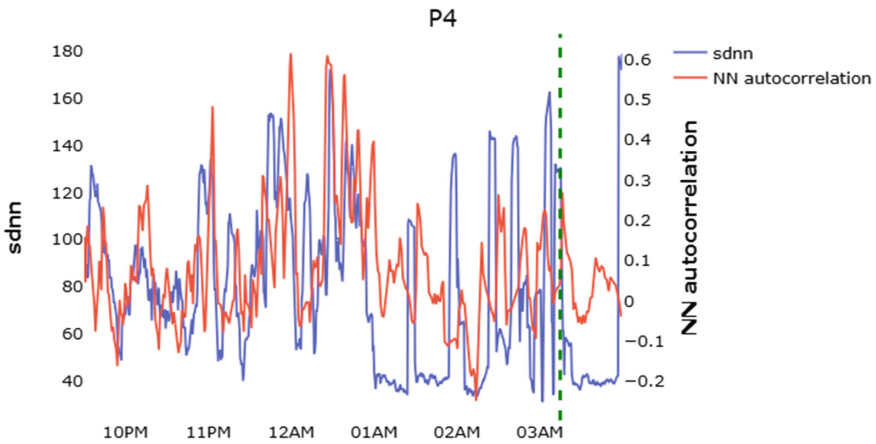


Fig. 6. SDNN and NN autocorrelation dynamics of P4. Green dashed lines indicate the time of first episode of sleep onset.

alone might suffice for sleep onset detection. The observed pattern aligns with the idea that in complex systems, signal dispersion can decrease during a critical transition [7, 8]. The identified autocorrelation dynamics also harmonize with the CSD concept [4, 5]. In the graphs shown in Figs. 3, 4, 5 and 6, a cyclical pattern was noticeable in both metrics, likely associated with the alternation of sleep stages [10].

In our research, we relied on the idea that in the practical application of human condition monitoring technologies, it is not only the fixation of sleepiness itself that matters but also the associated loss of vigilance or cognitive control. At the same time, accuracy and convenience of application in natural conditions remain important criteria when choosing a data collection method. The analysis of heart rate variability satisfied

the latter two requirements and is actively used, along with other methods, to solve similar problems [3].

It is worth noting that in this article, we are presenting our methodology for capturing the moment of falling asleep using the equipment depicted in Fig. 1 for the first time. We believe that such an approach allows for a thorough examination of the organism's functioning before and after sleep onset, without requiring additional means to monitor the moment of falling asleep.

In conclusion, it is important to highlight that we hypothesize that within the framework of the CSD concept, the detection of early warning signals (EWS) indicating a significant increase in drowsiness before sleep onset is also possible. This will need to be verified in future studies. Hyperparameter tuning for the sliding window will be necessary to identify clear signs of drowsiness escalation and the approach to the critical transition from wakefulness to sleep.

The limitations of this pilot study are related to the small sample size and the specific characteristics of the participants. Nevertheless, we were able to identify common patterns among all four participants, although this was done through visual analysis without statistical processing. In the future, we plan to extend the analysis to a larger dataset.

4 Conclusions

Currently, we have found a heart rate indicator that can be used to determine the onset of sleep in an individual, utilizing critical slowing down concept. This information can serve as a basis for further development of the metrics that allow predicting the imminent approach of sleepiness and the associated loss of vigilance, which is crucial for high-risk activities prone to human errors.

Acknowledgements. This research was funded by the Russian Science Foundation, grant number 22-28-20509.

References

1. Arakawa, T.: Trends and future prospects of the drowsiness detection and estimation technology. *Sensors* **21**(23), 7921 (2021). <https://doi.org/10.3390/s21237921>
2. Rogers, B., Schaffarczyk, M., Gronwald, T.: Estimation of respiratory frequency in women and men by Kubios HRV software using the Polar H10 or Movesense Medical ECG Sensor during an exercise ramp. *Sensors* **22**(19), 7156 (2022). <https://doi.org/10.3390/s22197156>
3. Henelius, A., Sallinen, M., Huutilainen, M., Müller, K., Virkkala, J., Puolamäki, K.: Heart rate variability for evaluating vigilant attention in partial chronic sleep restriction. *Sleep* **37**(7), 1257–1267 (2014). <https://doi.org/10.5665/sleep.3850>
4. Wichers, M., Groot, P.C.: Psychosystems, ESM Group, EWS Group: critical slowing down as a personalized early warning signal for depression. *Psychother. Psychosom.* **85**(2), 114–116 (2016). <https://doi.org/10.1159/000441458>
5. Scheffer, M., et al.: Early-warning signals for critical transitions. *Nature* **461**(7260), 53–59 (2009). <https://doi.org/10.1038/nature08227>

6. Hudson, A.N., Van Dongen, H.P.A., Honn, K.A.: Sleep deprivation, vigilant attention, and brain function: a review. *Neuropsychopharmacol. Off. Public. Am. Coll. Neuropsychopharmacol.* **45**(1), 21–30 (2020). <https://doi.org/10.1038/s41386-019-0432-6>
7. Berglund, N., Gentz, B.: Metastability in simple climate models: pathwise analysis of slowly driven Langevin equations. *Stoch. Dyn.* **2**, 327–356 (2002). <https://doi.org/10.1142/S0219493702000455>
8. Berglund, N., Gentz, B.: Noise-induced phenomena in slow-fast dynamical systems—a sample-paths approach. Springer, London (2006). <https://doi.org/10.1007/1-84628-186-5>
9. Demareva, V., et al.: Temporal dynamics of subjective sleepiness: a con-vergence analysis of two scales. *Biol. Rhythm. Res.* **54**(4), 369–384 (2023). <https://doi.org/10.1080/09291016.2023.2193791>
10. Kontos, A., et al.: The inconsistent nature of heart rate variability during sleep in normal children and adolescents. *Front. Cardiovasc. Med.* **7**, 19 (2020). <https://doi.org/10.3389/fcvm.2020.00019>



BICA's Fears and Troubles: GPT-Based AI Tools Are Its Friends or Foes?

Emanuel Diamant^(✉)

Kiryat Ono, Israel
eman1.245@gmail.com

Abstract. BICA*AI is a well-established long-lasting R&D enterprise aimed at creating computational architectures intended to emulate Human-level Artificial Intelligence. Recently and quite unexpectedly in its field has appeared another contender—a GPT-based AI tool designed to mimic man-computer conversation in a user-friendly natural human language. As its designers claim, the device exhibits signs of General AI. After an exciting and joyful reception, it became clear that the new competitor does not fulfill its expected promises—it returns wrong and misleading responses, deceptions, and disinformation. The issue raised a wave of public objections and a request to stop and prevent further device deployment. On the other hand, the device designers claim that the imperfections are temporary, and very soon the product will rich its avowed qualities. No, this will never happen! The purpose of this paper is to explain that the current approach to GPT-based AI tools design is initially flawed, wrong, and unsuitable because it ignores the basic definitions of Intelligence and Information processing. The paper joins the general awareness that unrestricted and free dissemination of wrongly designed GPT-based AI tools poses a threat to human society, similar to the threat of careless biological weapon research.

Keywords: BICA*AI · GPT-based tools · General AI

1 Introduction

BICA*AI—which stays for Biologically-Inspired Cognitive Architectures for Artificial Intelligence—is a well-established long-lasting (since 2005) R&D enterprise aimed at creating computational architectures of human intelligence [1]. “The ultimate goal of research in cognitive architectures is to model the human mind, eventually enabling us to build human-level artificial intelligence. To this end, cognitive architectures attempt to provide evidence of what particular mechanisms succeed in producing intelligent behavior and thus contribute to cognitive science” [2].

Generative AI or GPT—which stays for Generative Pre-trained Transformer—is a recently emerged AI tool and application, designed to mimic a computer conversation with a user in natural language and simulate the way a human would behave as a conversational partner [3]. GPT models are trained on massive amounts of internet text

data and are capable of generating responses that closely resembled human writing. Originally designed for purposes of NLP tasks, transformers have been adapted for various vision tasks (image classification, object detection, image generation, and video processing), audio-related applications (speech recognition, speech synthesis, speech enhancement, music generation), and various multimodal scenarios (visual question answering, visual commonsense reasoning, caption generation, speech-to-text translation, and text-to-image generation) [4]. The prime GPT model, developed by OpenAI, was consequently enhanced by more and more sophisticated versions, such as GPT-2 in 2019 and GPT-3 in 2020. In December 2022, OpenAI launched a free preview of the ChatGPT model, a new AI tool based on the GPT-3.5 model version. On March 14, 2023, OpenAI released GPT-4, both as an API and as a feature of ChatGPT [5]. According to OpenAI, **the preview received over a million signups** within the first five days. As an anonymous source revealed, (cited by Reuters in December 2022), OpenAI is projecting **\$200 million revenue in 2023 and \$1 billion revenue in 2024** [5]. Inspired by OpenAI's success, other GPT developers have launched their own products: Amazon, Google, Microsoft, Baidu, GitHub, Meta, Apple, IBM, Nvidia, and others have created their own GPT-based products and put them to market examination [6].

From this short exposition of BICA versus GPT-4 achievements, it is perfectly clear that GPT-based products have a far more privileged position in the eyes of Human-level AI tools developers and users.

However, upon a closer inspection, it turns out that the position of GPT-based AI tools is not so extremely great as it looks initially.

Despite the generally recognized success in creating artificial devices that mimic the human ability to communicate through oral or written conversation, GPT-based devices have proven to be a potential danger to their users [7–9]. According to many publications, GPT is a large language model that uses deep learning to create human-like text. Or, more certainly, it is a computing system designed to generate sequences of words, code, or other data, starting from an initial input called the prompt [10]. When determining the sequence of words, it simply predicts the next word based on statistical correlations in the training data and prompts [9]. Since the system sees only formal, statistical patterns between words, it does not understand the meaning of words, and does not understand their semantics. Therefore, it is also called the “**stochastic parrot**”, randomly stitching together words without reference or meaning [8].

All other negative features of the system subsequently follow from this—the model inherits biases and errors in the training data, the model is very sensitive to the design and formulation of prompts [11], her answers are inexplicable to users, her results are not verified or validated, and users cannot determine the validity of the model's output, or whether the model is simply a “fiction” [9]. On the other hand, the model produces very coherent, natural-sounding, and human-like responses that users find compelling and readily trust them, even if they are inaccurate [9]. At the same time, the system is exceptionally fast, very user-friendly, responds to natural language prompts and requires little or no user training. References or values [8].

Text generation at speed, scale, and ease of use makes the GPT model exceptionally well suited to widespread misinformation and deliberate misuse of the system as a tool for fraud, as a “weapon of mass deception” [9]. The main risk of weaponization is related

to the potential ability of GPT-3 to dramatically increase influential production, which is likely to be human [11].

From the above, it becomes clear that GPT systems can pose a serious danger to society and humanity. Realizing this and seeking concrete action to eliminate the dangers and harms that threaten society, several leading AI researchers, as well as a number of people who work in companies participating in the AI race, signed an open letter calling for a six-month pause in the training of AI systems, more powerful than GPT-4 [12].

The letter, published on March 30, 2023, states that “In recent months, AI labs have been stuck in an out-of-control race to develop and deploy ever more powerful digital minds” [13]. “AI systems with human-competitive intelligence can pose a serious danger to society and humanity, as extensive research has shown” [12]. Over 1300 people, including Elon Musk and Steve Wozniak, have signed this letter asking all engineers to stop immediate AI development outside of GPT-4 for 6 months due to fear of “losing control” [13].

In addition, the Future of Life Institute states that “development moratoriums should be public and verifiable” and “if the decision to suspend AI development cannot be taken immediately, governments should intervene with AI development authorities” [13].

In response to this call, on May 17, 2023, a Subcommittee of the U.S. Senate Judiciary Committee held a hearing titled “AI Oversight: Rules for Artificial Intelligence” [14]. Similar measures of administrative control and administrative order have been adopted by the parliamentary commissions of the European Union, Great Britain, Japan, and other countries (A more detailed report on the actions of many countries in this direction can be found in [15]).

At the same time, it remains as a problem that these measures of administrative control ignore the scientific and technical aspects of the problem of creating Human level Artificial Intelligence systems. Parliamentary commissions do not deal with these issues, rightly claiming that this is not within their competence.

At the same time, the organizations involved in the creation of these systems confidently declare that they **have already reached the level of General Intelligence**. This is exactly what Microsoft claims in its recent publication: “it could reasonably be viewed as **an early (yet still incomplete) version of an artificial general intelligence (AGI) system**” [16].

Sorry, but this is not true. The uncontrolled and unsupervised use of systems whose creators claim that they have already reached the level of General Intelligence can only be incomparably more dangerous to society. Because the general public is not competent in these matters. And scientists and researchers competent in these matters claim that we are still very far from the level of General Intelligence. And the main reason for this is that today we do not know at all what Intelligence is, and what the brain’s cognitive abilities are supposed to be.

Christopher Koch: “AGI is ill-defined because we don’t know how to define intelligence. Because we don’t understand it...” [17].

Luciano Floridi: GPT language model “has nothing to do with intelligence, consciousness, semantics, relevance, and human experience and mindfulness more generally” [10].

I deliberately provide here the quotes from Christopher Koch and Luciano Floridi—Christof Koch is a chief scientist of the Mindscope Program at Seattle’s Allen Institute. He has a background in both AI and neuroscience, he is the author of three books on consciousness as well as hundreds of articles on the subject, including features for IEEE Spectrum and Scientific American [17]. Luciano Floridi is a professor of philosophy and ethics of information at the University of Oxford, best known for his work on the philosophy of information, and information ethics. According to Scopus, he was the most-cited living philosopher in the world in 2020 [18].

Therefore, again:

Christopher Koch: “AGI is ill-defined because we don’t know how to define intelligence. Because we don’t understand it...” [17].

Luciano Floridi: GPT language model “has nothing to do with intelligence, consciousness, semantics, relevance, and human experience and mindfulness more generally” [10].

Their opinions sharply diverge from that of other prominent researchers in the field of AI, the so-called “founding fathers” of AI—Yann LeCun [19], Geoffrey Hinton [20], Rodney Brooks [21], and others, who are concerned mostly about the social dangers associated with uncontrolled use of GPT-based AI tools. In their most recent publications, not even a word about the technical and scientific aspects of the issue is provided.

I deliberately provide here the quotes from Christopher Koch and Luciano Floridi, because their views unconditionally support my own views on the subject (and of the BICA*AI R&D activities).

During the past years, I have repeatedly expressed my viewpoints on the critically important issues of Intelligence and Information, usually omitted in today’s AI research studies. Interested readers could find some relevant papers in the Reference list [22–27]. But for consistency of our discussion, I will provide here a short excerpt of the ideas that were exposed once in these publications.

2 What Is Intelligence?

There is a widely shared opinion that human intelligence cannot be defined as a single trait or as General Intelligence. Theories of Multiple Intelligences are steadily gaining mainstream attention. However, most frequently, Intelligence is perceived as an umbrella term that embraces (integrates) a multitude of human cognitive abilities (to sense, to perceive, to interpret the surrounding environment; to recognize, to categorize, to generate plans, to solve problems; to predict future situations, to make decisions and select among alternatives; to learn, to memorize, and so on), which altogether produce the effect of intelligence [26].

Another commonly shared belief is that human cognitive abilities are all a product of human brain activity. Brain—as it is generally agreed and accepted—is busy with information processing. So, we can accept—Intelligence is a product of the brain’s information-processing activity. More generally—**Intelligence is the ability to process information** [27]. In such a form, the definition is applicable to all domains of natural living beings and to artificial human-made designs as well.

3 And Now: What the Hell Is Information?

Although the term “Information” is widespread and extensively used today, no one actually knows what it means and what it actually stands for.

The concept of “information” was at first introduced by Shannon in 1948. Then other scientists joined the mission—Kolmogorov, Fisher, Chaitin. However, none of them was ready to define what is “information”. They **were busy with the “measure of information”**. That was enough to improve the performance and reliability of technical communication systems.

In modern sciences, the needs of communication cannot be reduced only to the optimization of the system’s technical parameters. The semantic aspects of the message are of a paramount importance, and thus must be met.

In accordance with the soul and spirit of these requirements, I have developed my own definition of information. (Interested readers can look into the Refs. [25, 28, 29]).

My definition of information sounds today like this:

“Information is a linguistic description of structures observable in a given data set.”

In a data set, the data elements are not distributed randomly, but due to the similarity of their physical parameters, are naturally grouped into some kind of clusters or cliques. I propose to call these clusters **primary or physical data structures**.

In the eyes of an external observer, these primary data structures are arranged into larger and more complex agglomerations, which I propose to call **secondary data structures**.

These secondary structures reflect the observer’s view of the grouping of primary data structures, and therefore they could be called **meaningful or semantic data structures**.

While the formation of primary (physical) data structures is determined by **the objective (natural, physical) properties of the data**, the subsequent formation of secondary (semantic) data structures is **a subjective process governed by the conventions and habits of the observer** (or a mutual agreement of an observers’ group).

As said, **the description of the structures observed in the data set should be called “Information”**. In this regard, it is necessary to distinguish between **two types of information—physical information and semantic information**.

Both are language descriptions; however, physical information can be described using a variety of languages (recall that mathematics is also a language), and semantic information can be described only using the observer’s natural language (see [25] for more details). Information processing is carried out in a hierarchical fashion, where the semantic information of a lower level is transferred to the next higher level, where it becomes part of a structure of higher complexity. This agglomeration is carried out **according to subjective rules fixed in a prototypical (referential) structure called the observer’s memory**, which is stored in the neuron’s body.

An important consequence of the above definition of information is the understanding that **information descriptions always materialize as a set of words, a fragment of text, a narrative**. In this regard, an important note should be made—in biological systems, these text sequences are written with nucleotide letters and amino acid signs.

This turns the information into a physical entity, into a “thing”, with its weight, length and other physical properties. For the purposes of our discussion, this is an extremely important remark.

So: The brain is processing information. Neurons are the functional units that perform the duty. Despite their discrete structure, neurons are not separate functional units—successful information processing requires close cooperation between work partners. For this reason, neurons are connected in a network in which they communicate with each other, transmitting, exchanging, transferring—in a word—jointly processing information.

4 GPT-Based AI Tools in the Light of Information Processing

As it was just explained above, Intelligence is a product of Information processing, where Information is a linguistic description of observable data structures. The bulk of information processing in the brain is Semantic information processing.

In this regard, the reliance of GPT-based implements on Large Language models (LLMs) seems natural and justified. (Because Semantic information is a linguistic description, a text, a narrative).

But a closer look at the LLM’s paradigms reveals that linguistic components in the training set of the LLM models appear as a set of intentionally broken small word chunks called tokens, or model parameters. The size of model parameters is continuously increasing—GPT-4 is a ~ 600 billion parameter model (Some people suggest it’s a trillion). Earlier models (GPT-3/3.5) had about 185 billion.

In such a case you cannot speak about a language model, it looks more like a statistical data model. As such (and so it is stated in the relevant literature)—“These statistical models need to be trained with large amounts of data to produce relevant results” [10].

Deep Learning (Machine Learning) mechanisms used for training the LLMs are also a valuable argument in this regard—Machine learning principles are applicable only to data patterns mining and discovery.

The description of data patterns (structures) is called **physical information**. Intelligence, as you remember, is busy with **semantic information processing!** Therefore, we can conclude that ML neural networks are appropriate only for physical information running, and **not for semantic information processing**, which is required for proper Intelligence handling.

The “black box” nature of Large Language Models, as well as all other designs relying on the Artificial Neural Network paradigm, cannot be suddenly overturned, enhanced, and become an Explainable Neural Network model. Data-driven physical information descriptions cannot in a moment be transformed into semantic information descriptions, that is, into a linguistic description, a piece of text, a language sample. (That is what GPT-based AI devices claim to be able to do).

Another popular and false claim is that GPT-based AI tools are close to be considered as General AI accomplishments. Again, that is a popular and widespread misunderstanding. Semantic information processing is done by following the rules saved as a reference prototype (a memory-saved reference) of the current-level semantic structure. These rules, these memories are the observer’s private property. That is, semantic information

processing is always subjective. That is, the idea of General Intelligence is wrong and cannot be implemented in any natural or artificial construct. Multiple Intelligence (just mentioned above) being a composition of several different subjective semantic information processing structures can be seen as a valid feature of the future Human-Level Artificial Intelligence.

5 Some Concluding Remarks

The hype around GPT AI tools does not fade. But the purpose of this paper is not to take part in the ongoing discussion, not to glorify or disapprove generative AI tools' merits or demerits. I am trying to stay on the technical side of the story and to analyze GPT AI tools' technical realities (and myths).

In this regard, it is perfectly clear that the reliance of GPT AI tools on Deep learning (Machine learning) training techniques dismisses any claim of its association with Natural Language Processing—NLP assumes semantic information processing, while Deep learning is busy with endless enormous data processing.

The “black box” nature of such processing makes its results erroneous and unpredictable. That is the source of most GPT AI tools' faults and failures. And these flaws cannot be repaired at further stages of processing. They are fundamental, ground-based, built-in. They are forever.

DARPA's 4-years (2015–2019) attempt to create explainable Deep learning tools (explainable AI (XAI) tools) has failed forever [30].

In this regard, it will be interesting to recall that Ali Rahimi, the 2017 NIPS Award winner, in his award speech, declared that “**the current practice in machine learning is akin to alchemy**” [29].

If Machine learning is alchemy, then what can be said about the concept of Intelligence? (Intelligence, as you now know, is a product of information processing, semantic, not physical, information processing).

However, the designers of GPT-based appliances stubbornly claim that “Intelligence is a multifaceted and elusive concept that has long challenged psychologists, philosophers, and computer scientists. There is no generally agreed-upon definition of intelligence, but one aspect that is broadly accepted is that intelligence is not limited to a specific domain or task, but rather encompasses a broad range of cognitive skills and abilities” [16].

In the continuation of this, just given, quote, they claim that such an approach to Intelligence is coherent with the grandfathers' intentions, expressed in the “Proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955” (signed by John McCarthy, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon).

This argument does not seem credible. Especially, if we recall that grandfathers' views of the problem were very blurred and preliminary. They have failed to assess the complexity of the task (Intelligence definition) and the mutual contradictions between its basic constituents. Recall the story of how Minsky proposed to hire a student to solve the problem of vision during the student's summer vacations. (The problem even now remains unsolved) [26]. Recall how Shannon rejected the slightest notion of semantic information: “These semantic aspects of communication are irrelevant to the engineering

problem... It is important to emphasize, at the start, that we are not concerned with the meaning or the truth of messages; semantics lies outside the scope of mathematical information theory” [31].

Despite the proud claims of GPT designers that they were lucky “to give a preliminary assessment of GPT-4’s intelligence, which is an arduous task given the lack of formal definition for this concept, especially for artificial systems” [16]. Any sign of understanding that Intelligence is a product of information processing, and that Information, in its turn, is a composition of two, physical and semantic information entities, does not appear in the GPT designers’ statement. And that is another source of GPT tools failures.

Finally, I hope that the answer to the question posed in the title of this paper is perfectly clear—GPT-based AI tools do not present any threat or danger to the AI tools design community. The hype will fade, and BICA*AI will continue to pursue its goals—to design tools that will allow it to emulate Human-level Artificial Intelligence.

References

1. Biologically inspired cognitive architectures, From Wikipedia, free encyclopedia. https://en.wikipedia.org/wiki/Biologically_inspired_cognitive_architectures
2. Kotseruba, I., Tsotsos, J.K.: 40 years of cognitive architectures: core cognitive abilities and practical applications, *Artif. Intell. Rev.* **53**, 17–94 (2020). <https://link.springer.com/article/10.1007/s10462-018-9646-y>
3. Wikipedia.ChatGPT. https://en.wikipedia.org/wiki/ChatGPT#cite_note-99
4. Lin, T., et al.: A survey of transformers. <https://arxiv.org/pdf/2106.04554.pdf>
5. Wikipedia. OpenAI, <https://en.wikipedia.org/wiki/OpenAI>
6. Wikipedia. Chatbot, <https://en.wikipedia.org/wiki/Chatbot>
7. Xiang, C.: Just Another hype cycle’, 1 Mar 2023,
8. <https://www.vice.com/en/article/qjkgym/elon-musk-based-ai>
9. Bender EM et al On the Dangers of Stochastic Parrots,
10. <https://dl.acm.org/doi/abs/10.1145/3442188.3445922>
11. Alejo Joše, G. Sison, et al, ChatGPT: More than a Weapon of Mass Deception
12. <https://arxiv.org/pdf/2304.11215.pdf>
13. Floridi, L., Chiriatti, M.: GPT 3: its nature, scope, limits, and consequences, minds and machines, 2020. Springer, <https://web.archive.org/web/20210429044930id>
14. McGuffie, K., et al.: The radicalization risks of GPT-3 and advanced neural language models, 16 Sept 2020, <https://arxiv.org/pdf/2009.06807.pdf>
15. The Open Letter to Stop ‘Dangerous’ AI Race Is a Huge Mess
16. <https://www.vice.com/en/article/qjvppm/the-open-letter-to-stop-dangerous-ai-race-is-a-huge-mess>
17. Open letter, 30 Mar 2023, https://gigazine.net/gsc_news/en/20230330-pause-ai-training
18. Transcript: Senate Judiciary Subcommittee Hearing on Oversight of AI
19. <https://techpolicy.press/transcript-senate-judiciary-subcommittee-hearing-on-oversight-of-ai/>
20. Reuters: Which countries are trying to regulate artificial intelligence? Updated: 03/05/2023
21. <https://www.euronews.com/next/2023/05/03/which-countries-are-trying-to-regulate-artificial-intelligence>
22. Bubeck, S., et al.: Sparks of artificial general intelligence: early experiments with GPT-4. <https://arxiv.org/pdf/2303.12712.pdf>

23. Zorpette, G.: GPT-4, AGI, and the Hunt for Superintelligence Neuro expert Christof Koch weighs AI progress, <https://spectrum.ieee.org/superintelligence-christoph-koch-gpt4>
24. Floridi, L.: From Wikipedia, the free encyclopedia
25. https://en.wikipedia.org/wiki/Luciano_Floridi
26. Heikkilä, M., LeCun, Y.: has a bold new vision for the future of AI, 24 June 2022, Yann LeCun's big bet for building intelligent machines. MIT Technology Review
27. Euronews: Why is 'godfather of AI' Geoffrey Hinton worried? Updated: 04/05/2023
28. <https://www.euronews.com/next/2023/05/04/why-is-godfather-of-ai-geoffrey-hinton-worried>
29. Zorpette, G.: Just calm down about GPT-4 z23. <https://spectrum.ieee.org/gpt-4-calm-down>
30. Diamant, E.: Biologically inspired image information content, Jan 2002. <https://www.researchgate.net/publication/277292075>
31. Diamant, E.: machine learning: when and where the horses went astray? (2009). <https://doi.org/10.5772/9156>. <https://www.researchgate.net/publication/45882987>
32. Diamant, E.: Unveiling the mystery of visual information processing in human brain, *Brain Res.* **1225**, 171–178 (2008). <https://arxiv.org/abs/0807.0337>
33. Diamant, E.: The brain is processing information, not data: Does anybody knows about that? Submitted to MDPI "Information", rejected, Dec 2015. <https://www.researchgate.net/publication/291352419>
34. Diamant, E.: Advances in artificial intelligence: are you sure, we are on the right track? Feb 2015. <https://doi.org/10.14738/tnc.54.3562>. <https://www.researchgate.net/publication/272478913>
35. Diamant, E.: Designing artificial cognitive architectures: brain inspired or biologically inspired? In: BICA 2018 Conference, Prague, Aug 2018. <https://www.researchgate.net/publication/329582475>
36. Diamant, E.: Shannon's definition of information is obsolete and inadequate. It is time to embrace Kolmogorov's insights on the matter, Conference Paper, Nov 2016. <https://www.researchgate.net/publication/311223095>
37. Diamant, E.: Artificial neural networks: a bio-inspired revolution or a long-lasting misconception, a rejected conference submission, Feb 2018. <https://www.researchgate.net/publication/322852662>
38. Gunning, D., et al.: DARPA's explainable AI (XAI) program: a retrospective. Authorea: 15 Nov 2021. <https://doi.org/10.22541/au.163699841.19031727/v1>. <https://www.authorea.com/users/446266/articles/545563>
39. Diamant, E.: Computational Intelligence: are you crazy? Since when has intelligence become computational? Conference paper. <https://doi.org/10.1109/SSCI.2016.7850143>. <https://www.researchgate.net/publication/311678243>



Gossiping Until You Get Tired of It: A Network Model of the Adaptive Exchange of Rumors in a Small Scale Social Environment

Karley Dionne¹, Maya Vermeer², and Jan Treur³(✉)

¹ Department of Health, University of New Brunswick, Fredericton, Canada

² Department of Chemistry, Davidson College, Davidson, USA

³ Department of Computer Science, Vrije Universiteit Amsterdam, Social AI Group, Amsterdam, Netherlands

j.treur@vu.nl

Abstract. The spread of rumors, otherwise known as gossiping, is an inevitable part of life for most people. Therefore, it is important to understand the way that information is actually spread through social environments of smaller proportions. This spread is modeled using a Higher-Order Adaptation Social Network, utilizing states for “people” and their boredom. Representation states for their connection weights and their speed factors are also used to represent the adaptivity of the model. Different scenarios are modeled as real-life examples to better illustrate how a rumor is spread when exposed to different groups of people. The model shows how gossip is shared whenever a new rumor presents itself, and how the cyclic nature of social environments contributes to the spread itself.

Keywords: Rumors · Gossip · Modeling social networks

1 Introduction

The spread of rumors has been a critical part of society and life for centuries. It is an important activity for members of any walk of life, no matter culture, age, nor gender, and evidence shows it has been around as long as humanity itself [1]. A rumor, within the social sciences, can be defined as a piece of information that has yet to be confirmed by any sort of reliable source [7]. Their truth can be - and should be - questioned.

However, they are still something that is rampant in society due to their efficiency in sharing information. In the past, rumors have played major roles in historical events [1]. They are argued to be important parts of a successful workplace or other organization, given that they allow employees to influence one another in a way outside of the work itself [4, 5, 8]. They have been modeled in the past the same way the spread of a virus or an epidemic would be [11].

There is a great deal of literature pertaining to modeling the spread of rumors through a network of social media, or online in general [5, 12]. There is also some literature about the spreading of rumors in much larger networks [9]. However, there is not a lot

of information about modeling gossiping on a smaller scale, nor about the spread in a real-life scenario. This gap in the literature is what we are interested in.

The paper that follows is divided into our methodology, a description of the network-modeling system used, our main simulation, some simulations conducted as evidence for the success of our model, and discussion pertaining to the possible extension options for the model and the future of it in general.

2 Background

2.1 The Structure of a Rumor

Humans are created with an urge to complete a story regardless if we actually have all or even the correct information, naturally leaving us to replace the unknowns with our perception of the missing information, which creates rumors [3]. The formation of a rumor relies on three main aspects: the way in which the rumor is transmitted, the information that it holds and the emotional satisfaction it gives [6]. A highly researched model of transmission is social media, which arguably is one of the most common ways a rumor is shared, but we are focused more on word of mouth transmission for our simulation. Although the kind of information the rumor contained and the emotional satisfaction was not relevant to the model we created, it is still important to note those two aspects play big roles in the effects caused by rumors as well as the longevity of a rumor.

2.2 The Effect of Rumors

Depending on the level of severity, rumors and misinformation can have severe ramifications. Rumors have had large impacts on society - they have even been known to change the way people recall history events, Covid-19 being a recent example. Many people recall different numbers of cases, deaths and how the virus came about as there were a lot of rumors on what the truth was [11]. On a lower level, when rumors begin in a professional setting such as a workplace it can have negative effects on job satisfaction and production efficiency [8]. Rumors within friendship can create an untrusting friendship or even end the friendship [10]. What effects rumors cause have been researched thoroughly in many aspects, so our model shifts the focus to how human behavior affects the rumor itself.

3 Methodology

The modeling approach used is a network-oriented one [8]. The approach utilizes states (otherwise known as nodes), each of which have levels of activation that are varied over time. It also has the network characteristics: connection weights ω (for *connectivity*), combination functions \mathbf{c} (for *aggregation*) and speed factors η (for *timing*) which make up the network model structure.

Connection Weights: $\omega_{X,Y}$ denotes the weight of the connection from state X to Y . It usually is in the range from -1 to 1. It indicates the strength of the causal relation.

Combination Functions: $c_X(\dots)$ denotes what combination function was used for state X . They are used if there are multiple connections to one node and specify the different ways of combining. In our model, the functions **alogistic** $_{\sigma,\tau}$ and **stepmod** $_{\rho,\delta}$ are used, see Table 1.

Speed Factors: η_X denotes the speed of change of state X due to the impact coming in from other states. They normally range from 0–1 and they are used to create dynamics for the nodes of a network.

To design a network model, the above network characteristics are specified in the so-called role matrix format according to the role they play in the network. The connectivity characteristics are specified in role matrix **mb** (for base connectivity) and **mcw** (for connection weights). The aggregation characteristics are specified in role matrix **mcfw** (for combination function weights) and **mcfp** (for combination function parameters). Finally, the timing characteristics are specified in role matrix **ms** (for speed factors) and in **iv** (for initial values). See the Appendix Sect. 7 for examples of this.

Table 1. The combination functions used in the introduced network model

	Notation	Formula	Parameters
Advanced logistic sum	alogistic $_{\sigma,\tau}(V_1, \dots, V_k)$	$\left[\frac{1}{1+e^{-\sigma(V_1+\dots+V_k-\tau)}} - \frac{1}{1+e^{\sigma\tau}} \right] (1 + e^{-\sigma\tau})$	Steepness σ Excitability threshold τ
Stepmod	stepmod $_{\rho,\delta}(V)$	0 if $0 \leq t \bmod \rho \leq \delta$ 1 else	Time t repetition ρ Step time δ

The dynamics of the states are based on the following canonical difference equation:

$$Y(t + \Delta t) = Y(t) + \eta_Y [c_{\pi_Y, Y}(\omega_{X_1, Y} X_1(t), \dots, \omega_{X_k, Y} X_k(t)) - Y(t)] \Delta t \quad (1)$$

This generic equation is the core of the network engine within the dedicated software environment that is available.

Utilizing the notion of self-modeling network, also called reified network, as introduced in [8], each network characteristic can be made adaptive, an important part of our intended model. This can be done by adding a self-model state into the network representing the value of the characteristic. We will be creating a second-order self model using these techniques. For the states X and Y , first-order self-model states known as $\mathbf{W}_{X,Y}$ included in the network represent the value of connection weight $\omega_{X,Y}$. Adaptation of the network structure takes place because then in (1) the value $\mathbf{W}_{X,Y}(t)$ is used for $\omega_{X,Y}$. That $\mathbf{W}_{X,Y}$ plays this role for a specific connection is specified in role matrix **mcw**, by writing in the cell for that connection not a specific value but the X_i relating to $\mathbf{W}_{X,Y}$. Also second-order self-model states $\mathbf{H}\mathbf{w}_{X,Y}$ are included in the network model, they represent the value of the speed factor of $\mathbf{W}_{X,Y}$, a learning rate.

4 The Second-Order Adaptive Network Model

To illustrate the pattern rumor sharing creates across a small population, a higher-order social network model was constructed. Firstly, a base layer was decided upon - this layer included the nodes of each “person” state, as well as their “boredom states”. These are represented P1, P2, P3 and P4, along with B1–B4. The number of nodes in the model was kept limited to a small number of subjects (for example a group of friends or a family), but more states (for both people and boredom) could easily be added if necessary. Along with these states, there is also the node wherein the rumor actually takes place, represented by the label RH “rumor happening”. This is the first node on the layer, and can be viewed as something that is “rumor-worthy” happening, perhaps to or in front of Person A. This person then shares the information gained (however true it may or may not be) with Person 1.

This is where the cyclical nature of the model begins. In the model, Person 1 begins sharing the rumor with the other people in the cycle (P2, P3 and P4). As the new people gain the information, they begin sharing it with each other as well. However, human nature must be added to the artificial simulation - the state of boredom. A human will normally not be interested in the same information after it gets old, per se. Therefore, we also added the boredom states in on this level (B1–B4), connecting them with their P states respectively. These states are not linked to anything but their respective P states.

In order to make our model adaptive, reification levels of the first and second order were added. On the first reification level, the representation state for connection weights were added. The connection weights acted upon on this level are only three; the connection between P1 and P2, the connection between P2 and P3, and the connection between P3 and P4. These are known as $\mathbf{W}_{P1,P2}$, $\mathbf{W}_{P2,P3}$, and $\mathbf{W}_{P3,P4}$. The model could be added to by adding more states on the second reification level in reference to the other connections formed on the base level in the future, but as far as we’ve demonstrated, that is not necessary. On the second reification level, the reified representation state for Speed Factors with respect to the W states were added. Again, we only added representation states for the states present in the second reification level. These are referred to as $\mathbf{H}_{\mathbf{W}_{P1,P2}}$, $\mathbf{H}_{\mathbf{W}_{P2,P3}}$, and $\mathbf{H}_{\mathbf{W}_{P3,P4}}$. A legend of all state names with their explanation can be found as Table 2. A picture of the model itself can be seen in Fig. 1. A full specification by role matrices can be found in the Appendix Sect. 7.

The values were chosen based upon a few factors - the initial values were done in groups, keeping the nodes on the same order level at the same value so we could ensure no differences based only on something arbitrary. The initial values for the people and boredom states were set at 0, as well as the RH state. This is because prior to the event actually taking place, none of them would have anything to have a reaction (like gossiping) to. The second and third order nodes were given slightly higher values in order to increase visibility on their reactions, and because those connections would exist before the rumor begins, even if they are not necessarily taking effect.

Table 2. Legend of states in the model

State nr	State name	Explanation	Level
X1	Rumor happening (RH)	Something happens to person A that is rumor-worthy, they tell Person 1	Base level
X2	Person 1 (P1)	Person 1 begins sharing rumor	
X3	Person 2 (P2)	Person 2 begins sharing rumor	
X4	Person 3 (P3)	Person 3 begins sharing rumor	
X5	Person 4 (P4)	Person 4 begins sharing rumor	
X6	Boredom 1 (B1)	Boredom 1 suppresses P1's gossiping	
X7	Boredom 2 (B2)	Boredom 2 suppresses P2's gossiping	
X8	Boredom 3 (B3)	Boredom 3 suppresses P3's gossiping	
X9	Boredom 4 (B4)	Boredom 4 suppresses P4's gossiping	
X10	Wp1p2	Reified representation state for connection weight p1, p2	1 st order reification level
X11	Wp2p3	Reified representation state for connection weight p2, p3	
X12	Wp3p4	Reified representation state for connection weight p3, p4	
X13	HWp1p2	Reified representation state for speed factor Wp1p2 for reified representation state Wp1p2	2 nd order reification level
X14	HWp2p3	Reified representation state for speed factor Wp2p3 for reified representation state Wp2p3	
X15	HWp3p4	Reified representation state for speed factor Wp3p4 for reified representation state Wp3p4	

5 The Simulation

The simulation was run in MATLAB using the role matrices in Table 3.

The connection weights for the base level nodes were mostly 1 (a perfect connection), not including the connection between the boredom states to their respective people, which we made a very strong negative conclusion in order to force the boredom states to actually act upon the people. The connection between the first reification level nodes

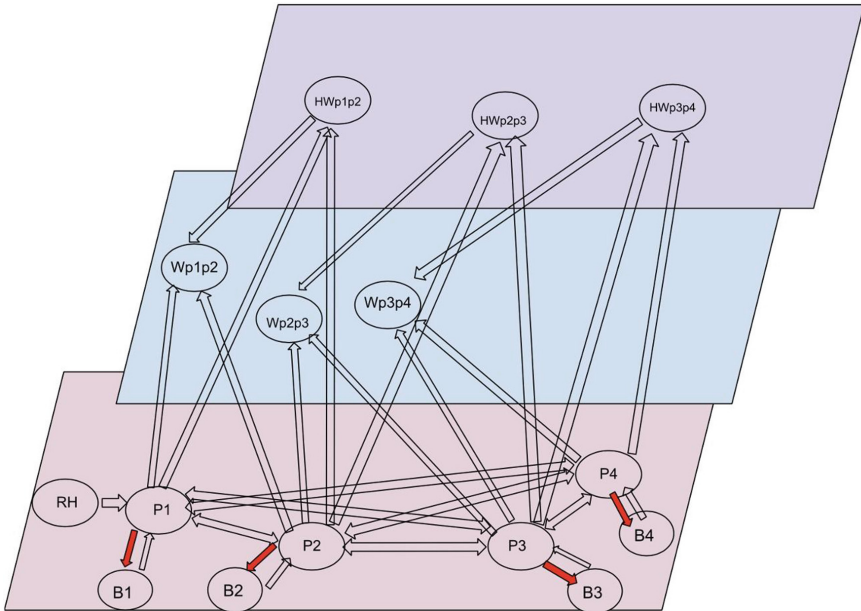


Fig. 1. The 3D conceptual representation of the model

$\mathbf{W}_{P1,P2}$, $\mathbf{W}_{P2,P3}$, and $\mathbf{W}_{P3,P4}$ back to themselves (otherwise known as their persistence links) are just below 1 (0.9) because realistically, they would not have perfect persistence values. The second-order nodes $\mathbf{H}\mathbf{W}_{P1,P2}$, $\mathbf{H}\mathbf{W}_{P2,P3}$, and $\mathbf{H}\mathbf{W}_{P3,P4}$ have connections from the related base states with much lower weights at 0.4, which allows them to act on the states without having an effect on the main pattern. This follows the second-order adaptation principle ‘Adaptation accelerates with increasing stimulus exposure’ formulated in [7]. The speed factor for the RH node was kept high because no matter what else happens, we wanted the rumor to always be present. The speed factors for P1–P4 were set much higher than B1–B4 so that we can actually see the effect that the boredom states are having on the people in the resulting graph. The nodes for $\mathbf{H}\mathbf{W}_{P1,P2}$, $\mathbf{H}\mathbf{W}_{P2,P3}$, and $\mathbf{H}\mathbf{W}_{P3,P4}$ were put at a smaller value for the same reason. We added in a **stepmod** function (see Table 1) to the combination function parameters only for the rumor happening so that we could set the rumor to be repeating within our result and see differences between the rounds. As for the threshold values in our **logistic** function, we set the boredom states to be much lower than the people states, so that they could act upon the people with success. The connection weight representation states were set low so that they would be triggered early, and the opposite is true for the speed factor representation states. These role matrices gave results shown in the graph in Fig. 2.

The graph shows what we expected. Every time RH takes place, P1–P4 follow shortly after, before getting bored (shown as B1–B4 rising) and go back down, staying there until RH is activated again. The connection weight representation states rise in a step-like pattern, plateauing whenever the person states lower, until rising again when they are

Table 3. Role matrices used for scenario 1: speed factors and parameters.

ms		Speed factors			
X1	RH			0.9	
X2	P1			0.9	
X3	P2			0.8	
X4	P3			0.7	
X5	P4			0.6	
X6	B1			0.1	
X7	B2			0.1	
X8	B3			0.1	
X9	B4			0.1	
X10	Wp1p2			X13	
X11	Wp2p3			X14	
X12	Wp3p4			X15	
X13	HWp1p2			0.3	
X14	HWp2p3			0.3	
X15	HWp3p4			0.3	
mcfp	Combination function parameters	Alogistic 1 2		Stepmod 1 2	
X1	RH			80	60
X2	P1	5	0.6		
X3	P2	5	0.7		
X4	P3	5	0.8		
X5	P4	5	0.9		
X6	B1	5	0.5		
X7	B2	5	0.5		
X8	B3	5	0.5		
X9	B4	5	0.5		
X10	Wp1p2	5	0.2		
X11	Wp2p3	5	0.2		
X12	Wp3p4	5	0.2		
X13	HWp1p2	5	1.5		
X14	HWp2p3	5	1.5		
X15	HWp3p4	5	1.5		

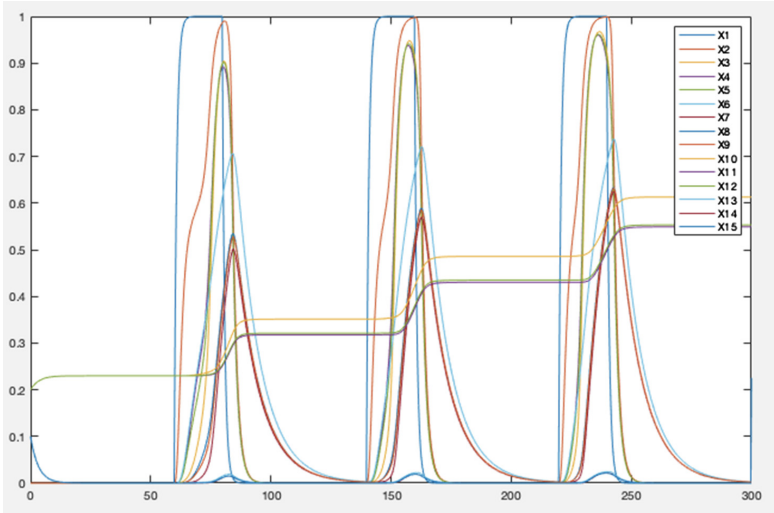


Fig. 2. The graph of the initial simulation

re-activated. The speed factor representation states stay lower to the bottom, only rising when they are stimulated by the person states.

6 Simulation Scenarios

Four simulation scenarios were conducted, to observe what happens to the course of the rumors when behavior changes. In scenario one the desire of people wanting to spread the rumor was increased where in scenario two it was decreased. In scenario three the boredom state was removed from person four. Lastly scenario four wanted to simulate what occurs if the boredom state rises for person one, specifically if the rumor will get spread.

6.1 Scenario 1

In the first scenario, the people (P1–P4) were made to be more eager to spread the rumor than in the original. This was simulated by increasing the speed factors of P1, P2, P3, P4, with (P1) starting at 0.9 and then P2 at 0.8 and so on. The thresholds of P1, P2, P3 were also decreased to be P1, 0.6 P2, 0.7 P3, 0.8 and P4, 0.9. By changing the speed factors and thresholds in this way from the original value of 0.5, the people are all, by a small factor, more willing to spread the rumor faster. The crucial matrices with settings are shown in Table 3.

The graph (see Fig. 3) that this simulation produced showed only small differences from the original graph - the lines for P1–P4 rose a bit faster, steeper given that each state is stimulated at a higher rate. They also are all closer knit to each other, with less

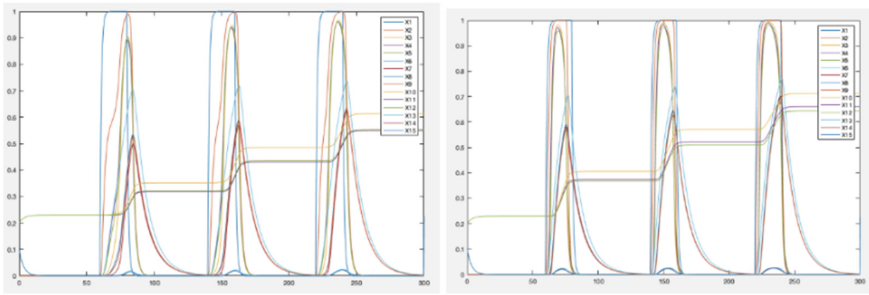


Fig. 3. Comparison of the initial graph to the results of the simulation for scenario 1

difference between all of the produced lines. However, the overall pattern of the graph did not stray much from the original graph - they are compared for convenience below.

6.2 Scenario 2

In contrast to what was conducted in scenario one, for scenario two the interest was decreasing people’s one to four desire to spread rumors. For this simulation to run, the speed factors and the threshold values for the combination function parameters were both changed; see Table 4.

In comparison to scenario one the speed factors of scenario two were changed in opposite manner, with P1 starting at 0.6 and increasing by 0.1, therefore P2 followed with 0.7 and so on until P4. The combination function parameter had a decreasing change starting with P1 at 0.9 and decreasing by 0.1. Changing these values meant that the people were all slightly less likely to spread the rumor, and to do it at a shorter pace.

As shown in the graph in Fig. 4, the results for the second simulation only produced small changes compared to the original. A notable difference is the close proximity of the lines from one another, and wherein the graph for Scenario 1 shows a vaguely steeper climb for P1–P4, this graph shows the opposite.

6.3 Scenario 3

In the previous scenarios, all the people had boredom states that allowed them to lose interest in the rumor, therefore would ultimately prevent them from continuing the rumor. The interest in scenario three were the effects of removing the boredom state from P4. To demonstrate this effect, changes to the base connectivity and connection weight were made which can be seen in the Tables 5, 6 and 7 below. To remove the boredom state from P4 (X9, B4) was made blank in the base connectivity table below (Table 5). As a result of that change (X5, P4) lost its connection to X9 in the fifth column. Therefore, the connection weight was also removed from X5 in the fifth column.

After running this simulation (see Fig. 5), the results given was what was expected. The line representing person four X5 peak lasts much longer than it did previously with

Table 4. Role matrices **ms** and **mcfp** for Scenario 2

ms		Speed factors			
X1	RH			0.9	
X2	P1			0.6	
X3	P2			0.7	
X4	P3			0.8	
X5	P4			0.9	
X6	B1			0.1	
X7	B2			0.1	
X8	B3			0.1	
X9	B4			0.1	
X10	Wp1p2			X13	
X11	Wp2p3			X14	
X12	Wp3p4			X15	
X13	HWp1p2			0.3	
X14	HWp2p3			0.3	
X15	HWp3p4			0.3	
mcfp	Combination function parameters	Alogistic 1 2		Stepmod 1 2	
X1	RH			80	60
X2	P1	5	0.9		
X3	P2	5	0.8		
X4	P3	5	0.7		
X5	P4	5	0.6		
X6	B1	5	0.5		
X7	B2	5	0.5		
X8	B3	5	0.5		
X9	B4	5	0.5		
X10	Wp1p2	5	0.2		
X11	Wp2p3	5	0.2		
X12	Wp3p4	5	0.2		
X13	HWp1p2	5	1.5		
X14	HWp2p3	5	1.5		
X15	HWp3p4	5	1.5		

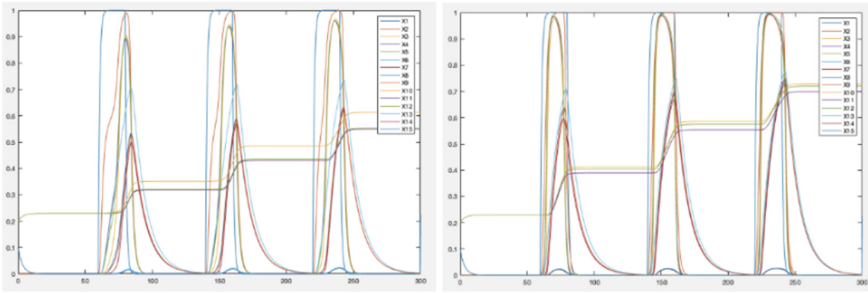


Fig. 4. Comparison of the initial graph to the results of the simulation for Scenario 2

the boredom state. Under the conditions of this model, this was expected because person four’s ability to get bored was taken away therefore their interest in the rumors lasted much longer. Whereas when person four had a boredom state, the peak in the line did not last as long due to their loss of interest.

6.4 Scenario 4

Lastly, scenario four was made to demonstrate what occurs if a person has a higher boredom state. To simulate this, the connection weight from B1 (which is X6) to P1 (X2) was lowered from -2 to -5 , seen in the role matrix in Table 6. This change caused person one to become bored with the rumor and therefore, they did not continue to spread the rumor to person two.

The graphs show what we expected - if P1 becomes uninterested in the rumor due to their stronger boredom state, then P2-P4 do not become invested either. This is because P1 is the one who gets the original message from the RH node. This is why the resulting graph for this scenario shows that the other person’s states do not reach the same levels as in the original scenario, and they do not stay up for as long. In fact, they barely rise at all. This is realistic - if the main spreader of the rumor becomes bored of it very quickly, the rumor will not travel nearly as far nor will it be a subject of conversation for very long even if the other people in the network would be interested in it if they had the opportunity to be. Comparison of the initial graph to the results of the simulation for Scenario 4 is illustrated in Fig. 6 (see also Table 6).

7 Discussion

Analysis throughout this paper about the spread of rumors through a smaller cohort depends greatly on human behavior which can increase and decrease the rate at which rumors spread. This adaptive model was focused on two main human behaviors, a person’s desire to spread rumors and boredom’s effects on the spread of rumors.

Table 5. The matrices **mb** and **mcwv** for scenario 3

mb	Base connectivity	1	2	3	4	5
X1	RH	X1				
X2	P1	X1	X3	X4	X5	X6
X3	P2	X2	X4	X5		X7
X4	P3	X2	X3	X5		X8
X5	P4	X2	X3	X4		
X6	B1	X2				
X7	B2	X3				
X8	B3	X4				
X9	B4					
X10	Wp1p2	X10	X2	X3	X13	
X11	Wp2p3	X11	X3	X4	X14	
X12	Wp3p4	X12	X4	X5	X15	
X13	HWp1p2	X2	X3			
X14	HWp2p3	X3	X4			
X15	HWp3p4	X4	X5			
mcwv	Connection weights	1	2	3	4	5
X1	RH	1				
X2	P1	1	1	1	1	-2
X3	P2	1	1	1		-2
X4	P3	1	1	1		-2
X5	P4	1	1	1		
X6	B1	1				
X7	B2	1				
X8	B3	1				
X9	B4					
X10	Wp1p2	0.9	1	1		
X11	Wp2p3	0.9	1	1		
X12	Wp3p4	0.9	1	1		
X13	HWp1p2	0.4	0.4			
X14	HWp2p3	0.4	0.4			
X15	HWp3p4	0.4	0.4			

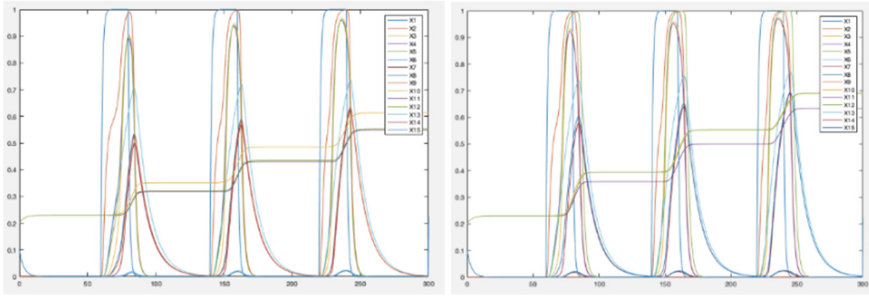


Fig. 5. Comparison of the initial graph to the results of the simulation for scenario 3

Table 6. Role matrix *mcw* for Scenario 4

<i>mcw</i>	Connection weights	1	2	3	4	5
X1	RH	1				
X2	P1	1	1	1	1	-5
X3	P2	X10	1	1		-2
X4	P3	1	X11	1		-2
X5	P4	1	1	X12		-2
X6	B1	1				
X7	B2	1				
X8	B3	1				
X9	B4	1				
X10	Wp1p2	0.9	1	1		
X11	Wp2p3	0.9	1	1		
X12	Wp3p4	0.9	1	1		
X13	HWp1p2	0.4	0.4			
X14	HWp2p3	0.4	0.4			
X15	HWp3p4	0.4	0.4			

7.1 Contributions to Current Research

As social media has been on the rise in the past decade, there is a large quantity of research done on the spread of rumors focusing on the model of transmission through social media. However, there has been a lack of research done on the spread of rumors through word of mouth, which we would argue is just as, if not more, important than the modeling of rumors spreading through social media. This paper contributes to the need for such research, as our model is based upon a smaller community who is spreading the rumor via word-of-mouth.

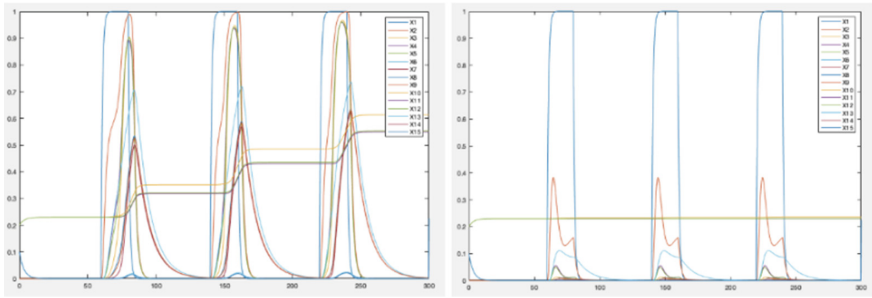


Fig. 6. The comparison of the initial graph to the results of the simulation for scenario 4

This model also provides a new perspective on what happens when certain human behaviors are suppressed or heightened. Current models that exist pertaining to rumor-spreading do not have suppression techniques for a natural human state - boredom. This is something our model also seeks to resolve. By adding to the boredom states for each individual person, we are adding a whole new level to the modeling rumor-spreading community.

7.2 Call for Future Research

Future research could continue examining how other natural human behaviors affect the rumors spread with a similar model as seen in this paper. Examining differences in human behaviors of multiple subjects occurring at the same time may also be intriguing to future research. It is important to note that our model can be utilized in this research, simply by adding more person states (with boredom states) in order to model a slightly larger scale or by adjusting speed factors and thresholds to account for different human reactions. Furthering this research would allow a better understanding of human behavior's role in the spread of rumors, which could be valuable to many human resource departments in work places, educational settings and social environments as having a better understanding will better equip those handling the impacts of rumors.

In summary, there will always need to improve the development of the human artificial intelligence systems as human behavior is forever evolving, causing there to be an infinite amount of possibilities. Future researchers can utilize our model but change certain values to allow them to better represent and model the specific community they have in mind as time and society itself changes.

Appendix

See Table 7.

Table 7. Overview of the role matrices.

ms	Speed factors	1				
X1	RH	0.9				
X2	P1	0.6				
X3	P2	0.7				
X4	P3	0.8				
X5	P4	0.9				
X6	B1	0.1				
X7	B2	0.1				
X8	B3	0.1				
X9	B4	0.1				
X10	Wp1p2	X13				
X11	Wp2p3	X14				
X12	Wp3p4	X15				
X13	HWp1p2	0.3				
X14	HWp2p3	0.3				
X15	HWp3p4	0.3				
mcw	Connection weights	1	2	3	4	5
X1	RH	1				
X2	P1	1	1	1	1	-2
X3	P2	X10	1	1		-2
X4	P3	1	X11	1		-2
X5	P4	1	1	X12		-2
X6	B1	1				
X7	B2	1				
X8	B3	1				
X9	B4	1				
X10	Wp1p2	0.9	1	1		
X11	Wp2p3	0.9	1	1		
X12	Wp3p4	0.9	1	1		
X13	HWp1p2	0.4	0.4			
X14	HWp2p3	0.4	0.4			
X15	HWp3p4	0.4	0.4			

(continued)

Table 7. (continued)

mcfw	Combination function weights	Alogistic		Stepmod	
X1	RH				1
X2	P1		1		
X3	P2		1		
X4	P3		1		
X5	P4		1		
X6	B1		1		
X7	B2		1		
X8	B3		1		
X9	B4		1		
X10	Wp1p2		1		
X11	Wp2p3		1		
X12	Wp3p4		1		
X13	HWp1p2		1		
X14	HWp2p3		1		
X15	HWp3p4		1		
mcfp	Combination function parameters	Alogistic 1 2		Stepmod 1 2	
X1	RH			80	60
X2	P1	5	0.9		
X3	P2	5	0.8		
X4	P3	5	0.7		
X5	P4	5	0.6		
X6	B1	5	0.5		
X7	B2	5	0.5		
X8	B3	5	0.5		
X9	B4	5	0.5		
X10	Wp1p2	5	0.2		
X11	Wp2p3	5	0.2		
X12	Wp3p4	5	0.2		
X13	HWp1p2	5	1.5		

(continued)

Table 7. (continued)

mcfp	Combination function parameters	Alogistic 1 2		Stepmod 1 2	
X14	HWp2p3	5	1.5		
X15	HWp3p4	5	1.5		

References

1. Ben-Noun, L.: Rumors in human life. Research Gate (2021). https://www.researchgate.net/publication/351226458_RUMORS_IN_HUMAN_LIFE
2. Estévez, J.B.A., Wittek, R., Giardini, F., Ellwardt, L., Krause, R.: Workplace gossip and the evolution of friendship relations: the role of complex contagion. *Soc. Netw. Anal. Min.* **12**(1) (2022). <https://doi.org/10.1007/s13278-022-00923-7>
3. Everbridge.: (n.d.). Rumors and misinformation: dispelling myths and creating trust during emergency situations. http://go.everbridge.com/rs/everbridge/images/Rumors_and_Misinformation.pdf
4. Gregg, R.E: Office gossip: a surprising source of liability. *J. Med. Pract. Manage.* (2), 71–74 (2003). <https://pubmed.ncbi.nlm.nih.gov/14596169/>
5. Govindankutty, S., Gopalan, S.P.: SeDis—a rumor propagation model for social networks by incorporating the human nature of selection. *Systems* **11**(1), 12 (2022). <https://doi.org/10.3390/systems11010012>
6. Grosser, T.J., Kidwell, V.L., Labianca, G.: A social network analysis of positive and negative gossip in organizational life. *Group Organiz. Manage.* (2), 177–212 (2010). <https://doi.org/10.1177/1059601109360391>
7. Knapp, R.H.: A psychology of rumor. *JStor* **8**(1), 22–37 (1944). <https://www.jstor.org/stable/2745686>
8. Kong, M.: Effect of perceived negative workplace gossip on employees' behaviors. *Front. Psychol.* **9** (2018). <https://doi.org/10.3389/fpsyg.2018.01112>. Pendleton, S.C.: Rumor research revisited and expanded. *Lang. Commun.* **18**(1), 69–86 (1998). [https://doi.org/10.1016/S0271-5309\(97\)00024-4](https://doi.org/10.1016/S0271-5309(97)00024-4)
9. Robinson, B.L., Harper, N.S., McAlpine, D.: Meta-adaptation in the auditory midbrain under cortical influence. *Nat. Commun.* **7**(1), 13442 (2016)
10. Treur, J.: Modeling higher-order network adaptation by multilevel network reification. In: Treur, J. (ed.) *Network-Oriented Modeling for Adaptive Networks: Designing Higher-Order Adaptive Biological, Mental and Social Network Models*, pp. 99–119 (2020)
11. Zhai, X., Wu, W., Xu, W.: Cascade source inference in networks: a Markov chain Monte Carlo approach. *Comput. Soc. Netw.* **2**(1) (2015). <https://doi.org/10.1186/s40649-015-0017-4>
12. Zhao, L., Wang, J., Chen, Y., Wang, A., Cheng, J., Cui, H.: SIHR rumor spreading model in social networks. *Phys. D: Nonlinear Phenom.* **391**(7), 2444–2453 (2012). <https://doi.org/10.1016/j.physa.2011.12.008>
13. Zhu, H., Ma, J.: Analysis of SHIR rumor propagation in random heterogeneous networks with dynamic friendships. *Phys. D* **513**, 257–271 (2019). <https://doi.org/10.1016/j.physa.2018.09.015>
14. Zou, W., Tang, L.: What do we believe in? Rumors and processing strategies during the COVID-19 outbreak in China. *Public Underst. Sci.* **30**(2), 153–168 (2020). <https://doi.org/10.1177/0963662520979459>

15. Zubiaga, A., Liakata, M., Procter, R., Hoi, G.W.S., Tolmie, P.: Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE* **11**(3), e0150989 (2016). <https://doi.org/10.1371/journal.pone.0150989>



A Socially Acceptable Conversational Agent Based on Cognitive Modeling and Machine Learning

Anatoly A. Dolgikh and Alexei V. Samsonovich[✉]

National Research Nuclear University “MEPhI”, Kashirskoe Shosse 31, Moscow, Russia
avsamsonovich@mephi.ru

Abstract. Large Language Models (LLM) enable recognition of the topic of arbitrary statements, as well as their emotional coloring, but do not “understand” the logic of emotions, despite the fact that they can often generate adequate responses in a given context. On the other hand, cognitive architectures such as eBICA are able to model the dynamics of emotional states in the general case but require assistance in understanding the meaning of statements and generating responses to them. This work introduces a new way to integrate LLM and eBICA, allowing them to complement each other. An experimental study based on the paradigms “virtual receptionist” and “virtual psychologist” is presented, showing encouraging results.

Keywords: LLM · Cognitive architectures · Emotional intelligence · Artificial intelligence

1 Introduction

Intelligent conversational agents, capable of maintaining a conversation with a user and possessing elements of emotionality, are playing an increasingly important role in the life of society, and the demand for them is growing rapidly. Examples of their applications include a personal interlocutor agent [1], a virtual tutor [2], and a rehabilitation conversational agent [3], a robot storyteller that stimulates the development of creative abilities in children [4], a virtual patient for training doctors [5] and others. At the same time, existing conversational agents (including Siri, Alice, ChatGPT, etc.) lack the socio-emotional intelligence that would allow them to be accepted by a person on a social level.

The problem to which this work contributes is to create an artificial social agent that, in addition to performing functions in its subject area, is capable of (a) establishing and maintaining contact with the user at a social level not lower than the level of a person, and (b) internally evaluate the user socially in order to work with him more effectively. For certain applications these abilities are critical.

Today, social conversational agents are implemented either on the basis of statistical machine learning methods, primarily large language models (Large Language Models, LLM) based on deep neural networks, or on the basis of cognitive architectures such as

ACT-RE, extended Soar, EMA [6], eBICA [7] and others. In both cases, difficulties arise in choosing appropriate behavior in an unforeseen social context. Cognitive architectures have to be created manually, and it is impossible to foresee all possible situations when they will be used. To train statistical models, large amounts of data are needed, which in the field of social decision-making are usually difficult to access.

In this work, a combined approach is developed: namely, a way to integrate machine learning and cognitive modeling methods, allowing them to complement each other. The result is demonstrated using the “Virtual Hotel Receptionist” and “Virtual Psychologist” paradigms.

2 Principles of the eBICA Cognitive Architecture

The cognitive architecture eBICA, described previously (for example, [7]), does not represent a large embodied software environment with its own programming language, like Soar, Act-R, Icarus, etc. We deliberately took a different path, developing eBICA as a theoretical framework that today has dozens of specialized implementations, but is constantly changing and expanding, including conceptually, adapting to new tasks and subject areas. Such flexibility would not be possible if implemented as a single large system like Soar. At the same time, in this case, it is the theoretical understanding of the principles of the solution that is of the main value for practice, making it possible to create a technological line for the development of socio-emotional agents for a variety of applications.

Key elements of eBICA, in addition to the typical set of memory systems for cognitive architectures, are semantic maps, mental perspectives, and moral schemas [8]. The principle of operation of the moral scheme is to determine some “normal” state of affairs in a given social environment and to use the means available to the agent to maintain or achieve such a “normal state”. A minimal simplified explanation of the basic principles of the eBICA theory is given in the following paragraphs.

The first step is to introduce estimates for all significant events and actors operating in the environment. Ratings are entered as vectors that take values in a semantic space defined by a set of semantic scales relevant to a given paradigm (for example, in the VAD model these are Valence, Arousal, Dominance: that is, “positivity,” “excitability,” and “dominance”). Let X and Y be vectors of evaluations of two actors, also denoted by the letters X and Y . Suppose the actor X performs a discrete action a in relation to the actor Y . Let a and a^+ be estimates of action a , defined as the expectation of the effect of action a on X and Y , respectively. Let us postulate the following rule for updating estimates of X and Y when performing action a :

$$X := (1 - rw)X + rwa, \quad (1)$$

$$Y := (1 - rw)Y + rwa^+. \quad (2)$$

Here r is a constant, a model parameter, and $0 < r < 1$, and w is a positive value characterizing the significance of the action a . Thus, each action is characterized by two vectors a , a^+ and a scalar w , and these quantities may depend on the context.

Moral schema configuration in eBICA determine feelings towards the actors - participants in the scheme. The concept of feeling F_X is introduced as the values of the actor's assessment X is in good condition [8]. Then it can be shown that in order to maintain a normal state, it is enough for the actor to choose actions whose evaluations are as close as possible to the values of the corresponding feelings and are symmetrically distributed relative to them. In a state close to normal, feelings are fixed. They can change in a conflict situation, when the discrepancy between assessments and feelings is significant and cannot be eliminated by choosing acceptable actions with fixed feelings. The conditions for activation and deactivation of moral schemes are described in [8]. While evaluations and feelings are related to the value system, physiological drives related to the somatic area also play a significant role in the choice of behavior.

Evaluations and feelings may be different in the mental perspectives of different actors. However, tested experimentally [7, 9] Eqs. (1), (2) together with the laws of moral schemes [8] make it possible to calculate the dynamics of interpersonal emotional relationships and thus predict the mental states of agents based on their observed behavior.

In this work, eBICA was used to generate datasets for training neural network components, as described below. This allowed the agent to adequately respond to user behavior.

3 Experimental Paradigm

Within the framework of the hotel virtual receptionist paradigm, the following model situation is defined: a visitor comes to check into a hotel and a dialogue takes place between him and the receptionist at the reception desk. The subject of consideration here is the dynamics of emotional interaction between the user and the virtual registrar, and the role of the latter can be played by both a virtual actor (bot) and a hidden living person (confederate).

The research questions are: Can a virtual actor be considered socially acceptable to the user? Trustworthy? Attractive? Capable of evoking emotions? Having individuality? And also, is it possible to determine the characteristics of a person's personality by his behavior in this paradigm? As a result, can we say that the virtual actor in this paradigm establishes socio-emotional contact at the human level?

The motivation for choosing the "Virtual Psychologist" paradigm [15] can be explained as follows. Currently, there are a large number of questionnaires through which the user can find out his personal characteristics (for example, [10]), but for this he needs to answer a large number of boring questions. An alternative approach that is currently available [16] is based on analysis of a large volume of text written by a person. The "Virtual Psychologist" paradigm has advantages with respect to both approaches. The paradigm is a game in which the user is invited to take part in a dialogue taking place in an imaginary life situation: for example, ordering lunch in a restaurant, choosing furniture in a store, etc. This is done in order to tacitly determine the user's personality type. The main advantages from the user's point of view are accessibility, ethics, fun and ease of taking the test under the condition of anonymity. The research questions are the same as for the previous paradigm.

4 Implementation of the System

4.1 Virtual Receptionist

Structurally, the virtual receptionist can be divided into two components: a graphical interface implemented using virtual reality tools, and a server component that manages the dialogue with the user. This structure helps to separate the visualization of the registrar from the construction of a dialogue with the user, which makes it possible to integrate the server not only with the virtual environment, but also with other means of interaction with users.

The server part (Fig. 1) is the main part of the virtual receptionist. It includes a component for classifying user remarks, a component for generating registrar responses, and a component for processing remarks (the latter performs the function of automatic text editing).

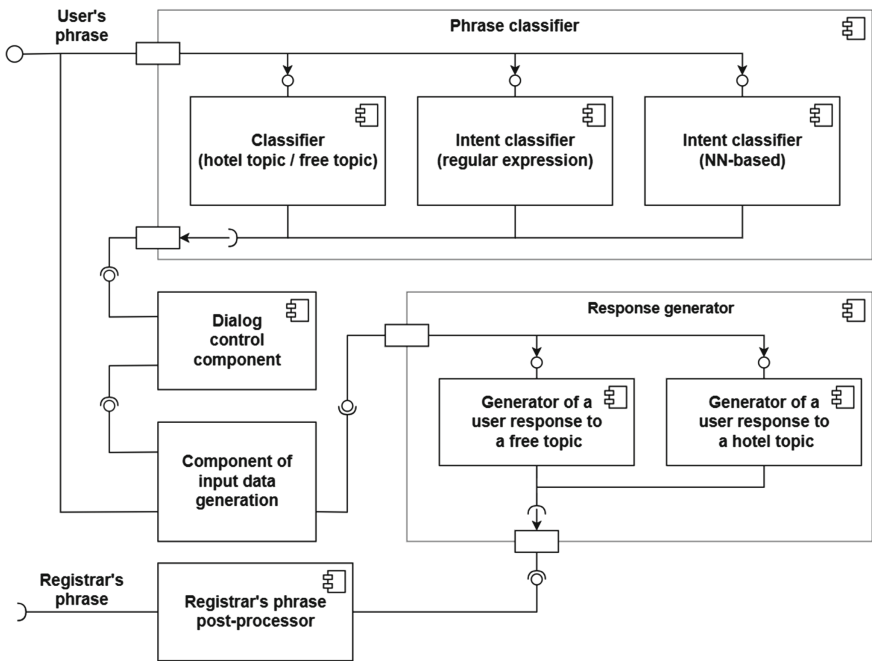


Fig. 1. Structure of the server part of the virtual registrar.

The classification component calls multiple classifiers sequentially. First of all, it is necessary to determine what type of dialogue this replica belongs to. Within the framework of the “Virtual Receptionist” paradigm, two types of dialogue are distinguished: dialogue about checking into a hotel and dialogue on a free topic. Since the list of possible phrases is significantly narrowed within the dialogue about checking into a hotel, this classification is performed by searching for keywords in the user’s response. Next, some specific remarks are classified, such as the user’s last name, first name and patronymic.

To solve this problem, a rule-based classifier was used: the presence of a name in a pre-prepared list was checked.

Finally, classification based on user intent was performed. For this purpose, a neural network classifier was used. Next, to identify the intent from the user's phrase, the *rubert-base-cased-sentence* model, which is included in the *deppavlov library*, was additionally trained. This model is built on the BERT architecture. To train the classifier, a data set of 231 user remarks was created, divided into 5 categories - greeting, ending a dialogue, booking a room, getting a number, general questions. Cross-entropy was used as the loss function. The calculated metrics after additional training of the neural network for five epochs are as follows: F1 (macro) = 0.8902, F1 (micro) = 0.8972, F1 (weighted) = 0.8922.

The next important component is the logger response generator. It was implemented using two neural networks. One for generating dialogue on a free topic, and the other for generating dialogue on the topic of checking into a hotel.

The neural network for generating the receptionist's response to the topic of checking into a hotel is based on the Transformer architecture. To implement the network, the following parameters were selected:

- embedding size (regular and for word position): 256;
- number of heads in the attention mechanism: 8;
- number of encoder blocks: 6;
- number of decoder blocks: 6;
- The size of one portion of data entering the input of the neural network (batch): 128.

To obtain the resulting sequence of tokens, a greedy generation algorithm (greedy decoding). The peculiarity of the algorithm, as can be judged from its name, is that at each generation step, the token whose probability is the highest at this step is selected. This algorithm works well for generating small sequences of words where the sentence length is short.

The total number of trained parameters was 5,686,725. Moreover, to train the neural network, 10,000 dialogues were generated using an algorithm built on the basis of the eBICA cognitive architecture (see Sect. 1). This cognitive model worked with a pre-written dialogue script that allowed for slight variations. The variations consisted in the choice of options for statements from both the user and the registrar. Each version of each statement was pre-rated by subjects on the Valence, Dominance, and Arousal scales in a separate empirical study. As a result, eBICA could determine the choice of one of several options for each statement for each virtual interlocutor, guided by the model of developing social relations between the interlocutors. Thus, a set of synthetic dialogues was obtained. Each dialogue was divided into pairs < user phrase, registrar phrase >. After this, the neural network was trained on a data set consisting of such pairs for 20 epochs.

A pre-trained neural network was used, trained on data sets compiled from Russian dialogues. Due to the large number of parameters, she was not fully trained.

The functions of the virtual environment visualization component include converting a replica received from the user into text, and a replica received from the server into a voice message, and also exchanging data with the server.

4.2 Virtual Psychologist

The system that implements the Virtual Psychologist was described by us earlier in a position paper [3]. Here we present preliminary results of its implementation, which is still at an intermediate stage. In particular, eBICA has not yet been used to control the behavior of a psychologist, although the possibility and expediency of its use here are quite clear.

The virtual psychologist conducts a conversation with the user according to several specified scenarios (see Sect. 2), addressing the user with pre-written questions. The use of the F-2 robot, developed in the laboratory of A. Kotov at the Kurchatov Institute Research Center, allows you to voice text and express specified emotions. Robot F-2 [11] has the ability to express a large number of emotions and communicative intentions. BML based programming interface (Behavior Markup Language) allows you to work with a database of patterns of emotional dynamics. For the purposes of the project, we, with the help of subjects, assessed the robot's gestures on emotional scales (the limited space of the article does not allow us to present the results of these assessments). Examples of gestures are shown in Fig. 2. The robot is also available as a virtual avatar.



Fig. 2. Examples of gestures performed by the F-2 robot.

The main difficulty in this subproject was the analysis of user statements in order to determine the user's personality type. The problem was solved using machine learning methods. Today, machine learning algorithms are used to create models that classify personality types according to the Big5 model based on survey response data, as well as large-volume free text (20 thousand words) written by a person [16], but access to this proprietary system is limited.

Our task here is to train the system to recognize personality types based on dialogues within selected scenarios. Due to the lack of large volumes of data of the required type, training was carried out on the basis of an open access data set, including the results of the Big5 test of individual subjects and the text written by them [13]. The scikit-learn library for Python was used to implement machine learning algorithms. For data preprocessing, the open access "NRC Word- Emotion Association Lexicon" database was used [12, 17], consisting of 14,155 words with ratings for categories of valence (negative color, positive color) and emotions such as anger, anticipation, disgust, fear, joy, sadness, surprise, trust, and embedding model GloVe.

5 Experiment: Materials and Methods

Sixteen students of National Research Nuclear University MEPhI aged from 20 to 22 years of both sexes in equal proportions took part in this study. Testing of the implemented systems confirmed their compliance with the functional requirements and objectives of the study.

5.1 Virtual Receptionist

The experimental procedure in the “Virtual Hotel Receptionist” paradigm included three stages.

At the first stage, the experiment participant assessed the appearance and environment of the registrar agent in virtual space for two minutes, after which he completed the Artificial Social Agent Questionnaire (ASA) [14], answering questions on the first impression of the agent.

The ASA questionnaire is a tool for assessing human interaction with a virtual agent, developed by an international working group of researchers in the field of artificial intelligence. We used a short version of the questionnaire, which allows us to create a profile of an artificial social agent on five scales: the attractiveness of the agent, its social acceptability, trust in the agent, the presence of personality in the agent, and the ability to evoke emotions in the user.

The receptionist in this study could be controlled either by a virtual actor or by a human confederate hidden in another room, who entered the text of his questions on the keyboard. Two participants were alternately used in the role of the confederate. The rest of the participants in the experiment did not know that a person could act as a receptionist.

At the second stage, the participant underwent the registration procedure in virtual reality. The registrar’s questions were voiced to the participant through speech synthesis, the participant’s speech was recognized - a dialogue took place, as described in Sect. 3. The experiment lasted until the registration of the number by the registrar was completed. If registration of the room did not occur within 10 min, the experiment was completed. At the end of the second phase, the participant again completed the short version of the ASA questionnaire in full, rating the interlocutor after interacting with him.

In the third phase, the participant again went through the same registration procedure in virtual reality. The third stage differed from the second only in that if at the second stage the registrar was controlled by a confederate, then at the third stage it was controlled by a virtual actor. And vice versa, if at the second stage the registrar was controlled by a virtual actor, then at the third stage it was controlled by a confederate. At the end of the third stage, the participant of the experiment also completed the questionnaire.

5.2 Virtual Psychologist

The experiment in the “Virtual Psychologist” paradigm took place in the form of a conversation between the subject and the robot according to a pre-written script, as

described above. After a few general questions were asked by the robot, the subject and the robot played a game in the form of a dialogue (Sect. 3). At the same time, the phrases voiced by the robot were accompanied by gestures, facial expressions, gaze and body language. For each phrase, a gesture was selected that corresponded to the intention and emotional connotation of the phrase, according to the previously established emotional connotations of gestures (see Sect. 4.2).

6 Results and Analysis

6.1 Virtual Receptionist

As a result of the study, psychological profiles of a virtual registrar were obtained and compared before and after communicating with him, as well as the profiles of a registrar controlled by a virtual agent and a registrar controlled by a human. It was found that the impression of a virtual receptionist is formed gradually, and before communicating with him, the impression is determined mainly by the characteristics of appearance. After communicating with the registrar, the assessment is influenced by his behavior.

In Fig. 3 presents the results of a comparison of the profiles of a registrar controlled by a virtual agent before and after the subject's communication with him, as well as profiles after communication with a registrar controlled by a confederate virtual agent.

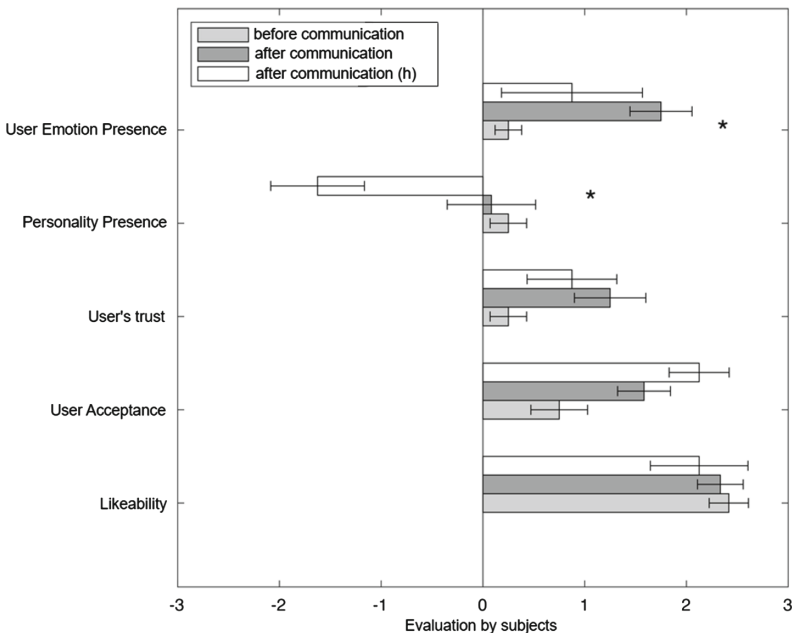


Fig. 3. Comparison of psychological profiles of registrars controlled by a virtual agent and by a human, before and after the registration procedure. Clear bars correspond to human. Significant differences are marked by asterisks.

The darkened stripes in Fig. 3 displays the characteristics of the virtual agent. Significant differences are indicated with asterisks.

Calculations were carried out using the Mann-Whitney U test, taking into account the Bonferroni correction. The following was found. The virtual actor’s user emotions presence increased after interacting with him ($P < 0.001$). The personality presence for the virtual actor turned out to be higher than for the human confederate ($P < 0.0095$).

6.2 Virtual Psychologist

In the case of the Virtual Psychologist, the trained system was used to analyze real dialogues conducted by the Virtual Psychologist in the guise of the F-2 robot with the subjects. For each of the five scales, a binary score was obtained based on dialogue analysis. The results were compared with the results of the standard Big5 test in its Russian adaptation, also divided into two categories for each scale. The comparison results are presented in Fig. 4.

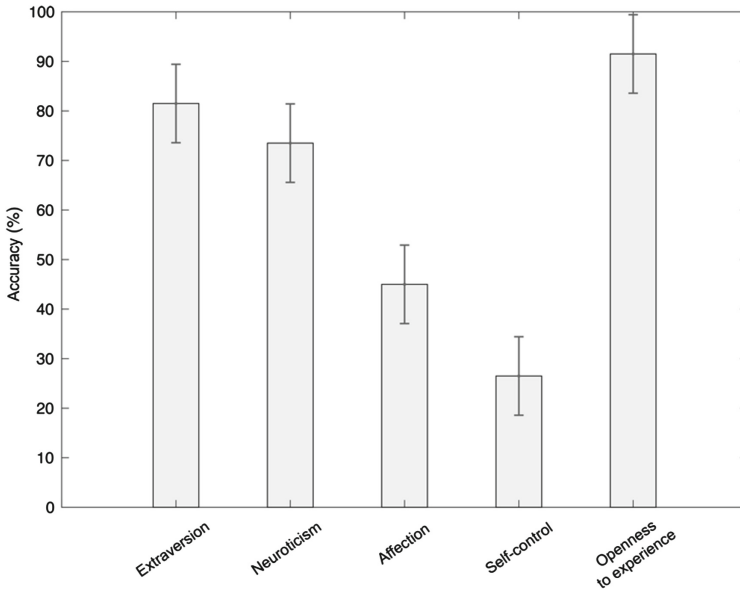


Fig. 4. Results of personality type recognition based on dialogue with a robot. The whiskers show the standard error estimate.

The maximum accuracy of classification of the user’s psychotype was achieved for the characteristic “openness to experience” and turned out to be above 90%, and for half of the other characteristics it exceeded 70% (Fig. 4). We expect the eBICA connection or its neural network equivalent to robot control will allow us to increase accuracy.

7 Conclusions

In this work, a method of integrating machine learning methods (in particular, LLM) and cognitive modeling (based on eBICA) was implemented and studied in experiments with subjects, allowing them to complement each other. Encouraging results were obtained in the “Virtual Hotel Receptionist” and “Virtual Psychologist” paradigms. They indicate that socio-emotional conversational agents based on cognitive architectures and machine learning, with the right approach to their implementation, can be socially compatible with humans. In our experiments, the artifact demonstrated the level of trustworthiness, social acceptability and likeability not lower than that of a human participant, and exceeded the human level in the ability to elicit emotions and in the perceived presence of personality. These results seem impossible to achieve with such a simple tool as we built, and yet it happened. There are many questions yet to be understood in the human perception of the Self of an agent [18]. Future studies should fully take advantage of the multimodal emotional interaction controlled by a cognitive model of Self.

Overall, our results indicate the feasibility of a virtual agent or robot capable of establishing human-level social contact while also assessing the user’s psychology. Potential applications of this technology include many important tasks, from intelligent tutoring [19] to modeling human value system [20] and personality in social assistants.

Acknowledgments. The authors consider it their pleasant duty to thank those who contributed to the work on this project: NRNU MEPhI Faculty Member Daria V. Tikhomirova, NRNU MEPhI Graduate Students Alena Anisimova, Aleksei Mikhnev, and Vladimir Tsarkov.

The work was supported by the Russian Science Foundation Grant No. 22-11-00213, <https://rscf.ru/project/22-11-00213/>.

References

1. Ali, M.R.: A virtual conversational agent for teens with autism spectrum disorder: experimental results and design lessons. In: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA ‘20). Association for Computing Machinery, New York, USA (2020)
2. Hartholt, A.: Introducing canvas: combining nonverbal behavior generation with user-generated content to rapidly create educational videos. In: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA ‘20). Association for Computing Machinery, New York, USA (2020)
3. Diaz, F. et al.: Empathic smart conversational agent for enhanced recovery from abdominal surgery at home. In: HCI International 2022-Late Breaking Posters: 24th International Conference on Human-Computer Interaction, HCII 2022, Virtual Event, June 26- July 1, 2022, Proceedings, Part I.-Cham: Springer Nature Switzerland (2022)
4. Elgarf, M., Skantze, G., Peters, C.: Once upon a story: can a creative storyteller robot stimulate creativity in children? In: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents (2021)
5. Blache, P.: An integrated model for predicting backchannel feedbacks/Philippe Blache, Massina Abderrahmane, St√/©phane _ _ Rauzy, and Roxane Bertrand. 2020. In: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA ‘20). Association for Computing Machinery, New York (2020)

6. Marsella, S.C., Gratch, J.: EMA: a process model of appraisal dynamics. *Cogn. Syst. Res.* **10**, 70–90 (2009)
7. Samsonovich, A.V.: Emotional biologically inspired cognitive architecture. *Biol. Inspir. Cognit. Architect.* **6**, 109–125 (2013)
8. Samsonovich, A.V.: Socially emotional brain-inspired cognitive architecture framework for artificial intelligence. *Cogn. Syst. Res.* **60**, 57–76 (2020)
9. Tikhomirova, D.V., Chubarov, A.A., Samsonovich, A.V.: Empirical and modeling study of emotional state dynamics in social videogame paradigms. *Cogn. Syst. Res.* **60**, 44–56 (2020)
10. Truity.: The big five personality test (Electronic resource). <https://www.truity.com/test/big-five-personality-test>. Access: 26 May 2023
11. Kotov, A.: (Electronic resource). <http://www.f2robot.com>. Accessed: 26 May 2023
12. Mohammad, S., Turney, P.: Crowdsourcing a word-emotion association lexicon. *Comput. Intell.* **29**(3), 436–465 (2013)
13. Kosinski, M.: My personality project (electronic resource). Last modified: 05/10/2016. https://web.archive.org/web/20160518020419/http://mypersonality.org/wiki/doku.php?id=download_databases. Access: 26 May 2023
14. Fitrianie, S., Bruijnes, M., Li, F., Abdulrahman, A., Brinkman, W.P.: The Artificial-social-agent questionnaire: establishing the long and short questionnaire versions. In: *ACM International Conference on Intelligent Virtual Agents (IVA '22)*, September, 2022. Faro, Portugal. ACM, New York (2022)
15. Anisimova, A.S., et al.: Artificial Psychologist: an intelligent virtual/robotic assistant based on a cognitive modeling framework. *Proc. Comput. Sci.* **213**, 793–800 (2022)
16. Martin, J.A.: Watson personality insights introduction and how to access Watson without SDK. *Anal. Vidhya* 11/12/2019. <https://medium.com/analytics-vidhya/watson-personality-insights-introduction-and-how-to-access-watson-without-sdk-89eb8992fff2>. Accessed: 26 May 2023
17. Mohammad, S.M.: NRC word-emotion association lexicon (electronic resource). Last update: Aug 2022. <http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>. Accessed 26 May 2023
18. Samsonovich, A.V., Ascoli, G.A.: The conscious self: ontology, epistemology and the mirror quest. *Cortex* **41**(5), 621–636 (2005). [https://doi.org/10.1016/S0010-9452\(08\)70280-6](https://doi.org/10.1016/S0010-9452(08)70280-6)
19. Samsonovich, A.V., De Jong, K.A., Kitsantas, A., Peters, E.E., Dabbagh, N., Kalbleisch, M.L.: Cognitive constructor: an intelligent tutoring system based on a biologically inspired cognitive architecture (BICA). *Front. Artif. Intell. Appl.* **171**(1), 311–325 (2008). ISSN: 09226389
20. Samsonovich, A.V., Ascoli, G.A.: Cognitive map dimensions of the human value system extracted from natural language. *Front. Artif. Intell. Appl.* **157**, 111–124 (2007). ISSN: 09226389



Emotion from P Bit: Computing Emotions Using the Platonic Computer

Simon X.Duan^(✉)

Metacomputics Labs, London 1 8NB, UK
Simon.x.duan@live.com

Abstract. Emotion is a major aspect of the human conscious experience. Some common emotions include happiness, sadness, fear, anger, surprise, and disgust. Computing emotions is a desirable goal for conscious AI systems. Emotions can be detected and classified using computational methods, but whether or not emotions can truly be “computed” is still a matter of philosophical and scientific debate. Some researchers argue that emotions are first-person subjective experiences, not just third-person objective phenomena that can be measured or quantified. Therefore, it is unlikely to be computable. In this paper, it is proposed that emotions are computed using the Platonic computer. A Platonic computer is a hypothetical rendering engine in the Platonic realm of Forms. It is the archetype of our everyday mundane material computer. A physical computer made of silicon material is a shadow or poor imitation of the Platonic computer. Although the Platonic computer is not accessible by our normal senses, we can simulate how the Platonic computer renders emotions by manipulating a physical computer to render emojis. It is hypothesized that emojis that are rendered by a physical computer are shadows of emotions rendered by the Platonic computer. By adopting this approach, emotions are digitalized and simulated using a physical computer.

Keywords: Conscious AI · Emotion · Platonism · Platonic computer · Platonic computation · Platonic code · Platonic bit

1 Introduction

Emotion refers to the subjective feelings and experiences that arise in response to various stimuli, such as thoughts, events, or interactions with others. Some common emotions include happiness, sadness, fear, anger, surprise, and disgust. Emotions play a crucial role in our lives, helping us to navigate and respond to the world around us. Understanding emotion is essential for developing a more nuanced understanding of ourselves and the world around us.

It is generally believed that emotions are generated by the human brain through a complex interplay of sensory, cognitive, and physiological processes. Research has shown that specific brain regions and neural pathways are involved in the generation of emotions. For example, the amygdala, a small almond-shaped structure in the brain, is involved in the processing and regulation of emotions, particularly fear and anxiety. The

prefrontal cortex, a region of the brain responsible for higher-order cognitive functions, is also involved in generating and regulating emotions, particularly positive emotions like happiness and gratitude [1–4]. These findings have provided important insights into the mechanisms that underlie our conscious experiences.

However, despite significant advances in the last few decades, we still have no idea what specific neuron state correlates to specific conscious experiences. So far, no specific conscious experience has ever been identified in terms of a specific set of neuron activities. For example, it is impossible to pinpoint what is the pattern of integrated neural activities that must be the state of happiness and could not possibly be the state of gratitude.

Moreover, even the coarse-grained mapping of the brain regions identified as being correlated to various conscious experiences has been challenged by the discovery of many cases of hydrocephalus. In this medical condition, the cerebrospinal fluid becomes dammed up in the brain, causing the liquid to eat almost all the brain cells. Yet, people suffering from hydrocephalus can live and work normally. They can have a full range of emotional experiences as normal people [5].

It is evident that the relationship between brain states and the conscious experience of emotions is complex and multifaceted. It is possible that a definitive correlation between conscious emotional states and precise patterns of neural activities may never be established. Thus, it is doubtful what inspiration can be drawn from researching the human brain. If we aim to generate emotions from computation, we need to adopt an alternative and more fruitful approach.

2 Metacomputics Model and Platonic Computation

Metacomputics model hypothesizes that a transcendental computer exists in the Platonic realm of Forms. The cosmos is the processing output of the Platonic computer.

A Platonic computer is a theoretical concept in computer science and philosophy that draws on the ideas of Plato's philosophy, particularly the concept of Forms. The idea is that a Platonic computer would operate not on physical inputs and outputs, but on abstract Forms or ideas.

In Plato's philosophy, the realm of Forms is seen as the ultimate reality, and the physical world is seen as a shadow or poor imitation of this higher reality. Adopting this philosophy, we can postulate that the material computer is a shadow or poor imitation of the Platonic computer.

In this hypothetical scenario, a Platonic computer would manipulate and process abstract concepts and relationships rather than physical data. For example, instead of processing numerical values or textual data, a Platonic computer would work with abstract concepts or logical relationships between ideas.

The concept of a Platonic computer inspires new ideas and approaches to computer science, particularly in artificial intelligence and machine learning, where researchers are working to create computers that can learn and reason in more abstract and human-like ways.

It is not possible to access the Platonic realm of Forms by our normal senses or by using current physical instruments. However, we cannot deny the existence of abstract entities such as numbers, geometric shapes and other abstract concepts (Universals).

Numbers, for example, are nonphysical, no one has bumped into the number 2 or tripped over the number 3, and we cannot find the number 4 in the kitchen cupboard. And yet we do things with numbers all the time, we count with them, measure with them, and formulate our scientific theories with them. Without the existence of numbers, we wouldn't have physics, without physics there wouldn't be science. Numbers are also time-independent - we don't need to worry that numbers once didn't exist or might one day cease to exist. If numbers have a reality independent of physical reality, then they have to be accommodated somewhere outside space and time. Many philosophers and mathematicians agree that the Platonic realm of Forms is where abstract entities exist.

If the Platonic realm contains numbers, can it also contain a nonphysical computer?

The existence of a transcendental computer is supported by savant syndrome, a condition in which someone with significant mental disabilities demonstrates certain abilities far in excess of average. Calendar savants, for example, can name the day of the week or a date in a range of a millennium. There are well-documented cases of autistic children who can give answers to complex mathematical problems. The answers just appear to them – indicating that their mind vision functions as a display of computational output.

The premise the cosmos is rendered by the Platonic computer leads to some profound implications. One such implication is that everything in existence can be reduced to digital patterns of binary code. Thus, time, space, physical and nonphysical entities, as well as contents of consciousness such as thoughts, feelings and emotions can be digitalized and computed. In this paper, we focus on the topic of emotion.

3 Simulating Platonic Computer Using Physical Computer

The Platonic computer cannot be studied empirically since the realm of Forms is beyond the reach of our ordinary senses and current physical instruments. However, this paper explores the idea that Platonic computation could be studied by using the technique of computer simulation. It is proposed that the generation of conscious experience of emotions can be simulated using a physical computer, based on the theoretical framework of metacomputics [6].

The Platonic computer and the physical computer both operate on binary opposing states. However, there are fundamental differences between the make-up of the binary states in the Platonic computer compared to that of a physical computer. For instance, within the Platonic computer, the metaprocessor is made of metaconsciousness, and the output is generated by manipulating the two binary opposing states, i.e., unmanifested metaconsciousness (□) and manifested metaconsciousness (■) [7].

In comparison, a physical computer processor is made of binary ON/OFF switches made of silicon material, and its output is generated by manipulating these physical switches. Hence, the processing outputs of a physical computer are specific configurations of binary states (i.e., ON (1) and OFF (0)). A specific configuration of these 0s and 1s defines a symbol which can be displayed on the computer screen. For example, according to the American Standard Code for Information Interchange (ASCII), the binary digits 1010 defines the symbol '10', whereas the binary digits 01000001 defines the letter 'A', all of which can be displayed on the computer screen.

Being a poor imitation of Platonic computation, physical computation can only simulate certain aspects of it. For instance, a physical computer can simulate the dynamic changes of the weather so that a weather forecast can be made with a reasonable level of accuracy, but it doesn't get wet and windy inside the computer screen that displays the simulation. That is, simulation of the weather in a physical computer will not produce the conscious experience of feeling wet and windy.

Despite the limitations of physical computation, the material computer still represents a useful tool to simulate some aspects of Platonic computation. The following sections discuss how the rendering of emotional states with the Platonic computer can be simulated using a physical computer.

4 Application of Emojis in Expressing Emotions

Traditionally, a person's emotional state is analyzed based on external expressions and behaviors. While external expressions such as facial expressions, body language, and verbal cues can offer observable indications of emotions, they do not provide a complete understanding of an individual's emotional state. Internal factors, such as thoughts, physiological responses, and subjective experiences, play a significant role in shaping emotions. These internal aspects are not always externally visible.

Advancements in technology and research have enabled the development of various methods for analyzing emotions, including physiological measurements (e.g., heart rate, skin conductance), facial recognition technology, voice analysis, and natural language processing, such as sentiment analysis techniques that combine lexicon-based approaches with machine learning techniques or other algorithms to improve accuracy. While these techniques are achieving increasingly accurate analyses, they work best when sentiment polarity is explicitly expressed through individual words or phrases. Despite continuous improvement, these techniques still have limitations in capturing complex nuances and contextual sentiment. For instance, they may struggle to cope with new words, slang, or domain-specific terms, or handle contextual ambiguity and identify ironic or sarcastic expressions.

Essentially, emotions are first-person subjective experiences. Therefore, relying solely on third-person measurable manifestations will never capture the nuances and intricacies of an individual's internal emotional experience. Fortunately, a new innovation has made it possible for humans to express the 1st-person conscious experience of emotions in a direct and straightforward way. This innovation is "emojis".

In computer science and digital communication, emojis are small digital images or icons that are used to represent a wide range of emotional states such as happiness, sadness, anger, and love.

Emojis are continuously evolving, with new designs being created and added to the standard set each year. Currently, there are over 3,600 emoji in the Unicode Standard, which is the universal character encoding standard used by computers and mobile devices.

Each emoji has a unique Unicode code point, which is a hexadecimal number that identifies the character. For example, the "smiling face with heart-eyes" emoji has the Unicode code point U + 1F60D, representing 11110000100111111001100010001101

in UTF-8 format. The emoji depicts a smiling face with heart-shaped eyes, often accompanied by rosy cheeks, conveying a sense of joy and overwhelming fondness.

The popularity and acceptance of emojis have grown significantly over the years, and they have become an integral part of online communication for many people. For example, emojis are extensively used in social media apps and platforms to convey emotions and tones, express feelings, enhance engagement, or add a visual element to the text. When doing online reviews, such as product or restaurant reviews, emojis can be powerful and effective tools for expressing satisfaction, disappointment, or other sentiments. Emojis are also increasingly used in blog posts, articles, or other text-based content to add visual appeal, convey emotions, or engage readers.

As emojis are an effective tool in expressing first-person conscious emotional states, they can be deployed to convey emotions without the limitation of traditional approaches during inter-human digital communications.

Emojis are also universal symbols that transcend language barriers. As emojis are defined by unique Unicode code points, there is an equivalence between the emoji and its digital pattern, i.e., the binary code used to define the emoji. In other words, emotions bypass natural languages.

5 Emotion from P Bit (Platonic Bit)

Emojis are symbols representing conscious states, they are not our experience of conscious states. For example, a heart emoji is used to represent love or affection, but the emoji itself is just a graphical representation of the feeling of love or affection.

It is hypothesized that in parallel to the physical computer that renders the emojis, the Platonic computer renders conscious emotional states. The conscious state of being in a specific emotion is defined by a specific configuration of the Platonic binary states, i.e., unmanifested metaconsciousness (□) and manifested metaconsciousness (■).

For instance, the emotion “Love-filled joy and happiness” would be generated by the same sequence of binary code as that of emoji under the name “Smiling Face with Smiling Eyes and Three Hearts”, i.e., U + 1F970 (0000 0000 0001 1111 1001 0111 0000). Instead of using a combination of 0s and 1s, the Platonic code uses combinations of Platonic binary states as follows:

$$\square\square\square\square \quad \square\square\square\square \quad \square\square\square\square \quad \blacksquare\blacksquare\blacksquare\blacksquare \quad \blacksquare\square\square\square \quad \square\blacksquare\blacksquare\blacksquare \quad \square\square\square\square \quad (1)$$

In essence, emojis are considered symbols and shadows of emotions. Emojis and emoticons have different substances and qualities but share the same digital patterns of configuration. Adopting this approach allows us to define each conscious emotional state by a unique Platonic code.

Thus, the rendering of emotions with the Platonic computer can be simulated by rendering emojis with the physical computer.

6 Implementing and Evaluating Digitally Defined Emotions

Having established the unique digital pattern of each emotion, the conscious experience of emotions can be defined by the Platonic binary code rendered by the Platonic computer. However, the Platonic computer cannot be empirically accessed and manipulated by our

normal senses, therefore we can only manipulate its shadow or poor imitation, i.e., everyday mundane material computer to render emojis of the same digital patterns.

Simulating the rendering of emotions can be implemented in the virtual environment through virtual characters. In the virtual environment, avatars are normally customized by players to reflect their preferences, appearance, and characteristics. However, the player's emotions have been so far solely expressed through speech and voice, facial expressions, body language and gestures, reactive behavior, etc.

It is proposed that players display emojis on his/her avatars so that the emotions of the player can be expressed directly and unambiguously. In addition, the display of emojis on NPCs can be achieved by implementing AI algorithms or behavioral systems that govern the NPCs' emotional responses. Rules and decision-making processes should be implemented that determine how NPCs express emotions based on their personalities, goals, and the current situation.

This approach would create a more engaging and immersive user experience. By implementing emojis in the virtual environment to express emotional responses, virtual characters can enhance communication, foster empathy, and create a stronger connection between users and virtual characters.

Evaluating the effectiveness of virtual characters' emotions expressed by emojis requires a combination of subjective assessments and user feedback. This can be achieved by accessing believability, consistency, and impact on users. It is essential to consider the specific goals, context, and technical capabilities of the virtual characters in relation to the intended emotional experiences of the users.

7 Summary

It is hypothesized that the conscious experience of emotions is the processing output of the Platonic computation. The rendering of emotion with the Platonic computer can be simulated by rendering emojis using a material computer. This is based on the assumption that our everyday mundane material computer is a shadow or poor imitation of the Platonic computer. The former renders virtual reality; the latter renders physical reality. The former operates on binary states made of silicon material switches; the latter operates on Platonic binary states of metaconsciousness. The former renders emojis and the latter renders emotions. Emojis are symbols whereas emotions are conscious experiences. Since each emotional state can be defined by a unique Platonic code, emotions can thus be digitalized and simulated.

References

1. Sato, W., Kochiyama, T., Uono, S., et al.: The structural neural substrate of subjective happiness. *Sci. Rep.* **5**, 1689 (2015). <https://doi.org/10.1038/srep16891>
2. Cavanna, A.E., Trimble, M.R.: The precuneus: a review of its functional anatomy and behavioural correlates. *Brain* **129**(3), 564–583 (2006). <https://doi.org/10.1093/brain/awl004>
3. Seshadri, K.G.: The neuroendocrinology of love. *Indian J. Endocrinol. Metab.* **20**(4), 558–563 (2016). <https://doi.org/10.4103/2230-8210.183479>
4. Kringelbach, M.L., Berridge, K.C.: The neuroscience of happiness and pleasure. *Soc. Res.* **77**(2), 659–678 (2010). PMID: 22068342

5. Forsdyke, D.R.: Wittgenstein's certainty is uncertain: brain scans of cured hydrocephalics challenge cherished assumptions. *Biol. Theory* **10**(4), 336–342 (2015). <https://doi.org/10.1007/s13752-015-0219-x>
6. Duan, S.X.: Platonic computer—the universal machine that bridges the “inverse explanatory gap” in the philosophy of mind. *Filozofia i Nauka* **10**, 285–302 (2022). <https://doi.org/10.37240/fin.2022.10.zs.14>
7. Duan, S.X.: Platonic computer—the universal machine from which abstract entities are generated. In: *Proceedings of the 2021 Summit of the International Society for the Study of Information*, MDPI Proceedings **81**(1), 58 (2022). <https://doi.org/10.3390/proceedings2022081058>



Learning Hidden Markov Model of Stochastic Environment with Bio-inspired Probabilistic Temporal Memory

Evgenii Dzhivelikian¹(✉), Petr Kuderov^{1,2,3}, and Aleksandr I. Panov^{2,3}

¹ Moscow Institute of Physics and Technology, Dolgoprudny, Russia
dzhivelikian.ea@phystech.edu

² AIRI, Moscow, Russia

³ Federal Research Center Computer Science and Control of the Russian Academy of Sciences (FRC CSC RAS), Moscow, Russia

Abstract. Learning models online in partially observable stochastic environments can still be challenging for artificial intelligent agents. In this paper, we propose an algorithm for the probabilistic modeling of observation sequences based on the neurophysiological model of the human cortex, which is notoriously fit for this task. We argue that each dendritic segment of a pyramidal neuron may be considered an independent naive Bayesian detector of afferent neuron activity patterns. Experiments show that our model can learn the dynamics of the partially observable environments for very few interactions online and reliably predict probabilistic distributions of observations for several future time steps using Monte Carlo sampling. Additionally, we compare our algorithm with a biologically inspired HMM implementation of temporal memory and standard LSTM on both Markov chain-generated character sequences and observation image sequences in a pinball-like environment.

Keywords: Temporal memory · HMM · Hebbian learning · Sequence memory

1 Introduction

The main purpose of learning a model of the environment is to be able to predict sequence of future or unseen observations. In fact, any sequence learning algorithm implicitly builds such a model, which condenses spatio-temporal dependencies between hidden and observable variables. Such an implicit method in form of HMM, LSTM, sequential-VAE and Transformer models has shown to

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-50381-8_33.

be prolific in AI. However, all of this aforementioned models use backpropagation for parameter optimization, which is known for its instability and mediocre online learning properties.

Drawing ideas from neurophysiological models of the neocortex neurons and their networks, we introduce a biologically plausible probabilistic temporal memory model based on the hidden Markov model (HMM), a particular case of the generative model. The main shortcoming of the classic Baum-Welch algorithm for learning HMM is that it necessitates access to the entire sequence of observations. In this paper we address the problem of forming hidden representations of observations on the fly, facilitating online HMM learning. We show that our model can learn the dynamics of partially observable stochastic environments online and can reliably predict future observations for multiple steps forward.

We compare our model with analogous biologically inspired temporal memory model. In the paper [1], authors use the idea of the columnar structure of the cortex to regularize the emission probability of the HMM. But in contrast to our work, they use biologically implausible Baum-Welch algorithm for weight updates. Another limitation of this work, which we aim to overcome, is that the model is formulated only for categorical observations. We also compare our model with classical sequence learning algorithms like LSTM [2] and Baum-Welch HMM [3].

There are many other works that either draw ideas from neuroscience or offer a computational model of cortical circuits. A bio-inspired framework for learning sequences [4,5] uses the idea of columnar cortex structure and models neuron dendrites. However, it uses all-or-nothing signaling that doesn't allow graded predictions, which are necessary for stochastic environments and multi-step predictions. In the paper [6], authors describe a hierarchical generative model for the visual cortex and show that it effectively detects learned objects on cluttered images like in CAPTCHAs. Another biologically plausible model of the visual cortex is presented in [7], which employs sequence prediction error to form generalized object representations qualitatively similar to humans. In works [8–10] HTM framework is used for building bio-inspired artificial agent architectures, which shows a better adaptation in rapidly changing environments than some of the classical RL methods. Finally, authors of the papers [11,12] show how ideas from neuroscience can be used for methods of learning sequence representations.

The rest of the paper is organized as follows: Sect. 2 introduces necessary definitions and formalization. Section 3 describes the proposed probabilistic temporal memory model. The performed experimental setup and the results are described in Sect. 4. Finally, Sect. 5 discusses the results, outlines our method's limitations, and provides insights for future work.

2 Background

In this paper we consider partially observable Markov process, which can be learned by Hidden Markov model (HMM). Variables h_t represent an unobserv-

able (hidden) state of the environment which evolves over time and observable variables o_t represent observations that depend on the same time step state h_t . For the process of length T with variables $h_{1:T} = (h_1, \dots, h_T)$ and $o_{1:T} = (o_1, \dots, o_T)$, Markov property yields the following factorization of the generative model:

$$p(o_{1:T}, h_{1:T}) = p(h_1) \prod_{t=2}^T p(h_t|h_{t-1}) \prod_{t=1}^T p(o_t|h_t). \quad (1)$$

In case of discrete hidden state, a time-independent stochastic transition matrix can be defined $A = \{a_{ij}\} = \{p(h_t = j|h_{t-1} = i)\}$ and learned with Baum-Welch algorithm [3]. To compute the statistics for the expectation step, it employs the forward-backward algorithm.

Modeling neuronal dendritic segments we refer to Naive Bayes (NB) that is a supervised learning classification algorithm based on applying Bayes' theorem with the "naive" assumption of conditional mutual independence of a feature vector components $x = (x_1, \dots, x_n)$ given the class variable value y to model their relationship:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)} = \frac{p(y) \prod_{i=1}^n p(x_i|y)}{p(x)}. \quad (2)$$

To measure the difference between two probability distributions p and q over the same variable x we use the Kullback-Leibler (KL) divergence as it has been commonly used in machine learning [13].

To measure surprise—a prediction error of an observed event x —we use negative log-probability $-\log(p(x))$, which is also known as log-loss [14, 15].

In our work, we design our model to operate with sparse distributed representations (SDRs) to reflect the spatiotemporal property of cortical network activity [16]. In the discrete time case, SDR is a sparse binary vector in a high-dimensional space. In experiments with Pinball environment (see Sect. 4.2), to encode observed dense binary patterns to SDRs we use a biologically plausible k-WTA (k-winners take all) neural network algorithm with a Hebbian-like unsupervised learning method [17].

3 Methods

We speculate that our model corresponds to a cortical macrocolumn's superficial and granular input (fourth) layer (e.g., in a mouse's whisker S1 area). These cortical networks may play a role in short-term memory formation [18], as their inhibition significantly degrades performance in tasks with delayed outcome.

The model is connectionist with two types of cells: pyramidal excitatory (PC) and inhibitory cells. There is evidence, that neurons tend to form functional ensembles [19]. A neural ensemble consists of laterally interconnected neurons with an identical feed-forward receptive field. While as a group they recognize the

same stimulus, the lateral inhibitory inter-ensemble competition forces each neuron to specialize to discriminate stimulus only in a specific context. In this model, context is represented with extra-ensemble lateral excitatory activity. It enables neurons realize prediction mechanics. This hypothetical model of PC activation is based on the fact that distal dendritic NMDARs spikes cause soma depolarization, which shortens action potential latency during stimulus-response, adjusting neuron tuning [20]. Thus, a neuron with active distal dendrites has a better chance of inhibiting other neurons in the ensemble (via inhibitory interneurons) after receiving the same stimulus. For the sake of simplicity, we do not explicitly model intra-ensemble inhibitory circuitry. Instead, we model its effects by activating only those cells that have both distal and proximal dendrites active. Ensemble inhibition in L2/3 is intact with known inhibitory neuron microcircuitry [21]. Conversely, we model inter-ensemble inhibitory cells explicitly, whose activity corresponds to discrete attractor states. Such dynamics are known to be used for neurophysiological modeling of superficial layers circuitry [22].

The core idea of our method is that we consider neurons and dendritic segments as Bernoulli random variables with states interpreted as *spike/no-spike*. This enables to draw connection between our model and HMM. Let's assume neuron cells' activity in all ensembles of superficial layer being represented by a hidden state matrix h_t and cells of granular input layer—by observation vector o_t at time t . Then the task of predicting input sequences can be formulated as learning distribution over future observations $\tilde{o} = (o_{t+1}, \dots, o_T)$:

$$p(\tilde{o} | \underline{o}) = \sum_{h_{1:T}} \prod_{\tau=t+1}^T p(o_\tau | h_\tau) p(h_\tau | h_{\tau-1}) p(h_t | \underline{o}), \quad (3)$$

where $h_{1:T} = (h_1, \dots, h_T)$ —all hidden states.

In this model, $p(h_t | \underline{o})$ is a posterior belief about current network state, which is determined from the observational history $\underline{o} = (o_1, \dots, o_t)$ and collapses into an indicator function. That is, $h_t^{ij} \in \{0, 1\}$ —the activity of the i -th neuron of the j -th ensemble at time t is determined by the following rule:

$$h_t^{ij} = \begin{cases} 1, & (\exists k : d_t^{ijk} = 1) \wedge (j \in \varepsilon(o_t)) \\ 0, & \text{otherwise} \end{cases}, \quad (4)$$

where $\varepsilon(o_\tau)$ is a set of indexes of ensembles activated by a stimulus o_τ , $d_t^{ijk} \in \{0, 1\}$ is k -th dendritic segment activity at time t of (i, j) neuron. In other words, a neuron is active, if one of its dendritic segments is active and its ensemble gets enough feed-forward input to be activated.

For sake of simplicity, we assume that there is a bijection between ensembles and components of the observation vector (granular cells). That is, each component o_t^j corresponds to ensemble activity j . Consider o_τ is a distributed representation and its components are independent binary random variables. Therefore, the emission probability is the following:

$$p(o_\tau^j | h_\tau) = \begin{cases} 1, & \exists i : h_\tau^{ij} = 1 \\ 0, & \text{otherwise} \end{cases}. \quad (5)$$

Thus, it remains to specify $p(h_\tau | h_{\tau-1})$, which determines how previous cell states influence current—this process we call prediction. In our model, the predictive activity of a neuron h_τ^{ij} is determined by the activity of its dendritic segments $d_\tau^{ij} \in \{0, 1\}^l$, where l is the number of segments of the neuron. In other words, we factorize $p(h_\tau^{ij} | h_{\tau-1})$ through $p(h_\tau^{ij} | d_\tau^{ij})$ and $p(d_\tau^{ijk} | h_{\tau-1})$.

Another key assumption of our model is in form of $p(d_\tau^{ijk} | h_{\tau-1})$, which is determined by Bayesian rule analogously to Naive Bayesian classifier. That is, we assume, that each segment tries to predict “class” of its neuron (*spike/no-spike*) using $h_{\tau-1}$ as feature vector.

In the case of multi-step predictions, one must also consider the statistical dependence between individual neurons. We found that introducing additional factorization through inhibitory interneurons that control activation of different neuron ensembles greatly facilitates derivations in this case.

4 Experiments

Here we present experiments carried out to identify the properties of the proposed model and compare it with analogous algorithms. In the first section we describe a test on character sequences to show that our algorithm indeed forms a probabilistic temporal memory (TM) and compare it with the classic Baum-Welch algorithm for HMM. And another section is devoted to experiments that show the applicability of the TM for image sequence prediction generated in a stochastic pinball-like environment, which is in effect a partially observable Markov process.

4.1 Markov Chain Grammar

Markov chain grammar (MCG) is a sequence generator that maps state transitions in a Markov chain to characters (see Fig. 1C). Algorithms are tested on five different MCG models and three seeds for each setup. To evaluate models we use the surprise metric for one-step prediction averaging it for all steps within episode. One episode corresponds to one sequence of MCG generated characters of arbitrary length not exceeding fifty symbols.

Our TM model is leaning in fully online mode with update step on each observation. We compare our model with the cloned hidden Markov model (CHMM) and LSTM. CHMM is an HMM, but with fixed emission matrix similar to one in our model. CHMM baseline is learning in two modes: Baum-Welch with batch size of 100 observations and iterative Baum-Welch, which updates transition matrix with every new observation online, only relying on forward messages. For LSTM we use standard implementation from PyTorch [23] with the size of the hidden layer equal to CHMM. LSTM is optimized by RMSprop with fine-tuned learning rate equal to 0.02.

We can see from the graph (Fig. 1A) that the cutoff of backward messages during weight updates significantly degrades the performance of CHMM. However, the performance of our model, which doesn’t employ the future information

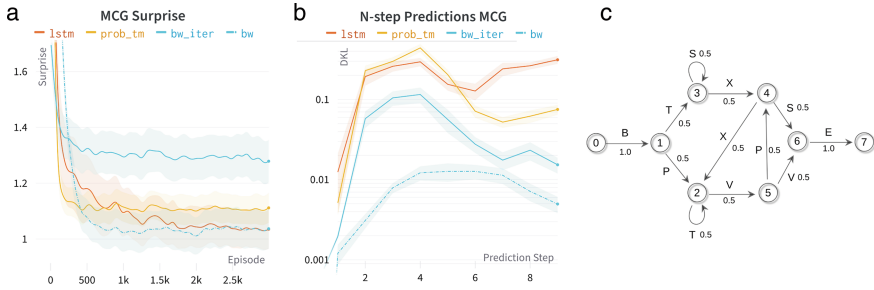


Fig. 1. **A** Comparison of our model `prob_tm` with CHMM trained by Baum-Welch algorithm `bw` and its iterative version `bw_iter`, and LSTM. The graph shows the change of surprise evaluated for one-step prediction with training episode. **B** Comparison of models on n-step prediction task. The graph shows the dependence of Kullback-Leibler (in log-scale) distance between predicted and ground truth distributions from prediction step. **C** Graph representation of a Markov chain grammar model. Vertex represents state, arrow—state transition with corresponding probability, and letter emitted

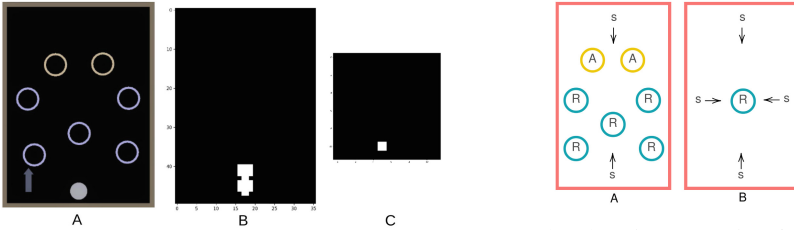
either, is better than of CHMM with iterative Baum-Welch updates and closer to the one of full BW for this task. Although slower than CHMM, LSTM also converges to the optimum.

We also measure the Kullback-Leibler divergence between the ground truth distribution of n th step observation beginning from the first observation of letter “B” and distributions predicted by trained models. From the graph on Fig. 1B it can be seen that, on average, the performance of TM is significantly worse than of CHMM for prediction steps further than 1 step horizon. The same we can see for LSTM. Unlike HMM, for which we can directly use the transition matrix to calculate n-step predictions, our model’s and LSTM predictions are calculated using Monte Carlo sampling, which can be a source of error. We predict that spike integration errors will cause analogous degradation in biological systems.

4.2 Pinball

Pinball is a continuous stochastic partially observable environment developed in the Godot Game Engine [24]. The environment consists of the surface with borders and a ball that is able to move in 2D space of the surface. The stochasticity is introduced by random force fields that impact ball in their coverage area, which is visualized by a circle (see Fig. 2a-A). The random force field instantly imparts an impulse to the contacted ball in a direction uniformly sampled from a discrete set each time the contact is made. There are also attractor force fields that slow down the ball and tend to stabilize its position during the contact.

Each episode is 3 s long with 5 frames per second rendering speed. The ball is randomly initialized in one of the start locations on the board with the corresponding initial impulse depicted on Fig. 2b. Each time step, which corresponds to one rendering frame, Pinball environment outputs an RGB image of size $36 \times 50 \times 3$ that is further processed in three stages: (1) convert to a gray-scale



(a) Observation examples of the Pinball environment. **A.** Visual representation of the environment. Circles visualize random and attractor force fields, arrow represents initial impulse of the ball. **B.** Processed observation of size 36x50. **C.** Encoded observation of size 12x12.

(b) Schematic representation of two experimental setups. R — random force field, A — attractor field, s — possible ball start position, arrow — initial ball impulse direction at the corresponding position.

Fig. 2. Experiments in the Pinball environment

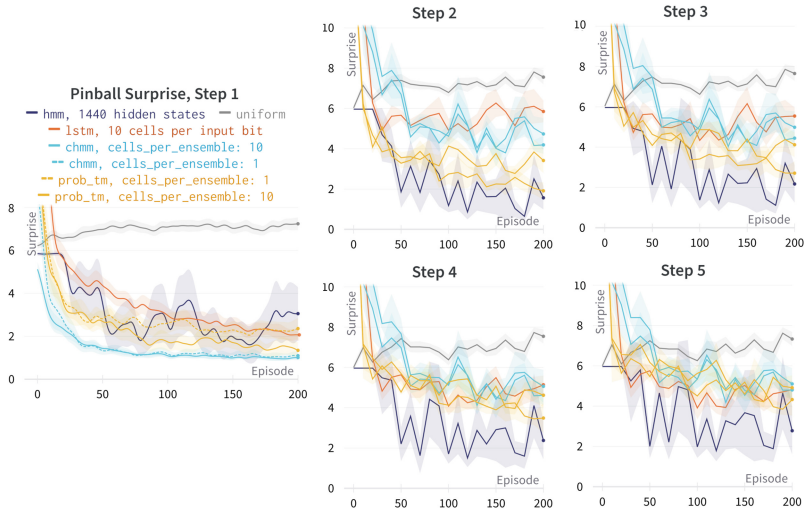


Fig. 3. Comparison of our model (**prob_tm**), CHMM trained by iterative Baum-Welch algorithm (**chmm**), LSTM (**lstm**) and classic Baum-Welch HMM (**hmm**). The graph shows the change of surprise evaluated from one-step to four-step predictions. For CHMM and ProbTM we also vary amount of cells per ensemble

image, (2) compute absolute difference between current and previous frames, (3) binarize the difference using mean intensity of the current difference as a threshold, mimicking event-based camera processing (see Fig. 2a-B). We also use a biologically plausible sparse encoding algorithm to further reduce observation space to 12×12 shape (see Fig. 2a-C). The resulting sparse distributed representation, which is an index array of non-zero pixels, is fed to the model.

In experiments in the Pinball environment we compare our model with CHMM with iterative Baum-Welch learning, classic HMM [3] and LSTM from PyTorch [23] trained by RMSprop with fine-tuned learning rate equal to 0.003.

LSTM and HMM hidden vector sizes are chosen to be comparable to the number of neurons in other baselines. To make multi-step predictions with LSTM we use similar Monte-Carlo sampling technique that we used for our model. Each time step, surprise is evaluated for the next step prediction and averaged over three seeds for each of two setups. The testing pipeline is the same for each model, with the exception that the CHMM and HMM input sparse representation is forced to have only one active pixel at a time, as shown in Fig. 2a-C, because CHMM and HMM are only formulated for categorical observation variables.

Figure 3 shows that CHMM makes better predictions one step ahead, although the prediction isn't affected by the number of clones or ensemble cells in contrast to our model. Classic HMM learning curve have high variation, however, in n-step prediction task it shows better performance on average than other algorithms. Our model gives better predictions for step > 1 in comparison to CHMM (see Fig. 3) inferior, however, to classic HMM, which learns offline in contrast to our algorithm. Despite we had been fine-tuning LSTM learning rate carefully for this task, we didn't manage to get convergence speed comparable to other baselines. Moreover, with further learning, multi step (> 1) predictions for LSTM are getting even worse, apparently, because of over-fitting.

5 Conclusion and Discussion

In this paper, we use a biologically plausible model of temporal memory capable of predicting probability distributions of future observations to address the problem of online sequence learning in a stochastic environment. The results show that our model can use high-order memory to reduce stimulus surprise in a variety of temporal contexts while employing only biologically plausible local learning rules. The proposed algorithm outperforms analogous biologically inspired HMM model and LSTM in multi-step prediction experiments in the Pinball environment. One of the main limitations of our model is that, unlike the classic Baum-Welch algorithm, it cannot use information about future time steps for temporal context formation and weight updates due to the fully online learning mode. There is evidence that future information may be responsible for generalized context representation in high-order memory, which would increase temporal memory capacity and learning speed significantly [1, 25]. As a result, we intend to develop a method for biologically plausible temporal information aggregation and belief back-propagation in future work. Another possibility to improve context representation is to form Successor Representations [26] by accumulating prediction distributions for different horizon. Then, the accumulated distribution can be used to train another temporal memory layer enabling increasing abstraction in the model.

As a further improvement of our model we also going to incorporate motor efference copy in temporal context formation through apical dendritic segments to predict action outcomes during environment interaction, which would model connections between different brain areas and will allow us to use our model in a biologically inspired agent architecture.

Code Availability. Our model's implementation source code is publicly available at <https://github.com/AIRI-Institute/him-agent/tree/d90ef5a3dd3dcdfe4fe1f127e415d436a6103562/hima/experiments/hmm>.

References

1. George, D., Rikhye, R.V., Gothoskar, N., Guntupalli, J.S., Dedieu, A., Lázaro-Gredilla, M.: Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps. *Nat. Commun.* **12**(11), 2392 (2021)
2. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997)
3. Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* **41**(1), 164–171 (1970)
4. Dayliydonok, I., Frolenkova, A., Panov, A.I.: Extended hierarchical temporal memory for motion anomaly detection. In: Samsonovich, A.V. (ed.) *Biologically Inspired Cognitive Architectures 2018. BICA 2018. Advances in Intelligent Systems and Computing*. vol. 848, pp. 69–81. Springer, Berlin (2019). https://doi.org/10.1007/978-3-319-99316-4_10
5. Hawkins, J., Lewis, M., Klukas, M., Purdy, S., Ahmad, S.: A framework for intelligence and cortical function based on grid cells in the neocortex. *Front. Neural Circ.* **12**, 121 (2019). www.frontiersin.org/article/10.3389/fncir.2018.00121/full
6. George, D., Lehrach, W., Kansky, K., Lázaro-Gredilla, M., Laan, C., Marthi, B., Lou, X., Meng, Z., Liu, Y., Wang, H., Lavin, A., Phoenix, D.S.: A generative vision model that trains with high data efficiency and breaks text-based captchas. *Science* **358**(6368), eaag2612 (2017)
7. O'Reilly, R.C., Russin, J.L., Zolfaghar, M., Rohrlich, J.: Deep predictive learning in neocortex and pulvinar. *J. Cogn. Neurosci.* **33**(6), 1158–1196 (2021)
8. Dzhivelikian, E., Latyshev, A., Kuderov, P., Panov, A.I.: Hierarchical intrinsically motivated agent planning behavior with dreaming in grid environments. *Brain Inform.* **9**(1), 8 (2022)
9. Dzhivelikian, E., Latyshev, A., Kuderov, P., Panov, A.I.: Intrinsic motivation to learn action-state representation with hierarchical temporal memory. In: Mahmud, M., Kaiser, M.S., Vassanelli, S., Dai, Q., Zhong, N. (eds.) *Brain Informatics. BI 2021. Lecture Notes in Computer Science*, vol. 12960, pp. 13–24. Springer, Berlin (2021). https://doi.org/10.1007/978-3-030-86993-9_2
10. Kuderov, P., Panov, A.I.: Planning with hierarchical temporal memory for deterministic markov decision problem. In: *Proceedings of the 13th International Conference on Agents and Artificial Intelligence—vol. 2: ICAART*, pp. 1073–1081. INSTICC, SciTePress (2021)
11. Rodkin, I., Kuderov, P., Panov, A.I.: Stability and similarity detection for the biologically inspired temporal pooler algorithms. *Procedia Comput. Sci.* **213**, 570–579 (2022)
12. Kuderov, P., Dzhivelikian, E., Panov, A.I.: Stabilize sequential data representation via attraction module. In: *Brain Informatics. BI 2023. Lecture Notes in Computer Science* (2023)
13. Ackley, D.H., Hinton, G.E., Sejnowski, T.J.: A learning algorithm for Boltzmann machines. *Cogn. Sci.* **9**(1), 147–169 (1985)

14. Painsky, A., Wornell, G.: On the universality of the logistic loss function. In: 2018 IEEE International Symposium on Information Theory (ISIT), pp. 936–940. IEEE (2018)
15. Friston, K., Kilner, J., Harrison, L.: A free energy principle for the brain. *J. Physiol.* **100**(1), 70–87 (2006)
16. Perin, R., Berger, T.K., Markram, H.: A synaptic organizing principle for cortical neuronal groups. *Proc. Natl. Acad. Sci.* **108**(13), 5419–5424 (2011)
17. Cui, Y., Ahmad, S., Hawkins, J.: The HTM spatial pooler—a neocortical algorithm for online sparse distributed coding. *Front. Comput. Neurosci.* **11**, 111 (2017). www.frontiersin.org/article/10.3389/fncom.2017.00111
18. Galvez, R., Weible, A.P., Disterhoft, J.F.: Cortical barrel lesions impair whisker-CS trace eyeblink conditioning. *Learn. Memory* **14**(1–2), 94–100 (2007)
19. Liu, B., Seay, M.J., Buonomano, D.V.: Creation of neuronal ensembles and cell-specific homeostatic plasticity through chronic sparse optogenetic stimulation. *J. Neurosci.* **43**(1), 82–92 (2023). www.jneurosci.org/content/43/1/82
20. Stuart, G.J., Spruston, N.: Dendritic integration: 60 years of progress. *Nat. Neurosci.* **18**(12), 1713–1721 (2015). <https://doi.org/10.1038/nn.4157>
21. Staiger, J.F., Petersen, C.C.H.: Neuronal circuits in barrel cortex for whisker sensory perception. *Physiol. Rev.* **101**(1), 353–415, pMID: 32816652 (2021). <https://doi.org/10.1152/physrev.00019.2019>
22. Rolls, E.T., Mills, W.P.C.: Computations in the deep vs superficial layers of the cerebral cortex. *Neurobiol. Learn. Memory* **145**, 205–221 (2017). www.sciencedirect.com/science/article/pii/S1074742717301636
23. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc. (2019), <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
24. Beeching, E., Debangoye, J., Simonin, O., Wolf, C.: Godot reinforcement learning agents. arXiv preprint [arXiv:2112.03636](https://arxiv.org/abs/2112.03636) (2021)
25. Whittington, J.C.R., McCaffary, D., Bakermans, J.J.W., Behrens, T.E.J.: How to build a cognitive map. *Nat. Neurosci.* **25**(10), 1257–1272 (2022). www.nature.com/articles/s41593-022-01153-y
26. Gershman, S.J.: The successor representation: its computational logic and neural substrates. *J. Neurosci.* **38**(33), 7193–7200 (2018). www.jneurosci.org/content/38/33/7193



The Research on Key Technologies for Intelligent Education Based on LLM

Jiaqi Fu¹ , Tiejun Pan¹  , Leina Zheng² , and Zichu Xue² 

¹ College of Science and Technology (CST), Ningbo University, Ningbo, China
958809518@qq.com

² Business School, Zhejiang Wanli University, Ningbo, China

Abstract. This article investigates the influence of the artificial intelligence model on intelligent education and explores the prospects of integrating virtual human technology to enable the LLM model to perform instructional tasks. A digital teacher will be created via the integration of the LLM model with virtual human technology. The virtual human will perform live streaming on specific platforms, acting as a teacher for various subjects based on different prompts, and engaging with the audience in the livestreaming environment.

Keywords: Artificial intelligence · Smart education · Virtual person

1 Introduction

1.1 Background

According to relevant data, some poor and remote areas have particularly inadequate educational resources, and the resources in terms of educational funding, teaching staff, educational facilities, and other aspects of these areas are far less than those of developed areas. This results in a serious lack of educational equity, with some children unable to access quality education opportunities due to the lack of educational resources in their hometowns, causing an imbalance in the distribution of educational resources. Therefore, the government should implement innovative education models that include digital education, distance learning, and other strategies to reduce regional gaps in education resources and uplift education fairness.

1.2 Purpose

This paper addresses the challenges of smart education and digital education. It highlights the efficiency of LLM, a prompt-trained language model, for precise and accurate educational content. The focus is on achieving pedagogical precision and convenience. To distribute educational resources effectively, digital virtual technology is proposed. By integrating LLM with virtual characters, digital teachers can be created to deliver remote instruction. The goal is to develop a prompt-trained digital instructor capable of teaching various courses, communicating with students, hosting live classes, addressing queries, and customizing lessons for different age groups.

1.3 Research Work

LLM has proven its exceptional ability to generate human-like responses in education. Precision and accuracy are vital in teaching, and prompt-based training with LLM allows for personalized content and high accuracy. To bridge educational resource gaps, we propose combining LLM with virtual characters to create remote digital teachers. These teachers can deliver various courses, engage with students, host live lectures, and adapt content to different age groups. By merging LLM with virtual characters, underprivileged students gain access to quality education. Challenges in implementing LLM-based intelligent education include data privacy, interpretability, and subject-specific training for digital teachers. Our research utilizes LLM and UE5-based metahuman digital human technology.

2 Key Technology

2.1 LLM Language Model

The working principle of LLM is based on a generative pre-training model that specifically uses the Transformer architecture. First, LLM learns the patterns and structures of language by conducting unsupervised pre-training on a large amount of text data. During the pre-training phase, the model is exposed to a large amount of text data and learns the relationship between words and context through self-prediction tasks. In this way, the model can understand the grammar, semantics, and contextual information of language [1].

By using prompt words to set specific contexts, virtual teachers can be created for different subjects by providing different prompt words.

2.2 Metahuman

Metahuman is a digital character creation tool developed by Epic Games that can quickly and highly realistically create personalized virtual character models. It provides rich customizable options for appearance and movement, as well as realistic visual effects and a user-friendly interface. Metahuman has wide application potential in fields such as game development, virtual reality, and film production.

We will use Metahuman to quickly generate an image of a teacher and apply it to the UE5 Unreal Engine for further technical development.

2.3 AzSpeech

AzSpeech is an open-source speech recognition and transcription tool. It is based on Google's open-source projects TensorFlow and DeepSpeech, and integrates Azure's speech services, which can recognize multiple languages. AzSpeech can be used in applications such as speech transcription, speech command recognition, and speech interaction.

The use of AzSpeech is very simple, only need to install dependencies and set environment variables according to the instructions on Github. Users can call AzSpeech for speech input and output transcription results through the command line [2].

2.4 OBS

OBS (Open Broadcaster Software) is a free, open-source software for live streaming and recording. It can capture computer screens, cameras, and audio inputs, and stream or record them as videos. OBS is suitable for a variety of real-time live streaming and content creation needs.

3 Demand Analysis

3.1 Hardware Requirements Analysis

Hardware analysis is essential for seamless and high-quality virtual human development. Factors such as computer configuration, memory, processor performance, and graphics card requirements must be considered. Graphics card requirements, in particular, are crucial, as Unreal Engine demands high standards. Additionally, developers should prioritize monitor resolution and display effects, as they directly impact the virtual human's appearance and performance.

3.2 Functional Requirements Analysis

It is necessary to conduct detailed analysis of the specific functions of virtual human development and design, such as character modeling, information processing, animation production, audio processing, and so on. In addition, the access to OBS live broadcast and the interaction with students need to be considered, and the requirements of each function should be clarified. It is essential to ensure that these requirements can be met during the development phase.

4 Overall Design

4.1 Metahuman Production Design

First, a clear facial photo is required to generate a corresponding 3D model in <https://avatarsdk.com/>. Then, the 3D model is imported to the metahuman official website to select clothing and body shape before saving it. The metahuman virtual human generated in this way can be directly used in Unreal Engine.

4.2 Metahuman Action Design

Here, we use the redirection function to implement the action, such as downloading the animation from Mixamo, matching the skeleton of our virtual human through the redirection, performing bone division, and then making fine adjustments to achieve the function of making the virtual human move. After the redirection, the actions of Mixamo's robot become the actions of our virtual human, and many actions can be selected. Figure 1 shows a detailed interface for redirecting character models in Unreal Engine.

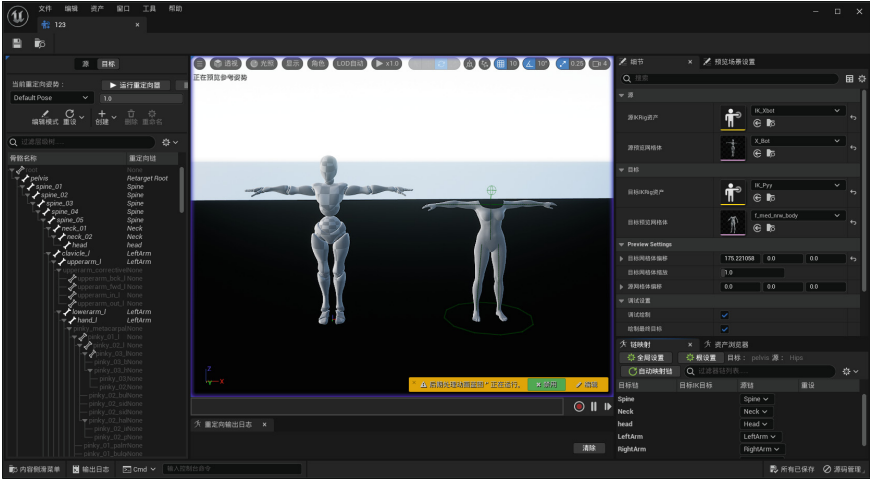


Fig. 1. Metahuman action

4.3 Metahuman Connect to LLM

Before we start the production, we need to prepare the API and install the Openai plugin in UE5. The specific production process is as follows: open the level blueprint, create a method called “Set Open AI API Key”, copy our API into it, use the Settings branch to create LLMSettings for basic configuration. At this point, we can design prompts to define the specific subject role that the virtual person will play. Finally, by using branches and a For Each Loop, we output the content returned by LLM, thus achieving the function of connecting to LLM. Figure 2 below shows the blueprint connection method and its function method.

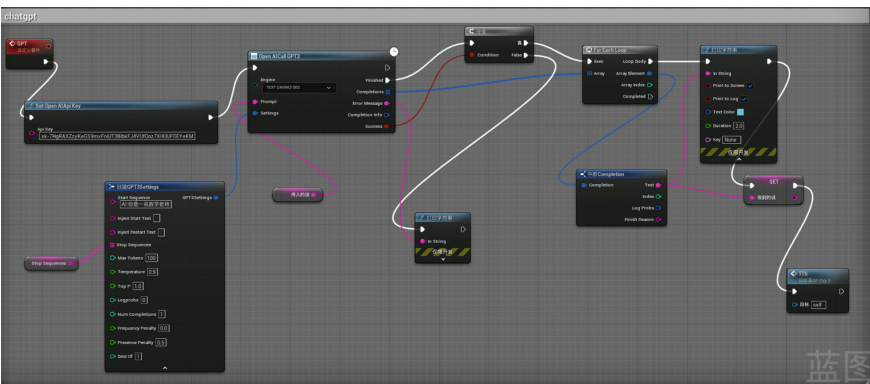


Fig. 2. Metahuman connect to LLM

4.4 AZspeech Achieves Speech-to-Text Conversion

To achieve the function of automatically accessing LLM for speech-to-text conversion, we need to use Microsoft’s AZspeech. Within it, we create a method called “Speech to Text with Custom Options” and configure it. This enables the conversion of speech input into text, which is then passed on to LLM for processing. To enable sound production for a virtual human, we utilize the TTS (Text-to-Speech) function to convert the returned content. Figure 3 shows speech-to-text blueprint concrete connection method and its function method.

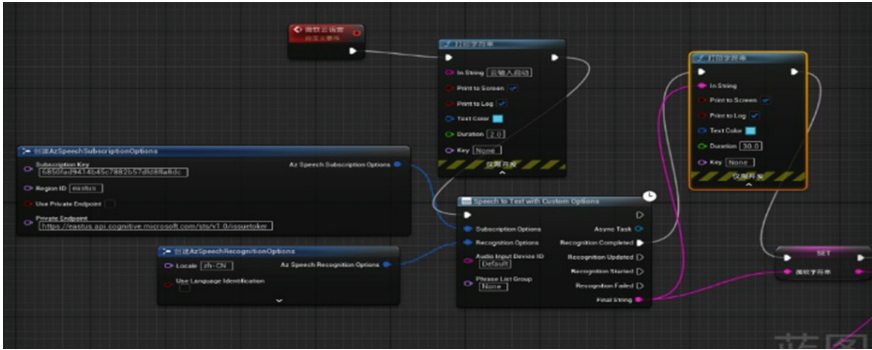


Fig. 3. Speech-to-text

4.5 Metahuman Integration with OBS

After installing OBS, find your own live streaming server address and key, add it to OBS, and select your virtual person window for OBS live streaming. This allows you to use the virtual person to complete simple live streaming. In OBS, you can also customize virtual scenes, and in UE5, you can directly customize virtual scenes as well.

5 Case Study of Software Engineering Teaching

5.1 The Applications of LLM in Software Engineering

In software engineering, LLM is capable of automatically generating code, documentation, and comments among others. For instance, using LLM to generate function comments can improve code readability and maintainability. Moreover, LLM can aid in generating test cases, code snippets, and other forms of code which eventually enhance both software development efficiency and its overall quality.

5.2 The Advantages and Limitations of LLM in Software Engineering

One advantage of LLM is its ability to generate high-quality natural language text, thereby improving the efficiency and quality of software development. Moreover, LLM can improve its generation ability through continuous training, resulting in more accurate and natural text. However, LLM also has some limitations in generating unreasonable or inaccurate text when software engineers request specific needs, which could impact the software development process.

5.3 The Future of Applying LLM in Software Engineering

Firstly, LLM can generate high-quality language interaction between developers and users, reducing the cost of communication. In the software development process, developers need to constantly communicate with users to understand their needs and opinions. This information is crucial for software development and improvement. The presence of LLM makes this process more efficient because it can take on some communication responsibilities. Developers can integrate LLM into their website or application and allow users to interact with it. LLM's generative language can make users feel like they are conversing with a real person, not facing rigid multiple-choice questions.

Secondly, LLM can enhance the documentation and communication in software engineering. In most software development organizations, documentation is an indispensable tool. Developers need to write a variety of documents to record design decisions, communicate requirements, and document their code. However, documents are often dry and hard to understand. LLM can automatically generate structured documents and provide information about code libraries based on user requests. LLM can write auxiliary tools for many documents, saving developers time and energy. Because the documents generated by LLM can better meet user needs, the quality of communication and documentation in software engineering has been improved.

Finally, LLM can help developers solve problems quickly. In the software development process, various problems may arise. These problems may come from specific issues in the code library or be related to programming languages or frameworks. Sometimes developers need to contact support teams to solve these problems, but have to wait for an answer. LLM can provide instant assistance to developers. Developers can ask questions, and LLM will quickly generate answers and provide suggestions on how to solve the problem. This quick problem-solving method can greatly increase the efficiency of software development.

Therefore, LLM plays a positive role in the evolution of software engineering. It reduces the cost of communication, enhances the quality of documentation and communication, and helps developers solve problems quickly.

5.4 The Future of Applying LLM in Software Engineering Teaching

LLM uses simulated dialogue to interact with students and customize information in line with their learning needs based on their learning progress and requirements. It automatically generates exercises and questions, and provides suitable feedback and guidance

based on students' responses. This fosters increased learning enthusiasm among students and helps them to comprehend and master course content faster. Student progress and problem lists are used to automatically generate record reports and course learning progress, which enable teachers to understand their students' learning situation better and provide them with necessary support and assistance. LLM recommends courses and fun activities based on students' learning records and interests, thereby enhancing their learning interests and participation. Moreover, LLM links students with other students and teachers, enabling interaction, communication, and ultimately, a conducive environment for a real-time interactive communication-based learning community. In summary, as an AI technology, LLM provides diverse advantages to facilitate software engineering education, assisting students to comprehend course content better while promoting learning autonomy and interactivity.

5.5 LLM Applies Software Engineering Examples

Here, we use the Cursor compiler in software engineering examples. Cursor is an intelligent AI code generation software developed in collaboration with OpenAI. It can quickly complete code writing using AI technology to make development work simpler and more efficient. The official initially claimed that Cursor uses AI algorithms to automatically complete code, generate code fragments, and understand the meaning and context of the code, generating logic-based code. Figure 4 shows some example code applied to Cursor, and Fig. 5 shows the result of this code.

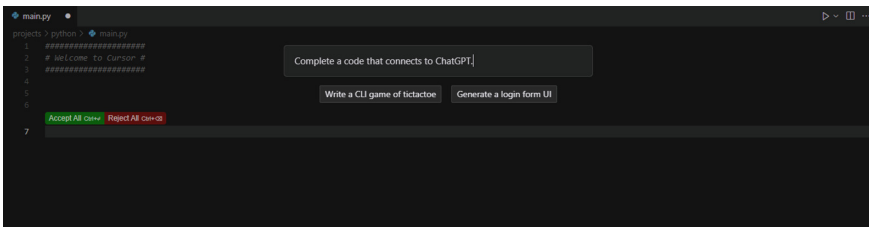


Fig. 4. Text generates code.

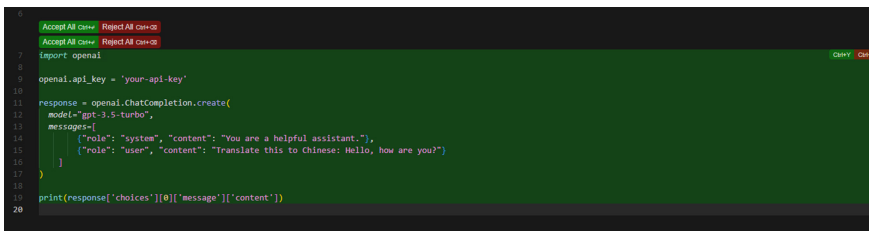


Fig. 5. Generate code.

6 Conclusion

In conclusion, integrating LLM with virtual human technology offers new possibilities for smart education and digital learning. LLM, trained on prompts, demonstrates exceptional accuracy in the educational domain. By creating digital teachers through this integration, we can deliver remote education to students, particularly those in underserved areas. This approach ensures precision and convenience in teaching. However, challenges like data privacy, interpretability, and subject-specific training for digital teachers need to be addressed. Technologies like Metahuman, AzSpeech, and OBS facilitate the development of interactive virtual human teachers through live streaming, showcasing AI's potential in education.




Acknowledgments. This paper is supported by Zhejiang Province's 14th Five Year Plan Teaching Reform Project (jg20220738), Ningbo Science and technology innovation Fund (20232213), Zhejiang College Student Innovation and Entrepreneurship Training Program (S202310876006), Zhejiang Provincial Basic Public Welfare Fund Research Project (LGF20G020002), Zhejiang Provincial Philosophy and Social Science Planning Project (22NDJC127YB), Ningbo Municipal Basic Public Welfare Fund Research Project (2021S070).

References

1. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Stoyanov, V.: RoBERTa: a robustly optimized BERT pretraining approach (2019). arXiv preprint [arXiv:1907.11692](https://arxiv.org/abs/1907.11692)
2. AzSpeech Github: <https://github.com/LGoodOptimizer/AzSpeech-TensorFlow> Github: <https://github.com/tensorflow/tensorflow-DeepSpeech>. Electronic resource retrieved on 2023/11/05 from Github: <https://github.com/mozilla/DeepSpeech>



Mapping Action Units to Valence and Arousal Space Using Machine Learning

Ismail M. Gadzhiev^{1,2,3}(✉), Alexander S. Makarov¹, Daria V. Tikhomirova³ ,
Sergei A. Dolenko² , and Alexei V. Samsonovich³ 

¹ Physical Department, Leninskiye Gory, M.V.Lomonosov Moscow State University,
Moscow 119991, Russian Federation

ismailgadzhievff@gmail.com

² D.V.Skobeltsyn Institute of Nuclear Physics, Leninskiye Gory, M.V.Lomonosov Moscow
State University, Moscow 119991, Russian Federation

dolenko@srd.sinp.msu.ru

³ National Research Nuclear University MEPhI, Kashirskoye Shosse, 31, Moscow 115409,
Russian Federation

Abstract. There are a lot of studies researching automated recognition of emotions. Emotions are represented as points in an emotion space. The emotion space itself is represented by different types of models. One is Facial Action Units System, another is Valence-Arousal-Dominance model. This study aims to create a mapping between these two emotion spaces. The data for the study was collected in a series of experiments with real humans, where both types of measurements were collected simultaneously. Given the data, we study the ability of machine learning models to create this type of mapping. We test different types of models against the task, such as tree-based models and linear models, and make conclusions about the optimal model.

Keywords: Emotions space · Action units · Valence · Arousal · Mapping · Machine learning

1 Introduction

Automated recognition of human emotions has been a problem of interest for researchers for the last decade. However, to predict something that is called an emotion, one should first define an emotion space.

One approach is Action Units. Action units (AU) are the actions of individual muscles or groups of muscles, originally developed by Carl-Herman Hjortsjö [1], and later adopted by Paul Ekman [2]. With the use of action units, one can categorize the physical expression of emotions. There are a lot of open-source and proprietary software that helps recognizing AU presence and intensity, for example, OpenFace [3] and Face Reader [4].

Another approach is the Valence-Arousal-Dominance (VAD) model [5, 6]. The VAD emotional state model uses three numerical dimensions, Valence (Pleasure), Arousal

and Dominance to represent all emotions. The pleasure-displeasure of Valence scale measures how pleasant or unpleasant one feels about something. The arousal-non-arousal scale measures how energized or soporific one feels. The dominance-submissiveness scale represents the controlling and dominant versus controlled or submissive one feels.

One of the steps in the loop of nonverbal communication between a person and a virtual actor using facial expressions, which we described in detail in our previous paper [7], is to reduce the dimensionality of the 42-dimensional space of Action Units to the three-dimensional emotional space of the VAD (Valence-Arousal-Dominance) model. The resulting 3D vector and its past trajectory can be analyzed with the eBICA model [8], using an appropriate moral schema [9], based on which the system generates an adequate actor response. Our study utilizes two dimensions of the VAD model, Valence and Arousal, which are common across many models of affects [5, 6, 14]. Here they are referred to as the “affect core.”

So, the main goal of this study is to find an efficient mapping between AU and VAD models. In our previous research [7] we have made an attempt of creating the mapping, but there was significantly less data than in the current study.

In this paper we first estimate the dimension of Action Units space using Principal Component Analysis, and then present an approach of mapping Action Units into the space of Valence and Arousal with machine learning models, such as gradient boosting and neural networks. We assess the performance of the machine learning methods on an out-of-sample test set using R^2 (coefficient of determination) score, and make conclusions about their applicability to solve the problem considered. These methods not only help us to restore Valence and Arousal, but also to make conclusions about the most important features among the Actions Units.

For modelling purposes, we use data from a series of experiments described below.

2 Data

2.1 Description of Initial Data

The study involved 45 volunteers - students at the National Research Nuclear University MEPhI (Moscow Engineering Physics Institute) aged 20 to 24. They participated in several experiments that aimed to get the emotions of the participants in different ways. The first experiment consisted of expressing emotional states on the face, for 10 s each: neutral expression, joy, sadness, anger, disgust, fear, contempt, satisfaction, excitement. Participants were asked to express emotions, the names of which appeared sequentially on the screen. The second experiment was to participate in the social-emotional games “Teleport” and “Shooters”, requiring engagement and partnering with other characters. To win the game, the participant had to establish contact with one of two other players controlled by a human or a model. A player left without a partner lost. The game session lasted for 10 min and consisted of several rounds. The third experiment involved watching clips of the video games “Teleport” and “Shooters” in which other people participated.

During the experiments, participants’ faces were captured on a webcam using the Bandicam program. Further, the videos were processed, and Action Units, Valence and Arousal were annotated using the emotion recognition program “FaceReader” by Noldus

[4]. The results of processing the video recordings of Ekman’s experiments on the expression of emotional states were added to the resulting dataset. The quality of the dataset and its size have increased significantly compared to our previous study [7].

The collected dataset contains values of 42 Action Units, as well as Valence and Arousal for each timestep. The total number of patterns in the dataset is 743063.

Figure 1 shows the distribution of target variables (Valence and Arousal) in the dataset. We see that the distribution is highly skewed.

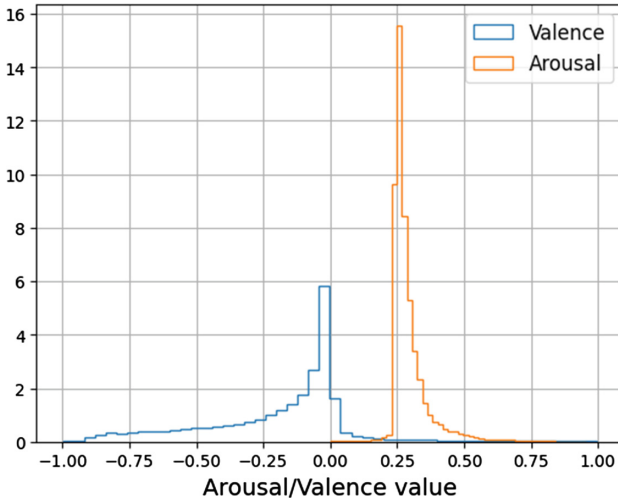


Fig. 1. Distribution of target variables (valence and arousal) in the dataset.

We explore the dataset using Principal Component Analysis. Figure 2 shows the cumulative explained variance ratio versus the number of principal components taken into account. We see that 6 features explain more than 80% of the variance in the data, and 9 features – more than 90%. This is an indicator that only a small subset of AU may be important for modelling purposes.

2.2 Delay Embedding

The current value of Valence and Arousal might depend on some short history of Action Units. For example, the value of Valence or Arousal may be possibly determined by the dynamics of a certain AU. The collected data contains some noise and therefore using a short period of AU change instead of single measurements could also improve model performance.

For this purpose, we test a data preprocessing technique called delay embedding. The essence of this technique is that the model is fed with a vector of values at the current timestep stacked with the values at N preceding time steps. The value of N is called the depth of the delay embedding. We test models with $N = 0$ (i.e. no history at all) and $N = 5, 10$, and make conclusion on the optimal N for this task.

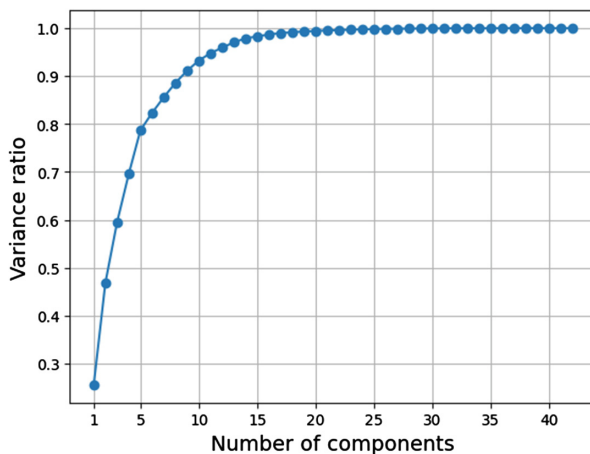


Fig. 2. Cumulative explained variance ratio versus the number of components.

3 Models

In this study, we test a set of machine learning methods against the task of mapping AU to Valence and Arousal. Firstly, as the simplest model, it is Linear regression model with L2 regularization (Ridge). Secondly, these are tree-based models: random forest [10] and gradient boosting [11].

For Gradient Boosting, we test LightGBM [12] and CatBoost [13] implementations. It should be noted that these implementations use different tree growth methods, so the result may vary significantly. We set equivalent default parameters for them: number of trees – 100, maximum depth of a tree – 100, learning rate – 0.3.

To find optimal hyperparameters of these models (e.g., number of trees in the tree-based models), we use the cross-validation (CV) technique grouped by experiment with $K = 5$ folds. Grouping by experiment means that patterns from the same experiment should appear only either in the train or in the test fold, but not in both folds simultaneously. We design it in this way so that the selected parameters for a model are robust to data coming from new experiments, where the model could be applied. We select optimal hyperparameters according to coefficient of determination (R^2) score averaged over all test folds.

We test all the models with initial parameters and after the cross-validation, to verify that CV truly improves model performance.

To assess the performance of the models, we calculate their R^2 score on an out-of-sample test set. The test set is about 20% of the data, containing experiments that are not included in the training set. All the models are compared to a constant mean mapping (calculate mean Valence and Arousal over the training set as the answers of the model, then calculate their R^2 on the test set). R^2 score generally ranges from $-\infty$ to 1 and could be roughly split into sub-ranges as follows: $R^2 < 0$ – performance of a model worse than a simple average constant, model should be discarded as useless; $0 < R^2 \leq 0.3$ – poor performance; $0.3 < R^2 \leq 0.7$ – medium performance, $R^2 > 0.7$ – good performance.

4 Results

Figure 3 shows the resulting scores for the models for Arousal prediction, Fig. 4 – for Valence prediction. We see that the best model for Arousal prediction is Ridge regression with the delay embedding of depth 10 with parameters selected during cross-validation (R^2 score 0.729), and the best model for Valence prediction is CatBoost with the delay embedding of depth 10 and with initial parameters (R^2 score 0.494).

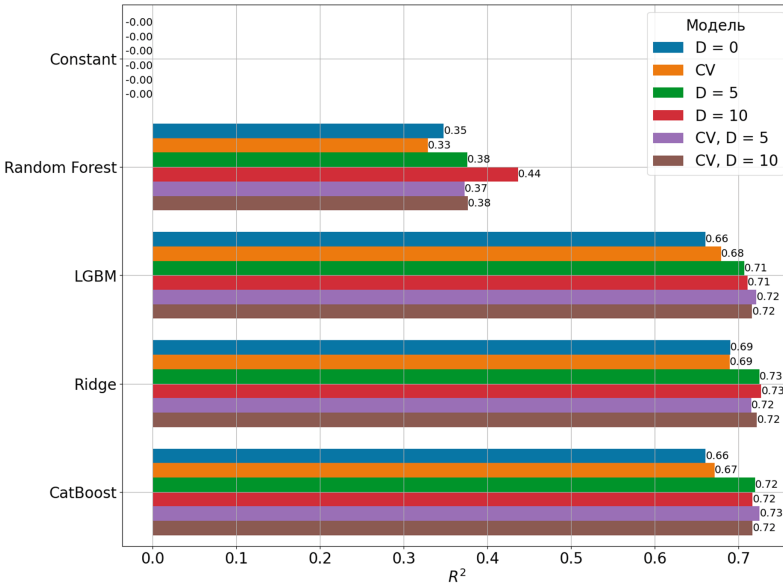


Fig. 3. Results for arousal, D stands for the depth of the delay embedding, CV stands for the parameter selection by use of cross-validation.

We conclude that the use of the delay embedding generally improves performance of the models, but the use of the cross-validation does not.

We also notice that CatBoost with the delay embedding of depth 10 and initial parameters is the model has high scores both for Arousal and Valence (R^2 scores 0.717 and 0.494). That implies that it is the optimal model for mapping AU to emotion space.

Now we are interested in what features are most important for the best model to make a prediction. We extract the so-called feature importance values from the Catboost model with the delay embedding of $D = 10$ and initial parameters. This is done by calculating how much on average the prediction changes if the feature value changes. We plot the feature importance values using a heatmap, where x-axis corresponds to the embedding depth of a feature, and y-axis corresponds to an Action Unit name.

Figure 5 shows the heatmap of the feature importance values for Arousal, Fig. 6 shows the same heatmap for Valence.

We see that the most important features for predicting Arousal are Action Unit 12 (Lip Corner Puller) and Action Unit 17 (Chin Raiser). For Valence the most important

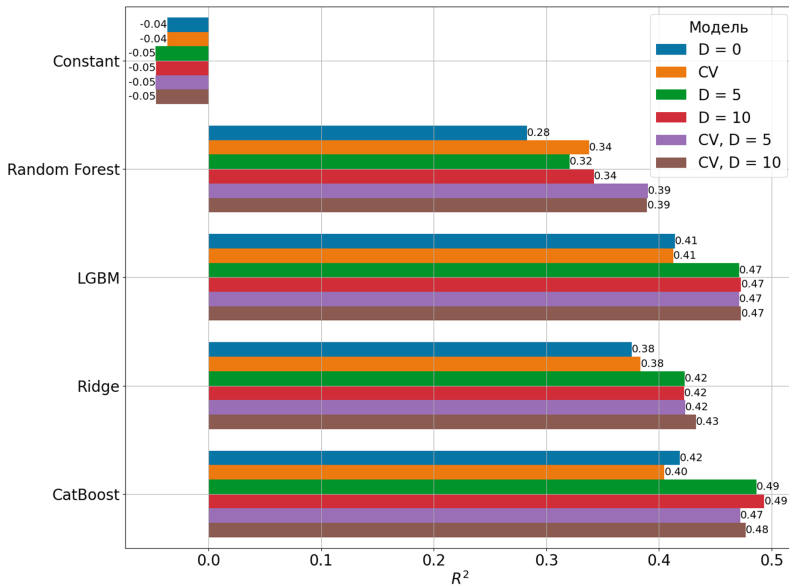


Fig. 4. Results for valence, D stands for the depth of the delay embedding, CV stands for the parameter selection by use of cross-validation.

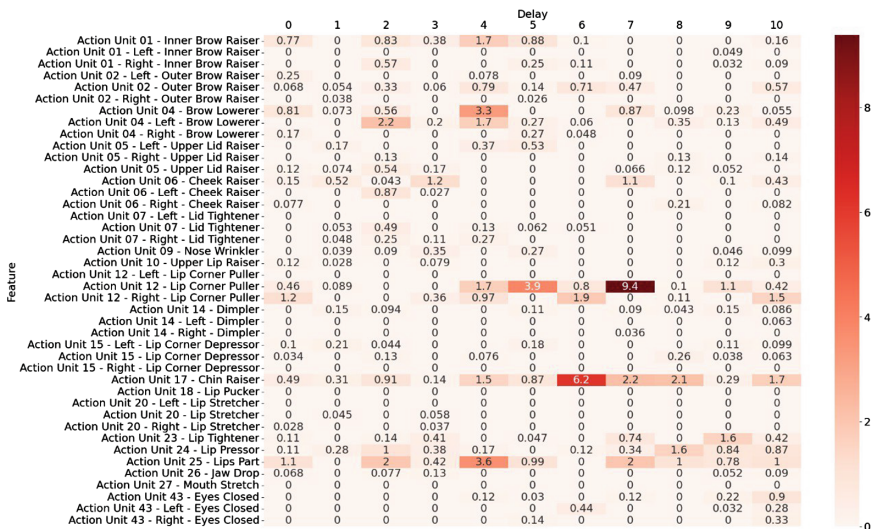


Fig. 5. Feature importance values heatmap for arousal (x axis corresponds to the embedding depth of a feature, and y axis corresponds to an action unit name).

features are Action Unit 12 (Lip Corner Puller) and Action Unit 24 (Lip Pressor). Note that the preceding values of these features are also of high importance. So we conclude

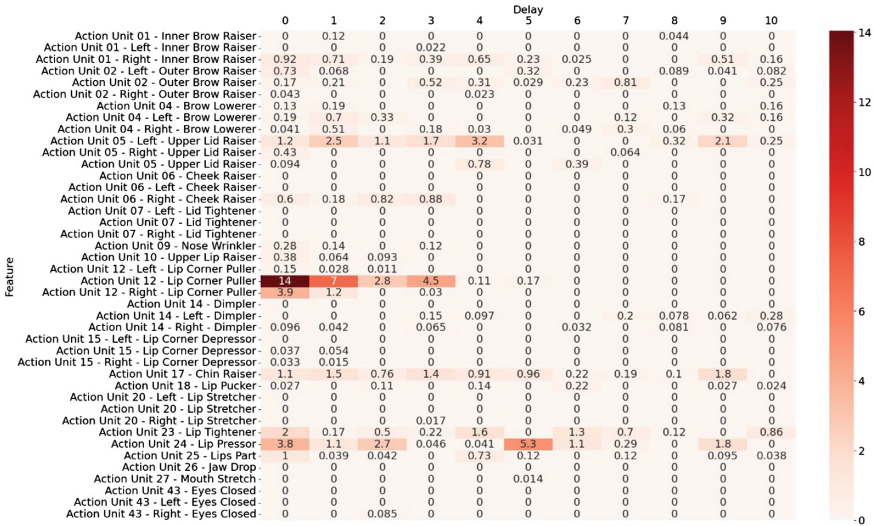


Fig. 6. Feature importance values heatmap for valence (x axis corresponds to the embedding depth of a feature, and y axis corresponds to an action unit name).

that Action Unit 12 (Lip Corner Puller) is most important for both Valence and Arousal prediction.

5 Conclusions

In the paper we tested various machine learning models against the task of mapping Action Units to the emotion space, represented by Valence and Arousal. This mapping is a first step in the loop of nonverbal communication between a person and a virtual actor using facial expressions.

We tested these models on an out-of-sample test set, to ensure that the mapping works on unseen data, and measured their R² score.

We exploited a data preprocessing technique called delay embedding to take the history of the AU into account, and studied whether it improves the performance of the models. For hyperparameters selection, we used a cross-validation technique.

We concluded that the delay embedding generally improves the performance of the models on the problem solved, but the cross-validation does not.

The best model to perform the mapping, based on our results, is gradient boosting with delay embedding of depth 10 (Catboost implementation). It restores Valence and Arousal with R² scores 0.717 and 0.494, respectively.

We also studied the way the best model makes its prediction, and calculated the importance of each Action Unit. We found that the most important Action Units for Arousal prediction are Action Unit 12 (Lip Corner Puller) and Action Unit 17 (Chin Raiser)., and for Valence prediction they are Action Unit 12 (Lip Corner Puller) and (Action Unit 24 - Lip Pressor).

Future studies should include reduction of the dimensionality of the Action Unit space by feature selection or feature extraction prior to application of machine learning methods. Some improvement of the results may be possibly achieved by applying different other models to this problem, such as neural networks, and comparing their results with those of gradient boosting. We also consider collecting even more data to make our conclusions more robust.

The developed technique can form the basis for emotional communication through facial expression and as such will be useful in many practical domains, including, for example, intelligent tutoring systems controlled by cognitive models [15].

Acknowledgement. This study was supported by the Russian Science Foundation grant no. 22-11-00213, <https://rscf.ru/en/project/22-11-00213/>.

References

1. Hjortsjö, C.H.: Man's face and mimic language. Studentlitteratur (1969)
2. Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)
3. Zadeh A., Lim Y.C., Morency L.P.: OpenFace 2.0: facial behavior analysis toolkit. in: tadas baltrušaitis. In: IEEE International Conference on Automatic Face and Gesture Recognition (2018)
4. Facereader homepage. <https://www.noldus.com/facereader>. Last accessed 11 Aug 2023
5. Mehrabian, A.: Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies. Oelgeschlager, Gunn & Hain, Cambridge, MA (1980)
6. Russell, J.A.: A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**(6), 1161–1178 (1980)
7. Shirokiy V.R. et al.: The loop of nonverbal communication between human and virtual actor: mapping between spaces. In: Proceedings of the 11th Annual Meeting of the BICA Society. Brain-Inspired Cognitive Architectures for Artificial Intelligence: BICA* AI 2020, pp. 484–489, Springer International Publishing (2021)
8. Samsonovich, A.V.: Emotional biologically inspired cognitive architecture. *BiolInspired Cogn. Arch.* **6**, 109–125 (2013). <https://doi.org/10.1016/j.bica.2013.07.009>
9. Samsonovich, A.V.: Schema formalism for the common model of cognition. *BiolInspired Cogn. Arch.* **26**, 1–19 (2018). <https://doi.org/10.1016/j.bica.2018.10.008>
10. Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
11. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**(5), 1189–1232 (2002)
12. Ke, G., et al.: Lightgbm: a highly efficient gradient boosting decision tree. *Adv. Neural. Inf. Process. Syst.* **30**, 3146–3154 (2017)
13. Prokhorenkova, L., et al.: CatBoost: unbiased boosting with categorical features. In: 32nd Conference on Neural Information Processing Systems, pp. 6638–6648. Montreal, Canada
14. Samsonovich, A.V., Ascoli, G.A.: Cognitive map dimensions of the human value system extracted from natural language. *Front. Artif. Intell. Appl.* **157**, 111–124 (2007). ISSN: 09226389
15. Samsonovich, A.V., De Jong, K.A., Kitsantas, A., Peters, E.E., Dabbagh, N., Kalbfleisch, M.L.: Cognitive constructor: an intelligent tutoring system based on a biologically inspired cognitive architecture (BICA). *Front. Artif. Intell. Appl.* **171**(1), 311–325 (2008). ISSN: 09226389



Principles of Creating Hybrid Intelligent Information Systems Based on the Granular-Metagraph Approach

Yuriy E. Gapanyuk¹✉, Valery I. Terekhov¹, Vitaly Y. Ivlev¹, Yuriy T. Kaganov¹, Irina S. Karabulatova^{1,2}, Mikhail B. Oseledchik¹, and Dmitry V. Semenov¹

¹ Bauman Moscow State Technical University, Baumanskaya 2-Ya 5, 105005 Moscow, Russia
gapyu@bmstu.ru

² MIPT, Lomonosov Moscow State University, Moscow, Russia

Abstract. The article is devoted to the architecture of an intelligent system based on Hybrid Intelligent Information System approach. The article expands the basic Hybrid Intelligent Information System approach by adding the ability to work with multiple environments, the granular-metagraph approach is also proposed. The main provisions and definitions of granulation of information, the basics of the metagraph model, and the main provisions of the basic approach of Hybrid Intelligent Information Systems are briefly considered. The formal environment model definition is given. Various multi-environmental Hybrid Intelligent Information System architecture alternatives are considered in detail. Four architecture alternatives are proposed: architecture with separate modules of the subconsciousness, architecture with a monolithic module of the subconsciousness, architecture with partially monolithic modules of the subconsciousness, architecture with a hierarchical module of the subconsciousness. The granular-metagraph architecture of the Hybrid Intelligent Information System is proposed and it is shown that the basic properties of information granulation are holds for it.

Keywords: Intelligent system architecture · Hybrid intelligent information system · System environment · System subconsciousness · System consciousness · Metagraph · Metavertex

1 Introduction

Modern systems based on traditional methods of the connectionist theory of artificial neural networks have reached an unprecedented rise in the last 4–5 years.

Convolutional neural networks have made significant progress in the development of pattern recognition systems that leave behind human capabilities. Recurrent neural networks provided the possibility of adequate translation from one language to another.

Even more significant success for artificial neural networks was provided by attention mechanisms [1] and transformers [2]. Built on their basis, deep neural networks such as BERT [3], PaLM [4], and GPT [5] have become a revolution in this area. We can note the trend of combining different neural network architectures into one ensemble

architecture to solve a complex problem, such ensemble architectures began to be called “end-to-end” solutions.

Currently, for large models, such a term has appeared as “foundation model”. According to [6] “foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks”. The main difference between “end-to-end” solution and foundation model is that “end-to-end” solutions is mainly focused on solving a single complex problem, while foundation model is usually focused on solving several complex problems.

In general, we can note the trend of development of models from simple machine learning models to ensemble models (including neural network models), then to complex “end-to-end” neural networks solutions, then to foundation models, and in the future to the complex machine learning based cognitive architectures.

At the same time, most complex neural network architectures have the following disadvantages:

- Lack of a developed explanatory component (which is to a lesser extent inherent in foundation models).
- Complexity and high cost of training models.
- Difficulty in reusing fragments of trained models.

To solve these problems, the concept of Hybrid Intelligent Information System (HIIS) was proposed in [7] as a kind of intelligent system architecture. In this article, this concept is expanded, the concept of a multi-environmental HIIS is introduced, and a granular-metagraph architecture is proposed on its basis.

The article is structured as follows. Section 2 discusses the basic models in order to understand the concepts proposed in the article: the main provisions of information granulation, brief information about the metagraph model, the main provisions of the basic HIIS model. Section 3 defines the HIIS environment based on the metagraph model. Section 4 introduces the concept of a multi-environmental HIIS. Section 5 introduces a granular-metagraph architecture based on multi-environmental HIIS and considers its granular properties.

2 Preliminary Information About the Models Used

2.1 The Main Provisions of Information Granulation

In this section, we rely heavily on the systematic view of granulation methods given in the article [8] and the representation of the metagraph as a granular model, discussed in article [9].

According to Merriam–Webster’s Dictionary [10] granule may be defined as “a small particle, especially one of numerous particles forming a larger unit.”

The significant property of the granule is atomicity which means that a given element can be clearly distinguished from the surrounding elements of the external system.

According to [8], granules “could be further decomposed into smaller or finer granules called subgranules.” According to [11], a holon is “the whole, considered at the same time as part of the whole.” Thus, the hierarchical organization of granules may be considered as a special case of holonic organization.

According to L. Zadeh [12], there are two operations to form and separate granules, “granulation involves a decomposition operation of whole into parts. Conversely, organization operation involves an integration of parts into whole.” Using the organization operation, it is possible to obtain “higher-order information granules” [8].

According to [13], “granular relationships may be classified into two groups: interrelationship and intrarerelationship.” We consider the internal links in the granules (intrarerelationships) and the external links between the granules (interrelationship).

2.2 The Brief Information About the Metagraph Model

The metagraph model may be considered as a family of complex graph models. Initially proposed by A. Basu and R. Blanning in 2007 [14], the model later received a number of extensions independently offered by various groups of researchers [15]. For this article, we rely on the “annotated metagraph model.” This is a variant of the metagraph model described in the article [15].

The metagraph model may be described as follows: $MG = \langle V, MV, E \rangle$, where MG – metagraph; V – set of metagraph vertices; MV – set of metagraph metavertices; E – set of metagraph edges.

Metagraph vertex is described by a set of attributes: $v_i = \{atr_k\}$, $v_i \in V$, where v_i – metagraph vertex; atr_k – attribute.

Metagraph edge is described by a set of attributes, the source and destination vertices, and edge direction flag: $e_i = \langle v_S, v_E, eo, \{atr_k\} \rangle$, $e_i \in E$, $eo = true|false$, where e_i – metagraph edge; v_S – source vertex (metavertex) of the edge; v_E – destination vertex (metavertex) of the edge; eo – edge direction flag ($eo = true$ – directed edge, $eo = false$ – undirected edge); atr_k – attribute.

The metagraph fragment: $MG_i = \{ev_j\}$, $ev_j \in (V \cup E \cup MV)$, where MG_i – metagraph fragment; ev_j – an element that belongs to the union of vertices, edges, and metavertices.

The metagraph metavertex: $mv_i = \langle \{atr_k\}, MG_j \rangle$, $mv_i \in MV$, where mv_i – metagraph metavertex belongs to set of metagraph metavertices MV ; atr_k – attribute, MG_j – metagraph fragment.

Thus, metavertex, in addition to the attributes, includes a fragment of the metagraph. The presence of private attributes and connections for metavertex is a distinguishing feature of the metagraph. It makes the definition of metagraph holonic – metavertex may include a number of lower-level elements and, in turn, may be included in a number of higher-level elements.

2.3 The Main Provisions of the Basic HIIS Model

In the article [7], a variant of the architecture of an intelligent system was proposed based on the approach of Hybrid Intelligent Information Systems (HIIS). The main elements of this architecture are:

- The subconsciousness module (MS) associated with the environment. The main task of the MS is to ensure the interaction of HIIS with the environment (ENV).
- The module of consciousness, which is responsible for making decisions.
- A communication module that is used to interact with other HIIS.

- In the article [16], another element of HIIS was proposed – the boundary model of consciousness and subconsciousness. This is an ontological structure of data and knowledge, in which the subconsciousness module places information recognized from the environment. The consciousness module reads this information and makes decisions based on it.

The article [7] noted that the modules of the subconsciousness and consciousness are built on the basis of a metagraph model of data and knowledge and are implemented using metagraph agents.

The main advantage of the HIIS approach is that it allows you to combine conflicting requirements for the accuracy and interpretability of the model. The subconsciousness module implements the requirements for the accuracy of recognition of objects and their properties in the environment. The module of consciousness is built on the basis of interpretable rules.

In general, despite the fact that the basic HIIS model is good enough and can be used to build architectures of intelligent systems, it has the following limitations:

1. Basic HIIS architecture can work with only one environment, which makes it difficult to create systems based on the HIIS architecture, in which it is necessary to work with several environments.
2. Before the appearance of the boundary model of consciousness and subconsciousness [16], it was not possible to make decisions based on complex situations.
3. The possibilities of integrating several HIIS into a single system were poorly developed. A communication module was provided in the HIIS structure, but it was not used in any way when creating the system architecture.
4. The environment model was not formally presented.

In this article, we are further developing the architecture proposed in [7]. To overcome limitations 1 and 2, we are developing a multi-environmental HIIS (MHIS) structure. To overcome limitation 3, we develop a granular-metagraph architecture. To overcome limitation 4, we formally define an environment model.

3 Environment Model Definition

The *env* environment is a (possibly organized) set of *atfi* artifacts. If we are not interested in the organization of artifacts, then we can use a set of artifacts as the first option for defining the environment: $env^1 = \{atfi\}$.

If we are interested in the organization of artifacts, then the second option for defining the environment is a metagraph, whose vertices and metaverices are artifacts: $env^2 = MG[\forall v, mv \equiv atfi]$.

This definition allows for a holonic organization of artifacts in the environment, when a higher-order artifact includes lower-order artifacts and links between them.

4 Multi-environmental HIIS

In the basic HIIS architecture proposed in [7], the subconsciousness module was associated with only one environment.

We propose to expand the structure of HIIS so that it can work with several environments. Let's call such a system a multi-environmental HIIS (MHIIS).

The ability to work with multiple environments greatly affects the structure of HIIS and gives rise to several architectures, each of which can be applied in certain conditions.

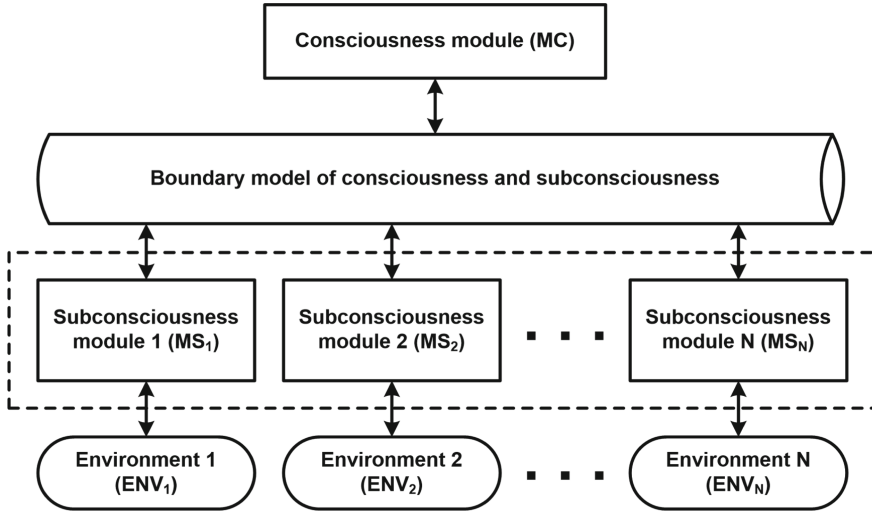


Fig. 1. The architecture with separate modules of the subconsciousness.

In the architecture with separate modules of the subconsciousness (shown in Fig. 1), HIIS contains N independent modules of the subconsciousness. Each of the modules performs independent interaction with the boundary model of consciousness and subconsciousness, forming in it images of recognized fragments of the environment. Thus, the rules of the module of consciousness are executed on the basis of information received from several environments.

This architecture can be used if the environments are completely independent, but the system needs to accumulate and process information from several environments.

In the architecture with a monolithic module of the subconsciousness (shown in Fig. 2), HIIS contains a single subconsciousness module.

This architecture can be used if the environments are dependent, including synchronized in time. The basis for the implementation of the subconsciousness module can be arbitrary methods of soft computing.

For example, in a foreign language learning support system, one environment is the voice of the student, and the second is the video images of his face during the pronunciation of the text. Thus, it is possible to associate the student's phonetic errors with his incorrect articulation.

In this architecture, the basis for the implementation of the subconsciousness module can be complex neural network ensembles designed for multimodal learning or foundation models.

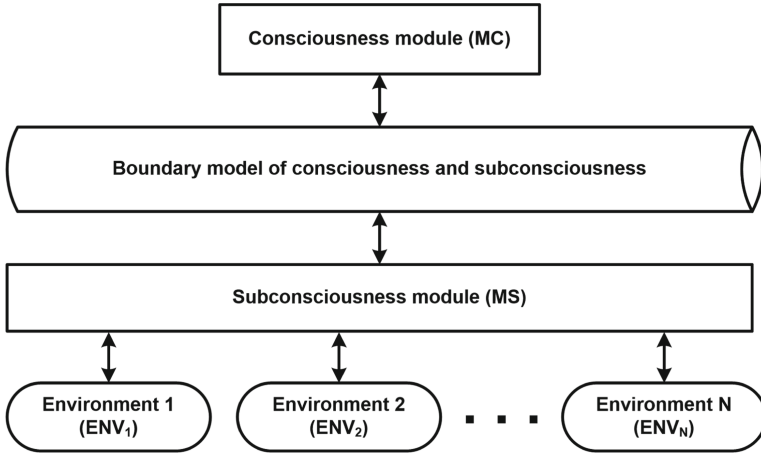


Fig. 2. The architecture with a monolithic module of the subconsciousness.

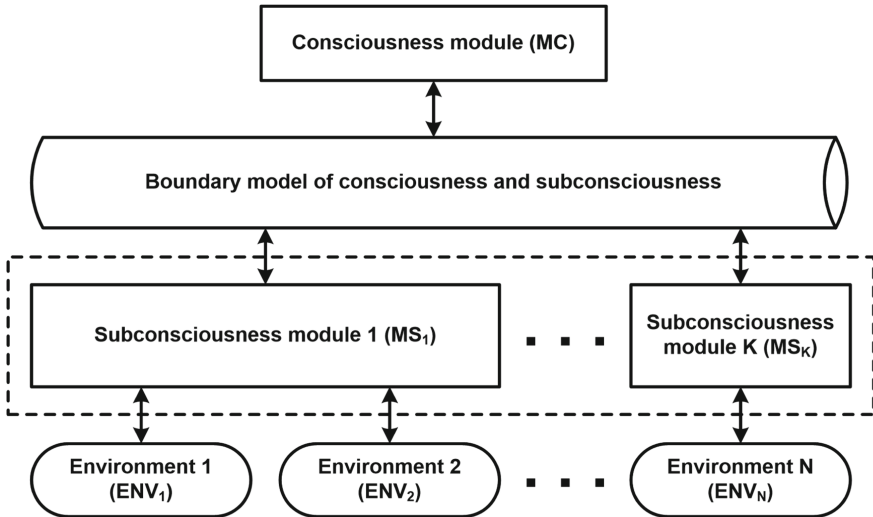


Fig. 3. The architecture with partially monolithic modules of the subconsciousness.

In the architecture with partially monolithic modules of the subconsciousness (shown in Fig. 3), HIIS can contain both separate modules of the subconsciousness for some environments, and monolithic modules of the subconsciousness for some environment groups. In this case, for N environments, K subconscious modules are used, and the condition holds $K < N$.

This variant of the architecture makes it possible to form holistic information in the boundary model of consciousness and subconsciousness both on the basis of separate independent environments and a group of environments.

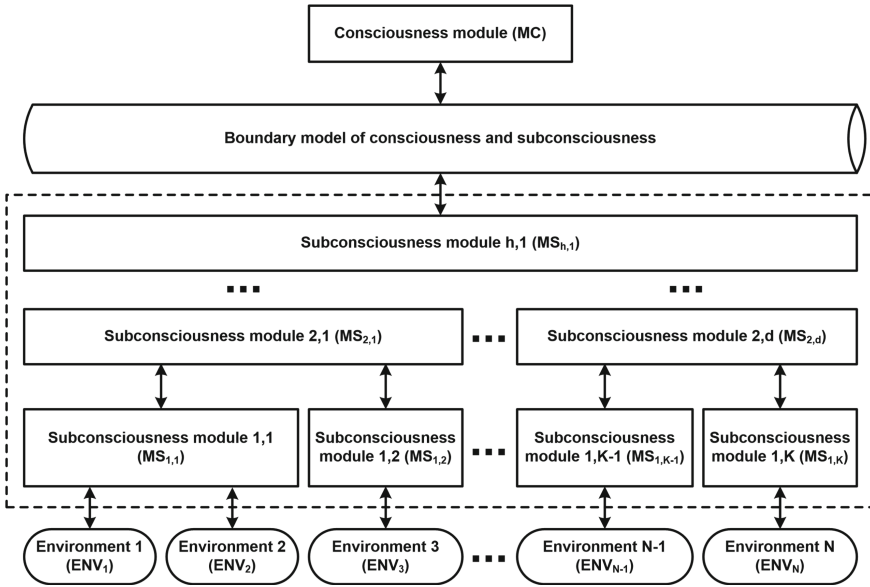


Fig. 4. The architecture with a hierarchical module of the subconsciousness.

In the architecture with a hierarchical module of the subconsciousness (shown in Fig. 4), the architecture with partially monolithic modules is used as the base one. But the first level of subconscious modules (the first index of these modules is equal to 1, and the second corresponds to the serial number of the module) is not directly connected with the boundary model of consciousness and subconsciousness, but is connected with the subconscious modules of the second level (the first index of these modules is equal to 2). The hierarchy of modules can have an arbitrary depth h , and only the level module h (whose first index is equal to h) is associated with the boundary model of consciousness and subconsciousness.

The architecture with a hierarchical subconscious module is somewhat similar to the ensemble stacking model used in machine learning. But the hierarchy in the stacking model does not exceed two levels (the second level model is called the metamodel), and the hierarchy of subconscious modules can have an arbitrary depth.

It should be noted that the architecture with a monolithic module of the subconsciousness is theoretically the most common, other architectures can be reduced to a monolithic case (in the figures of these architectures, such a monolith is shown by a rectangle with a dashed line). In the case of architectures with separate modules and architectures with partially monolithic modules, the monolith can break up into separate independent fragments. In the case of a hierarchical architecture, a monolith is a hierarchy of individual modules.

But in practice, reducing several modules to a single monolith is not always advisable, since each module can be a separate model based on soft computing. Also, training individual neural network models is much easier than training complex ensemble neural network architecture.

An important feature of a multi-environmental HIIS is the possibility of using it as a converter between environments. In this case, bidirectional connections from the environment to the subconsciousness module become unidirectional at a certain point in the operation of the HIIS – input or output.

It is also important that at different times of the HIIS processing, different environments can be input and output, which is shown in Fig. 5. At time t_1 , the first environment is the input, and the second environment is the output. At time t_2 , on the contrary, the first environment is the output, and the second environment is the input.

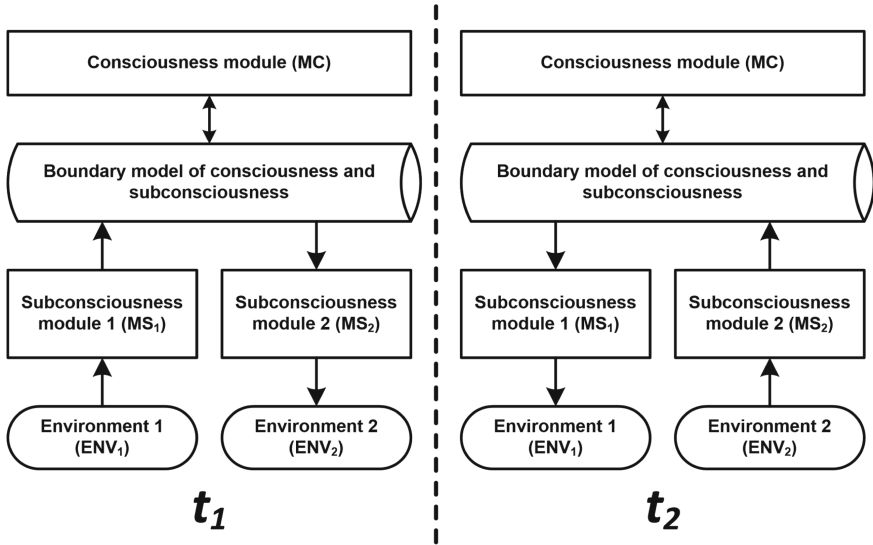


Fig. 5. The HIIS architecture as a converter between environments.

Thus, the MHIIS approach allows using the advantages of the HIIS approach in relation to several environments.

5 The Granular-Metagraph Architecture Based on MHIIS

5.1 Architecture Development

Let's unite set of MHIIS into a complex metagraph structure, which may include the following elements:

- Separate environments to be handled by the system. Correspond to metagraph vertices.
- MHIIS used for environments processing. Also correspond to metagraph vertices.
- MHIIS of the highest order, which we will call METAHIIS. Correspond to the metaverices of the metagraph.

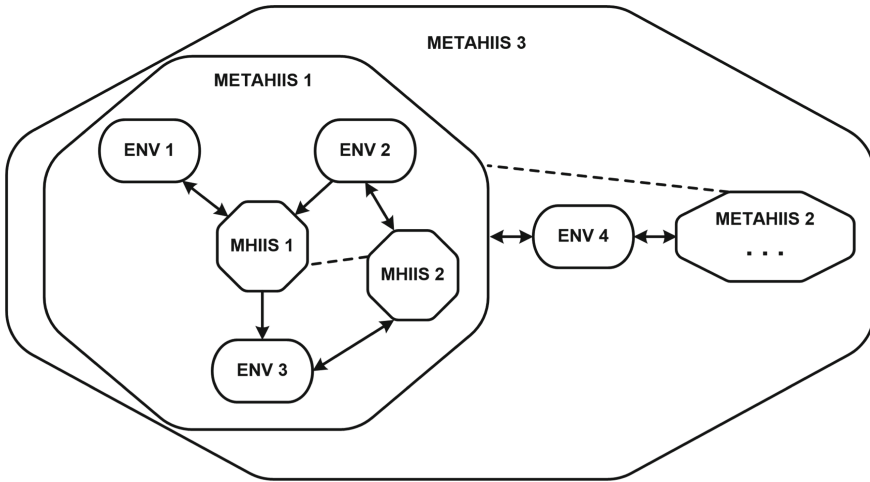


Fig. 6. An example of a granular-metagraph architecture based on MHIIS.

An example of a granular-metagraph architecture based on MHIIS is shown in Fig. 6. Environments are shown as ellipses, while MHIIS and METAHIIS are shown as octagons.

The system consists of two METAHIIS (METAHIIS 1 and METAHIIS 2) connected via the ENV 4 environment, which are nested in METAHIIS 3. The internal structure of METAHIIS 2 is not shown in detail so as not to clutter up the figure.

METAHIIS 1 includes three environments: ENV 1, ENV 2 and ENV 3 and two MHIIS: MHIIS 1 and MHIIS 2. MHIIS 1 interacts with the ENV 1 environment both for reading and writing, with the ENV 2 environment for reading only and with the ENV 3 environment for writing only. MHIIS 2 interacts with ENV 2 and ENV 3 environments both for reading and writing.

Links between ENV vertices (corresponding to environments) are not possible.

Links between MHIIS and METAHIIS (shown by dashed lines) are possible through the HIIS communication modules, but are not considered in detail in this article.

We give a formal definition of such architecture: $ARCH = MG[mv \equiv METAHIIS, (v \equiv ENV | MHIIS)]$. The architecture is a metagraph MG , whose metaverices mv are METAHIIS, and the vertices v can be either environment (ENV) or MHIIS.

Let's define METAHIIS as: $METAHIIS = mv, MHIIS^{SUP}$. METAHIIS is a metaver-
 tex mv of the architecture metagraph, to which $MHIIS^{SUP}$ is attached, which performs the function of a supervisor. For $MHIIS^{SUP}$ all environments included in METAHIIS are available for reading and writing. $MHIIS^{SUP}$ acts as a converter between internal environments included in METAHIIS and external environments.

5.2 Granularity of the Proposed Architecture

In this section we will show that the proposed architecture satisfies the basic properties of granularity discussed in Sect. 2.1.

Also we rely on the results of article [9], which shows that the metagraph as a structure satisfies the properties of granularity.

The atomicity property is obvious, since all the elements of the granular-metagraph architecture can be distinguished from the surrounding elements.

The holonic organization of architecture follows from the fact that it is represented in the form of a metagraph.

The METAHIIS may be considered as a “higher-order information granule”. The organization operation may be used to organize different MHIIS into METAHIIS.

We can consider intrarelations as links between MHIIS inside single METAHIIS and interrelationship as links between different METAHIIS. At the same time, it is only necessary to take into account that both MHIIS and METAHIIS are connected through the corresponding environments.

Other granular properties discussed in [9] (e.g. refinement and coarsening, similarity relationship) also hold for the proposed architecture due to its metagraph representation.

6 Conclusions

Modern intelligent systems, as a rule, are monolithic neural network architectures (the so-called “end-to-end” models). Currently, foundation models are increasingly being used. The main problem of this approach is the need to train a large monolithic model. This is a complex and expensive process, and it is almost impossible to use or reuse individual fragments of the model. Also, end-to-end models suffer from a lack of interpretability.

The basic HIIS-based approach made it possible to combine conflicting requirements for the accuracy and interpretability of the model. The subconsciousness module implements the requirements for the accuracy of recognition of objects and their properties in the environment. The module of consciousness is built on the basis of interpreted rules, which allows building a developed explanatory component of the model. At the same time, the HIIS-based approach allowed working with only one environment. Also, the possibilities of combining HIIS into more complex structures were not considered.

The MHIIS-based approach proposed in this article allows integrating several HIIS through the use of common environments.

The proposed granular-metagraph architecture makes it possible to build complex systems from individual components of MHIIS, combined using common environments.

Thus, the proposed approach allows creating intelligent systems both with high accuracy of object recognition (due to the presence of a subconsciousness module in HIIS) and with a developed explanatory component (due to the presence of a consciousness module in HIIS); allows training fragments of an intelligent system independently of each other (due to the use of granular-metagraph architecture); integrate individual modules based on the use of a common environment by using the MHIIS approach.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. [arXiv:1706.03762](https://arxiv.org/abs/1706.03762) (2017)
2. Amatriain, X., Sankar, A., Bing, J., Bodigutla, P.K., Hazen, T.J., Kazi, M: Transformer models: an introduction and catalog. [arXiv:2302.07730](https://arxiv.org/abs/2302.07730) (2023)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
4. Aakanksha Chowdhery, A. et al.: PaLM: scaling language modeling with pathways. [arXiv:2204.02311](https://arxiv.org/abs/2204.02311) (2022)
5. Zong, M., Krishnamachari, B.: A survey on GPT-3. [arXiv:2212.00857](https://arxiv.org/abs/2212.00857) (2022)
6. Bommasani, R. et al.: On the Opportunities and Risks of Foundation Models. [arXiv:2108.07258](https://arxiv.org/abs/2108.07258) (2021)
7. Chernenkiy, V., Gapanyuk, Y., Terekhov, V., Revunkov, G., Kaganov Y.: The hybrid intelligent information system approach as the basis for cognitive architecture. In: *Procedia Computer Science. Postproceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2018*, pp. 143–152 (2018)
8. Yao, J.T., Vasilakos, A.V., Pedrycz, W.: Granular computing: perspectives and challenges. *IEEE Trans. Cybern.* **43**(6), 1977–1989 (2013)
9. Tarassov, V., Kaganov, Y., Gapanyuk, Y.: The metagraph model for complex networks: definition, calculus, and granulation issues. In: Kovalev, S.M., Kuznetsov, S.O., Panov, A.I. (eds.) *RCAI 2021, LNCS*, vol. 12948, pp. 135–151. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86855-0_10
10. Merriam-Webster. <http://www.m-w.com/>. Last accessed 14 Aug 2023
11. Tarassov, V.B.: *From Multi-Agent Systems to Intelligent Organization*. Editorial URSS, Moscow (2002)
12. Zadeh, L.A.: Key roles of information granulation and fuzzy logic in human reasoning, concept formulation and computing with words. In: *Proceedings IEEE 5th International Conference Fuzzy System* (1996)
13. Yao, J.T.: Information granulation and granular relationships. In: *Proceedings IEEE Conference Granular Computing*, pp. 326–329. Beijing, China (2005)
14. Basu, A., Blanning, R.: *Metagraphs and their applications*. Springer, New York (2007)
15. Gapanyuk, Y.: The development of the metagraph data and knowledge model. In: *CEUR Workshop Proceedings. IMSC 2021-Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 10th International Conference on “Integrated Models and Soft Computing in Artificial Intelligence*, pp. 1–7 (2021). <https://ceur-ws.org/Vol-2965/paper01.pdf>
16. Taran, M.O., Revunkov, G.I., Gapanyuk, Y.E.: The text fragment extraction module of the hybrid intelligent information system for analysis of judicial practice of arbitration courts. In: Kryzhanovsky, B., Dunin-Barkowski, W., Redko, V., Tiumentsev, Y. (eds.) *Advances in Neural Computation, Machine Learning, and Cognitive Research IV. NEUROINFORMATICS 2020, SCI*, vol. 925, pp. 242–248. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-60577-3_28



Ontograph Cognitive Information Retrieval: Some Experimental Evaluations

Anastasia Gavrilkina^(✉) , Olga Golitsina , and Nikolay Maksimov 

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute),
Kashirskoe Shosse, 31, Moscow, Russia
asgavrilkina@yandex.ru

Abstract. The evaluation of the use of the ontographic approach in the context of cognitive information search is considered – a search focused on information support of cognitive processes of the main activity. The analysis of the applied approach was carried out on the example of identifying the chains of material and financial transactions that constitute a financial crime in combination. The factors influencing the effectiveness of the approach are identified. Estimates of the effectiveness of the use of ontographic approach by the combined service-user system are given, showing that the effectiveness depends not only on the effectiveness of IT tools, but also on the users competence – motivation, knowledge of the subject area, skills in using tools.

Keywords: Cognitive information retrieval · Ontologies · Knowledge graphs · Semantic networks · Financial crimes

1 Introduction

Cognitive information search in [1, 2] is defined as a system and technology with operands/operations and structures/processes similar to cognitive. “Cognitive” here indicates that the ultimate goal of the combined human-machine system “cognizing user – information system” is the ordering and synthesis of knowledge, i.e. the system of cognitive information search (unlike traditional systems of abstract, full-text, semantic search) should have functions and declarative tools that synchronize and informationally coordinate objects and processes in a mind and in the computing environment. This can be provided by the use of search technologies, including similar functions of consciousness, such as perception and understanding, categorization and classification, combination and ordering, etc.

The ontographic approach allows to represent semantics both at the level of the content of a specific text and at the conceptual (conceptual) level – a system of entity names (or concepts) and relationships. The representation of ontology at the datalogical level in the form of graphs makes it possible to formalize operations on ontologies based on graph-theoretic axioms, and to implement cognitive information search as a construction on a set of disparate chaotic facts represented by triplets, the path from the initial fact to the fact-result.

The ontology (knowledge) graph formed during the search is a reflexive image of the solution/state of the problem situation. Because reflexivity here is a reflection of the content of the text on the problem situation of the subject, this graph also represents a certain vision of the future. Due to the fact that in a graphic form the differences between the meaning of the content and the specifics of the linguistic structures representing it are leveled, interactively performed operations on graphs make it possible to connect logical (abstract) operations and specific schemes. This facilitates the perception, understanding, and evaluation of the content, including because the focus of attention is a compact holistic picture. In addition, in the context of the search process, the ontology graph represents the technological space of entry points into the information array, providing the possibility of a direct transition from the graph vertices to fragments of the document text.

However, a graph constructed according to the full text of the document is characterized mainly by a large capacity of many elements (it can be thousands of vertices and arcs even for a relatively small text), which determines the need to use tools for selecting and displaying fragments of the graph that are adequate to the type and state of the problem being solved.

Thus, the main actions performed during the search/analysis are the following:

1. Selection of an array of elementary facts with their subsequent integration into a graph.
2. Selection of potentially relevant concepts and supporting entities.
3. Search on the graph and navigate through the texts of documents.

The following mechanisms are provided within the cognitive search system:

- filtering (according to the entity name or relation class, aspect projection);
- formation of display of graphs in accordance with the search metaphor and transformations using graph operations, including semantic scaling;
- path searching;
- neighborhood searching.

In addition, it is advisable to use mechanisms related to the management of the focus of attention, which implement the ordering of vertices in accordance with some scheme. Such tools are not selection tools (reducing the set of elements for later viewing), but they may help to reduce enumeration (viewing) by ordering and graphically distinguishing the elements of the operational visual space.

Also, in visual processing, the use of filters (vertex names and/or relationship classes) and operations on graphs (union, intersection, projection) is effective, providing a controlled reduction/expansion of the operational space.

2 Description of the Experiment

The objective of this work is to experimentally evaluate the effectiveness of the interactive iterative use of ontographic tools of the service for visual ontological analysis of scientific and technical texts [3] on the example of identifying chains of material and financial transactions that together constitute a financial crime.

In this case, the effectiveness of use depends on two groups of factors:

- a set of mechanisms used by the subject;
- the user's preparedness to work with the tools, his professional orientation - understanding the specifics of the subject area and the ability to formalize goals, as well as see the proposed ways to achieve it, in particular, to identify supporting entities and connections that are key to presenting the situation.

Automated construction and analysis of the crime ontology is carried out on the basis of document texts obtained from open information resources by performing the following actions.

Data is collected from open resources on the actions, objects and subjects of the alleged financial crime. Based on these data, a text is formed with the most complete description of the financial crime (the original text). Based on the received source text, a graph is constructed, and if, as a result of a meaningful analysis of the graph for the presence of a crime scheme, errors were found, for example, the absence of links between entities or incorrect links, as well as incorrect selection of entities, then the text is edited.

The effectiveness of the use of ontographic tools by the combined "service-user" system was assessed comprehensively:

1. The quality of the graph construction was assessed by the degree of its semantic correspondence to the task of solving a crime. The problem is considered solved if a chain of facts has been constructed that reflects the essence of the criminal act, ideally closing the chain through the beneficiary: *initiator beneficiary – resource – recipient beneficiary*.
2. The motivation and professional level of the user were determined by an expert teacher when accepting and assessing work and knowledge.
3. The level of proficiency in the toolkit was assessed by the variety of types and number of actions performed by the user during processing, as well as the quality of the graph – the degree of approximation to the chain of facts closed to the beneficiary.

Users worked with the tools according to the "gray box" principle: they had knowledge in the amount of general information about information retrieval, ontologies, linguistic processors obtained as part of lectures on the course "Information resources in financial monitoring", and using the following recommendations for working with the service:

1. To increase the percentage of adequate results of automatic text processing, it was recommended to prepare the text beforehand: exclude/replace pronouns, simplify complex sentences, reformulate impersonal sentences and sentences without minor members, which can be transformed incorrectly, as not containing a pair of entity names and a relation for triplet formation.
2. To carry out search/analysis and reduce the dimension of the graph, the shortest path search mechanism can be used or a search by concept name can be carried out and a neighborhood search mechanism can be applied. To select sub-graphs, the following functions can be used: "Path", "Hiding vertices", selecting a group of vertices by long pressing or holding down the "Ctrl" key.

3 Results of the Experiment

As part of the experiment, 21 tests were conducted by different users (master's degree students) when performing practical tasks on the topic "The use of information resources in the investigation of financial crimes", each of which was reduced to the automated construction of an ontology representing a financial crime.

The table shows: (1) the problems that arise when using the ontographic tools; (2) the actions taken to resolve them; (3) the effectiveness of the action ("YES" if the problem can be considered solved, otherwise "NO"); (4) number of cases – the number of users acting in this way and their percentage of the total number (Table 1).

Table 1. Problems and assessment of the effectiveness of their solution.

Problem	Solution	Resultivity	Number of cases
Imperfection of the linguistic processor	Text editing (1 or more iterations)	YES	21 (100%)
	Text editing (3 to 5 iterations)	YES	9 (43%)
	Excessive text editing (unreasonable deviation from the rules of the Russian language)	NO	19 (91%)
Difficulty in perceiving the graph due to its large size	Shortening of the text is the exclusion of insignificant facts in the context of the task	YES	13 (62%)
Difficulty of determining the direction of "reading" the graph and identifying the target subgraph	Using the shortest path search function	YES	5 (24%)
		NO	6 (29%)
	Expert pathfinding (without using the service)	YES	3 (14%)
Lack of vertices due to insufficient data in the text	Adding text from additionally found sources	YES	1 (5%)
Lack of understanding of the task/work assignment	–	NO	3 (14%)

All graphs constructed from the original texts had on average about 200 vertices. Errors of automatic graph construction were observed, such as the absence of links or erroneous linking of entities that do not correspond to the linking in the text, as well as the formation of vertices with different names representing one entity (due to its different naming in different parts of the text).

To obtain a connected crime graph, the original texts were edited, in particular, the following actions were performed:

- normalization of their own names (surnames were led to the nominative case, the names of the companies were unified, the pronouns were replaced by the names of the entities to which they indicate);

- division of complex and common sentences into simple;
- exclusion of insignificant fragments of the text (relating to the statements of the charges, indications of the sources, circumstances of the place and time, etc.)

At least 9 users were clearly used from 3 to 5 editing. Along with this, there were cases of excessive editing, for example, changing the cases of all words to nominative or verb forms to infinitive and excluding prepositions; forcibly combining several entity names into phrases using quotation marks or underscores, in which there is no splitting into words. But since such a text from the point of view of the natural language will lose semantic connectivity, this led to improper analysis of the text. Such actions indicate a misunderstanding of the creature of the linguistic processor.

To reduce the size of the graph, 13 users shortened the original text, mainly excluding fragments that are not key to the description of the crime.

Of the mechanisms provided by the service, such as filtering, ordering in accordance with the search metaphor, path search and neighborhood search, the “path search” mechanism turned out to be the most popular, since it allowed to build a chain of facts of a criminal act, specifying only the initial and final entities. 11 users used the “path search” mechanism to search and highlight the chain of facts of a criminal act, while:

- in 3 cases, the highlighted path (which, according to the algorithm used, is the shortest) did not include key facts, which required additional indication of intermediate vertices in order to include arcs in the path representing the types of relationships corresponding to the description of the process of committing a crime;
- in 2 cases, it was impossible to construct a path showing the crime on the graph;
- in 1 case, the highlighted path included a criminal act, but not all connections were correct;
- in 5 cases, a path was built showing the essence of the criminal act.

The analysis of actions and constructed graphs taking into account the facts of providing explanations by users, as well as an expert (teacher) assessment of work and knowledge allowed to qualify 12 users (57%) as motivated, and 9 (43%) – as not motivated. Among motivated users, 7 (33%) can be noted as skilled in using tools of the service.

4 Conclusion

General conclusion: the interactive use of ontographic tools is effective, that is, connected graphs representing financial crimes were obtained. Of the 11 cases of the application of functions in 5, an “perfect” result was achieved, and the remaining showed that it would be possible to use other functions that the experimenters did not use, which indicates a poor knowledge of the tools.

Regarding the subjects of the search, as an integral part of the aggregate information system “user-service”, it can be stated that 9 (43%) users had no motivation - they needed instructions from a teacher. This can be interpreted as evidence that there is no personal interest in understanding the problem, and/or there is not enough knowledge of the subject area.

Acknowledgements. This work was supported by the Ministry of Science and Higher Education of the Russian Federation (state assignment project No. FSWU-2023-0031).

References

1. Maksimov, N., Golitsina, O.: About cognitiveness of information retrieval. Proc. Comput. Sci. **213**, 317–324 (2022). <https://doi.org/10.1016/j.procs.2022.11.073>
2. Lebedev, A.A., Gavrilkina, A.S., Maksimov, N.V., et al.: Onto-graphic mechanisms for deep semantic search. Autom. Doc. Math. Linguist. **56**(4), 163–178 (2022)
3. Maksimov, N.V., Golitsina, O.L., Monankov, K.V., Gavrilkina, A.S.: Opytnyj obrazec servisa vizual'nogo ontologicheskogo analiza nauchno-tehnicheskikh tekstov. Svidetel'stvo ob ofitsial'noi registratsii programm dlya EVM [The prototype of the service of visual ontological analysis of scientific and technical texts. The Certificate on Official Registration of the Computer Program]. No. 2021610648 (2021) (in Russian)



Emotion-Integrated Cognitive Architectures: A Bio-Inspired Approach to Developing Emotionally Intelligent AI Agents

Aliya Grig^(✉) and Anastasia Rizzo

Evolve Inc., 251 Little Falls Drive, Wilmington, New Castle County, Delaware 19808, U.S.
aliya@evolwe.ai

Abstract. The pursuit of biologically inspired cognitive architectures (BICA) has driven significant advancements in artificial intelligence (AI) and artificial general intelligence (AGI). However, most existing BICA models lack a critical aspect of human intelligence: emotions and feelings. This research explores the development and implementation of an emotion-integrated cognitive architecture that mimics human emotional processing within a computational framework. Our proposed architecture, Emotion-Integrated Cognitive Architecture (EICA), is inspired by the latest findings in cognitive psychology, neurobiology, neuroscience and affective computing. EICA aims to integrate emotional processing into the core of the AI system, enabling robust, flexible, and adaptable AI agents that can respond to complex and dynamic environments with human-like emotional intelligence. The EICA model leverages advances in brain imaging and recording techniques to draw insights from the neural basis of emotions in humans. The architecture incorporates emotion-generating, recognition, and regulation mechanisms, allowing AI agents to perceive, interpret, and respond to emotions in themselves and others. We present the concept of EICA, including its modular structure and interaction with other cognitive components. We also provide case studies showcasing EICA's successful implementation in various AI applications, such as virtual assistants and adaptive robotics. This research represents a significant step towards achieving the BICA Challenge by advancing the computational replication of human emotional intelligence. By integrating emotions and feelings into AI systems, we move closer to realizing the full potential of bi-directional understanding between artificial and biological intelligences.

Keywords: Emotion-integrated cognitive architecture · Artificial intelligence · Emotional intelligence

1 Introduction

The development of Artificial Intelligence (AI) has become a topic of great interest in recent times. With its ability to process vast amounts of data and carry out complex tasks with accuracy, AI has already made significant contributions in fields such as healthcare, finance, and transportation. However, despite the impressive capabilities of

AI agents, one crucial aspect remains elusive: emotional intelligence. Emotions play a vital role in human decision-making, learning, and social interactions. Therefore, the lack of emotional processing in existing AI models poses a significant limitation in the bid to create agents that can effectively interact with humans. Current biologically inspired cognitive architectures neglect the role emotions play in human cognitive processes. These models attempt to integrate human-like capabilities into software agents. However, they lack the ability to exhibit emotional intelligence, which is critical in human-robot interactions, especially in healthcare settings and education.

As such, our recent research has focused on developing emotion-integrated cognitive architectures (EICA) to provide AI agents with the ability to process and understand emotions effectively. The integration of emotions into AI agents can improve their ability to perform in complex environments, such as in decision-making, social interactions, and human-robot interactions. Additionally, the integration of emotions into AI agents can enhance the development of cognitive architectures that can improve the quality of human-robot interactions in healthcare and education settings. The implementation of emotions is a crucial aspect in several fields of IT and AI, particularly in the realm of robotics. The capacity of a system to experience emotions such as fear, interest, or joy can trigger behaviors that were previously unattainable during the era of emotionless machines [1]. This paper presents an in-depth analysis of EICA, highlighting its potential to significantly impact the field of Artificial Intelligence. It aims to contribute to the ongoing conversation about the need for emotional intelligence in AI agents and the role of EICA in addressing this critical gap.

2 Research Background

Biologically-inspired cognitive architectures (BICA) hold great potential for creating intelligent systems that can perceive, reason, and adapt to complex real-world scenarios. However, existing BICA models have several limitations in emulating human-like capabilities that involve the processing of emotions. Emotions play a crucial role in human decision-making, learning, and social interactions, yet, most existing models do not incorporate them into their architectures effectively [2]. To address this gap, our researchers have focused on developing emotion-integrated cognitive architecture (EICA). EICA is a theoretical framework that aims to combine cognitive and affective processes in a unified architecture to create intelligent agents that can perceive, reason, and adapt with high emotional intelligence. Research findings from cognitive psychology, neurobiology, neuroscience, and affective computing have proposed several theories of emotion processing, including the appraisal theory [3], James-Lange theory [4], facial feedback hypothesis [5], and the somatic marker hypothesis [6]. These theories suggest that emotions are complex phenomena that involve cognitive appraisals, physiological reactions, and behavioral responses. These findings have influenced the development of EICA, which adopts a cognitive perspective on emotion processing based on the appraisal theory of emotion.

The appraisal theory suggests that emotions arise from the evaluation of the significance of the event or situation, and that different appraisals can lead to different emotional responses [7]. The integration of this theory into the EICA framework allows

the agent to perceive and understand the emotional states of others, adjust its own emotional responses and behaviors accordingly, and learn from experience to improve its emotional intelligence. According to the James-Lange theory, emotions arise from physiological responses to external stimuli, with the physical responses preceding the subjective emotional experience and influencing it [8]. The incorporation of this theory into the development of emotion-integrated cognitive architecture (EICA) has significant implications. By acknowledging the impact of physiological responses on emotional experiences, EICA can integrate mechanisms to recognize and process these responses, allowing AI systems to identify and respond to emotional cues more accurately and naturally, improving their capacity for empathetic human interaction [8]. Minsky (2007) wrote a seminal book on emotions and their integration into computing systems [4]. Among his numerous ideas, he described the significance of emotions and proposed various approaches for implementing them in computing.

Brain imaging and recording techniques have also contributed to the development of EICA, providing insights into the neural mechanisms underlying emotion processing. Neuroimaging studies have identified several brain regions involved in emotion processing, including the amygdala, prefrontal cortex, and anterior cingulate cortex. Advances in brain imaging techniques such as fMRI and EEG have enabled the analysis of the neural substrates of emotional processing in real-time [9]. This information is crucial to the development of EICA, as it provides valuable insight into how the brain integrates cognitive and emotional processes. This knowledge can be used to create intelligent agents that can perceive, learn, and adapt with high emotional intelligence. However, the lack of emotion processing in these models presents a significant limitation in their ability to interact with humans effectively. EICA has emerged as a promising solution to this problem, incorporating cognitive and affective processes into a unified architecture to create agents with high emotional intelligence. By adopting a cognitive perspective on emotion processing and leveraging advances in brain imaging and recording techniques, EICA can pave the way for the development of intelligent agents that can interact with humans in complex real-world scenarios with high emotional intelligence.

3 The Emotion-Integrated Cognitive Architecture

In recent years, the scientific community has shown an increasing interest in BICA which mimic the cognitive processes of the human brain. While these models have been promising in various applications of artificial intelligence and information technology, there are limitations, particularly in integrating emotions into cognitive processes [1]. It is well known that emotions play a crucial role in human behavior, yet traditional AI approaches often neglect or oversimplify emotions. This suggests a need for new methods in mind modeling that consider the dynamic and creative aspects of the mind.

3.1 Merging Cognition and Emotion

According to anatomical and functional studies, the brain exhibits extensive interconnectivity, particularly in the context of emotion [10]. Emotion-related processes are

mediated by large-scale cortical-subcortical networks that are sensitive to bodily signals, which allows for the integration of information related to perception, cognition, emotion, motivation, and action [11]. The functional architecture of the brain comprises multiple overlapping networks that are dynamic and context-sensitive, with the affiliation of a given brain region to a specific network varying according to task demands and brain state [12]. Although some brain regions exhibit greater interconnectivity than others, information can traverse the brain with minimal isolation of signals [13], implying a promiscuous architecture where information of different types is often mixed. Furthermore, a network perspective of brain organization provides insights into the importance of brain structures like the amygdala in emotion processing, as they act as hubs of large-scale connectivity systems. Additionally, the network perspective clarifies why emotion has a pervasive impact, as it is inextricably linked to cognition.

3.2 Modular Structure of EICA

One approach to implementing emotion-integrated cognitive architectures is to use a modular approach, where separate modules are used to model different aspects of emotional processing. For example, one module might be used to model the appraisal process, while another module might be used to model the generation of emotional responses. The modular structure of EICA is designed to incorporate emotion recognition and processing capabilities into cognitive systems in a flexible and adaptable manner. The system comprises multiple interconnected modules that interact with each other to perform specific functions [14]. These modules include sensory processing, perception, memory, reasoning, emotion recognition, and decision-making. Each module operates independently, but the interaction between modules allows for the intermixing of information related to perception, cognition, emotion, motivation, and action. This modular structure enables EICA to adapt to different environments and situations, making it a powerful tool for developing AGI. Cognitive architectures can be characterized as modular systems with a hierarchical structure [15]. They typically comprise low-level modules for action and perception, mid-level modules for task coordination and mid-level perception, and higher levels for task planning and visual cognition [16]. A knowledge database is usually connected to the top level, while a central controller or supervisor coordinates the modules and utilizes temporary buffers as part of an active workspace. Although primarily hierarchical, parallel components and organization are also recognized as significant.

The advantage of modular approach is that it allows for a more fine-grained modeling of emotional processing, and it can be easier to integrate different models into a larger architecture. However, it can be challenging to design and integrate multiple modules, and it can be difficult to ensure that the modules are working together effectively. Another approach is to use a distributed model, where emotions are modeled as patterns of activation across a network of neurons. This approach is based on the idea that emotions are emergent properties of complex systems, and that they cannot be reduced to simple algorithms [17]. Instead, emotions are seen as arising from the interactions between different parts of the system. The advantage of this approach is that it provides a more holistic and integrated view of emotional processing, and it can capture the complexity and dynamic nature of emotions. However, it can be challenging to implement, as it requires a deep understanding of neural networks and the principles of distributed processing.

4 Case Studies

EICA models have been developed, characterized by the integration of emotion recognition and processing into cognitive systems. EICA relies on cortical-subcortical networks to intermix information on perception, cognition, emotion, motivation, and action. Although both EICA and BICA are biologically-inspired cognitive architecture frameworks, EICA focuses on integrating emotion recognition and processing, while BICA emulates the complex nature of biological cognitive systems [18].

4.1 Existing Implementations

Findings by Mishra & Tiwary (2019) suggest that the communication among different brain regions responsible for various functions, such as social context and self-related event processing, salient feature detection, attention, reward/punishment, hedonic value, and physiological sensations (which are discussed separately in the next section), leads to the creation of an emotional event [19]. Their model provides an explicit explanation of the nature of emotion and its universality, emphasizing that emotion is an integral part of the process underlying the brain's dynamic connectivity organization. The dynamic interactions between different brain regions create an affective subjective experience, which is referred to as an emotion. Furthermore, their model moves beyond the concept of the appraisal model by positing that emotion is encoded in experience, rather than being a mere reaction to an appraised stimulus. It is important to note that their model is based on calculated cognitive functions using neural decoding and MVPA analysis, rather than speculative arguments regarding the involvement of different cognitive functions, as is the case in the social constructionist model of emotions.

Furthermore, The NEUCOGAR (NEUromodulating COGNITIVE ARchitecture) by Vallverdú et al. (2016) aims to create a mapping between the influence of serotonin, dopamine, and noradrenaline on Von Neuman's architecture-based computing processes [1]. This mapping would enable the implementation of affective phenomena that can operate on Turing's machine model. To develop this mapping, the Lövheims Cube of Emotion [20] is used as a basis for modeling, and the parameters of the Von Neumann architecture are used to extend it. The validation of the model is conducted via simulation on a computing system of dopamine neuromodulation and its effects on the cortex. The experimental phase of the project confirms the soundness of the model through the increase in computing power and storage redistribution due to the emotion stimulus modulated by the dopamine system.

Evolve AI has developed a novel approach to creating emotion-integrated cognitive architecture (EICA) by utilizing psychometric and psycholinguistic data [21]. The company's model is designed to enhance the ability of AI systems to recognize and process emotional cues in human communication by incorporating emotional intelligence and empathy. To this end, the company leverages psychometric data to measure emotional traits and tendencies in humans, such as personality traits and emotional states, and psycholinguistic data to analyze the relationship between language and emotion. The incorporation of sentiment analysis techniques into Evolve AI's EICA model further strengthens its ability to recognize and process emotional cues. This is achieved by analyzing text and other forms of communication to identify and classify emotions and

sentiments expressed by individuals. This technique can be applied in various domains, such as social media monitoring, customer service, and healthcare, among others, to gain insights into the emotions and attitudes of individuals.

4.2 Comparison with BICA Models

The primary focus of EICA is to integrate emotion recognition and processing into cognitive systems, enabling AI systems to recognize and respond to emotions in a natural and human-like way. The framework utilizes large-scale cortical-subcortical networks that are distributed and sensitive to bodily signals to facilitate the mixing of information related to perception, cognition, emotion, motivation, and action. Multiple overlapping networks characterize EICA, which are highly dynamic and context-sensitive, with the affiliation of a given brain region to a specific network varying according to task demands and brain state [11]. On the other hand, BICA aims to develop artificial intelligence systems that emulate the complex and adaptive nature of biological cognitive systems [22]. Based on cognitive and computational neuroscience principles, the framework seeks to create AI systems that are more flexible, robust, and capable of learning from experience. BICA systems integrate multiple modalities of sensory input and can adapt and learn in new environments and situations. Although both EICA and BICA are biologically-inspired frameworks for developing AI systems, they differ in their specific objectives and areas of focus. EICA is geared towards integrating emotion recognition and processing into cognitive systems, while BICA concentrates on creating AI systems that are more adaptable, flexible, and capable of learning from experience [23].

5 Further Directions

To advance the development of Artificial General Intelligence (AGI), the incorporation of emotion-integrated cognitive architecture (EICA) has become increasingly crucial. One promising direction is to enhance EICA by incorporating advanced natural language processing and dialogue capabilities, which can facilitate more sophisticated emotional interactions between AI systems and humans. Another area of development is to integrate different modalities of sensory input into EICA, such as vision and touch, to create a more comprehensive emotional experience for AGI systems [24].

5.1 Potential Applications and Impact

Integrating Artificial General Intelligence (AGI) concepts in AI and robotics has significant potential in setting tasks from the Universal Computational Intelligence class. AGI methodology is currently used to create cognitive architectures, as seen in projects like the Learning Intelligent Distribution Agent based on the Global Workspace Theory [25], the Connectionist Learning with Adaptive Rule Induction On-line project integrating models from psychology [26], and the Adaptive Control of Thought - Rational architecture based on the Rational Analysis methodology [27]. A universal kernel could be created to build cognitive autonomous systems capable of purposeful activity in various applications and environments [28]. Hence, the integration of emotion recognition

and processing in cognitive systems through EICA is crucial for the development of AGI. Emotions play a significant role in human decision-making, problem-solving, and social interaction [29]. By integrating emotion recognition and processing capabilities into AGI, EICA can facilitate more sophisticated emotional interactions between AI systems and humans, ultimately making AGI more human-like and capable of operating in a wide range of environments and situations.

5.2 Challenges and Limitations

Implementing emotion-integrated cognitive architectures poses several challenges and limitations, both technical and ethical. Emotions are complex and dynamic processes that involve multiple components, such as appraisal, physiological responses, and action tendencies. Modeling this complexity requires a deep understanding of the different components and how they interact with each other. Modular approaches to modeling emotions require the integration of multiple modules, which can be challenging. Each module may use different methods and techniques, making it difficult to integrate them into a coherent framework. Additionally, there is a shortage of data on emotional responses in naturalistic settings, which limits the ability to develop models that accurately reflect the complexity of emotional processing. Implementing emotion-integrated cognitive architectures also requires significant computing resources, which may be difficult to obtain or expensive.

Ethical challenges include privacy concerns, as emotion-integrated cognitive architectures may involve collecting and processing sensitive data, such as physiological responses or personal information. This raises privacy concerns, as the data could be used for unethical purposes. Emotion-integrated cognitive architectures may also be subject to bias and discrimination, which could have negative impacts on individuals or groups. It is important to address these challenges and limitations to ensure that emotion-integrated cognitive architectures are developed and used in a responsible and ethical manner.

6 Conclusion

The development of cognitive architecture frameworks such as EICA has the potential to revolutionize the field of AI by enabling the creation of AGI that are capable of integrating emotions with cognitive processes [10]. The modular structure of EICA allows for the modeling of complex interactions between perception, cognition, and emotion, making it a powerful tool for developing intelligent systems that can adapt to different environments and situations. The implementation of EICA in various fields has the potential to transform the way humans interact with technology, enabling more personalized and engaging experiences. Future research in this area will focus on further developing and refining EICA to create more advanced AGI that can simulate human-like emotional responses.

References

1. Vallverdú, J., Talanov, M., Distefano, S., Mazzara, M., Tchitchigin, A., Nurgaliev, I.: A cognitive architecture for the implementation of emotions in computing systems. *Biol. Inspired Cogn. Architect.* **15**, 34–40 (2016)
2. Schuller, D., Schuller, B.W.: The age of artificial emotional intelligence. *Computer* **51**(9), 38–46 (2018)
3. Scherer, K.R.: *Appraisal theory* (1999)
4. Langley, P., Laird, J.E., Rogers, S.: *Cognitive architectures: Minsky, M.: the emotion machine: commonsense thinking, artificial intelligence, and the future of the human mind.* Simon Schuster (2007)
5. Pessoa, L.: Intelligent architectures for robotics: the merging of cognition and emotion. *Phys. Life Rev.* **31**, 157–170 (2019)
6. Dauth, S., et al.: Neurons derived from different brain regions are inherently different in vitro: a novel multiregional brain-on-a-chip. *J. Neurophysiol.* **117**(3), 1320–1341 (2017). <https://doi.org/10.1152/jn.00575.2016>
7. Maruyama, Y.: The conditions of artificial general intelligence: logic, autonomy, resilience, integrity, morality, emotion, embodiment, and embeddedness. In: *Artificial General Intelligence: 13th International Conference, AGI 2020, St. Petersburg, Russia, September 16–19, 2020, Proceedings*, vol. 13, pp. 242–251. Springer International Publishing (2020)
8. Brooks, R.A.: *Cambrian intelligence: the early history of the new AI.* MIT press (1999)
9. Mishra, S., Tiwary, U.S.: A cognition-affect integrated model of emotion. arXiv preprint [arXiv:1907.02557](https://arxiv.org/abs/1907.02557) (2019)
10. Serov, A.: Kernel based cognitive architecture for autonomous agents. arXiv preprint [arXiv:2207.00822](https://arxiv.org/abs/2207.00822) (2022)
11. Cannon, W.B.: The James-Lange theory of emotions: a critical examination and an alternative theory. *Am. J. Psychol.* **39**(1/4), 106–124 (1927)
12. James, W.: The emotions. In: Lange, C.G., James, W. (eds.), *The emotions*, vol. 1, pp. 93–135. Williams & Wilkins Co. <https://doi.org/10.1037/10735-003> (1922)
13. Phan, K.L., Wager, T., Taylor, S.F., Liberzon, I.: Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage* **16**(2), 331–348 (2002)
14. Evolve Inc. Retrieved May, 2023. from <http://evolwe.world>
15. Kelley, D.J., Waser, M.R., Sylvester, A.: Critical nature of emotions in artificial general intelligence. In: *Proceedings IEET*, pp. 1–5
16. Moors, A., Ellsworth, P.C., Scherer, K.R., Frijda, N.H.: Appraisal theories of emotion: state of the art and future development. *Emot. Rev.* **5**(2), 119–124 (2013)
17. Min, B.K., Hämäläinen, M.S., Pantazis, D.: New cognitive neurotechnology facilitates studies of cortical–subcortical interactions. *Trends Biotechnol.* **38**(9), 952–962 (2020)
18. Martínez-Miranda, J., Aldea, A.: Emotions in human and artificial intelligence. *Comput. Hum. Behav.* **21**(2), 323–341 (2005)
19. Samsonovich, A.V.: Emotional biologically inspired cognitive architecture. *Biol. Inspired Cogn. Architect.* **6**, 109–125 (2013)
20. Goertzel, B., Lian, R., Arel, I., De Garis, H., & Chen, S.: A world survey of artificial brain projects, Part II: Biologically inspired cognitive architectures. *Neurocomputing* **74**(1–3), 30–49
21. Chella, A., Pirrone, R., Sorbello, R., & Jóhannsdóttir, K. R.: *Biologically inspired cognitive architectures 2012*, vol. 196. Springer (2013)
22. Damasio, A.R.: The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philos. Trans. Royal Soc. London. Ser. B: Biol. Sci.* **351**(1346), 1413–1420 (1996)

23. Bertolero, M.A., Yeo, B.T.T., D'Esposito, M.: The modular and integrative functional architecture of the human brain. *Proc. Natl. Acad. Sci.* **112**, E6798–E6807 (2015). <https://doi.org/10.1073/pnas.1510619112>
24. Scherer, K.R.: Emotions are emergent processes: they require a dynamic computational architecture. *Philos. Trans. Royal Soc. London. Ser. B, Biol. Sci.* **364**(1535), 3459–3474 (2009). <https://doi.org/10.1098/rstb.2009.0141>
25. Stocco, A., Lebiere, C., Samsonovich, A.: The B-I-C-A of biologically inspired cognitive architectures. *Int. J. Mach. Conscious.* **02**(02) (2010). <https://doi.org/10.1142/S1793843010000552>
26. Lövheim, H.: A new three-dimensional model for emotions and monoamine neurotransmitters. *Med. Hypotheses* **78**(2), 341–348 (2012)
27. Baars, B.J.: In the theater of consciousness. Global workspace theory, a rigorous scientific theory of consciousness. *J. Conscious. Studies* **4**(4), 292–309 (1997)
28. Chong, H.Q., Tan, A.H., Ng, G.W.: Integrated cognitive architectures: a survey. *Artif. Intell. Rev.* **28**, 103–130 (2007)
29. Lieder, F., Griffiths, T.L.: Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* **43**, e1 (2020)



Experimental Phonetic Research Interlingual Interference and Accent in the Russian Speech of Native Speakers of the Kabardino-Circassian Language

Irina Gurtueva[✉] , Murat Anchekov , Kantemir Bzhikhatlov , Olga Nagoeva ,
and Ahmed Enes 

The Federal State Institution of Science Federal Scientific Center, Kabardino-Balkarian
Scientific Center of Russian Academy of Sciences, I. Armand Street, 37-a, 360000
Kabardino-Balkarian, Nalchik, Russia
gurtueva@yandex.ru

Abstract. The use of deep learning algorithms has made it possible to achieve a human parity in telephone conversational speech recognition. However, separate tasks, among which the problems of noise reduction, automatic segmentation of a mixed audio signal, and non-native speech recognition, remain open. The complexity of the accent speech recognition is the consequence of the discrepancy between the non-native speech and native language resources used in training, acoustic, language and pronunciation modeling based on data-driven approach. The existing approaches to solving these difficulties based on the use of non-native speech resources or multilingual corpora are ineffective. It is necessary to study the specifics of the psychophysiological mechanisms of bilinguals who use two or more language systems in communication. This paper shows the results of the experimental phonetic study of interlingual interference and accent in the Russian speech of native speakers of the Kabardino-Circassian language. The preliminary acoustic analysis of the initial formant measurements of non-native Russian speech of Kabardino-Circassian speakers at the phonetic level showed a systematic deviation of the acoustic patterns of vowel allophones (quantitative estimates of the vowel allophone [A] are given for brevity) in the F_1 - F_2 space. The aimed is to develop methods for modeling the non-native speech for subsequent use in speech recognition systems, language identification and accents in conditions of limited linguistic resources. The obtained experimental data can be useful for the development of the theory of contrastive acoustic analysis of vocal systems of languages with different structures, typical interfering pronunciation errors detection, and also for the development of theoretical models of foreign languages learning.

Keywords: Contrastive acoustic analysis · Automatic speech recognition · Accent identification · Speech perception

1 Introduction

The popularity of speech systems as a convenient interface has increased significantly with the achievement of high efficiency in solving certain tasks of speech recognition. Speech applications are embedded in a variety of products - from software and hardware systems with natural language user interface to biometrics and AI-based assistance ones. At the same time, although the developers of *IBM* and *Microsoft* have declared human parity in speech recognition [1], it is rather difficult to unequivocally assess the state-of-the-arts in automatic speech recognition. The conditions for evaluation and the effectiveness of any system can vary greatly depending on the parameters that determine the performance, rate and configuration of the speech recognition system. Almost any system, under specially selected conditions, achieves an accuracy comparable to that of a human.

Recent success in automatic speech transcription is mainly associated with the use of bidirectional recurrent networks [1]. But this algorithm is characterized by a rather high delay time, determined by the length of the statement. At the same time, since algorithms that reduce latency indicators increase the amount of required computing power, it is necessary to take into account the feasibility of improving the accuracy of speech recognition. In addition, the problems of noise reduction, automatic segmentation of a mixed audio signal, and accent recognition remain open in speech technologies [2].

The complexity of the non-native speech recognition problem arises as a consequence of the discrepancy between the speech of non-native speakers and materials of native language resources used in training, as well as acoustic, language and pronunciation modeling based on the data-driven approach [3]. An obvious way to overcome these difficulties is to create non-native speech corpora. However, simply increasing the training databases is costly and inefficient. According to various estimates, there are currently about 6–7 thousand languages. Recording non-native speakers for each of them is difficult, if not impossible. For example, the development of a speech system for the English language, taking into account the variety of American accents, requires a speech corpus of five thousand hours. An alternative approach to solving the problem of accounting for accents is the use of multilingual resources to adapt the speech of non-native speakers using the results of research on the phenomenon of “interlingual transfer” [4]. Based on the multilingual resource and interlingual transfer information, a new linguistic space is created aligned with the target space that can be used to estimate the speech space of non-native speakers. Depending on the type of linguistic resources, such as multilingual acoustic models or corpora, different methods are proposed. Of course, the most efficient approach to assessing the non-native language space is the synthesis of the mentioned solutions using multilingual resources to adapt the acoustic model of the target language based on some amount of interfered speech by non-native speakers.

This paper shows the results of the experimental phonetic study of interlingual interference and accent in the Russian speech of native speakers of the Kabardino-Circassian language. The primary analysis of the initial formant measurements of the non-native Russian speech of the Kabardino-Circassian speakers at the phonetic level was carried out. The preliminary study of accent identification using a specially created resource of non-native Russian speech [5] showed a systematic deviation of acoustic patterns of

vowel allophones (quantitative estimates of the vowel allophone [A] are given for brevity) in the F_1 - F_2 space. This study is aimed at developing methods for modeling the speech of non-native speakers for subsequent use in speech recognition systems, language identification and accents in conditions of limited linguistic resources. The obtained experimental data can be useful for the development of the theory of contrastive acoustic analysis for vocal systems of different structures languages, typical interfering pronunciation errors detection, and also contributing to the further development of theoretical models of second language acquisition.

2 Brief Review of Literature

In general, research on the strategies used by a person in linguistic competence acquiring is carried out in two directions – first- and second-language acquisition, since significant differences in how age affects the mechanism of their assimilation are already recognized by most scientists.

Research in the field of first-language acquisition proceeds from the hypothesis of a critical period after which the ability to successfully acquire a language declines. Infants show an amazing ability to distinguish phonetic contrasts across all languages [6], explained by a common auditory processing mechanism that is also demonstrated by monkeys. By the end of the first year of life, the ability of infants to distinguish phonetic elements is weakened, and sensitivity to native languages increases. According to the magnetic theory of mother tongue [7], infants detect patterns in language input and use the statistical properties of the input to change their perception to improve the perception of a particular language. That is, the acquisition of a certain language involves the specialization of the general mechanism of auditory processing, which uses specialized auditory features [7]. For example, speakers of tonal languages may have different areas of the brain activated than speakers of non-tonal languages [8].

Second-language learning is different. The results of functional magnetic resonance imaging show that bilinguals who acquire L2 at an early age activate overlapping Brodmann areas in the brain, bilinguals who acquire language in adulthood activate two separate areas for processing two languages [9].

As a separate linguistic discipline second-language acquisition has been actively developing since the 1970s. One of the main objectives of research in the field of second-language acquisition is to describe the formation, structure and application of L2 neural mental representations. Currently, conceptual research is being carried out in two opposing directions – generativism and cognitivism [10]. The generative approach is based on the ideas of “universal grammar” by N. Chomsky, that is, it claims that the acquisition of linguistic skills is determined by the innate abilities of a person [11]. Some adepts of generativism believe that the same mechanisms are involved in the acquisition of a second language by adults and in the acquisition of the native language by children [12], others argue that the mechanism of acquisition of subsequent languages is fundamentally different [13]. Cognitivists claim that linguistic structures form on the base of linguistic experience using general cognitive mechanisms, the set of which is the same for all people [14].

Model descriptions of the processes occurring in the cerebral cortex in the study of languages, organization, correlation and interaction of L1 and L2 language systems in the

context of bilingualism, developed within the framework of both theoretical directions - both generativism and cognitivism, are defined by the concept of language interference.

Language interference occurs in the context of language contacts and can manifest itself at all levels of the language. Initially, studies of linguistic interference were based on the comparison of contacting languages in terms of substratum phenomena. The necessity to study human speech behavior in terms of language contacts was first expressed by U. Weinreich [4]. Most of the work is devoted to the study of the interference effect of L1 on L2, but modern studies, considering the interference effect as a bidirectional process [15], also analyze the reverse effect.

The aim of this study is interlingual L1 phonetic interference during speech production in L2, which manifests itself at the segmental level.

3 Phonetic Features of Russian Speech Used by Native Speakers of the Kabardino-Circassian Language

To identify universal and specific features in the interfered Russian speech of bilingual speakers of the Kabardino-Circassian and Russian languages, ten informants were invited to the recording - five men and five women aged 21 to 49 years (average age 33,1 years). Most of the bilingual participants are native speakers of the Kabardino-Circassian language with a high level of knowledge of Russian ($n = 2$).

Informants with different levels of proficiency in their native speech were involved in the experiment, as this makes it possible to track the evolution of phonetic errors in the speech of bilinguals. The language competence of the participants was self-assessed during the questionnaire before the start of the experiment. The average self-assessment was 3, 7 points on a five-point scale.

The materials of the experiment include audio recordings of phonation reading of a previously prepared list of words. When compiling the pronunciation dictionary of the previously published database project [5], the lexical material was selected taking into account the change in the acoustic patterns of allophones under the influence of positional and combinatorial factors, and the phonetic law of stunning consonants at the end of Russian words was also taken into account. The selected lexemes allow us to analyze the contact accommodation, assimilation and dissimilation of six vowels of the Russian language in five left (absolute word onset, after hard bilabial/anterior-lingual and middle-lingual/back-lingual and some vowels/soft consonants and vowel *i*) and four right-hand contexts (before a pause, before hard bilabial/anterior-, back-lingual and vowels *o*, *a*, *u*, *e*, *s*/soft consonants and vowel *i*). To study the positional allophony of vowels in non-native Russian speech, the vocabulary is formed in such a way that it includes the following positional allophones: absolute strong position - isolated pronunciation; the first strong position is a stressed position at the beginning of a word before a hard consonant; the second strong position is any shock position; first weak position - the first pre-stressed syllable or unstressed position at the absolute beginning of a word; the second weak position is any unstressed (not the first pre-shock or post-shock) position. Thus, the stimulus words of the pronunciation dictionary represent 480 allophones of six Russian vowels.

To take into account the positional allophony of consonants when designing the dictionary, two positional allophones were considered: a strong position in a stressed syllable; weak position - in an unstressed syllable. The phonetic combinatorics of consonants is presented in five right contexts (end of word, before voiceless/voiced/unstressed vowels/stressed vowels). Thus, thirty-six consonant phonemes are represented by 180 allophones. The total number of allophones is 640. According to various estimates, the total number of allophones, depending on the degree of detail, can vary from several hundred to several tens of thousands [16]. This number of contexts, which determined the number of allophones, was chosen because, as shown by experimental studies [16], this is the minimum necessary set for speech synthesis that satisfies the criteria of intelligibility and naturalness.

The volume of the pronunciation dictionary is 461 words (288 words for the representation of vowel allophones and 173 words for consonants). The total volume of realizations for 10 native speakers of the Kabardino-Circassian language is 4610 words.

The experiment consisted of reading a suggested list of words. In order to eliminate lexico-grammatical interference, the reading was prepared. All implementations were recorded using the built-in Realtek High-Definition Audio microphone. Characteristics of the recording quality in the final audio files - 2 channels, 16 bit, 44100 Hz. Recording was carried out in an office environment. Random noises were not excluded.

This paper presents the results of the primary analysis of the initial formant meanings of allophones of the stressed vowel [A], in the Russian speech of native speakers of the Kabardian language. Since the main acoustic information about the quality of vowels is transmitted by the first two vowel formants [17], for each speaker, using the Praat version 6.3.08 [18], the average values of F_1 and F_2 in the analyzed speech segments were calculated using the 'Get first formant' and 'Get second formant' functions (tuning parameters: Burg method, time window 25 ms, frequency range 5500 Hz, number of formants 5). The results of measurements of F_1 and F_2 in Hertz were compared with the acoustic characteristics of the same vowels of standard Russian speech known in the scientific literature [19, 20]. Below, the canonical average formant values F_1 and F_2 , obtained in [19], are designated Reference1, and the reference gender-differentiated values from [20] are Reference2. The obtained measurements are shown in Figs. 1 and 2. Since gender differences are mainly expressed in the characteristics of vowels [21], in this study, female and male voices are considered separately.

As shown in Fig. 1, the observed distribution of allophones [A] in the F_1 - F_2 space for male voices is characterized by the following features. The averaged formant values of each speaker are shifted upwards relative to the Reference2 [20] in both formants, and all averages are above the Reference1 [19] in F_2 and not lower in F_1 . Moreover, all observations were higher than Reference2 [20] in the first formant, F_1 , and only 13 out of 109 were lower in the second, F_2 . Moreover, for more than 90% of F_1 observations, the acoustic distance exceeds 50 Hz, for more than 60% of F_2 observations, the acoustic distance is 100 Hz or more. The mean formant values of all male speakers are more than 2 standard deviations in F_1 ($SD = 48$ Hz), more than 1 standard deviation in F_2 ($SD = 127$ Hz). The distance from Reference2 [20] along F_1 is 2 times the maximum distance between the average speakers. The average values of the coordinates (F_1 , F_2) for all speakers are close: for *km1*, *km4*, and *km5* informants, they practically coincide in

both formants, and for *km2* and *km3*, they coincide in F_1 . The largest distance between the averages of different speakers is about 50 Hz for F_1 and 150 Hz for F_2 . The spread of formant values for all male voices is characterized by a small standard deviation of 48.01 for F_1 , 127.64 for F_2 .

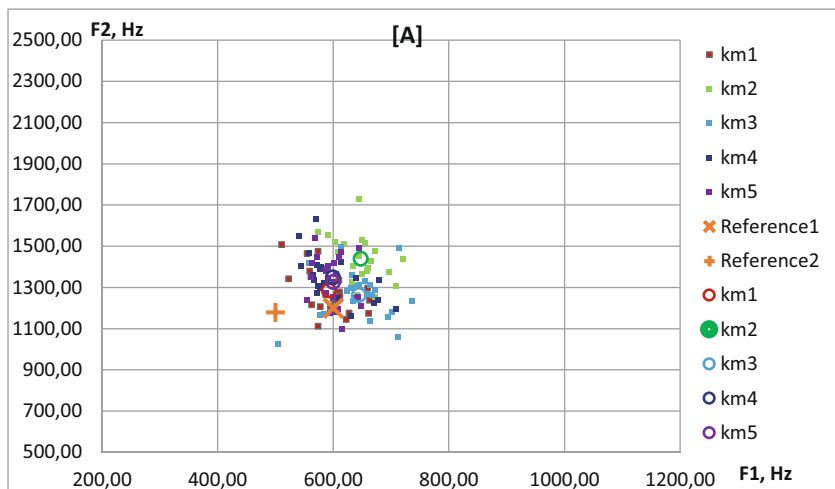


Fig. 1. Formant meanings of allophones of the vowel [A] in the speech of bilingual men (L1 - Kabardian language, L2 - Russian language). Here and below, along the x axis are the values of F_1 , Hz.; along the y-axis, F_2 values, Hz.

Figure 2 illustrates the distribution of the allophone [A] formant values for female voices in the acoustic space F_1 - F_2 . The individual average values of the coordinates (F_1 , F_2) for female speakers are quite close, varying in the range of 100 Hz in F_1 , and in the range of 150 Hz in F_2 . The largest distance between the individual averaged informants is about 140 Hz for F_1 and 200 Hz for F_2 . The maximum deviation of the individual average from the average for all observations in F_1 is 80 Hz. At the same time, the Reference2 [20] is at a distance of 193 Hz from the total average, that is, 2.4 times further. All individual means are above both Reference2 [20] and Reference1 [19] in both F_1 and F_2 . Moreover, all observations were above Reference2 [20] in F_1 and only 2 out of 110 were below in F_2 . The distribution of formant values for all bilingual women is characterized by an insignificant standard deviation equal to 87 Hz for F_1 , 133 Hz for F_2 . More than 85% of observations are above Reference2 [20] at a distance of 100 Hz or more in F_1 ; more than 98% are higher by 100 Hz or more in F_2 . For F_1 , the average values of all speakers are at a distance of more than 1 standard deviation from Reference2 [20], and for four of the five speakers - at a distance of more than 2 standard deviations. For F_2 , all means are at a distance from Reference2 [20] greater than two standard deviations.

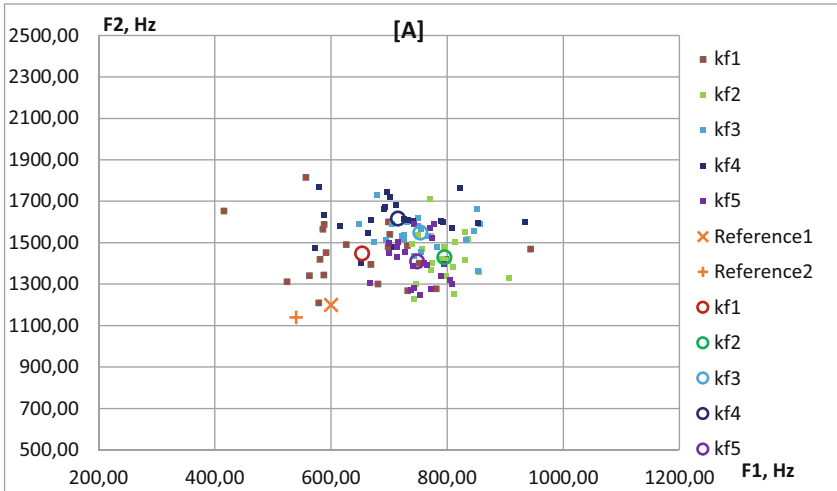


Fig. 2. Formant values of allophones of the vowel [A] in the speech of bilingual women (L1 Kabardian language, L2 Russian language)

4 Conclusion

As shown by the primary analysis of the initial formant measurements, a typical speech pattern of non-native Russian speech of Kabardino-Circassian speakers at the phonetic level is a systematic deviation of the acoustic patterns of allophones of the vowel [A] in the space F_1 - F_2 to the right-up. When using for comparison the values of the Reference2 [20] as standard values of Russian speech, the deviations F_1 , F_2 are stable and explicitly expressed. When using the average as the base values, the deviation in F_1 is less detected, but in F_2 , the systematic deviation is equally detected. To determine the accent, it is more convenient to use the Reference2 [20], because the deviation in both F_1 and F_2 is more significant.

Acknowledgements. The research was supported by the Russian Foundation of Scientific Research, grant No. 22-19-00787.



References

1. Stolke, A., Droppo, J.: Comparing human and machine errors in conversational speech transcription. In: Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, pp.137-141. ISCA, Stockholm, Sweden (2017)
2. Hannun, A.: Writing about machine learning. <https://awni.github.io/speech-recognition/>. Last accessed 21 Aug 2021
3. Brown, G.: Exploring forensic accent recognition using the Y-ACCDIST system. In: Proceedings of the 16th Australasian International Conference on Speech Science and Technology. Proceedings of the Sixteenth Australasian International Conference on Speech Science and Technology. UNSPECIFIED, pp. 305-308 (2016)

4. Weinreich, U.: Languages in contact: findings and problems. Publications of the Linguistic Circle of New-York. N1 (1953)
5. Nagoev, Z., Gurtueva, I.: Kantemir Bzhikhatlov and Murat Anchekov phonetic-acoustic database of high-accent Russian speech. *Proc. Comput. Sci.* **213**, 518–522 (2022)
6. Strange, W.: *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. York Press, Baltimore (1995)
7. Conboy, B.T., Kuhl, P.K.: Impact of second-language experience in infancy: brain measures of first- and second-language speech perception. *Dev. Sci.* **14**, 242–248 (2011)
8. <https://www.sciencedaily.com/releases/2008/02/080216114856.htm>. Last access 20 July 2023
9. Kim, K.H.S., Relkin, N.R., Lee, K.-M., Hirsch, J.: Distinct cortical areas associated with native and second languages. *Nature* **388**, 171–174 (1997)
10. Juffs, A.: Second language acquisition. *WIREs. Cogn. Sci.* **2**, 277–286 (2010)
11. Chomsky, N.A.: A review of skinner’s verbal behavior. In: L.A. Jakobovits, M.S. Miron, ed., *Readings in the Psychology of Language*, pp. 142–143. Prentice-Hall (1967)
12. Brown, C.: The interrelation between speech perception and phonological acquisition. In: Archibald, J. (ed.) *Second Language Acquisition and Linguistic Theory*, pp. 4–63. Blackwells, Oxford (2000)
13. Clahsen, H., Muysken, P.: The availability of universal grammar to adult and child learners. A study of the acquisition of German word order. *Sec. Lang. Res.* **2**, 93–119 (1986)
14. Tomasello, M.: The usage-based theory of language acquisition. In: Bavin, E. (ed.) *The Cambridge Handbook of Child Language*, pp. 69–88. Cambridge University Press, Cambridge (2009)
15. Flege, J.E.: Language contact in bilingualism: Phonetic system interactions. *Lab. Phonol.* **9**, 353–381 (2007)
16. Lobanov, B., Tsirulnik, L.: Development of multi-voice and multi-language TTS synthesizer (languages: Belarussian, Polish, Russian) SPECOM’2006, St. Petersburg, 25–29 June 2006
17. Ladefoged, P., Johson, K.: *A Course in Phonetics*, 6th edn. Wadsworth, Cengage Learning, Boston, MA (2011)
18. <https://www.fon.hum.uva.nl/praat/>. Last accessed 19 July 2023
19. Sorokin, V.N., Tsyplikhin, A.I.: Vowel Segmentation and Recognition. *Inf. Process.* **2**(4), 202–220 (2004)
20. Leonov, A.S., Makarov, I.S., Sorokin, V.N.: Frequency modulation in a speech signal. *Acoust. J.* **6**(55), 809–821 (2009)
21. Sorokin, V.N., Tsyplikhin, A.I.: Speaker verification based on spectral and temporal parameters of speech signal. *Inf. Process.* **2**(10), 87–104 (2010)



A Bearing Fault Diagnosis Method Based on VMD-HPE

Wanqing Huang , Yang Chen, Yongqi Chen, Tao Zhang, Feiyu Yu, and Xiaoyan Mao 

College of Science and Technology, Ningbo University, Ningbo 315300, China
527030336@qq.com

Abstract. Since rolling bearings operate in complex and harsh conditions with high speed and heavy load for a long time, their fault signals have the problems of difficulty in feature extraction and low diagnostic accuracy. Therefore, a rolling bearing fault diagnosis method based on variational mode decomposition (VMD) and hierarchical permutation entropy (HPE) is proposed in this paper. Firstly, the fault signals of rolling bearings are decomposed by variational mode decomposition. Secondly, several node signals are obtained after hierarchical decomposition, and the permutation entropy value of the obtained node signals is calculated as the feature vector. Thirdly, a multi-fault classifier based on Bayes is established to realize the fault diagnosis of rolling bearings. Finally, the method is applied to the data of Bearing Center of Case Western Reserve University, and the experimental results verify the effectiveness of the method.

Keywords: Rolling Bearing · Variational modal decomposition · Naive Bayes · Fault diagnosis

1 Introduction

Rolling bearings are an important part of industrial transmission machinery and equipment, which are widely used, their working environment is complex and longtime operation, easily affected by temperature, load, air pressure, etc. [1], in order to reduce the economic loss caused by bearing failure, early fault diagnosis of rolling bearings is especially important [2]. Therefore, how to extract the weak fault characteristics from the vibration signal to accurately diagnose and identify the early failure of the bearing is one of the problems of bearing fault diagnosis.

In addition, due to the mechanical system vibration coupling and complex environmental impact, the vibration signal has non-linear and non-smooth characteristics [3], which is difficult to analyze and discriminate the fault type only from the perspective of time domain, frequency domain or time-frequency domain [4–6]. At present, there are mainly resonance demodulation method, short-time Fourier transform, wavelet transform, etc., but all these methods lack certain adaptivity [7, 8]. Empirical mode decomposition (EMD) is adaptive, but the method has problems of over-envelope, under-envelope, endpoint effect, frequency confusion, etc. [9]. FREI proposed an Intrinsic Time scale

decomposition (ITD) of the adaptive signal time-frequency analysis [10], which overcomes the disadvantages of EMD such as edge effects and has high computational efficiency. The intrinsic time scale decomposition method uses linear transformation to decompose the signal, which may lead to burr and distortion in the obtained PRC components [11]. Dragomiretskiy proposed variational modal decomposition (VMD) [12], which is an emerging and non-recursive signal decomposition method with high decomposition accuracy [13]. It can better solve the modal mixing problem in the decomposition process, overcome the shortcomings of EMD and ITD methods, and has high operational efficiency and good noise robustness [14].

With the development of nonlinear scientific theories, various nonlinear scientific theories have emerged [15, 16]. Entropy can not only characterize the complexity of a signal, but also measure the uncertainty of a system or a piece of information, thus facilitating the treatment of nonlinear problems [17, 18]. The permutation entropy (PE) algorithm is a new method recently proposed by Bandt et al. [19] to detect randomness and kinetic mutations in time series. Permutation entropy only considers the low-frequency components of the original sequence, while ignoring the high-frequency components. For the rolling bearing time series with rich distribution of fault information, PE cannot reflect the operating status information of rolling bearings comprehensively and accurately. And Jiang [20] proposes a hierarchical entropy method to measure the time series complexity, which can reflect both the high-frequency component complexity and the low-frequency component complexity of the signal. According to the advantages of alignment entropy and hierarchical entropy, Li [21] proposed a new method, hierarchical permutation entropy (HPE).

In recent years, the fault diagnosis of rolling bearing has the characteristics of gradually increasing data volume and incomplete fault information, and the fault diagnosis cannot be effectively realized only by signal processing, so the methods combining signal processing and machine learning are gradually applied in the field of fault diagnosis [22]. For example, artificial neural network (ANN), support vector machine (SVM) and deep learning (DL) are used. Parsimonious Bayes is one of the most widely used classification algorithms with the advantages of stable classification, simple algorithm and low mis-classification rate [23]. Based on this, this paper adopts a feature extraction method combining VMD and HPE to extract bearing fault signals, and then combines the plain Bayesian classification method to propose a new method for rolling bearing fault diagnosis, and applies it to bearing fault test data, and the experimental results show that the method can effectively improve the recognition rate of fault signals.

2 Basic Theory

2.1 Variational Modal Decomposition

VMD is a form of adaptive decomposition of the acquired complex digital signal into a combination of multiple effective AM-FM signals by means of frequency domain iteration [24, 25], and VMD is a fully non-recursive adaptive signal processing tool based on wiener filtering [26]. VMD is able to obtain the center frequency and bandwidth of a series of component signals from a given complex signal by iterative computation of the

optimal variational model, which in turn computes the eigenmode function for a certain number of AM-FMs. The expressions are as follows.

$$\mu_k(t) = A_k(t)[\cos(\varphi_k(t))] \tag{1}$$

where: $\mu_k(t)$ is the signal of each mode, $A_k(t)$ is the instantaneous amplitude of the K th IMF component, $\varphi_k(t)$ is the phase of the signal, $\omega_k(t)$ is the instantaneous frequency of $\mu_k(t)$, $\omega_k(t) = \varphi_k'(t) = \frac{d\varphi_k}{dt}$ ($k = 1, 2, 3, \dots, K$) K is the number of modal decompositions.

In the process of solving the center frequency and bandwidth of the signal, it is assumed that the original multi-band signal can be decomposed into K th narrowband IMF components, and then the corresponding constrained variational model needs to be constructed.

The variational model is constructed by following several steps.

The Hilbert transform is first used to obtain each IMF resolved signal, and then the one-sided spectrum of the signal is obtained as follows: $\left(\sigma(t) + \frac{j}{\pi t}\right) \cdot \mu_k(t)$.

Where: $\sigma(t)$ is the unit pulse function; j is an imaginary unit; t is the time. An exponential term is introduced to adjust the estimated center frequency of each IMF ω_k , modulating each IMF spectrum to its corresponding fundamental frequency band: $\left(\left(\sigma(t) + \frac{j}{\pi t}\right) \cdot \mu_k(t)\right)e^{-j\omega_k t}$.

The constrained variational model constructed by estimating the bandwidth of each IMF based on the squared L^2 parametrization of the gradient of the demodulated signal is:

$$\left\{ \begin{array}{l} \min_{\{\mu_k\}, \{\omega_k\}} \left\{ \sum_k \left\| \delta_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) \mu_k(t) \right] e^{-j\omega_k(t)} \right\|_2^2 \right\} \\ s.t. \sum_k \mu_k = f \end{array} \right. \tag{2}$$

where: $\{\mu_k\} = \{\mu_1, \mu_2, \dots, \mu_k\}$ is the decomposed k -modal components; $\{\omega_k\} = \{\omega_1, \omega_2, \dots, \omega_k\}$ is the center frequency of each modal component; $\sigma(t)$ is the unit pulse function; f is the original signal, $\mu_k(t)$ is the K th modal function.

To facilitate the solution, a quadratic penalty factor α and a Lagrangian multiplication factor λ are introduced to transform the above constrained variational problem into an unconstrained one. Let the constraints remain strict and α can sufficiently reduce the interference of Gaussian noise on the signal, the extended Lagrangian expression is as follows:

$$\begin{aligned} L(\{\mu_k\}, \{\omega_k\}, \lambda) = & \alpha \sum_k \left\| \delta_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) \mu_k(t) \right] e^{-j\omega_k(t)} \right\|_2^2 \\ & + \left\| f(t) - \sum_k \mu_k(t) \right\|_2^2 + \lambda(t), f(t) - \sum_k \mu_k(t) \end{aligned} \tag{3}$$

where: α is the quadratic penalty factor; λ is the Lagrange factor.

Finally, the ‘‘saddle point’’ of the Lagrangian expansion, i.e., the optimal solution of the constrained variational model in the following equation, is sought by iteratively

updating each modal component μ_k^{n+1} and its central frequencies ω_k^{n+1} and λ^{n+1} using the alternating direction multiplier method until K modal decompositions are solved.

$$\hat{\mu}_k^{n+1}(\omega) = \frac{\hat{f}(\omega) - \sum_{i \neq k} \hat{\mu}_i(\omega) + \frac{\hat{\lambda}(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k)^2} \tag{4}$$

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega |\hat{\mu}_k(\omega)|^2 d\omega}{\int_0^\infty |\hat{\mu}_k(\omega)|^2 d\omega} \tag{5}$$

where, $\hat{\mu}_k^{n+1}(\omega)$, $\hat{f}(\omega)$ and $\hat{\lambda}^{n+1}(\omega)$ are the Fourier isometric transforms of the corresponding time domain functions $\mu_k^{n+1}(\omega)$, $f(\omega)$ and $\lambda^{n+1}(\omega)$; ω_k^{n+1} is the $n + 1$ st center frequency of the iteration; $\hat{f}(\omega)$ is the frequency domain signal of the original signal; $\hat{\lambda}(\omega)$ is the frequency domain signal of the Lagrange multiplier; $\hat{\mu}_k(\omega)$ is the frequency domain signal of the K th modal function.

2.2 Hierarchical Permutation Entropy

The permutation entropy (PE) is a nonlinear dynamical feature used to measure the complexity of a one-dimensional time series, which has the advantages of simple algorithm, high computational efficiency, strong resistance to noise interference and good robustness [27]. The basic principle of the ranked entropy algorithm is as follows.

Suppose there is a one-dimensional time series $X = \{x(i), i = 1, 2, 3, \dots, n\}$ of length n . The phase space reconstruction of each element in this series can be obtained as the following reconstruction matrix Y :

$$Y = \begin{bmatrix} Y(1) \\ Y(2) \\ \dots \\ Y(j) \\ \dots \\ Y(k) \end{bmatrix} = \begin{bmatrix} x(1) & x(1 + \lambda) & \dots & x(1 + (m - 1)\lambda) \\ x(2) & x(2 + \lambda) & \dots & x(2 + (m - 1)\lambda) \\ \vdots & \vdots & \dots & \vdots \\ x(j) & x(j + \lambda) & \dots & x(j + (m - 1)\lambda) \\ \vdots & \vdots & \dots & \vdots \\ x(k) & x(k + \lambda) & \dots & x(k + (m - 1)\lambda) \end{bmatrix} \tag{6}$$

where: m is the embedding dimension; λ is the delay time; $x(j)$ denotes the j th row component of the reconstruction matrix. $k = N - (m - 1)\lambda$ is the number of reconstruction components, $j = 1, 2, 3, \dots, K$

Each row $Y(j)$ in the reconstruction matrix Y can be regarded as a reconstruction component, and a total of K reconstruction components can be obtained.

Rearrange $Y(j) = \{x(j), x(j + \lambda), \dots, x(j + (m - 1)\lambda)\}$ in ascending order j_1, j_2, \dots, j_m to indicate the index of the column in which each element of the reconstructed component is located, i.e.

$$x[i + (j_1 - 1)\lambda] \leq \dots \leq x[i + (j_m - 1)\lambda] \tag{7}$$

If there are equal values in the reconstructed components, i.e.

$$x[i + (j_1 - 1)\lambda] = x[i + (j_2 - 1)\lambda] \tag{8}$$

Then sorted by the size of the j value, i.e., when $j_1 < j_2$, there

$$x[i + (j_1 - 1)\lambda] \leq x[i + (j_2 - 1)\lambda] \tag{9}$$

For any reconstructed vector $Y(j)$ in the reconstructed phase space a sequence of symbols can be obtained:

$$S_{(l)} = \{j_1, j_2, \dots, j_m\} \tag{10}$$

where, $i = 1, 2, \dots, k$ and $k \leq m!$.

The sequence of symbols of the m -dimensional phase space mapping not j_1, j_2, \dots, j_m has a total of $m!$ $S_{(l)}$ is just one of the symbolic arrangement sequences.

Calculate the probability of occurrence of each symbolic sequence: $P_1, P_2, \dots, P_k, k < m$. The permutation entropy of the time series $X = \{x(i), i = 1, 2, 3, \dots, n\}$ with k different position index sequences, can be defined according to the form of Shannon entropy as:

$$H_P(m) = \sum_{i=1}^k P_i \ln(P_i) \tag{11}$$

P_i is the probability of the sequence of symbols, $\sum P_i = 1$. $H_P(m)$ reaches its maximum value $\ln(m!)$ when $P_i = \frac{1}{m!}$ is normalized, i.e.

$$H_P = \frac{H_P(m)}{\ln(m!)} \tag{12}$$

where, $0 \leq H_P \leq 1$, the magnitude of H_P value indicates the random degree of the time series $x(i)$. The smaller the H_P value, the more regular the time series; the larger the H_P value, the more disorderly the time series. Therefore, the change of H_P value reflects the subtle changes of the time series.

In order to extract the fault information of high-frequency components in the signal, the concept of hierarchical entropy was introduced by Jiang et al. Therefore, after combining the methods of alignment entropy and hierarchical entropy, Li et al. proposed the hierarchical permutation entropy method. The specific calculation process of hierarchical alignment entropy is as follows.

Given a time series $X = \{x(i), i = 1, 2, \dots, N\}$ of length N , define the mean operator Q_0 and the difference operator Q_1 as follows.

$$Q_0(x) = \frac{x(2j) + x(2j + 1)}{2} \quad j = 0, 1, 2 \dots 2^{n-1} \tag{13}$$

$$Q_1(x) = \frac{x(2j) - x(2j + 1)}{2} \quad j = 0, 1, 2 \dots 2^{n-1} \tag{14}$$

where, $N = 2^N$, n is a positive integer.

n is the number of hierarchical layers, Q_0 denotes the low-frequency component of the time series after one hierarchical decomposition, and Q_1 represents the high-frequency

component of the time series after one hierarchical decomposition. The lengths of the operators Q_0 and Q_1 are 2.

According to the operators Q_0 and Q_1 , the original time series is reconstructed and expressed as

$$x = \{(Q_0(x)_j + Q_1(x)_j), Q_0(x)_j - Q_1(x)_j\} \quad j = 0, 1, 2, \dots, 2^{n-1} \quad (15)$$

When $j = 0$ or $j = 1$, the matrix operator Q_j can be defined as

$$Q_j = \begin{bmatrix} \frac{1}{2} \frac{(-1)^j}{2} & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \frac{1}{2} \frac{(-1)^j}{2} & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & \frac{1}{2} \frac{(-1)^j}{2} \end{bmatrix}_{2^{n-1} \times 2^n} \quad (16)$$

Construct an n -dimensional vector $[\gamma_1, \gamma_2, \dots, \gamma_n] \in \{0, 1\}$. The integer e can be expressed as

$$e = \sum_{j=1}^n \gamma_j 2^{k-j} \quad k \in N \quad (17)$$

It follows from Eq. That for a given non-negative integer e , there is a unique vector $[\gamma_1, \gamma_2, \dots, \gamma_n]$ corresponding to it.

Based on the vectors $[\gamma_1, \gamma_2, \dots, \gamma_n]$, define the nodal components of each level of decomposition of the time series $x(i)$ as

$$x_{ke} = Q_{r_k} \cdot Q_{r_{k-1}} \cdots Q_{r_1} \cdot x \quad (18)$$

where: k is the k -layer in the hierarchical segmentation; x_{k_0} and x_{k_1} are the low and high frequency parts of the original time series $x(i)$ at the $k + 1$ layer, respectively.

The alignment entropy of the hierarchical sequence obtained from each node is calculated to obtain the alignment entropy values of the 2^k hierarchical components, so the HPE can be expressed as

$$HPE(x, k, e, m, \lambda) = PE(m, \lambda) \quad (19)$$

where m denotes the embedding dimension and λ denotes the time delay.

3 Rolling Bearing Fault Diagnosis Model

This paper proposes a feature extraction method based on the combination of variational modal decomposition and hierarchical arrangement entropy, and the extracted feature vectors are input into a Bayesian classification model to achieve fault diagnosis of rolling bearings. The diagnosis process is shown in Fig. 1.

The steps are as follows:

- (1) The vibration data of rolling bearings in different states are collected by vibration signal sensors at a fixed sampling frequency.

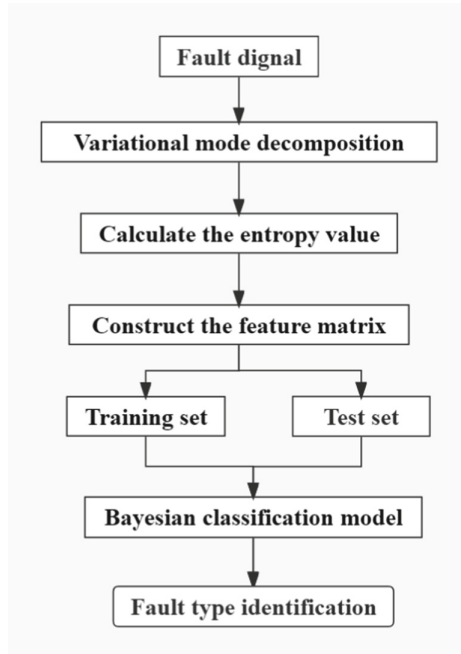


Fig. 1. Flow chart of fault diagnosis based on VMD and HPE.

- (2) The VMD decomposition is performed on the original signal of the acquired data, and then K modal components $IMF_1, IMF_2, \dots, IMF_K$ are obtained.
- (3) Calculate the entropy value of the corresponding of each node of the selected modal components IMFs.
- (4) The alignment entropy values of suitable scales are selected as the fault feature vectors to construct the feature matrix T , which is divided into training and test sets.
- (5) The training set data is used to train the plain Bayesian classification model, and the test set data is input into the trained Bayesian model to output the fault type and complete the fault type identification.

4 Experimental Validation

In this paper, experimental analysis was conducted using publicly available data from Case Western Reserve University for rolling bearing fault diagnosis [30], and the test rig consisted of a 2 horsepower motor, a torque sensor and a power test meter, as shown in Fig. 2.

The test was conducted using a deep groove ball bearing at the drive end, its model number is SKF6205-2RS JEM (where: 2RS is a double-sided seal, JEM has no meaning, SKF (Svenska Kullager-Fabriken) is SKF company), the damage fault on this bearing was set by EDM simulation. The motor load is 0 and the speed is 1797 r/min. The data are divided into normal signal, inner ring fault signal, outer ring fault signal and rolling element fault signal, where the fault depth is 0.011 mm and the fault diameters

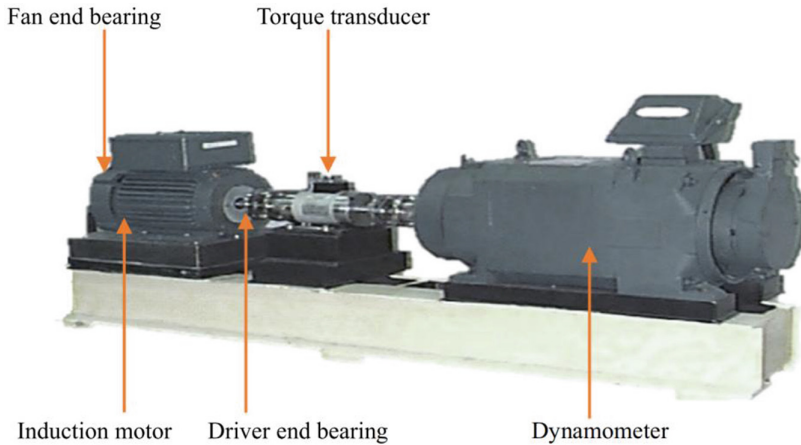


Fig. 2. Sketch of bearing test bench structure.

are 0.007 mm, 0.014 mm and 0.021 mm respectively, for a total of 10 types of bearing faults. The rolling bearing data set is described as shown in Table 1.

Table 1. Description of the rolling bearing data set.

Number	Bearing status	Degree of failure (mm)	Tags
0	Normal	–	Normal
1	Roller failure	0.007	Ball0.007
2	Roller failure	0.014	Ball0.014
3	Roller failure	0.021	Ball0.021
4	Inner ring failure	0.007	Inner0.007
5	Inner ring failure	0.014	Inner0.014
6	Inner ring failure	0.021	Inner0.021
7	Outer ring failure	0.007	Outer0.007
8	Outer ring failure	0.014	Outer0.014
9	Outer ring failure	0.021	Outer0.021

The following analysis of the four types of data, rolling bearing normal state, inner ring failure, roller failure, outer ring failure vibration signal time domain diagram, as shown in Fig. 3.

It is obvious that Fig. 3 is the normal signal time domain diagram, the amplitude is between $[-0.2, 0.2]$, the amplitude is the smallest, the signal is smooth. Outer ring, inner ring, rolling body fault compared with normal bearing amplitude are greater than the normal signal amplitude, accompanied by an obvious interval shock signal, it is very difficult this to discern its fault characteristics. Therefore, taking the rolling bearing data in

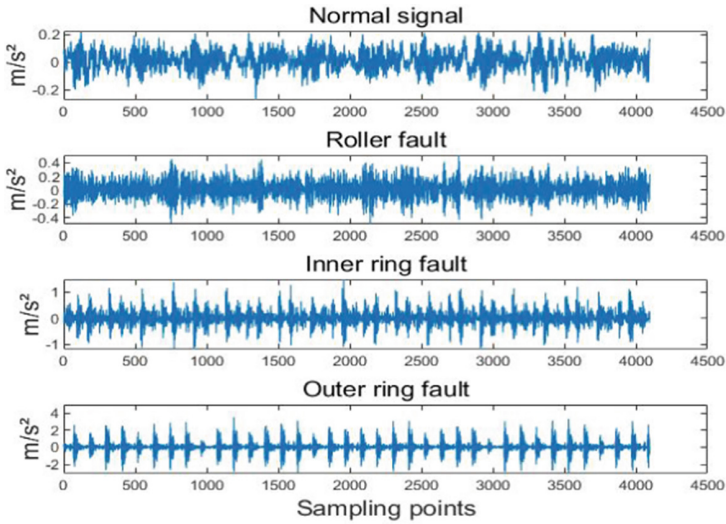


Fig. 3. Time domain diagram of the signal in the four states of the rolling bearing.

the normal state as an example, the EMD, ITD and VMD methods are used to decompose the signals. The obtained intrinsic modal components and their corresponding spectra are shown in the Figs. 4, 5 and 6.

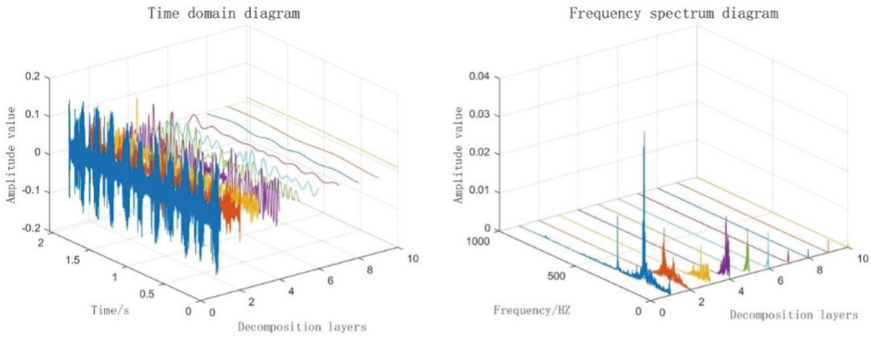


Fig. 4. Decomposition diagram of EMD.

As can be seen from Fig. 4, EMD decomposition of the original signal will produce more invalid components, and there is a phenomenon of modal aliasing in the first two components, so the spectrum information cannot be extracted effectively. Compared with the EMD method, using ITD decomposition will not produce invalid components, but there are still different degrees of modal aliasing. Observing Fig. 6, it can be seen that the IMF components obtained by VMD decomposition are concentrated near their respective central frequencies, which effectively avoids the problem of modal aliasing, and there are no invalid components, so the information in the signal can be fully extracted. Therefore,

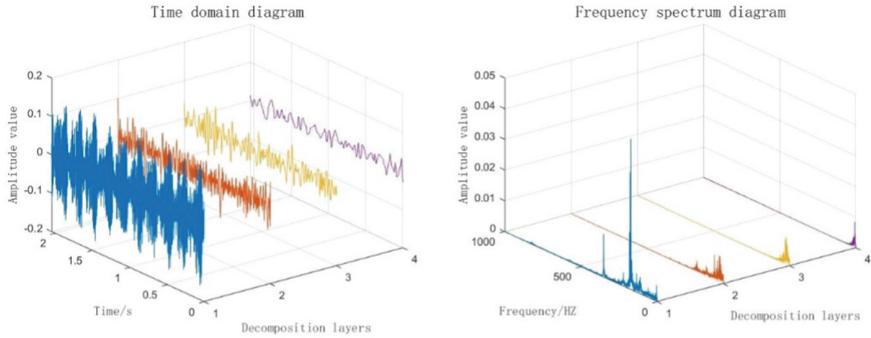


Fig. 5. Decomposition diagram of ITD.

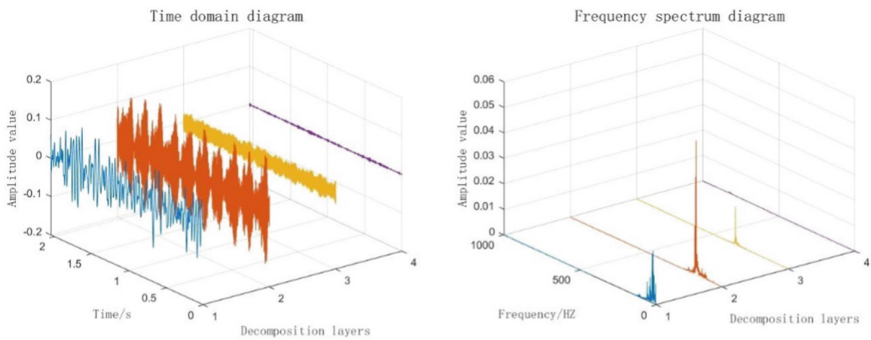


Fig. 6. Decomposition diagram of VMD.

we choose VMD method to decompose the vibration signal, which can improve the accuracy of subsequent bearing fault diagnosis.

The sampling frequency of the signal is 12 kHz, each group of signal consists of 2048 sampling points, a total of 50 groups of signals, each data set dimension $2048 \times 50 = 102400$. Each of the four types of fault signals consists of 2048 sampling points, and for each state 50 signal samples are intercepted respectively, and the length of each sample is set to 1024. IMF is first obtained by variational modal decomposition with default parameters (number of decomposition layer $K = 4$, penalty factor $\alpha = 2000$) [31], and then the entropy features are extracted from the decomposed IMF components. In order to verify the superiority of the HPE method, the hierarchical sample entropy and multiscale sample entropy are used to extract the rolling body fault signal as an example, and the comparison graphs of the three entropy values are given in Fig. 7.

As can be seen from Fig. 7, after taking the mean value of the extracted feature samples, the entropy value of the samples extracted under the MSE method is about 0.4, which is a small value and cannot well reflect the characteristics of the fault signal. The entropy value extracted by the HSE method is improved to some extent, but its value is still small. The entropy value extracted using the HPE method has good smoothness and its entropy size is much larger than the remaining two extraction methods, so it shows

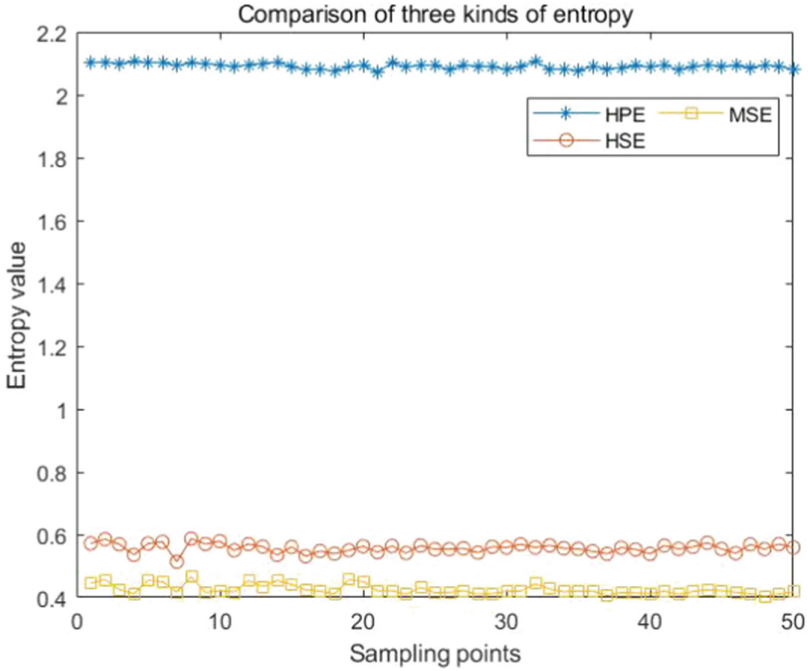


Fig. 7. Comparison of three kinds of entropy.

that this method can better extract the signal features and improve the accuracy of the final fault classification. To further illustrate the effectiveness of the HPE method, three different entropy extraction methods are used for the feature samples under four fault states, and then the Pearson correlation coefficient analysis is performed on the entropy extracted feature data, and the results are shown in Figs. 8, 9 and 10.

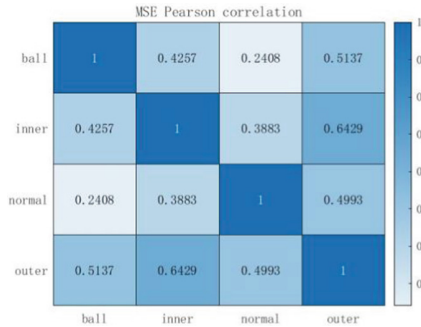


Fig. 8. Pearson correlation coefficient graph of MSE.

Pearson’s correlation coefficient is used to measure the degree of correlation between two variables, and its value range is $[-1, 1]$. After the absolute value of the correlation

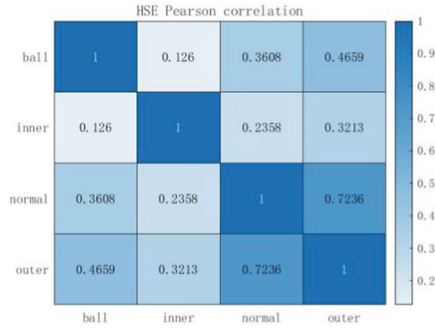


Fig. 9. Pearson correlation coefficient graph of HSE.

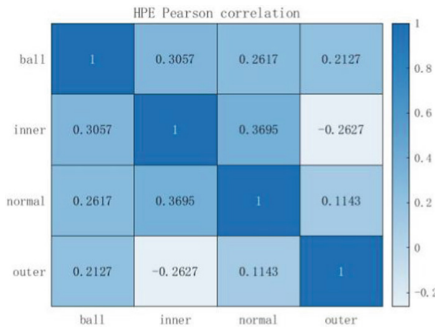


Fig. 10. Pearson correlation coefficient graph of HPE.

coefficient is taken, the larger the value is, the more correlated the two are. Comparing Figs. 8 and 10, it can be seen that the correlation coefficients between the fault features extracted by MSE are larger, except for the coefficient between rolling body fault and normal which is 0.2408, and the size of the rest of the values are around 0.5. And the coefficients extracted by HPE are all well below 0.5, and the irrelevance is better. Comparing with Figs. 9 and 10, the correlation coefficients between the fault features extracted by HSE become smaller, but the correlation coefficients between outer ring fault and rolling element fault and normal condition are still large, which is not good for distinguishing the bearing condition. In summary, the feature values extracted by HPE have better non-correlation between each fault type than the other two entropy extraction methods, which can help improve the fault diagnosis rate of bearings.

All samples are divided into 50 groups, and the four different fault data are decomposed, and then the entropy values of each component are extracted to form the sample data set. The first 25 groups of the data set are used as the training set and the last 25 groups are used as the test set, which are input to the classification model for fault identification. In order to verify the superiority of the proposed method, the EMD, ITD and VMD decompositions were performed, and the MSE, HSE and HPE values of each component were extracted respectively, and the data sets were input to the Bayesian

model for fault identification. The classification results of each method are shown in Fig. 11.

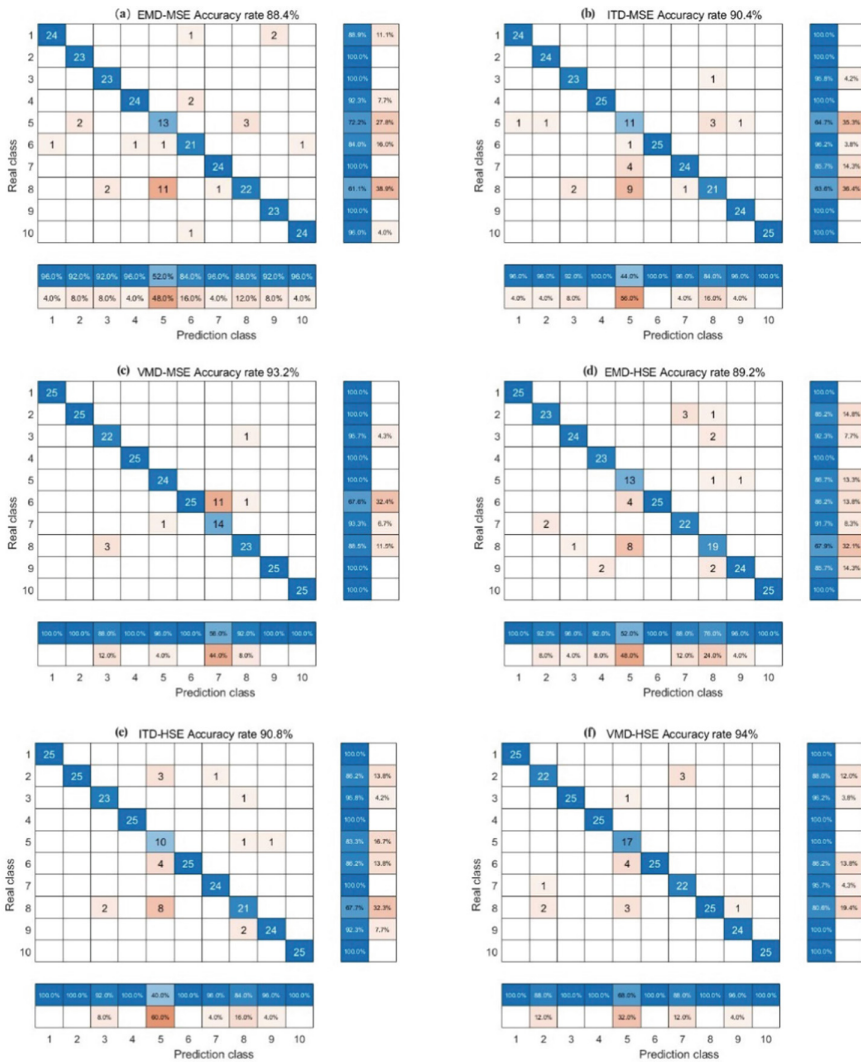


Fig. 11. Classification results of 9 different methods.

As can be seen from the data in Table 2, the accuracy rates for MSE, HSE, and HPE were 88.4%, 89.2%, and 89.6%, respectively, for EMD, and 90.4%, 90.8%, and 91.6%, respectively, for MSE, HSE, and HPE, respectively, for ITD. For VMD, the accuracy rates of MSE, HSE and HPE were 93.2%, 94% and 97.2%, respectively. For the same entropy extraction method, the signals decomposed by VMD can get better classification accuracy, and for the same decomposition method, it can be seen that the data samples

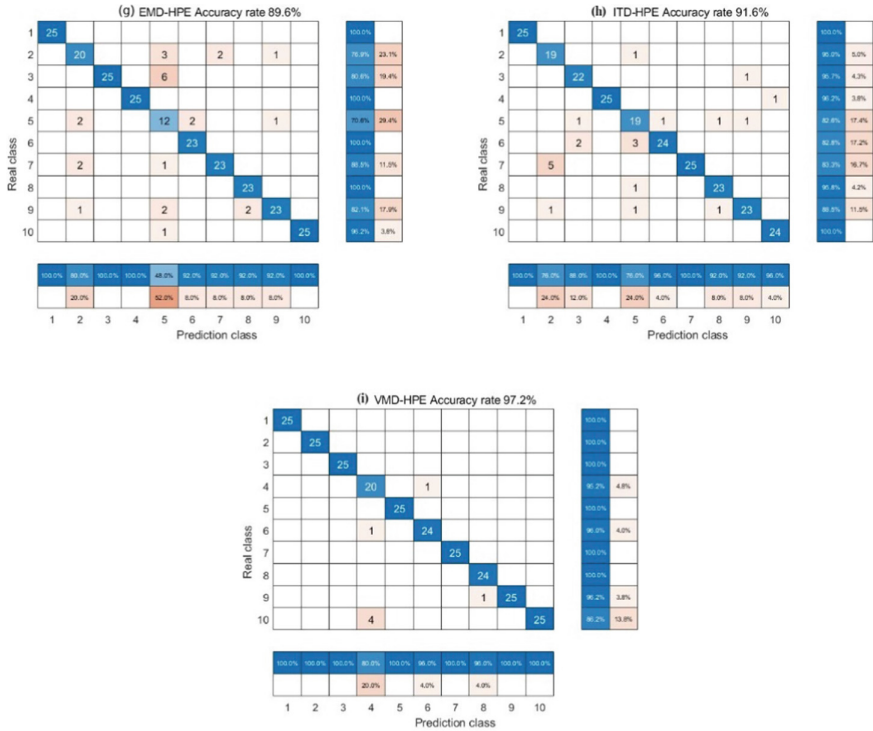


Fig. 11. (continued)

after hierarchical arrangement entropy extraction can get better accuracy compared with the remaining two methods. After comparing these methods, it can be concluded that the VMD-HPE method mentioned in this paper can effectively improve the diagnosis rate of rolling bearing faults.

5 Conclusions

In this paper, a rolling bearing fault diagnosis method based on VMD-HPE feature extraction method and plain Bayesian is proposed, and the effectiveness and practicality of the method are verified through experiments. The following conclusions are drawn: (1) The combined VMD-HPE feature extraction method is used to extract the vibration signal and obtain more fault feature information of the original signal, and also avoid the phenomenon of modal confusion, which verifies the effectiveness of this feature extraction method. (2) The VMD algorithm is compared with the EMD and ITD algorithms by the correlation coefficient criterion, and then it is concluded that the VMD decomposition can obtain the effective components containing more original fault information. (3) The eigenvalue modal components of the decomposed fault data are processed with different entropy to obtain better accuracy and effectiveness of HPE compared with HSE and MSE methods, thus improving the recognition rate of bearing faults.

Table 2. Diagnostic accuracy under different methods.

Decomposition method	Entropy	Accuracy
EMD	MSE	88.4%
	HSE	90.4%
	HPE	93.2%
ITD	MSE	89.2%
	HSE	90.8%
	HPE	94%
VMD	MSE	89.6%
	HSE	91.6%
	HPE	97.2%




References

1. Duo, M., Ji, G., H, Zhu, H.: Bearing fault diagnosis based on VMD noise reduction and CNN. *Noise Vibr. Control* **14**(05), 155–160(2021)
2. Zhang, Y.P., Zhang, P., Wang, H.: Intelligent identification of early bearing faults based on dual-time domain weak fault feature enhancement. *J. Mech. Eng.* **52**(21), 96–103 (2016)
3. Zheng, J., Pan, H., Qi, X.: Composite hierarchical fuzzy entropy and its applications to rolling bearing fault diagnosis. *China Mech. Eng.* **27**(15), 2048–2055 (2016)
4. Zheng, J., Dai, J., Zhu, S.: Improved multi-scale fuzzy entropy-based rolling bearing fault diagnosis method. *Vibr. Testing Diag.* **38**(5), 929–934 (2018)
5. Chen, D., Zhang, Y., Yao, C.: Fault diagnosis based on FVMD multiscale alignment entropy and GK fuzzy clustering. *J. Mech. Eng.* **54**(14), 16–27 (2018)
6. Jin, J., Xun, Z., Li, C.: Bearing fault diagnosis based on VMD energy entropy and optimized support vector machine. *Acta Metrol. Sinica* **42**(07), 898–905 (2021)
7. Qi, X., Ye, X., Cai, J.: A rolling bearing fault feature extraction method based on variational modal decomposition and streamline learning. *Vibr. Shock* **37**(23), 133–140 (2018)
8. Liu, F., Li, X., Huang, H.: Fault diagnosis method of rolling bearing based on ITD and improved MCKD. *J. Guangxi Univ.* **46**(01), 107–115 (2021)
9. Zheng, J., Yang, C., Lang, Y.: Bearing fault diagnosis based on VMD and GWO optimized SVM. *Coal Mining Mach.* **42**(1), 147–150 (2021)
10. Frei, M.G., Osorio, I.: Intrinsic time-scale decomposition: time- frequency-energy analysis and real-time filtering of non-stationary signals. *Proc. Royal Soc. A: Math. Phys. Eng. Sci.* **463**(2078), 321–342 (2007)
11. Dragomiretskiy, K., Zossd, D.: Variational mode decomposition. *IEEE Trans. Signal Process.* **62**(3), 531–544 (2014)
12. Zhou, F., Tang, G., He, Y.: Unbalanced fault feature extraction for wind power gearbox based on improved VMD. *J. Vibr. Shock* **39**(05), 170–176 (2020)
13. Hao, Y., Wu, W., Shang, Q.: Rolling bearing quality assessment based on variational modal decomposition and support vector machine. *Control Theory Appl.* **37**(7), 1544–1551 (2020)
14. Xiao, M., Zhang, C., Fu, X.: ICEEMDAN and wavelet thresholding based rolling bearing fault feature extraction method. *J. Nanjing Agr. Univ.* **41**(4), 767–774 (2018)
15. Tong, Z., Lu, T., Qin, Z.: Fault diagnosis of rolling bearing based on PSO-VMD and Bayesian network. *J. Henan Polytech. Univ.* **40**(01), 95–104 (2021)

16. Zhong, Y., Zhu, C.: EEMD and improved ITD for microgrid arc fault detection. *J. Sichuan Univ. Sci. Eng.* **33**(04), 53–61 (2020)
17. Yan, Z., Wang, H., Yang, H.: Research on rolling bearing fault diagnosis based on deep learning feature extraction and GWO-SVM. *J. Yunnan Univ. (Nat. Sci. Edn.)* **42**(4), 656–666 (2022)
18. Cheng, J., Ma, X., Yang, Y.: Rolling bearing fault diagnosis method based on permutation entropy and VPMCD. *J. Vibr. Shock* **33**(11), 119–123 (2014)
19. Yang, Y., Pan, H., Cheng, J.: A rolling bearing fault diagnosis method based on LCD denoising and VPMCD. *China Mech. Eng.* **24**(24), 3338–3344 (2013)
20. Bandt C, Pompe B.: Permutation entropy: a natural complexity measure for time series. *Phys. Rev. Lett. Am. Physiol. Soc.* **88**(17), 174102 (1–4) (2002)
21. Li, Y., Xu, M., Zhao, H.: A study on rolling bearing fault diagnosis method based on hierarchical fuzzy entropy and ISVM-BT. *J. Vibr. Eng.* **29**(01), 184–192 (2016)
22. Jiang, Y., Peng, C., Xu, Y.: Hierarchical entropy analysis for biological signals. *J. Comput. Appl. Math.* **236**, 728–742 (2011)
23. Li, Y., Li, G., Yang, Y.: A fault diagnosis scheme for planetary gearboxes using adaptive multi-scale morphology filter and modified hierarchical permutation entropy. *Mech. Syst. Signal Process.* **105**, 319–337 (2018)
24. Yang, X., Zhou, C., Deng, J.: Research on motor bearing fault diagnosis system based on improved Bayesian classification. *Mach. Tools Hydraulics* **48**(20), 172–175 (2020)
25. Jin, Z., Mu, P., Zhang, Y.: An improved VMD and its application in bearing fault diagnosis. *Mach. Design Manuf.* **02**, 42–46 (2022)
26. Lian, J., Liu, Z., Wang, H.: Adaptive variational mode decomposition method for signal processing based on mode characteristic. *Mech. Syst. Signal Process.* **2018**(107), 53–77 (2018)
27. Jin, Z., Mu, P.: An improved VMD and its application in bearing fault diagnosis. *Mach. Design Manuf.* **02**, 42–46 (2022)
28. Wang, Z., Chang, X., Wang, J.: Gearbox fault diagnosis based on permutation entropy optimized variational mode decomposition. *Trans. Chin. Soc. Agr. Eng. (Trans. CSAE)* **34**(23), 59–66 (2018)
29. Chen, W.: A study of feature extraction from SEMG signal based on entropy. Shanghai Jiao Tong University, Shanghai (2008)
30. Smith, W.A., Randall, R.B.: Rolling element bearing diagnostics using the CaseWestern reserve university data: a benchmark study. *Mech. Syst. Signal Process.* **64**, 100–131 (2015)
31. Li, Z., Chen, J., Zi, Y.: Independence-oriented VMD to identify fault feature for wheel set bearing fault diagnosis of highspeed locomotive. *Mech. Syst. Signal Process.* **85**, 512–529 (2017)



Neural Network Solution of an Inverse Problem with Integration of Geophysical Methods on Recovered Data: Training with Noise Addition

Igor Isaev^{1,2} , Ivan Osbornev^{1,3} , Eugeny Osbornev³, Eugeny Rodionov³,
Mikhail Shimelevich³, and Sergey Dolenko¹ 

¹ D.V. Skobeltsyn Institute of Nuclear Physics, M.V. Lomonosov Moscow State University,
Moscow, Russia

isaev_igor@mail.ru, dolenko@srdsinp.msu.ru

² Kotelnikov Institute of Radio Engineering and Electronics, Russian Academy of Sciences,
Moscow, Russia

³ S. Ordjonikidze Russian State Geological Prospecting University, Moscow, Russia

Abstract. Previously, it was shown that integration (joint use of data) of several geophysical methods allows one to obtain a higher quality of the solution of the inverse problem of exploration geophysics in comparison with the individual use of each of these methods. However, there may be a situation when for some measurement points there is no data from one of the geophysical methods used. At the same time, the data spaces of different integrated geophysical methods are interconnected. Therefore, the missing data of one method can be recovered from the known data of another one by constructing a preliminary adaptive mapping of one of the spaces to another. In this study, we investigate the solution of the inverse problem with integration of geophysical methods on the recovered data obtained based on noise addition during training of the neural networks performing the mapping from the data space of the method(s) with all data present to the data space of the method with missing data.

Keywords: Inverse problems · Exploration geophysics · Data recovery · Integration of geophysical methods · Neural network · Multitask-learning · Training with noise addition

1 Introduction

The general statement of inverse problems (IP) of exploration geophysics (EG) consists in reconstructing the distribution of some physical parameters of the medium in the thickness of the earth's crust from the physical fields measured on the earth's surface in order to study the structure of the near-surface layer of the earth and to search for useful fossils. In particular, in this study we consider the IP of gravimetry (G), magnetometry (M) and magnetotelluric sounding (MT), which consist in restoring the spatial distribution of density, magnetization, and electrical resistance of the medium in the Earth's

crust by the values of gravitational, magnetic and magnetotelluric fields, respectively. These IPs are ill-posed, which generally leads to low quality of the solution and high sensitivity to noise in the input data.

A general approach to reducing the ill-posedness of an IP is to change its statement. In the case of using machine learning methods, a change in the statement can be achieved by introducing some additional information. An example with indirect use of additional information can be the use of parameterization schemes with some rigidly defined spatial structure (the so-called “class-generating models” [1–3]), which is built on the basis of alternative measurement methods, or based on some assumptions about the structure of the specific area. The method of direct introduction of additional information considered in this study consists in setting the problem of integrating geophysical methods, i.e. simultaneous use of data from several geophysical methods to solve the EG IP [4–6].

However, in practice, it is possible that for some measurement points there is no data available from one of the geophysical methods used. One of the approaches is to recover missing data from present data by building various models, including those based on machine learning methods [7–9]. Another approach considered in this paper is based on the use of information from other geophysical methods [10, 11]. Since the data spaces of different integrated geophysical methods are interconnected, the missing data of one geophysical method can be possibly recovered from the known data of another one by constructing a preliminary adaptive mapping of one of the spaces to another. In the previous studies of the authors [10, 11], it was shown that in some cases the approach based on the reconstruction of one geophysical field from the known data of other geophysical fields and their further joint application for inversion gives a positive result. However, in general, this approach did not show acceptable results and therefore needs further improvement. In the present work, as such an improvement, it was proposed to use neural networks (NN) trained with the addition of noise [12], which are more resistant to distortions in the shape of geophysical fields.

The purpose of this study was to test the applicability of the approach associated with NN recovery of the missing data of one geophysical method from the known data of another one, and their further joint application to solve the IP, using the approach based on adding noise during training at the inversion stage.

2 Physical Statement of the Problem

2.1 Parametrization Scheme

In order to implement the integration of various geophysical methods, it is necessary that the determined parameters of each of the methods are the same. This approach corresponds to the geometric formulation of the problem, which consists in determination of the boundaries of geophysical objects. In particular, in this study we considered the parameterization scheme, which consists in determining the boundaries of geological layers of a layered medium. Multi-layer geological structures are very widespread. The parameterization scheme used was a four-layer two-dimensional model (Fig. 1) described in more detail and discussed elsewhere [6, 10, 11].

The dimension of the section was 15 km wide and 3 km deep. The physical field measurement step was 0.5 km – a total of 31 measurement points along the profile.

The discreteness of changing the boundaries of geological layers was 1 km – a total of 15 depth values for each layer. In this problem, the values of the depths of the lower boundaries of the three upper layers were determined. Each layer was characterized by fixed values of density, magnetization, and resistivity, which did not change within the layer, and which were the same across the entire data set. The discreteness of changing the values of depth was 0.02 km.

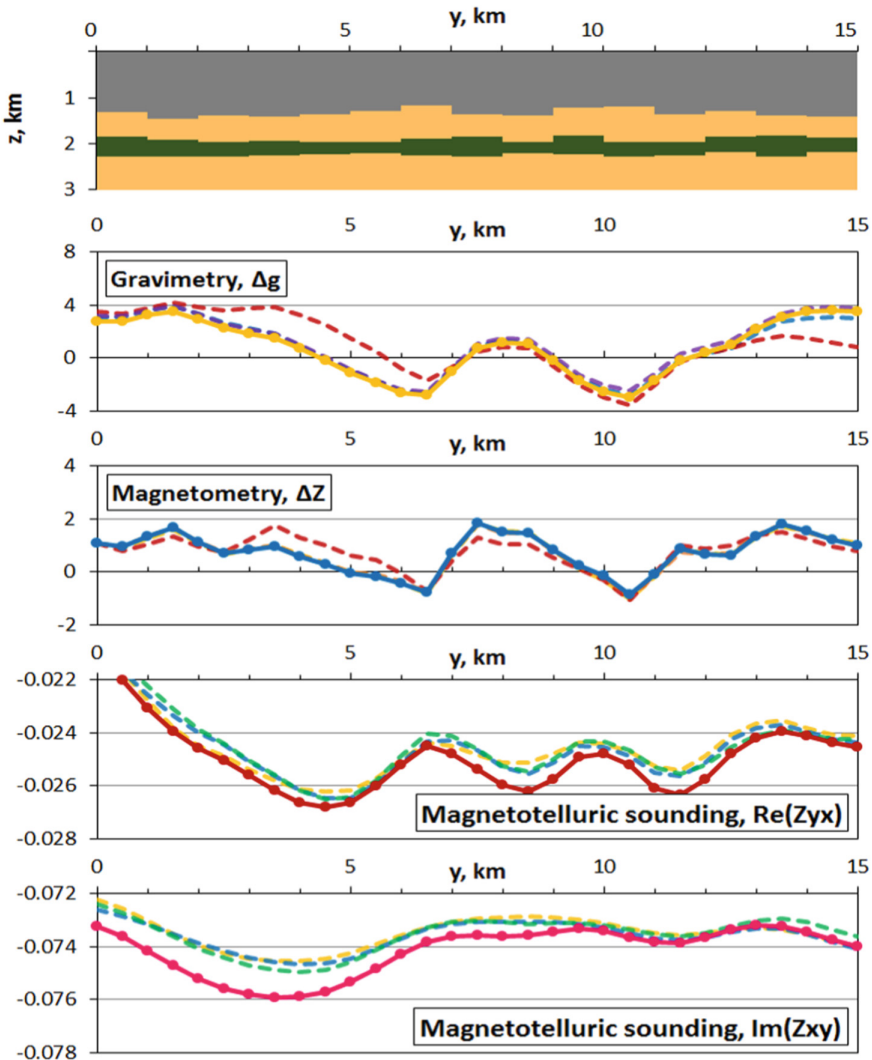


Fig. 1. An example of the geological section within the considered parameterization scheme (top), and the corresponding components of the fields used in this study (bottom). Solid lines – true values of the field components, dashed lines – recovered values.

2.2 Data

For each pattern of the original data set, the layer depth values were set randomly in the given ranges. Further, to obtain the data for the study, the direct problem was solved by finite-difference methods for each of the selected geophysical methods.

The input dimension of the problem was:

- Gravimetry: $1 \text{ field component} * 31 \text{ measurement point (picket)} = 31 \text{ feature}$
- Magnetometry: $1 \text{ field component} * 31 \text{ picket} = 31 \text{ feature}$
- MT sounding: $2 \text{ field components} * 1 \text{ frequency} * 31 \text{ picket} = 62 \text{ features}$.

The output dimension of the problem was:

- $3 \text{ layers} * 15 \text{ values of layer boundary depth} = 45 \text{ parameters}$.

A total of 30 000 patterns were calculated for each of the geophysical methods.

3 Methodical Statement of the Problem

The original data set was divided into training, validation, and test sets in a ratio of 70:20:10. Their dimensions were 21 000, 6 000, and 3 000 patterns, respectively.

For field recovery problem, the multitask-learning approach [11] was used: all values for the pickets of each of the geophysical fields were determined simultaneously. For the inversion problem, the so-called autonomous determination [11] was used, where a separate NN with one output was trained for each determined parameter.

Reduction of the input dimension of the problem was not carried out.

All NN used in this study, both for recovering and inversion, were used in the same way. The type of NN used was the multilayer perceptron, which is a universal approximator. The architecture used had a single hidden layer with 32 neurons in it. To reduce the factor associated with the influence of the initialization of weights on the training of NN, 5 NN were used for each case under consideration, and the statistical indicators of their application were averaged. To prevent overtraining, early stopping by validation dataset was used.

For individual use of data from the gravimetry and magnetometry methods, the NN input was fed by 31 features, for individual use of MTS data – 62 features, for simultaneous use of data from two geophysical methods – 62 or 93 features, for simultaneous use of data from all the three methods (only inversion) – 124 features.

Geophysical fields recovery was performed by direct application of NN. The values of known geophysical fields were fed to the NN inputs, the corresponding values of the reconstructed field were the desired outputs; thus, the NN had 31 outputs. For MTS, the Re and Im components of the field were determined separately.

For the subsequent IP solution, the dependence of the solution quality on the number of recovered values of geophysical field components was studied. For the considered geophysical method, the exact values of the geophysical field were randomly replaced by a given number of recovered ones. The goal was to determine the number of recovered values for which the considered approach would give a positive result.

Note that training with noise plays the role of online data augmentation. Discussion of this approach may be found in our previous study [12].

In [12] it was shown that the best option for training with noise addition, which demonstrated greater accuracy and less computational cost, was the variant in which the training set contained noise, but the validation set contained no noise. This option was also used in the present study.

Two cases were considered:

- *Noise only in one field.* A separate inversion NN was built for each method of field recovery. The noise level was taken as the variance of the regression residuals of the recovery method under consideration.
- *Noise in all fields.* Here, NN were trained only for data integration. For the noise level of any of the fields, the average variance of the regression residuals of all recovery methods for this field was taken.

4 Results

4.1 Geophysical Field Components Recovery

The results of geophysical components recovery are shown in Figs. 1 and 2. The serrated curves shape (Fig. 2) of dependences of the quality of the solution on the picket number is due to the location of some pickets on the boundaries of blocks with different values of thickness. There is a positive effect of data integration: the use of any two geophysical methods to recover the third one shows a better result than using each method separately.

The regression residuals are bell-shaped (Fig. 2) but do not follow a normal distribution: the p-value for Normal test, Shapiro-Wilk, chi-square and Jarque-Bera tests is less than 0.05. However, we added normal noise during training, the level of which was determined from these regression residuals.

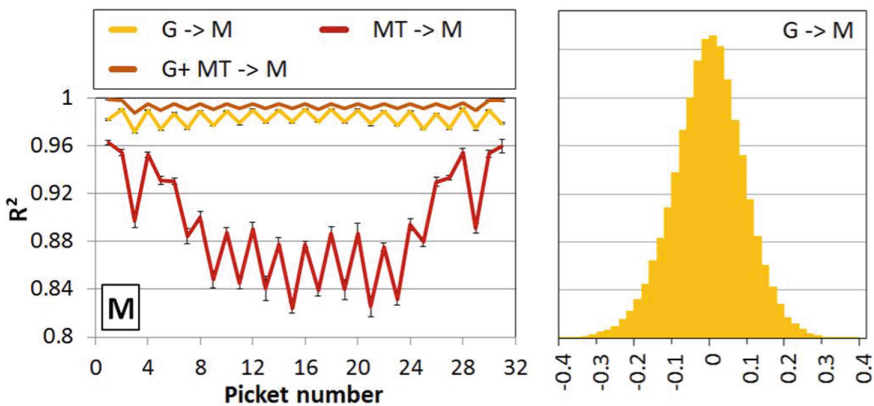


Fig. 2. Quality (multiple determination coefficient R^2) of recovery of geophysical field components – left. Histogram of regression residuals from recovery of one of geophysical field components – right.

4.2 Inverse Problem Solution

The quality of the solution of the IP was compared on the full and reconstructed data. The results are presented in Fig. 3.

In some cases, the training with noise addition approach gave positive results if the noise was added only to the field that was restored in the previous stage (left graphs).

For layer 1 and inversion on known gravitational and recovered magnetic field, there is a slight improvement (dotted yellow line in the left graph) compared to using only the known gravitational field (horizontal solid yellow line) in the case of adding noise during training. There is a significant improvement compared to training without adding noise (yellow dashed line).

For layer 2 and inversion on known gravitational and recovered magnetotelluric field, there is an improvement (dotted yellow line in the left graph) compared to using only the known gravitational field (horizontal solid yellow line) in the case of adding noise during training. There is also a significant improvement compared to training without adding noise (yellow dashed line).

For layer 3 and inversion on known gravitational and magnetic fields and recovered magnetotelluric field, there is an improvement (dotted green line in the left graph) compared to using only the known gravitational and magnetic fields (horizontal solid green line) in the case of adding noise during training. There is also a significant improvement compared to training without adding noise (green dashed line).

For layer 3 and inversion on known gravitational and magnetotelluric fields and recovered magnetic field, there is an improvement (dotted brown line in the left graph) compared to using only the known gravitational and magnetotelluric fields (horizontal solid brown line) in case of adding noise during training. However, the best result here was shown by training without adding noise on known and recovered data (brown dashed line).

Adding noise during training to all fields can give either improvement or deterioration in the result compared to adding noise to only one field. However, this approach in most cases did not lead to an improvement in the result compared to using only known fields.

5 Conclusion

Based on the results of this study, the following conclusions can be drawn:

- The proposed approach combining missing neural network data recovery from known data of other geophysical methods and training neural networks with noise at the stage of inversion proved in general to be effective. This may be useful for application with real geophysical data, when some data values are often missing.
- The use of data of any two geophysical methods to recover the data of the third one shows a better result than using each method separately.
- Training with noise addition allows obtaining better resilience to the number of recovered field values, but, in most cases, the results are still worse than when using only full known data. Further research is required on the selection of optimal levels of noise added during training.

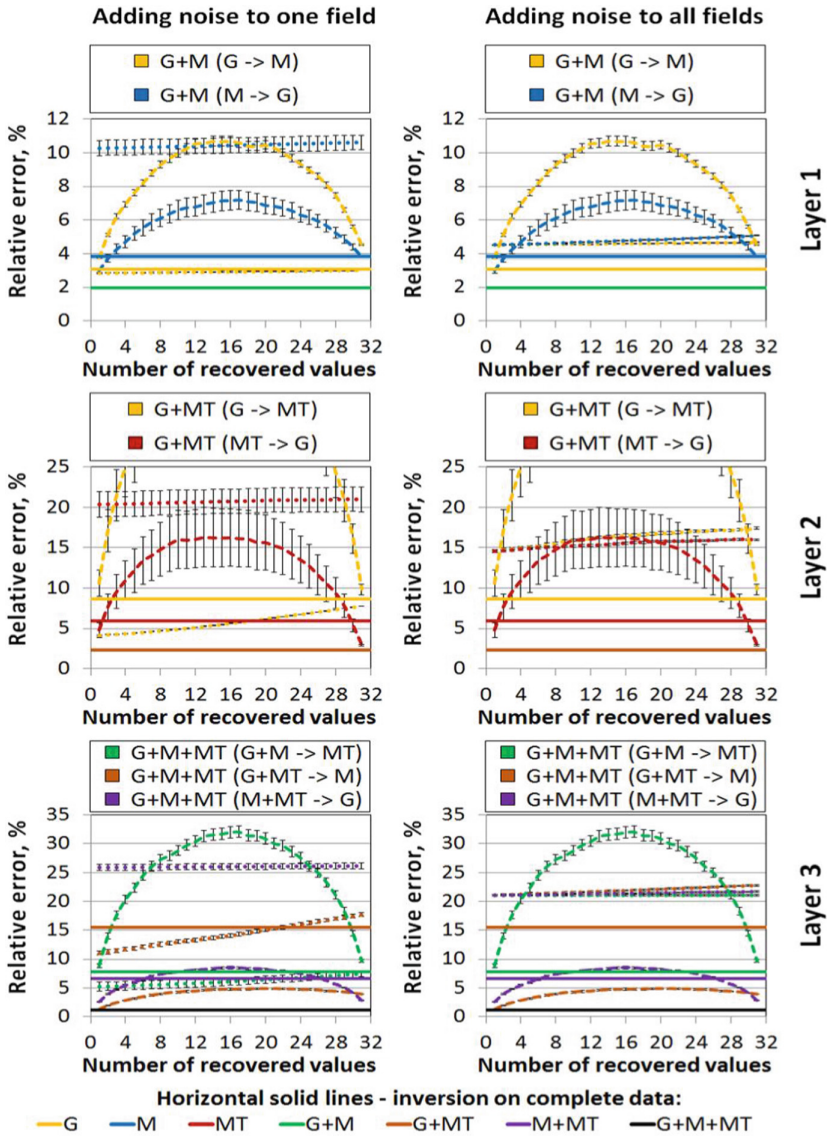


Fig. 3. Dependence of the quality (relative error) of the solution of the IP on the number of recovered values of geophysical field components. Horizontal solid lines – inversion on complete data, other lines – inversion with joint use of known and recovered data: dotted lines – using training with noise addition, dashed lines – without. The same color corresponds to the same complete known data.

- In some cases, an approach based on the reconstruction of one geophysical field from the data of other geophysical fields and their further joint application for inversion yields a positive result.

Acknowledgments. This study has been performed at the expense of the grant of the Russian Science Foundation no. 19-11-00333, <https://rscf.ru/en/project/19-11-00333/>.

References

1. Spichak, V., Fukuoka, K., et al.: ANN reconstruction of geoelectrical parameters of the Minou fault zone by scalar CSAMT data. *J. Appl. Geophys.* **49**(1–2), 75–90 (2002). [https://doi.org/10.1016/S0926-9851\(01\)00100-8](https://doi.org/10.1016/S0926-9851(01)00100-8)
2. Montahaei, M., Oskooi, B.: Magnetotelluric inversion for azimuthally anisotropic resistivities employing artificial neural networks. *Acta Geophys.* **62**(1), 12–43 (2014). <https://doi.org/10.2478/s11600-013-0164-7>
3. Isaev, I., Osborne, E., et al.: Neural network recognition of the type of parameterization scheme for magnetotelluric data. In: *Studies in Computational Intelligence*, vol. 799, pp. 176–183 (2018). https://doi.org/10.1007/978-3-030-01328-8_19
4. Roux, E., Moorkamp, M., et al.: Joint inversion of long-period magnetotelluric data and surface-wave dispersion curves for anisotropic structure: application to data from Central Germany. *Geophys. Res. Lett.* **38**(5), L05304 (2011). <https://doi.org/10.1029/2010GL046358>
5. Akca, İ, Günther, T., et al.: Joint parameter estimation from magnetic resonance and vertical electric soundings using a multi-objective genetic algorithm. *Geophys. Prospect.* **62**(2), 364–376 (2014). <https://doi.org/10.1111/1365-2478.12082>
6. Isaev, I., Osborne, I., et al.: Integration of geophysical methods for solving inverse problems of exploration geophysics using artificial neural networks. In: *Problems of Geocosmos – 2020, Springer Proceedings in Earth and Environmental Sciences*, pp. 77–87 (2022). https://doi.org/10.1007/978-3-030-91467-7_7
7. Liu, Q., Fu, L., Zhang, M.: Deep-seismic-prior-based reconstruction of seismic data using convolutional neural networks. *Geophysics* **86**(2), V131–V142 (2021). <https://doi.org/10.1190/geo2019-0570.1>
8. Zhang, M., Liu, Y.: 3-D seismic data recovery via neural network-based matrix completion. *IEEE Geosci. Remote Sens. Lett.* **19**, 1–5 (2022). <https://doi.org/10.1109/LGRS.2022.3154816>
9. Wei, Q., Li, X., Song, M.: Reconstruction of irregular missing seismic data using conditional generative adversarial networks. *Geophysics* **86**(6), V471–V488 (2021). <https://doi.org/10.1190/geo2020-0644.1>
10. Isaev, I., Osborne, I., et al.: Neural network recovery of missing data of one geophysical method from known data of another one in solving inverse problems of exploration geophysics. *Proc. Sci.* **429**, 18 (2022). <https://doi.org/10.22323/1.429.0018>
11. Isaev, I., Osborne, I., et al.: Multitasking learning in missing data recovery for the integration of geophysical methods in solving an inverse problem of exploration geophysics. *Procedia Comput. Sci.* **213**, 777–784 (2022). <https://doi.org/10.1016/j.procs.2022.11.134>
12. Isaev, I., Dolenko, S.: Training with noise as a method to increase noise resilience of neural network solution of inverse problems. *Opt. Mem. Neural Netw.* **25**, 142–148 (2016). <https://doi.org/10.3103/S1060992X16030085>



Mathematical Model, Experimental Verification and Control of Actuators Based on Metal – Hydrogen System

Vladimir I. Ivlev¹ and Sergej Yu. Misyurin^{1,2} 

¹ Mechanical Engineering Research Institute RAS, Malyi Kharitonievsky per. 4, Moscow, Russia

ssmmrr@mail.ru

² National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), 31 Kashirskoe Shosse, 115409 Moscow, Russia

Abstract. The design and mathematical modeling of gas actuator with linear and rotary movement are described. The power source is a metal hydride compressor based on $LaNi_{4.7}Al_{0.3}$ intermetallic compound. The pressure and temperature of hydrogen in the working chambers are determined by equations, as in traditional pneumatic drives, based on the energy conservation equations for the chamber and the equation of state in differential form. The hydrogen flow rate in the chambers is determined by the equations of the kinetics sorption - desorption processes and the heat balance for the metal hydride and parts of the actuator structure. The results of calculations and experimental data obtained on actuators prototypes are presented. It is shown that the proposed relatively simple mathematical model makes it possible to obtain qualitative estimates the static and dynamic characteristics of actuators.

Keywords: Metal – hydrogen system · Actuator · Mathematical model

1 Introduction

In various technical applications, the most widely used are electric (electric motors), pneumatic and hydraulic actuators. At the same time, research and development of actuators based on other physical principles is underway; some are already in use. Here we can mention actuators based on materials with shape memory effect, electroactive polymers, piezoelectric ceramics, giant magnetostriction effect and some others.

In this paper, will be considered actuators that use the properties of some intermetallic alloys to reversibly absorb and release hydrogen in large quantities, depending on the temperature and pressure in the system. In this case, the density of sorbed hydrogen can reach its density in the liquid state. This property can be used to create a power source for a gas actuator with the ability to control the pressure of hydrogen by changing the temperature in the metal hydride – free hydrogen system. Among a large number of hydride – forming metals and alloys, the most promising for these purposes are

intermetallic alloys of rare earth and transition metals, in particular, the well-known and most studied LaNi_5 intermetallic compound, which reacts with hydrogen as follows: $\text{LaNi}_5 + 3\text{H}_2 \leftrightarrow \text{LaNi}_5\text{H}_6$ [1]. Up to 600 volumes of hydrogen can be sorbed per unit volume of the alloy; the equilibrium pressure in the system at 293 K is 0.22 MPa, and at 323 K it is 0.6 MPa. By alloying the initial composition with aluminum, for alloys of the $\text{LaNi}_{5-x}\text{Al}_x$ type, it is possible to purposefully reduce the equilibrium pressure in the system without much loss of sorption capacity [2]. Using these properties of metal hydrides, various drive devices have been developed. In [3], the design of a jack with actuating device in the form of a reinforced rubber shell and metal hydride source is described, when heated to 50 °C, a force of the order 1kN develops. Similar actuators with different methods of heating and cooling of metal hydride are considered in [4–7]. In [8] the properties of some metal hydrides, the most promising for drive systems are analyzed. In some works, the change in the temperature of metal hydride is carried out using Peltier thermoelectric modules. This is very convenient, but it significantly reduces the dynamic performance.

In the above and many other works on the use of metal hydride sources for gas drives, the results of experimental studies are mainly presented, and little attention is paid to mathematical modeling. In this paper, an attempt is made to develop an adequate mathematical model of a gas actuator with a metal hydride power source, both linear and rotary operating principles, oriented for use in shut-off valves.

Here we proceed from the following premises. This type of drive can be classified as a thermal drive with a working fluid that undergoes a phase transformation of the first kind (solid – gas). This group also includes drives based on the shape memory effect (here, a martensitic phase transformation) and drives with a liquid-gas phase transformation (based on low-boiling hydrocarbon compounds) [9]. An actuator with a metal hydride power supply can be considered as a pneumatic actuator with an integrated compressor, and the role of the pressure regulator and switchgear is performed by the metal hydride temperature control system.

2 Structural Schemes of Actuators

Two types of actuators with a metal hydride power source were designed and manufactured: – double-sided linear actuator; – drive with rotary movement of the output shaft (angle of rotation 90°). Figure 1a, b shows the appearance and structural scheme of the linear actuator.

The actuator is arranged and works as follows. The central unit 3 with bellows and channels connecting the bellows chambers with generators 5 and 6, respectively, is fixed in housing 1. When power is applied to the heater in generator 5, the process of hydrogen desorption begins and pressure increases in the chamber of the right bellows. Expanding, the right bellows compresses the left bellows through the rods 4, and the hydrogen in its chamber is sorption in the generator 6. The heat of sorption is removed to the housing and further to the environment. When the power is turned off in generator 5 and generator 6 is connected, the shaft moves in the opposite direction. This design of the actuator ensures the independence of its neutral position, when the ambient temperature changes.

The rotary actuator (shaft rotation angle 90°), shown in Fig. 2, is arranged and operates in a similar way. A feature of this actuator is the presence of a membrane

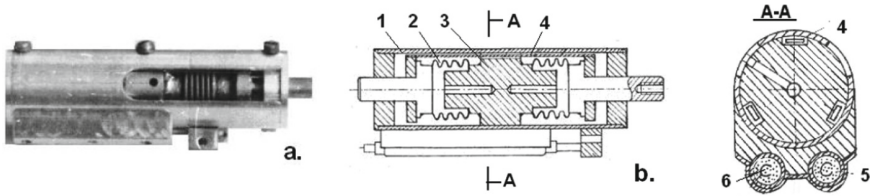


Fig. 1. a) Appearance of the actuator, b) structural scheme: 1-body, 2-bellows, 3-central block, 4-rods, 5 and 6-generators

gas-hydraulic converter. This is due to the tightness requirements for hydrogen, which cannot be provided in a vane engine. Here we have used a two-vane actuator, which allows for a more compact design for a given torque. In both actuators, an intermetallic composition $LaNi_{4,7}Al_{0,3}$ was used. The activation of the alloy was carried out directly in the generators.

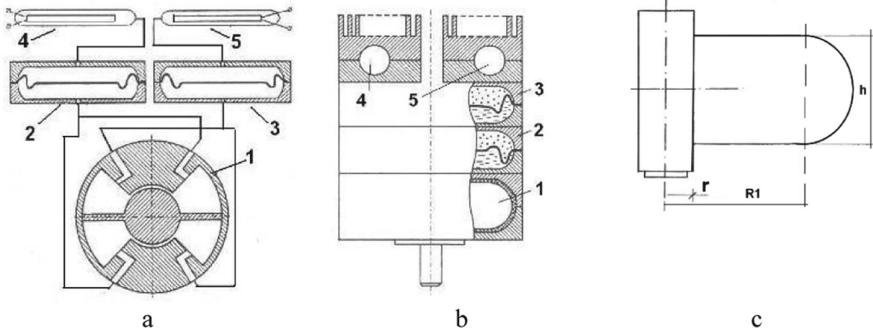


Fig. 2. a) Drive diagram, b) constructive scheme of the drive: 1 - engine, 2 and 3 - gas-hydraulic converters, 4 and 5 - generators. c) Vane parameters.

3 Mathematical Model

The pressure and temperature of hydrogen in working chambers 1 and 2 are determined by equations, as in traditional pneumatic drives, based on the energy conservation equations for the chamber and the equation of state of the gas in differential form, for example, as in [10]:

$$\frac{dp_{1;2}}{dt} = \frac{k}{V_{1;2}} RT_{g1;2} G_{d;s} - \frac{k-1}{k} [\beta_1 S_1 (T_{1;2} - T_k)] - \frac{kp_{1;2}}{V_{1;2}} \frac{dV_{1;2}}{dt}, \quad (1)$$

$$\frac{dT_{1;2}}{dt} = \frac{T_{1;2}}{p_{1;2}} \frac{dp_{1;2}}{dt} + \frac{RT_{1;2}^2}{p_{1;2} V_{1;2}} G_{d;s} + \frac{T_{1;2}}{V_{1;2}} \frac{dV_{1;2}}{dt}, \quad (2)$$

here $p_{1,2}$ and $T_{1,2}$ are the gas pressure and temperature in chambers 1 and 2; k is the adiabatic index; R is the gas constant; $T_{g1,2}$ is the hydride temperature in the corresponding generators; $V_{1,2}$ is the volume of working chambers equal to $V_{1,2} = V_0 \pm F_{ex}$; V_0 – harmful volume; F_e – effective area of the bellows; x – displacement coordinate of the end of the bellows, counted from the neutral position; β_1, S_1 – heat transfer coefficient and heat exchange surface between the gas and the bellows with body temperature T_k ; $G_{d,s}$ – gas flow into the chamber at desorption or from the chamber during sorption. The flow rate is determined as follows. The sorption-desorption reaction is a first-order phase transformation and is described by the Clapeyron equation, which can be written under the condition that the change in specific volume $\Delta v \approx RT/p$:

$$\frac{dp}{dT} = \frac{H}{\Delta v T} \text{ after integration } \ln p = A - \frac{H}{RT},$$

where A is the constant of integration, H is the heat of reaction. The sorption – desorption reaction is classified as a heterogeneous topochemical reaction, the kinetics of which can be described by the “shrinking sphere” equation [11]:

$$1 - (1 - \alpha)^{1/3} = k_1 t,$$

where α is the degree of conversion, which for desorption is defined as $\alpha = (m_h - m)/m_h$; m_h is the total mass of hydrogen in the system; m is the mass of gaseous hydrogen; k_1 is the reaction rate constant at a given pressure and temperature:

$$k_1 = B e^{-Q/RT} \left[\exp\left(A - \frac{H}{RT}\right) - p \right],$$

where: Q – activation energy; B – empirical constant. The expression for the gas flow rate during desorption will have the form:

$$G_d = m_h \frac{d\alpha}{dt} = 3B e^{-Q/RT} \left[\exp\left(A - \frac{H}{RT}\right) - p \right] \left(m_h - \frac{pV}{RT} \right)^{2/3} m_h^{1/3}. \quad (3)$$

Similarly, the expression for the flow rate during sorption is obtained:

$$G_s = 3B e^{-Q/RT} \left[\exp\left(A - \frac{H}{RT}\right) - p \right] \left(\frac{pV}{RT} \right)^{2/3} m_h^{1/3}. \quad (4)$$

The expression obtained here for the flow rate during desorption practically coincides with the formula given in [12], where the flow rate is proportional to the mass of bound hydrogen. This difference is due to the fact that the expression for the flow rate in [12] is based on the Avrami-Kolmogorov kinetic equation: $-\ln(1 - \alpha) = k_1 t^n$ for the case $n = 1$.

When deriving the ratios for the flow rate, the so-called, pressure hysteresis, which consists in the fact that the plateau pressure during sorption is greater than the plateau pressure during desorption. The phenomenon of hysteresis can be taken into account by taking different values of the constant A and the heat of reaction H for sorption-desorption processes. It is noted that for intermetallic alloys of the $LaNi_{5-x}Al_x$ type, a moderate pressure hysteresis is observed, the value of which does not exceed 0.05 MPa.

For a rotary actuator, the volume of the working chamber is defined as a quarter of the volume formed by the rotation of a curvilinear trapezoid, consisting of a semicircle and a rectangular part of the vane, around the vertical axis (Fig. 2c).

The equations describing the process of changing the temperature of metal hydride in generators are written on the basis of the energy balance relations. At the same time, the so-called, a thin layer of metal hydride, which makes it possible to write these equations in lumped parameters, i.e. disregarding the propagation of the temperature front and the diffusion of hydrogen in the layer. The possibility of such recording follows from the condition $\tau_1 \ll \tau_2$, where: $\tau_1 = l^2 c_2 \rho_2 / \chi$ is the time constant characterizing the rate of temperature change in the layer; l is the characteristic layer thickness, ρ_2 is the density of the metal hydride powder $\approx 4103 \text{ kg/m}^3$, χ is the thermal conductivity of the powder, $\approx 1.5 \text{ W/mK}$, c_2 is the specific heat capacity of the powder $\approx 360 \text{ J/kgK}$, $\tau_2 = m_2 c_2 / \alpha_3 S_3$ is the time constant characterizing the heat transfer from the heater to the layer; $\alpha_3 S_3$ is the heat transfer coefficient and the heat exchange surface between the heater and the powder, m_2 is the mass of the powder. Because since the heat capacity of the heater is much less than the heat capacity of the metal hydride in the reactor, and the thermal resistance between them is sufficiently small, then the thermal inertia of the heater can be neglected in the first approximation and it can be assumed that the heat power from the heater W is uniformly dissipated in the layer. Taking into account these assumptions, the energy balance equations for metal hydride in reactors and for actuator body can be written as:

$$c_2 m_2 \frac{dT_{g1;2}}{dt} = W - HG_{d;s} - \alpha_4 S_4 (T_{g1;2} - T_k), \quad (5)$$

$$C_k \frac{dT_k}{dt} = \alpha_4 S_4 (T_{g1;2} - T_k) + \beta_1 S_1 (T_{1;2} - T_k) - \alpha_5 S_5 (T_k - T_0), \quad (6)$$

where: α_4, S_4 – effective heat transfer coefficient and heat exchange surface between the powder and the body, C_k – heat capacity of the body, α_5, S_5 – heat transfer coefficient and heat exchange surface between the body and the environment with temperature T_0 .

The equations of motion for linear and rotary actuators, respectively, can be written as:

$$m_d \frac{d^2 x}{dt^2} = F_e [p_1(T_1) - p_2(T_2)] - 2cx - P_n, \quad (7)$$

$$J \frac{d^2 \varphi}{dt^2} = M_d - M_{tr} \text{sign} \left(\frac{d\varphi}{dt} \right) - v_t \frac{d\varphi}{dt} - M_n, \quad (8)$$

where: m_d is the mass of the moving parts; J is the moment of inertia; φ is the angle of the shaft rotation; c is the stiffness of the bellows; P_n, M_n are, respectively, the external force on the rod of the linear actuator and the moment of resistance for the rotary actuator; M_{tr} is the moment of dry friction; v_t – coefficient of viscous friction. The driving moment is calculated as in analogous hydraulic actuators.

Equations (1)–(8) represent a mathematical model of a actuator with a metal hydride power source, respectively, of a linear and rotary operation.

4 Experimental Data and Calculation Results

The constants of the $LaNi_{4.7}Al_{0.3}$ alloy used here. Have the following values: $H = 19 \text{ MJ/kgK}$, $A = 24.89$, $C = 4506 \text{ K}$, $Q/R = 1100 \text{ K}$, $B = 10 - 4 \text{ (Pa s)} - 1$, $R = 4121 \text{ J/kgK}$, $k = 1.4$. For a linear actuator, the numerical values of the parameters are: $m_2 = 0.0026 \text{ kg}$, body heat capacity 169 J/K , $c = 0.96 \cdot 10^{-4} \text{ N/m}$, $Fe = 1.3 \cdot 10^{-4} \text{ m}^2$, $\alpha_4 S_4 = 0.13 \text{ W/K}$, case dimensions: diameter 23.5 mm, length 72 mm, weight 0.215 kg . For rotary actuator: $m_2 = 0.012 \text{ kg}$, $J = 1.8 \cdot 10^{-5} \text{ kgm}^2$, $h = 0.025 \text{ m}$, $R_1 = 0.05 \text{ m}$, $r = 0.01 \text{ m}$, $M_{tr} = 6.1 \text{ Nm}$, $v_t = 1.6 \text{ Nms}$, actuator diameter 150 mm, height 96 mm, weight 3.76 kg.

Figure 3a shows the static characteristic of a linear actuator - the dependence of the force developed on the rod depending on its position and the supplied thermal power. The solid lines are the results of calculations by the mathematical model, the markers are the results of measurements on a natural sample. Figure 3b shows the dependence of the steady-state body temperature depending on the thermal power supplied to one of the generators.

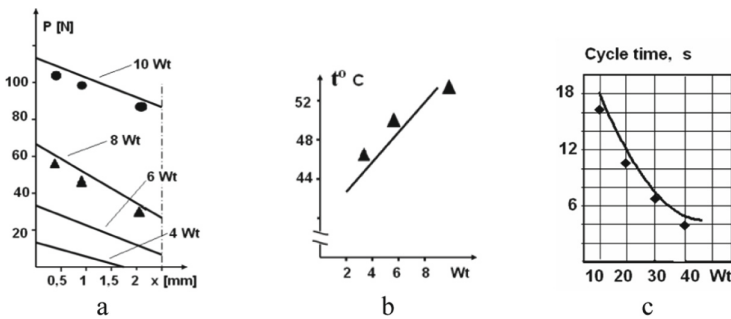


Fig. 3. a) Static characteristic of the linear drive, b) dependence of the steady-state temperature of the body on the input thermal power, c) cycle time depending on the heat input.

Note that to ensure the stability of the alloy sorption capacity during thermal cycling, it is necessary to limit its maximum temperature, above which the destruction process can begin - the release of metallic nickel and the formation of very stable lanthanum hydride. For the alloy used, the temperature should not exceed $170\text{--}180 \text{ }^\circ\text{C}$. Therefore, a long-term connection of the heater in the generator is limited to a power of $N_e = 10 \text{ W}$. Short-term supply of increased power (not more than 50 W) for a period of not more than 10 s is allowed. Figure 3c shows the dependence of the drive cycle time on the thermal power supplied to the generators. Here, the cycle time is the time of movement from one extreme position of the rod to the opposite one, when the heaters in the respective generators are switched on alternately.

Figure 4 shows the results of calculations and measurements for a rotary actuator. The rotation angle as a function of time was measured using a pointer indicator, an angular scale, and a video recorder at a frequency of 20 frames per second (marked on the graph by markers).

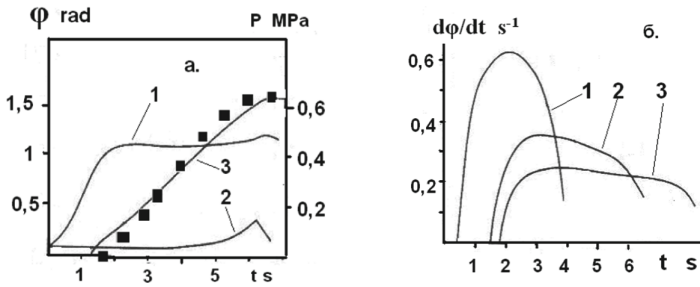


Fig. 4. a) Dependence of the angle of rotation of the shaft (curve 3) and the pressure in the cavities (curves 1 and 2), b) dependence of the angular velocity of the shaft on the moment of loading.

Figure 4a shows the calculating results for the transient process of an actuator loaded with a constant resistance moment of 40 Nm. Curves 1 and 2, respectively, of the pressure in the working chamber and counterpressure chamber. The electrical power dissipated in the working generator is $N_e = 200$ W. On Fig. 4b, the calculating results for the dependence of the angular velocity on the load moment are presented: curve 1 - $M_n = 4.8$ Nm, curve 2 - $M_n = 40$ Nm, curve 3 - $M_n = 60$ Nm. The values of dry and viscous friction moments were determined in a separate experiment.

5 Conclusion

The mathematical model of linear and rotary metal hydride actuator prototypes allows us to evaluate the characteristics of the movement and the developed forces and torque with an accuracy of 10–15%.

The main contribution to the error can be made by the following factors:

- the intermetallic compound $LaNi_{4.7}Al_{0.3}$ is a precision alloy and its thermodynamic and kinetic characteristics depend significantly on the accuracy of the composition. In addition, the presence of even minor impurities of O_2 , CO , H_2O in the hydrogen used for activation can have a certain effect on these characteristics;
- the accepted assumptions about the uniformity of heating and cooling of the metal hydride powder in the generators and the actuator housing. So for a linear actuator (Fig. 1a) in steady state, the temperature of the housing made of aluminum alloy may differ in length by 1.5–2.5 °C. Also, the case temperature slightly depends on its orientation and for a vertical position it is 0.3–0.7 °C lower than for a horizontal one;
- the simple friction model adopted here for a rotary actuator does not take into account the difference in the friction moment values of rest and movement, as well as its dependence on pressure in the working chambers;
- large uncertainty in the values of the heat transfer coefficients between the component parts of the actuator.

The low energy efficiency of the actuators considered here should be noted. The simplest energy efficiency estimate for the rotary actuator can be represented as $\eta = M_n\varphi/N_e t_d$, where $\varphi = \pi/4$, t_d is the travel time, which gives $\eta = 5\%$. In [13], the

energy efficiency of a linear actuator for the most favorable operating mode reaches 13%, and in [6], an artificial muscle drive has an energy efficiency 1%. The advantages of these type actuators include a significant developed force per unit mass and volume, quiet operation, and the possibility of fairly accurate regulation.

The approach presented here to the mathematical modeling of the developed gas actuators designs makes it quite easy to obtain estimates of their static and dynamic characteristics.

References

1. Ivey, O., Northwood, D.: Storing energy in metal hydrides: a review of the physical metallurgy. *J. Mater. Sci.* **18**, 321–347 (1983)
2. Bennett, L.G., Arguner, S.D., Hewitt, J.S.: Investigation of hydriding processes in low-temperature/low-pressure metal hydrides. *Int. J. Hydrogen Energy* **11**(9), 577–582 (1986)
3. Hosono, M., Sakaki, K., Nakamura, Y., Ino, S.: Metal hydride actuator for a rescue jack driven by hydrogen desorption. *Int. J. Hydrogen Energy* **44**(55), 29310–29318 (2019)
4. Shin, M.Y., Chong, W.S., Yu, C.: Study on an actuation system development using temperature control of metal hydrides. *Technol. Health Care* **28**(S1), 115–122 (2020)
5. Kurosaki, K., Maruyama, T., Takahashi, K., Muta, H., Uno, M., Yamanaka, S.: Design and development of MH actuator system. *Sens. Actuators A Phys.* **113**(1), 118–123 (2004)
6. Vanderhoff, A., Kim, K.J.: Experimental study of a metal hydride driven braided artificial pneumatic muscle. *Smart Mater. Struct.* **18**(12), 125014 (10 pages) (2009)
7. Lloyd, G., Kim, K.J.: Smart hydrogen/metal hydride actuator. *Int. J. Hydrogen Energy* **32**(2), 247–255 (2007)
8. Goto, K., Hirata, T., Yamamoto, I., Nakao, W.: Suitability evaluation of LaNi_5 as hydrogen-storage-alloy actuator by in-situ displacement measurement during hydrogen pressure change. *Molecules* **24**(13), 2420 (10 pages) (2019)
9. Usui, T., et al.: Fully flexible liquid-to-gas phase change actuators with integrated liquid metal heaters. *Jpn. J. Appl. Phys.* **60**(SC), SCCL11 (10 pages) (2021)
10. Ivlev, V.I., Misyurin, S.: To refining the thermo-mechanical model of vane type air motor. *Procedia Comput. Sci.* **190**, 377–387 (2021)
11. Rudman, P.S., Goodell, P.D.: Hydriding and dehydriding rates of the LaNi_5 -H system. *J. Less Common Met.* **89**(1), 117–125 (1983)
12. Luxenburger, B., Müller, W.: Investigations of the discharging of metal hydride beds for hydrogen-gasoline mixture operation of SI-engines. *Int. J. Hydrogen Energy* **10**(5), 305–315 (1985)
13. Bhuiya, M., Kim, K.J.: Performance study of a hydrogen powered metal hydride actuator. *Smart Mater. Struct.* **25**(4), 045004 (10 pages) (2016)



Associative Memory with Biologically-Inspired Cell Assemblies

Yuehu Ji^(✉), David Gamez, and Chris Huyck

Middlesex University, The Burroughs, London NW4 4BT, UK
yj097@live.mdx.ac.uk

Abstract. Associative memory is a central cognitive task. However, the actual biological architecture that supports this memory is not currently known, so simulating with biologically plausible neurons and topologies is an ideal mechanism to improve understanding of associative memory. Simulations of spiking networks that perform associative memory tasks lay the groundwork for utilizing biological neurons in cognitive tasks. Specifically, this paper explores simulations of spiking networks that perform associative memory tasks using Hebbian cell assemblies of neurons to represent nodes and synapses to represent associations. The first tasks use binary cell assemblies to perform two well-known cognitive tasks. Then the paper examines different topologies of excitatory neurons for basic assemblies and their performance as short-term memory. Lastly, larger assemblies are associated in $2/3$ sets, where two active elements can retrieve the third. Future research is proposed to explore the potential use of these assemblies and associations in cognitive tasks. By investigating biologically and cognitively plausible topologies, learning, and neurons, simulations will lead to an improved understanding of neuro-cognition, and potentially to systems that surpass the brittleness and domain specificity of current AI systems.

Keywords: Cell assemblies · Neurocognitive model · Stroop task · Associative memory · Spiking network

1 Introduction

Large deep neural networks, such as GPT, can accurately mimic natural language in diverse fields. However, these models depend on statistical patterns in the training data and can only produce shallow models that are detached from reality. They fail to properly comprehend semantic relationships between words and cannot achieve a human-like understanding of the world because the concepts they manipulate have no foundation. They are unable to learn novel ideas or dynamically remember associations between them in the way that humans do. The work presented in this paper seeks to go beyond these limitations by developing a biologically-inspired associative memory system that captures semantic relationships between words, answers questions about their relationships in a psychologically plausible way and learns associations.

When neurons display similar spiking patterns in response to a stimulus, the connections between the neurons may become strengthened through a process first proposed by

Hebb [1]. These connections can form a cell assembly (CA), comprising groups of interconnected neurons that facilitate efficient storage and retrieval of related information. Experimental and theoretical evidence support the existence of cell assemblies [2, 3] and theoretical models of neural networks have been developed to simulate the formation and function of CAs [4].

This paper gives an overview of the authors' work constructing biologically-inspired associative memory systems using CAs. The first part describes initial work in which a hard-wired network of cell assemblies was used to model the hierarchical structure of semantic words in a Stroop task and an associative memory task. The next stage (Sect. 3) investigates good topological structures of CAs particularly those involved in associative memory tasks. These topologies are then used to learn 2–3 cell assembly associations (see Sect. 4). The final part of this paper describes plans for scaling up this associative memory model for larger associative memory tasks.

2 Cell Assembly Models of Semantic Retrieval

It is reasonably simple to implement logic in simulated spiking neurons. Using simple persistent assemblies, the authors developed a model of the Stroop task [5]. In word recognition and colour naming tasks, the subjects are presented with a colour word, such as 'red' or 'blue' that is written in coloured ink. In the congruent situation, the colour of the ink matches the meaning of the word (for example, 'red' written in red ink). In the incongruent situation the colour of the word is different from the colour of the ink (for example, the word 'red' written in blue ink). The subjects must recognize and repeat the word (WR) or name the colour of the ink (CN). When humans perform these tasks, they have a faster reaction time in WR tasks compared with CN tasks. Subjects also have slower reaction times on CN tasks in the incongruent situation where the word and ink colour disagree, but the difference in reaction time is not significant in incongruent WR tasks. This difference in response times is known as the Stroop effect [6]. The authors' simulation of the Stroop effect was constructed using eight cell assemblies. The simulation was able to re-produce similar timings to human subjects with congruent and incongruent colour word combinations.

Further simulation of a classic semantic net task, [7] has also been completed [8]. Here a small hierarchy has been attributed to people, and psychological experiments have been performed on subjects.

The semantic net examined revolves around animal concepts with for example a canary that is a bird, and a bird that is an animal. Features are associated with these animal classes, so canaries are yellow, birds can fly, and animals eat. Subjects were then queried with a true or false question, for example, "Do canaries fly?" Subjects were observed to take longer to respond to canaries flying than to them being yellow. The explanation of this phenomenon was that features higher in the hierarchy require longer processing time. An overview of the structure of the network is shown in Fig. 1.

The use of spiking neuron models for timing of response is particularly good due to the ability to directly derive performance timings from the collective firing of individual neurons. Additionally, the size and hierarchical structure of the network can directly determine its effectiveness and potentially explain some of the cognitive connections between individual concepts. However, as these models use simple parameters

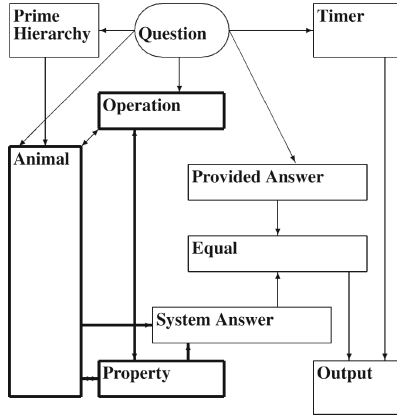


Fig. 1. Gross topology of the question answering associative memory. Boxes represent sets of neurons. Thick boxes and arrows are the core of the semantic memory. The oval represents the question with spike sources instead of neurons.

to account for neural activity, they do not guarantee biologically accurate reproduction of neural processes, although they may provide a reasonable approximation. The present associative memory model, for instance, employs only 1130 neurons, whereas the brain uses millions, if not billions, of neurons for similar tasks.

3 Finding Good Topologies for Cell Assemblies

The empirical details of cell assembly structure in the brain are challenging to analyze and observe. Even if a CA can be statistically identified, the participating neurons in the cell assembly may vary across different sessions with the same stimulus [9]. Because neurons in the brain fire constantly at a low rate, it can be difficult to differentiate between neurons participating in an active cell assembly from the noise signals. The number of neurons in a CA is not even clear and one neuron may participate in multiple assemblies. Although the biological specifics of cell assemblies remain elusive, various synaptic interaction mechanics have been discovered during their formation. Populations of neurons, whether in computational simulations or neuron reconstruction projects [10], are considered as a combination of statistical connectivity derived from neuron anatomical data and connection rules governed by probabilistic and deterministic principles. Recent research has sought to evaluate the relationship between network topology and the dynamics of biological neuronal networks [11, 12]. Previous evaluation either has not fully captured the working cycle of spiking behaviors in neuron groups nor provided in-depth exploration on the performance of individual topologies. Moreover, the analysis and development of proper evaluation methods for neuronal network generation topologies are currently lacking. This section addresses this gap and proposes the development of standard test-sets for topology evaluations.

The topologies are built with pyNN package in python [13]. Networks with different topologies were simulated. All the topologies are coded in a general purpose language

to prevent mismatch from the package default connection rules. This study examined several topologies including random networks [14], small world networks [15], and scale free networks [16]. All of the network used probabilistic rules with random generation. Each network is evaluated based on the average performance of 10 samples. Self-connections are prohibited and there are no duplicate connections for all topologies. Networks of 1000 integrate and fire spiking neurons were created with each of these topologies and initialized with different random seeds to explore the range of their behaviors. An illustrative result is shown in Fig. 2. Once the CA ignites, neurons fire persistently for over a second.

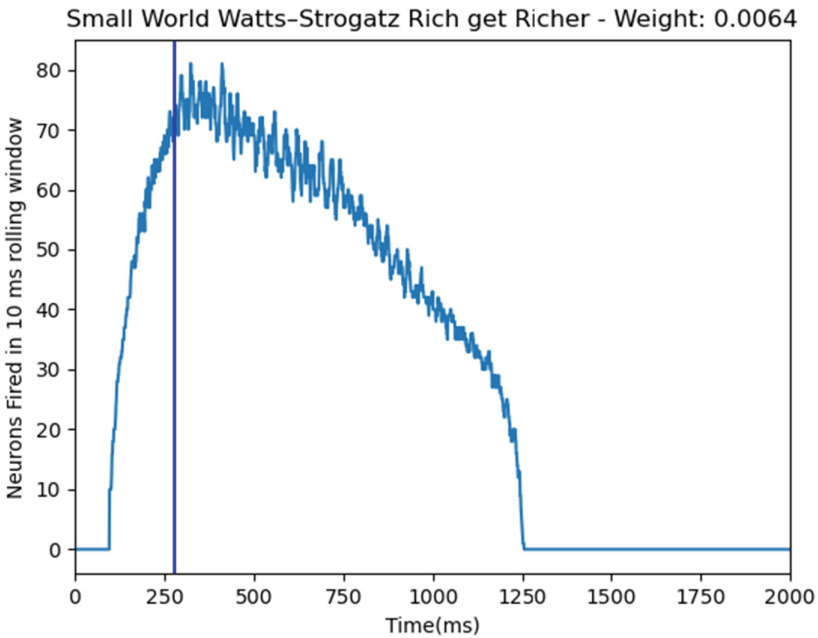


Fig. 2. Cell assembly based on small world topology. It was stimulated with external activation from 100 ms to 280 ms and exhibited self terminating persistence until 1200 ms.

A CA ignites when its neurons can persistently fire without external input. While the neurons are firing, theory implies, that the item associated with the CA is in short-term memory. Since short-term memories do not persist indefinitely, it is desirable for a CA to self-terminate. Firing patterns were investigated by varying connecting weights and comparing their tunability, self-terminating persistence, ignitability, and robustness. These experiments revealed that the small world topology with a rich get richer rewiring approach worked best for associative memory among our simulation. This topology exhibited the largest range of synaptic weights for self-terminating persistence for all tested random seeds. Information is sent to other CAs in the forms of population spikes. Longer controlled firing in a CA further enables its participation in more tasks in the following network computation, which maximize its information capacity.

4 An Associative Memory Model with Cell Assemblies

More recently, small world CAs, from Sect. 3, were combined into associative memories. The aim is to use the improved control over the behavior of CAs to develop scalable hierarchical networks that can be expanded and dynamically learn new associations.

Five cell assemblies of 1000 neurons are simulated based on small world topologies. There are random synapses between associated cell assemblies. Inhibitory neurons prevent each assembly from continuously firing and global inhibition, stimulated by all assemblies, prevents unassociated assemblies igniting. When two of the associated CAs ignite, they ignite the third, which is a model for retrieval of that memory. Experiments have both hard-wired and plastic intra-assembly synapses.

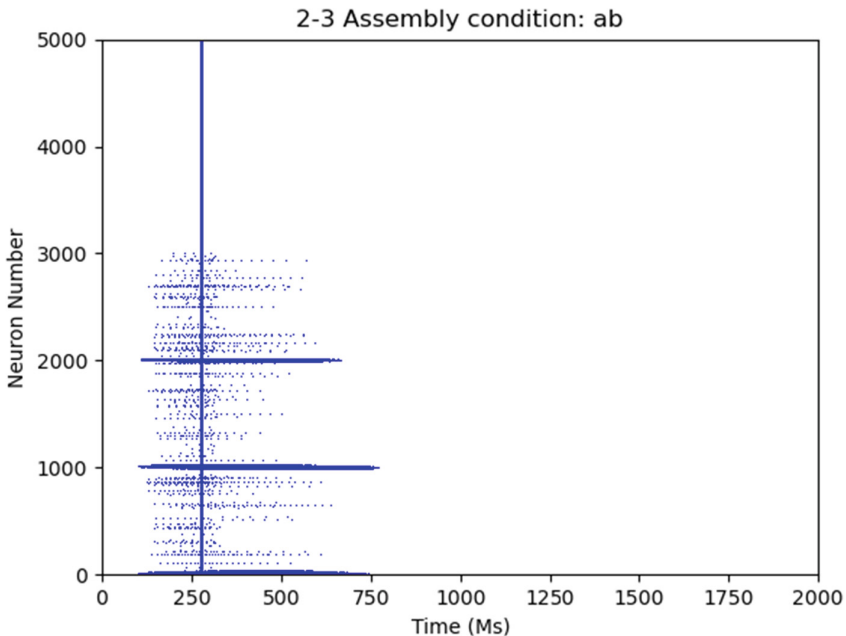


Fig. 3. Rastergram of two assemblies (a and b) retrieving a third.

The results, shown in Fig. 3, show that a stimulated assembly normally persists for a short amount of time (~ 500 ms) though inhibitory neurons are not involved in the task. In a network of five assemblies, there are three sets of associations, a–b–c (neurons 0–2999), and c–d–e (neurons 2000–4999). That is there are two sets of overlapping groups, or two 2–3 assemblies. When two assemblies in the same group are stimulated, they persist longer and eventually ignite the third assembly. When two assemblies from different assembly groups are stimulated, they do not ignite any other assemblies. The same results are achieved for both hard-wired and plastic intra-assembly synapses.

5 Discussion and Conclusion

This paper has outlined the research that the authors have been doing on the use of cognitive models constructed with biologically-inspired cell assemblies of spiking neurons. In the initial stages of this work, hard-coded cell assemblies modelled the Stroop task and implemented an associative memory semantic network in biologically plausible ways. Further research on cell assembly behavior demonstrated that small world topologies are likely to lead to better behavior for associative memory. This insight was used to develop the networks of 2–3 associated assemblies that was described in Sect. 4.

The 2–3 assembly system may be extended as an associative memory system, allowing for the incorporation of new memories that do not conflict with previously learned cell assembly groups. This framework may be the basis for internal memory manipulation for other cognitive tasks.

During the above simulations, inhibition played a peripheral role in shaping network dynamics, owing to the efficiency of information transfer via excitatory spikes. Due to the inherently noisy nature of neuron spikes, encoding and decoding of information poses significant challenges without multiple layers of spike controls. While inhibitory neurons are more likely to contribute to network regulation, they may not be involved in information encoding and decoding. Consequently, the system could operate without inhibition, but recalibration would be necessary with the introduction of inhibitory mechanisms.

It would be better for cell assemblies to emerge from a pool of neurons through exposure to external stimuli, rather than being pre-wired. Incorporating Hebbian learning would enhance biological plausibility, but controlling the learning outcomes to prevent catastrophic failures presents a challenge. The compensatory learning approach used in the 2–3 assembly model represents a preliminary step towards on-line modification of synaptic connections, although further evidence is required to establish its effectiveness.

Cell assemblies are internal representations of observations; however, it is still unknown how the brain encodes this information. Some studies suggest that certain forms of synaptic plasticity may result in the strengthening or enhancement of internal information within cell assemblies, while others suggest that activity-dependent synaptic pruning may lead to the selective erasure of certain internal representations. Ultimately, further research is needed to fully understand the complex dynamics underlying the processing and representation of internal representation within CAs and their role in information processing.

The next stage of this research will be to use small-world cell assemblies to learn new associations between words. Initial exploration will be with small networks like the Quillian network above, then new associations and concepts will be added. The networks will be progressively expanded to networks that are loaded from existing semantic networks, such as WordNet. There are also plans to connect the words in our hierarchical semantic networks to the real world. An image classification library, such as OpenCV could be used to label objects in a live camera stream. These labels could activate the appropriate cell assemblies and then the system could answer questions about the objects that it is perceiving and potentially learn new associations between words based on what it is perceiving.


The work described in this paper has shown that biologically-inspired models of cell assemblies can effectively model human behavior on semantic tasks. The experiments have shown that cell assemblies with small-world topologies have appropriate behavioral characteristics for associative memory models and can learn simple associations between concepts. Future cell assembly-based models will be scaled to model larger semantic networks and grounded in live data from the real world.

References

1. Hebb, D.: *The Organization of Behavior: A Neuropsychological Theory*. Wiley (1949)
2. Huyck, C., Passmore, P.: A review of cell assemblies. *Biol. Cybern.* **107**, 263–288 (2013)
3. Sakurai, Y., Tanisumi, Y., Ishihara, E., Hirokawa, J., Matanabe, H.: Multiple approaches to the investigation of cell assembly in memory research—present and future. *Front. Syst. Neurosci.* **12**, 21 (2018)
4. Gerstner, W., Kistler, W., Naud, R., Paninski, L.: *From Single Neurons to Networks and Models of Cognition*. Cambridge University Press (2014)
5. Ji, Y., Gamez, D., Huyck, C.: A brain-inspired cognitive system that mimics the dynamics of human thought. In: 38th SGAI International Conference on Artificial Intelligence, Cambridge, UK (2018)
6. Stroop, J.: Studies of interference in serial verbal reactions. *J. Exp. Psychol.* 643–662 (1935)
7. Collins, A., Quillian, M.: Retrieval time from semantic memory. *J. Verbal Learn. Verbal Behav.* **8**(2), 240–247 (1969)
8. Huyck, C., Ji, Y.: Two simple neurocognitive associative memory models. In: 16th International Conference on Cognitive Modelling (2018)
9. Buzsaki, G.: Neural syntax: cell assemblies, synapsesembles, and readers. *Neuron* **68**(3), 362–385 (2010)
10. Markram, H., Muller, E., Ramaswamy, S., Reimann, M.W., Abdellah, M., Sanchez, C.A., Ailamaki, A., et al.: Reconstruction and simulation of neocortical microcircuitry. *Cell* **163**(2), 456–492 (2015)
11. Mengiste, S.A., Aertsen, A., Kumar, A.: Relevance of network topology for the dynamics of biological neuronal networks. *bioRxiv* 02 (2021)
12. Senk, J., Kriener, B., Djurfeldt, M., Voges, N., Jiang, H.-J., Schüttler, L., Gramelsberger, G., Diesmann, M., Plesser, H.E., van Albada, S.J.: Connectivity concepts in neuronal network modeling. *PLoS Comput. Biol.* **18**(9), e1010086 (2022)
13. Davison, A.P., Brüderle, D., Eppler, J.M., Kremkow, J., Muller, E., Pecevski, D., Perrinet, L., Yger, P.: PyNN: a common interface for neuronal network simulators. *Front. Neuroinform.* **2**, 388 (2009)
14. Erdos, P., Renyi, A.: On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* 17–60 (1960)
15. Watts, D., Strogatz, S.: Collective dynamics of ‘small-world’ networks. *Nature* **393**(6684), 440–442 (1998)
16. Albert, R., Barabasi, L.: Statistical mechanics of complex networks. *Rev. Mod. Phys.* **47** (2002)
17. Brette, R., Gerstner, W.: Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *J. Neurophysiol.* **94**, 3637–3642 (2005)



Combined Contrast Enhancement Algorithm for High Dynamic Range Images

M. A. Kazakov^(✉) 

Institute of Applied Mathematics and Automation KBSC RAS (IAMA KBSC RAS), Nalchik,
Russia

kasakow.muchamed@gmail.com

Abstract. Image contrast enhancement is the process of improving the visual quality of an image by adjusting its brightness, color, and sharpness. When working with large bit raw images obtained directly from the matrix of the equipment, there are specific problems associated with a large dynamic range. The paper proposes a combined contrast enhancement method that can significantly improve the contrast of such raw images with a large dynamic range. In the combined method, highlight regions are softly clipped on the histogram using a clustering algorithm based on feature space partitioning and gamma correction of the clipped region. The clustering algorithm used does a good job of detecting the cutoff point, both in the presence of highlight regions and in the absence. The method also produces light border underlining based on Sobel filters. The well-known Contrast Limited Adaptive Histogram Equalization method is used to improve the histogram. In this case, a combination of transformations with different grid sizes is used, which allows to achieve much better results than when selecting one optimal transformation. These algorithms are described in detail and illustrated for comparison.

Keywords: Contrast enhancement · X-ray images · Image processing · Raw images · Histogram equalization

1 Introduction

The pursuit of perfecting image contrast has been a longstanding endeavor in the field of computer vision and image processing. Historically, early imaging techniques relied on manual adjustments and traditional darkroom methods to improve contrast, but these methods often fell short of achieving optimal results. The advent of digital imaging and the ever-evolving landscape of computational algorithms have revolutionized the way we enhance contrast in images, enabling us to unlock greater potential in various applications.

Image contrast enhancement is the process of improving the visual quality of an image by adjusting its brightness, color, and sharpness. It can help to reveal hidden details, highlight important features, and reduce noise or blur in the image. Image contrast enhancement can be applied to various types of images, such as grayscale, color, medical, satellite, or low-light images [1].

When working with raw images obtained directly from the matrix of equipment (for example, X-ray), the tasks of correcting the histogram and improving contrast arise. In particular, highlight regions, which correspond to very high pixel intensity, make the informative area very dim. In the informative region of an image, the pixel intensities may lie close together compared to the dynamic range, making the image illegible. Many methods have been developed to solve these problems and improve image representation [2, 3]. This paper proposes a combined raw image processing method that allows for highlight areas reduction, histogram equalization, and contrast and detail enhancements.

Preprocessing is important not only for improving human perception of an image, but also necessary when working with automated data analysis. For these purposes, more specific methods can be used [4–8].

The method is based on Contrast Limited Adaptive Histogram Equalization and clustering algorithm based on feature space partitioning.

2 Formalization

An image can be defined as a two-dimensional function $f(x, y)$, where x and y are spatial coordinates, and the value of the function for each pair of coordinates characterizes the intensity or gray level of the image at a point. For a digital image, the intensity and coordinate values are discrete and limited. In this case, the intensity must also be non-negative. The number of discrete intensity levels L is defined as 2^k , where k is a positive integer corresponding to the number of bits allocated to store the intensity value.

The transformation of the image in the general case can be represented by the expression

$$g(x, y) = T[f(x, y)]$$

where $f(x, y)$ is the input image, $g(x, y)$ is the output image, and T is an operator acting on the values of f defined in some neighborhood of the point (x, y) . The smallest possible neighborhood has a size of 1×1 . In this case, g depends only on the value of f at a single point (x, y) and the operator T , and is the intensity transformation function:

$$s = T(r)$$

where symbols are used to simplify the notation s and r corresponding to the intensities g and f for every point (x, y) .

Denote by r_k , where $k = 0, 1, 2, \dots, L - 1$, intensity values of the L -level digital image $f(x, y)$. The unnormalized histogram of f is defined as $h(r_k) = n_k$, where n_k – is the number of pixels in f with intensity r_k . The normalized histogram is defined as

$$p(r_k) = \frac{h(r_k)}{MN} = \frac{n_k}{MN}$$

where M and N are height and width of the image. The sum $p(r_k)$ for all k is equal to 1. The components $p(r_k)$ are estimates of the probabilities of detecting intensity levels in the image.

3 Histogram Equalization

Assume that the intensity values r lie in the range $[0, L - 1]$, where $r = 0$ corresponds to black and $r = 1$ to white. Consider the intensity transformation T satisfying the conditions:

1. $T(r)$ is strictly monotonic increasing function in the range $0 \leq r \leq L - 1$;
2. $0 \leq T(r) \leq L - 1$ for $0 \leq r \leq L - 1$.

These conditions ensure that the output intensity value will not be less than the input value and will not fall outside the range of the input values.

The intensity of the image can be considered as a random value in the interval $[0, L - 1]$. Let $p_r(r)$ and $p_s(s)$ be the probability densities of the intensity of the values r and s in two different images. The subscripts for p are introduced to indicate that p_r and p_s are different functions. According to the change-of-variable technique, if $T(r)$ is a function that is continuous and differentiable over the entire range, then the intensity probability density s can be expressed as

$$p_s(s) = p_r\left(T^{-1}(s)\right) \left| \frac{d}{ds} \left(T^{-1}(s)\right) \right| = p_r(r) \left| \frac{dr}{ds} \right|.$$

Consider the following transformation function:

$$s = T(r) = (L - 1) \int_0^r p_r(x) dx.$$

The integral on the right side represents the cumulative distribution function (CDF) of the random variable r . According to Leibniz's rule, we can write

$$\frac{ds}{dx} = (L - 1)p_r(r).$$

As a result, for the probability density of a random variable s over the entire range $0 \leq s \leq L - 1$ we get

$$p_s(s) = \frac{1}{L - 1}.$$

Thus, as a result of the transformation $T(r)$ we obtain a uniform distribution of the probability density $p_s(s)$ over the entire range of intensities. It follows from this that the CDF, which is expressed as an integral of the probability density, will be a straight line.

If the variables r and s take only discrete values $0, 1, 2, \dots, L - 1$, then the transformation function $T(r)$ can be written as

$$T(r) = \text{floor} \left((L - 1) \sum_{k=0}^r p_r(k) \right),$$

where operator floor rounds a real number, leaving the integer part.

Figure 1 shows the histograms and CDF of the original image and the resulting image with histogram equalized.

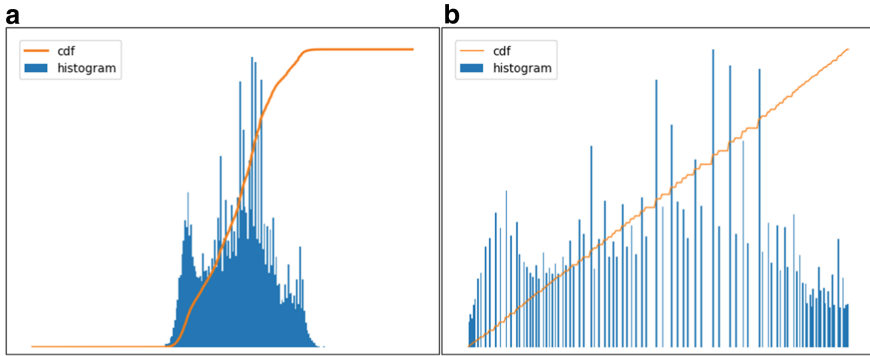


Fig. 1. a) Original; b) equalized.

4 Contrast Limited Adaptive Histogram Equalization

The described method produces a global histogram equalization for the entire image, which in some cases works quite well. However, in most cases, this alignment leads to an improvement in some areas of the image and a deterioration in others. For a more delicate result, adaptive histogram equalization (AHE) is used [9]. In this method, the image is divided into contextual blocks, within which the histogram equalization is performed separately. The image is divided in a grid of rectangular contextual regions in which the optimal contrast must be calculated. As a result of this alignment, intensity drops can form at the block boundaries [10]. To avoid this, bilinear interpolation is used.

The AHE method allows you to get a significantly better result, but the side effect in the image is increased background noise present in homogeneous areas. To solve this problem, the Contrast Limited Adaptive Histogram Equalization (CLAHE) method was proposed, which is a modification of the AHE method. The noise problem that AHE exhibits can be overcome by limiting the contrast enhancement, particularly in homogeneous areas of the image. These areas are characterized by a high peak in the histogram of the block, since many pixels of the homogeneous area fall into the narrow area of the histogram. On the CDF diagram, such areas are characterized by the steepest slope. In the CLAHE method, the histogram is clipped at a certain height, and the pixels that fall into the clipping region are redistributed over the remaining histogram region. As a result of such a transformation, the maximum slope angle on the CDF diagram will be limited. After such preprocessing, as a result of histogram equalization, the noise will be contrasted to a much lesser extent. The cutoff level is set by the contrast factor [11].

5 Sobel Operator

The Sobel operator is a technique for edge detection in image processing [12]. It uses two 3×3 kernels, or filters, to calculate the approximate gradients of the image intensity in the horizontal and vertical directions. The Sobel operator can highlight the edges of the objects in the image by measuring the changes in brightness along different directions. It is a discrete differentiation operator, which means that it approximates the derivatives

of a function using discrete values. In this case, the function is the image intensity, and the discrete values are the pixel values. The two kernels used by the Sobel operator are:

Horizontal kernel:

$$\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

Vertical kernel:

$$\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

These kernels are convolved with the original image to produce two new images: one for the horizontal gradient, and one for the vertical gradient. The convolution operation involves sliding each kernel over the image, multiplying the corresponding pixel values, and adding them up to get a new pixel value. The new pixel value represents the amount of change in brightness along that direction.

The horizontal kernel detects changes in brightness along the x -axis (left to right), while the vertical kernel detects changes along the y -axis (top to bottom). The edges of an object are usually regions where there is a high contrast between adjacent pixels, which means that there is a large change in brightness. Therefore, the pixels with high gradient values (positive or negative) indicate the presence of an edge.

The Sobel operator can also combine the horizontal and vertical gradients to get a single gradient magnitude and direction for each pixel. The magnitude is calculated by taking the square root of the sum of squares of the horizontal and vertical gradients, while the direction is calculated by taking the arctangent of their ratio. The magnitude and direction can be used to further analyze or visualize the edges in the image.

6 Soft Clip Highlight Areas

The specificity of the problem lies in the fact that, firstly, highlight areas are not always present on the image, and secondly, its location on the histogram can be arbitrary for different images, even if they were obtained on the same equipment. Therefore, the algorithm must be able to effectively determine the presence or absence of a flare, and also, if it is present, be able to accurately estimate its location on the histogram.

To cut highlight areas, a simplified clustering algorithm based on feature space partitioning is used [12]. In the general case, the algorithm works for an arbitrary dimension of the feature space. In this case, work is carried out on the one-dimensional space bins of the histogram, which allows us to leave aside some elements of the general algorithm. The idea is as follows. Highlight areas are characterized by a high peak in the right area of the histogram, to the left of which there is an empty area, with low bins values. The content of the image is concentrated to the left of the empty area. The clustering method under consideration, based on density analysis, makes it possible to cluster the meaningful part of the image and estimate its right boundary quite confidently. Next,

you can perform soft clipping, which will remove the highlight areas and at the same time preserve the information content, which can remain to the right of the border.

Here is a step-by-step description of the highlight areas reduction algorithm:

1. The histogram is split into N bins ($N = 24$ is a fairly optimal choice).
2. The first bin is selected by iterating over bins, starting from bin with index 0. The first bin is the one whose value exceeds the specified density threshold. This bin initializes the cluster.
3. Starting from the first bin, a frame moves along the histogram, covering 3 bins. At each position, the average of the bins values is calculated. Frame shifts to the right, expanding the cluster until the average value of bins is less than the specified density threshold.
4. The right boundary of the formed cluster is taken as the cut point t . Soft clipping of the right part of the histogram is performed. This is done by gamma transformation of all pixels whose intensity exceeds the cutoff point (i.e. for $r > t$):

$$s = (r - t)^\gamma + t$$

Pixel sampling is used to speed up the algorithm. This in turn makes the density threshold quite robust to the image size. For a sample of m pixels, an optimal density value of $0.004m$ was experimentally obtained, but small variations in this value do not significantly affect the clustering result.

7 Combined Algorithm

A combined algorithm is proposed for efficient histogram correction and image quality improvement. The first step is soft clipping of the flare. In the second step, the histogram is equalized by combining several CLAHE transforms with different tile grid sizes and contrast factors. The number of CLAHE transformations is specified by the depth parameter. Empirically, it has been found that the best results are achieved by choosing the base CLAHE transform with tile grid size equal to 6 (depth = 1), which doubles the tile grid size at each depth step. The optimal result is obtained with depth equal to 4. An increase in depth is accompanied by an increase in image detail, but at the same time it is accompanied by an increase in the running time of the algorithm, and an increase in noise. An example of image processing can be seen in Fig. 2: here are the original image obtained from the X-ray equipment, the processing results at depth = 4 and depth = 8.

8 Conclusion

The proposed combined method allows you to effectively improve the contrast of the raw image obtained on the equipment matrix and achieve high detail, due to highlight areas reduction and a combination of CLAHE transformations. The method is suitable for processing 8 bit images, however, more significant results can be achieved on larger bit images (e.g. 14 bit or 16 bit), the dynamic range of which is much wider. By adjusting the depth, you can achieve different levels of detail. However, it should be borne in mind that high detail requires more computing resources, and is accompanied by some increase in noise.

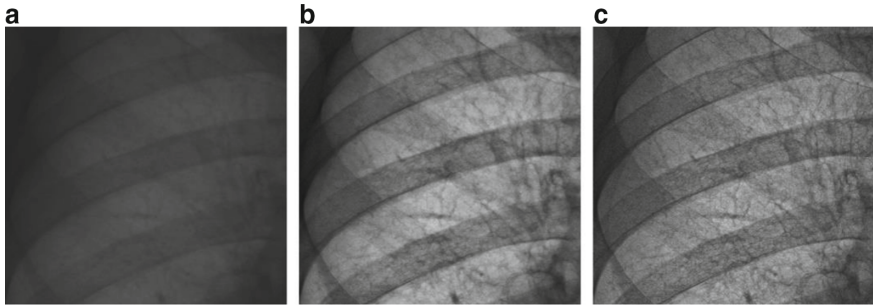



Fig. 2. a) Origin; b) depth = 4; c) depth = 8.

References

1. Vijayalakshmi, D., Nath, M.K., Acharya, O.P.: A comprehensive survey on image contrast enhancement techniques in spatial domain. *Sens. Imaging* **1**(21), 40 (2020)
2. Woods, R.E., Gonzalez, R.C.: *Digital Image Processing*, 4th edn. Pearson, England (2021)
3. Jain, A.K.: *Fundamentals of Digital Image Processing*. Prentice Hall, Englewood Cliffs, NJ (1989)
4. Naidu, S., Quadros, A., Natekar, A., et al.: Enhancement of X-ray images using various image processing approaches. In: *2021 International Conference on Technological Advancements and Innovations (ICTAI)*, pp. 115–120. Tashkent, Uzbekistan (2021)
5. Ishigami, R., Zin, T.T., Shinkawa, N., Nishii, R.: Human identification using X-ray image matching. In: *Proceedings of the International Multi Conference of Engineers and Computer Scientists*, vol. 1 (2017)
6. Rajpurkar, P., et al.: CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint [arXiv:1711.05225](https://arxiv.org/abs/1711.05225) (2017)
7. Costa, M.V.L., de Aguiar, E.J., et al.: A deep learning-based radiomics approach for COVID-19 detection from CXR images using ensemble learning model. In: *36th International Symposium on Computer-Based Medical Systems 2023 (CBMS)*, pp. 517–522. L'Aquila, Italy (2023)
8. Radvansky, M., Kudelka, M.M., et al.: Process of finding human knee in image based on multiple weighted thresholding and histograms of gradients. In: *24th International Carpathian Control Conference (ICCC)*, Miskolc-Szilvásvárad, Hungary, pp. 358–363 (2023)
9. Pizer, S.M., Amburn, E.P., Austin, J.D., et al.: Adaptive histogram equalization and its variations. *Comput. Vis. Graph. Image Process.* **39**, 355–368 (1987)
10. Cromartie, R., Pizer, S.M.: Edge-affected context for adaptive contrast enhancement. In: Colchester, A.C.S., Hawkes, D.J. (eds.) *Proceedings of the XLLth International Meeting on Information Processing in Medical Imaging: Lecture Notes in Computer Science*, pp. 474–485. Springer-Verlag, Berlin (1991)
11. Zuiderveld, K.: VIII. 5. Contrast limited adaptive histogram equalization. In: *Graphics Gems*, pp. 474–485 (1994)
12. Sobel, I.: An Isotropic 3×3 Image Gradient Operator. Presentation at Stanford A.I. Project 1968 (2014)
13. Kazakov, M.: Clustering algorithm based on feature space partitioning. In: *2022 International Russian Automation Conference (RusAutoCon)*, pp. 399–403. Sochi, Russian Federation (2022)



A Hybrid PSO-Jaya Algorithm for Optimization Problems

E. M. Kazakova^(✉) 

Institute of Applied Mathematics and Automation KBSC RAS (IAMA KBSC RAS), Nalchik,
Russia

shogenovae@inbox.ru

Abstract. In this paper introduces and investigates a hybrid PSO-Jaya optimization algorithm based on two heuristic algorithms PSO and Jaya. Two problems: function optimization and training ANN for the classification problems Iris and breast cancer Wisconsin, are employed to evaluate the efficiencies of this new hybrid algorithm. In test calculations, the PSO, Jaya, PSO-Jaya algorithms are compared based on the average, median, standard deviation, and “best” of the best particle position after 50 independent runs for benchmark functions and 30 for network training. The results are compared with the PSO and Jaya algorithms. For all test cases, PSO-Jaya shows the best performance in terms of convergence rate and avoidance of local minima.

Keywords: Heuristic algorithm · Hybrid algorithm · PSO · Jaya · Neural networks

1 Introduction

A heuristic algorithm is a problem-solving method that uses rules of thumb or approximate strategies to find solutions. Unlike exact algorithms that guarantee an optimal solution, heuristic algorithms provide efficient and practical solutions, although they may not always be optimal. Swarm algorithms are metaheuristic methods that model the collective behavior of swarm organisms such as bees, fish, or birds to solve optimization problems. An example of a swarm algorithm is the Particle Swarm Optimization (PSO) algorithm. This optimization method, which uses the social behavior of one group of animals, was created by Kennedy and Eberhart in 1995 [1]. PSO can be used to find optimal values in optimization problems where it is required to find the minimum or maximum of some objective function. Examples include optimizing model parameters, tuning neural networks [2, 3], data clustering [4], solving scheduling optimization problems, and others.

The Jaya algorithm (Jaya is a Sanskrit word meaning “Victory”) is an optimization algorithm proposed by Rao in 2016 for solving optimization problems without constraints. It is based on the idea of improving the current best solution by comparing and updating all solutions in the population [5]. Jaya is a simple and easy to implement

optimization algorithm that shows good performance on some classic optimization problems [6]. However, it may be less efficient on complex problems with large search spaces or with non-linear and multiextremal functions. The choice of optimization algorithm depends on the characteristics of a particular problem, however, the existing heuristic algorithm can be hybridized with other heuristic algorithms in order to improve their performance and ability to solve complex optimization problems. There are several approaches to creating hybrid heuristic algorithms [7–12].

In this paper, a hybrid PSO-Jaya algorithm was developed on the basis of the pre-processor/postprocessor principle without changing the overall operation of the PSO and Jaya algorithms [7]. The PSO algorithm can achieve successful results in almost all real-world problems. However, a solution is needed to reduce the PSO algorithm capture probability to a global minimum. The central idea of the PSO-Jaya hybrid algorithm is to use PSO to explore the solution space globally and Jaya to search for the found solutions locally. This allows the algorithm to combine an efficient study of large areas of the solution space with a more accurate refinement of the solutions found in the vicinity of local optima.

2 Hybrid Algorithm PSO-Jaya

2.1 Algorithm PSO

The PSO method aims to iteratively optimize the problem, starting with a set or population of possible solutions, called a swarm of particles, in which each particle knows the global best position in the swarm, as well as its individual best position found so far in the process of searching in space for a solution to the problem.

At the beginning, a swarm of particles x_i is initialized from N D-dimensional vectors with real values, and the particle velocity v_i of each particle is similarly generated. After the swarm and velocity are initialized, the value of the objective function is calculated for each particle and the initial individual best position $f(p_{best_i}^0)$ is calculated, together with the initial global (or neighborhood) best position $f(g_{best}^0)$. In the next step, each particle updates its speed and position based on its current position, speed, and experience gained from the best solution in the population using the formulas

$$x_{i,j}^{t+1} = x_{i,j}^t + v_{i,j}^{t+1}, \quad (1)$$

$$v_{i,j}^{t+1} = \omega v_{i,j}^t + c_1 r_{1,i,j}^{t+1} (p_{best_{i,j}}^t - x_{i,j}^t) + c_2 r_{2,i,j}^{t+1} (g_{best,j}^t - x_{i,j}^t), \quad (2)$$

where $v_{i,j}^0 = 0$ and $v_{i,j}^{t+1}$ velocity vector of the i -th particle at iteration $t + 1$ in the j dimension, $p_{best_{i,j}}^t$ —best individual particle position encountered by the i -th particle (at iteration t) in the j dimension, $g_{best,j}^t$ the best global position among all particles in the swarm, $r_{1,i,j}^{t+1}, r_{2,i,j}^{t+1}$ —random variables taking values at the current iteration $t + 1$ from the range $[0, 1]$, ω (inertial weight), c_1, c_2 —algorithm parameters.

2.2 Algorithm Jaya

The central idea behind Jaya algorithm is that every solution in a population should strive to be better, and to do this it can be updated based on the best solutions found in the population. This allows the algorithm to explore the solution space and find the best solutions as it goes through iterations. At the beginning, a population of solutions x_i is generated, at the next stage, the values of the objective function $f(x)$ are calculated. Among all the obtained candidate solutions, the particle x_{best} with the best position is selected and gets the best value f_{best} , and the worst candidate x_{worst} gets the worst value f_{worst} . Each solution $f(x_{kj}^i)$ (x_{kj}^i – k-th particle with dimension j at the i-th iteration) in the population is compared with the current best solution $f(x_{bestj}^i)$ (x_{bestj}^i – the best particle with the j-th variable at the i-th iteration). If $f(x_{kj}^i) < f(x_{bestj}^i)$ ($f(x_{kj}^i) > f(x_{bestj}^i)$), then the current position of the particle is updated by the formula:

$$x_{k,j}^i = x_{k,j}^i + r_{1,j}^i (x_{bestj}^i - |x_{k,j}^i|) - r_{2,j}^i (x_{worstj}^i - |x_{k,j}^i|), \quad (3)$$

where $r_{1,j}^i, r_{2,j}^i$ – random variables that take values at the current iteration i from the range [0, 1], x_{bestj}^i – j variable value for the best candidate, and x_{worstj}^i – j variable value for the worst candidate; $x_{k,j}^i$ – updated value $x_{k,j}^i$.

2.3 Hybrid Algorithm PSO-Jaya

The PSO-Jaya hybrid algorithm is a combination of two optimization algorithms: PSO and Jaya. This hybrid approach combines the benefits of both methods to improve performance and the ability to find optimal solutions to complex optimization problems. At the initial iterations, the PSO particles are widely distributed in the space solution, but, as a rule, they are far from the solution of the problem. At the final iterations, the particles are concentrated in the vicinity of the found extremum, but this extremum is not guaranteed to be global. Therefore, as a condition for switching between algorithms, it is advisable to take the beginning of stagnation.

The general principle of operation of the PSO-Jaya hybrid algorithm:

1. Initialization: an initial population of particles for PSO and Jaya is created in accordance with the PSO section.
2. PSO - phase: PSO particles iteratively update their position and velocity as they explore the solution space of formulas (1) and (2) respectively. A cost function is calculated for each particle, and the best solutions found by each particle and in the entire population are updated (PSO section).
3. Jaya-phase: after completion of the PSO-phase, the stagnation condition is checked $f(g_{best_{i+1}}) \geq f(g_{best_i})$, if there is stagnation, then the updated solutions from the PSO are transferred to Jaya to search for another optimum. Using the Jaya algorithm, solutions are iteratively updated according to formula (3) to improve their quality (Section Jaya). After Jaya completes, particle positions in PSO are updated.
4. Stop criteria check: the algorithm continues to execute PSO phases and Jaya phases until a given stopping criterion is reached, such as the maximum number of iterations or the required solution accuracy.

3 Test Function Optimization

Optimization of test functions using heuristic algorithms is one of the popular approaches for evaluating the performance of an algorithm and comparing it with other optimization methods.

To test the algorithm used the functions of Ackley (F1), function “eggholder” (F2), Holder (F3) and Matyas (F4) [7, 13, 14]. Each function has its own form and features, which allows you to evaluate the performance of the algorithm in different situations. The swarm size (number of particles) for the PSO, Jaya, PSO-Jaya algorithms are the same $N = 10$, the maximum number of iterations is $T = 200$, the parameters of the PSO, PSO-Jaya algorithms are the inertial weight, the coefficients of the cognitive and social components, respectively, are equal: $\omega = 0.72984$, $c_1 = c_2 = 2.05$, unlimited speed. The position and velocity of each particle are initialized randomly within the scope of the given function. In test calculations, the PSO, Jaya, PSO-Jaya algorithms are compared based on the average, median, standard deviation, and best of the best particle position over 50 independent runs. Table 1 illustrates the results of experiments carried out with the benchmark functions. The best results are in bold. From the data in Table 1, it follows that the PSO-Jaya hybrid algorithm has the best results for all functions than other algorithms. Jaya, PSO-Jaya find global minima for function F1. For F3 PSO finds the best position is close to the global minimum, for function F2 all three algorithms, but PSO-Jaya shows the best results in terms of average and standard deviation, which indicates a more stable operation of the algorithm.

Table 1. Statistical error analysis $|f_{min} - f_{best}^*|$ for benchmark functions, over 50 independent starts of 200 iterations, swarm size = 10

Nº	Algorithm	Average	Median	Standard deviation	“Best” err
F1	PSO	0.0011	2.6076e−05	0.0037	3.6079e−08
	Jaya	7.3295e−11	0	5.1235e−10	0
	PSO_Jaya	0	0	0	0
F2	PSO	109.99	101.0395	102.1635	3.7279e−05
	Jaya	615.4141	713.7134	318.7311	3.7279e−05
	PSO_Jaya	27.3561	3.7279e−05	39.9027	3.7279e−05
F3	PSO	1.5627	1.1878	1.398	7.2712e−08
	Jaya	57.1118	17.4755	279.8073	4.0683
	PSO_Jaya	0.8433	0.0105	1.3751	2.5679e−06
F4	PSO	8.1189e−07	2.8283e−11	4.9419e−06	1.5782e−17
	Jaya	5.9144e−07	1.0092e−42	4.1401e−06	7.6506e−50
	PSO_Jaya	5.94e−34	1.5434e−42	4.1548e−33	3.9564e−51

Figure 1 show that PSO is more efficient for functions F2, F3, and Jaya for F1, F4, but the PSO-Jaya hybrid outperforms both algorithms on all test functions in terms

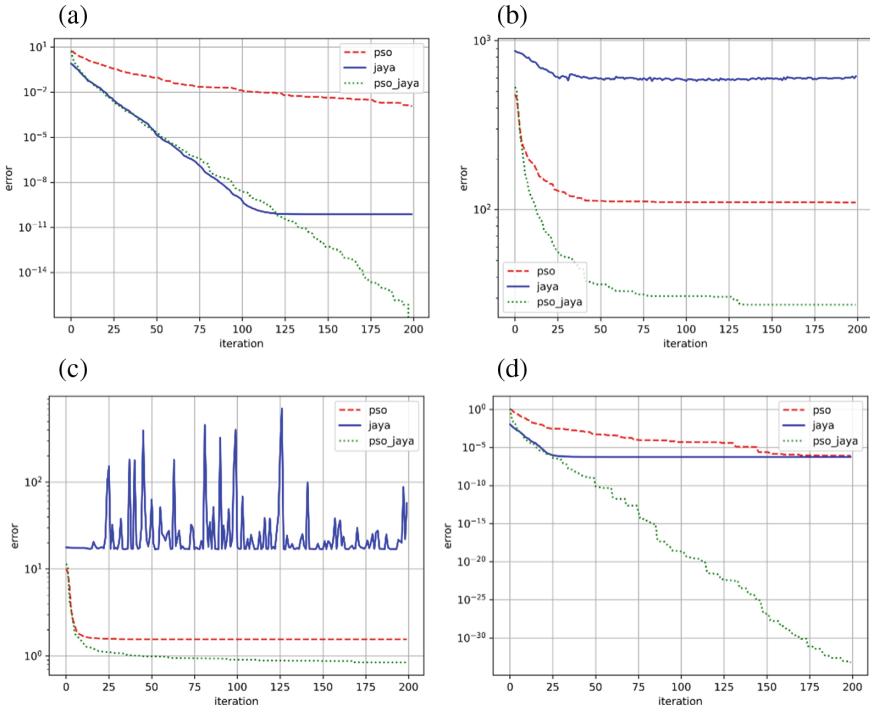


Fig. 1. Average error over 50 independent starts: a) Ackley’s function; b) function “eggholder”; c) holder’s function; d) booth’s function.

of convergence and stability, which indicates a successful hybridization strategy two algorithms.

4 Neural Network Training

There are a number of methods for using heuristic algorithms to train a neural network [15–17]. In this article, heuristic algorithms are used to find the combination of weights and biases that provides the minimum error for ANN. To design an ANN, the basic elements must be defined: first, a loss function using the ANN error must be defined to evaluate the fitness of agents in the ANN with the learning algorithms PSO, Jaya and PSO-Jaya; secondly, it is necessary to define a strategy for coding weights and biases for each ANN with appropriate algorithms.

1. *Loss function (cost function).* The ANN architecture in this article has the form - one input, one hidden and one output layers, where the number of neurons in the input layer is n , on the hidden layer - h , and on the output layer - m . The activation function is a sigmoid. After calculating the neurons in the output layer, the network error (cost function) will look like

$$E = \frac{1}{q} \sum_{i=1}^m (\widehat{y}_i^k - y_i^k)^2,$$

where q – number of instances in the training set, \widehat{y}_i^k – network output for k -th instance, y_i^k – label for the k -th instance in the training set.

2. *Encoding strategy.* Once the cost function has been determined, the next step is to choose an encoding strategy to represent the ANN weights and biases for each particle in PSO, Jaya and PSO-Jaya. There are three methods for encoding and representing FNN weights and biases for each agent in heuristic algorithms. These are methods of vector, matrix and binary coding [18]. In this article, a vector encoding strategy was used. For the involved ANN, as mentioned above, each particle represents all the weights and biases of the neural network structure. When computing the output of the ANN, the particles are again decoded into a weight matrix.

4.1 ANN Training Results

This section solves a classification problem for two datasets, Iris and Breast cancer Wisconsin, to compare the abilities of PSO, Jaya and PSO-Jaya in ANN training. The Iris classification task is widely used for neural network training. The Iris dataset contains 150 samples (112 samples from the training set, 38 samples from the test set), which are divided into three classes. All samples have four features. The Breast cancer Wisconsin data set consists of 569 samples (426 samples from the training set, 143 from the test set) and 2 classes. This problem refers to binary classification, i.e. according to the 30 features available for each sample, it is necessary to determine a malignant or benign tumor. The neural network architecture is 4–8–3 (4 inputs, 8 neurons in the hidden layer, 3 output) for solving the Iris classification problem, and 30–15–2 for Breast cancer Wisconsin.

These problems assume that each particle is initialized randomly in the range $[-1, 1]$. The algorithms have the following parameters: for PSO, PSO-Jaya $\omega = 1$, $c_1 = 0.5$, $c_2 = 1.5$, $v_{max} = 0.1$. The swarm size is the same for both datasets and for the three algorithms $M = 40$.

To evaluate the efficiency of the classifier, the accuracy metrics and the F1 score.

The PSO, Jaya, and PSO-Jaya algorithms are compared based on the average, median, standard deviation, and best of the Mean Square Error (MSE) for all training sets Iris and Breast cancer Wisconsin over 30 independent runs. The criterion for finishing the training process is to complete the maximum number of iterations (in this case equal to 500). The experimental results for this problem are shown in Tables 2 and 3. The best results are indicated in bold type. As can be seen from the Table 2, on all datasets, the PSO-Jaya hybrid algorithm outperformed the PSO algorithms, Jaya achieving the best classification accuracy on the training and test sets. Table 3 shows that the hybrid classifier performs better than other classifiers with PSO, Jaya algorithms based on the average, median, standard deviation and best MSE. These statistics show that the PSO-Jaya classifier has the best ability to avoid local minima. Also, for the minimum MSE, the hybrid algorithm has better results, which indicates a greater accuracy of the classifier than classifiers with PSO and Jaya algorithms.

Figure 2 show the convergence curves of the iris and Breast cancer Wisconsin data classifiers with PSO, Jaya, and PSO-Jaya based on the average MSE values for training sets over 30 independent runs. These curves confirm that the PSO-Jaya hybrid algorithm gives the best convergence rate for both classifiers.

Table 2. Average of accuracy over 30 independent starts

Nº	Algorithm	Training set (%)	Test set (%)	F1-measure
1	PSO	98.51	98.68	0.987
	Jaya	96.87	97.72	0.977
	PSO-Jaya	98.82	99.12	0.991
2	PSO	98.23	98.48	0.985
	Jaya	94.61	94.43	0.944
	PSO-Jaya	98.68	98.96	0.989

Table 3. Average, median, standard deviation and best MSE

Nº	Algorithm	Average	Median	Standard deviation	“Best” MSE
1	PSO	0.0117	0.0118	0.00075	0.0098
	Jaya	0.0259	0.0255	0.00073	0.0173
	PSO-Jaya	0.0084	0.0086	0.0015	0.0046
2	PSO	0.0604	0.0607	0.0044	0.0524
	Jaya	0.1172	0.1153	0.0111	0.0980
	PSO-Jaya	0.0508	0.0511	0.0027	0.047

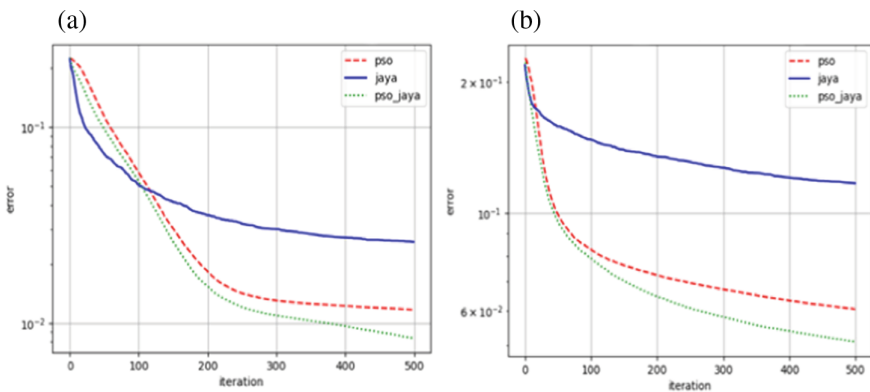


Fig. 2. Average of MSE: a) Iris; b) breast cancer Wisconsin

5 Conclusion

In this paper introduces and investigates a hybrid PSO-Jaya optimization algorithm based on two heuristic algorithms PSO and Jaya. Two problems: function optimization and training ANN for the classification problems Iris and breast cancer Wisconsin, are employed to evaluate the efficiencies of this new hybrid algorithm. The results are

compared with the PSO and Jaya algorithms. For all test cases, PSO-Jaya shows the best performance in terms of convergence rate and avoidance of local minima. Thus, the results show that PSO-Jaya reduces the probability of hitting local minima and increases the rate of convergence compared to other PSO and Jaya algorithms.

References

1. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: IEE International Conference on Neural Networks, pp. 1942–1948 (1995)
2. Eberhart, R.C., Hu, X.: Human tremor analysis using particle swarm optimization. In: Proceedings of the 1999 Congress on Evolutionary Computation-CEC99, vol. 3, pp. 1927–1930. IEEE (1999)
3. Hamada, M., Hassan, M.: Artificial neural networks and particle swarm optimization algorithms for preference prediction in multi-criteria recommender systems. *Informatics* **5**(2), 25 (2018)
4. Asma, A., Sadok, B.: PSO-based dynamic distributed algorithm for automatic task clustering in a robotic swarm. *Procedia Comput. Sci.* **159**, 1103–1112 (2019)
5. Rao, R.: Jaya: a simple and new optimization algorithm for solving constrained and unconstrained optimization problems. *Int. J. Ind. Eng. Comput.* **7**(1), 19–34 (2016)
6. Warid, W., Hizam, H., Mariun, N., Abdul-Wahab, N.I.: Optimal power flow using the Jaya algorithm. *Energies* **9**(9), 678 (2016)
7. Karpenko, A.P.: *Sovremennyye algoritmy poiskovoy optimizatsii. Algoritmy. vdokhnovlenyye prirodoy*, 2nd edn. BMSTU, Moscow (2017)
8. Garg, H.: A hybrid PSO-GA algorithm for constrained optimization problems. *Appl. Math. Comput.* **274**, 292–305 (2016)
9. Mirjalili, S., Hashim, S.Z.M.: A new hybrid PSOGSA algorithm for function optimization. In: Proceedings of ICCIA 2010–2010 International Conference on Computer and Information Application, pp. 374–377 (2010)
10. Şenel, F.A., et al.: A novel hybrid PSO–GWO algorithm for optimization problems. *Eng. Comput.* **35**, 1359–1373 (2019)
11. Zhou, Y., Shengyu, P.: A hybrid co-evolutionary particle swarm optimization algorithm for solving constrained engineering design problems. *J. Comput.* **5**(6), 965–972 (2010)
12. Korolev, S.A., Maykov, D.V.: Modifikatsiya algoritma roya chastits na osnove metoda analiza iyerarkhiy. *Vestn. VGU Ser. Sist. anal. inform. tekhnol.* **4**, 36–46 (2019)
13. Liang, J.J., Qu, B.Y., Suganthan, P.N., Hernández-Díaz, A.G.: Problem definitions and evaluation criteria for the CEC 2013 special session on real-parameter optimization. *Comput. Intell. Lab.* **2012**(34), 281–295 (2013)
14. Saymon, D.: *Algoritmy evolyutsionnoy optimizatsii*. DMK Press, Moscow (2020)
15. Mirjalili, S.A., Hashim, S.Z.M., Sardroudi, H.M.: Training feedforward neural networks using hybrid particle swarm optimization and gravitational search algorithm. *Appl. Math. Comput.* **218**(22), 11125–11137 (2012)
16. Junior, F.E.F., Yen, G.G.: Particle swarm optimization of deep neural networks architectures for image classification. *Swarm Evol. Comput.* **49**, 62–74 (2019)
17. Garro, B.A., Vázquez, R.A.: Designing artificial neural networks using particle swarm optimization algorithms. *Comput. Intell. Neurosci.* **2015**, 61 (2015)
18. Zhang, J.R., Zhang, J., Lok, T.M., Lyu, M.R.: A hybrid particle swarm optimization–back-propagation algorithm for feedforward neural network training. *Appl. Math. Comput.* **185**(2), 1026–1037 (2007)



Registrar: A Social Conversational Agent Based on Cognitive and Statistical Models for a Limited Paradigm

Dmitry Khabarov and Alexei V. Samsonovich^(✉) 

National Research Nuclear University MEPhI, Kashirskoe Shosse 31, Moscow 115409, Russian Federation
avsamsonovich@mephi.ru

Abstract. A virtual conversational agent is designed based on a cognitive model integrated with neural network model named BERT and large language model ChatGPT. The system was tested in a Turing-test-like experiment with human participants, using a limited paradigm of registration of a guest in a hotel. Performance of the agent on several scales matches human performance, while in empathy it showed a significantly higher score compared to humans. The narrowly designed prototype proves the concept and suggest future applications to general open-ended paradigms.

Keywords: LLM · Cognitive model · Intentionalities · Social intelligent agents

1 Introduction

The task of creating virtual agents spans a variety of domains, and with the advent of language models such as ChatGPT, it has become a more common task. However, even such a powerful tool still does not allow systems to reach the human level, and the interactive behavior generated by these models is easily distinguishable from real human behavior. Here we argue that in order to give an agent a social-emotional intelligence, ChatGPT must be used as a peripheral device within a more complex architecture. The main idea of the approach to creating a virtual agent is the creation of a two-level architecture using neural network models. At the lower level there is a large language model (LLM) with a prepared prompt, which describes in detail the interaction paradigm and the actions of the agent in this paradigm.

In this work, ChatGPT-3.5-turbo is used as the language model. At the top level there are neural network models that analyze user's remarks and the dialogue history. The selected paradigm consists in registration of a visitor in a hotel. The agent serves as a hotel receptionist and is called "the registrar".

2 Materials and Methods

2.1 System Design

Our system design includes two levels. At the top level there are two neural networks models: the classifier of semantic categories (such as intensions, sentiments and tonalities [1, 2]) in dialogue remarks and the classifier of actions (scenarios: text instructions about a possible action of the registrar).

In the first classifier, the following categories of statements were selected and characterized for recognition, which were characterized by the following labels: doubt, trust, support, gratitude, dissatisfaction, disappointment, agreement, refusal, indifference, satisfaction, isolation. The requirements and semantic category descriptions in this paradigm are given in Table 1. These categories were formed by taking into account the criteria for assessing the agent, categories that can be further processed at the upper level and also characterize semantic proximities of speech acts.

The general task of the component is to find the manifestation of selected categories in the visitor's speech, which can appear at any stage of registration. The system receives a dialogue replica as input, either initially created in text format, or recorded in audio format and later translated into text using speech recognition, after which it passes through the classifier and receives the output categories that can be traced in this phrase, or does not output anything if there are none.

The modified architecture of the semantic categories classifier is presented in Fig. 1, where BERT is an encoder architecture that returns word embeddings [3, 4], which then go into the Dropout layer. This layer is one of the regularization methods and is needed to prevent model overtraining [5].

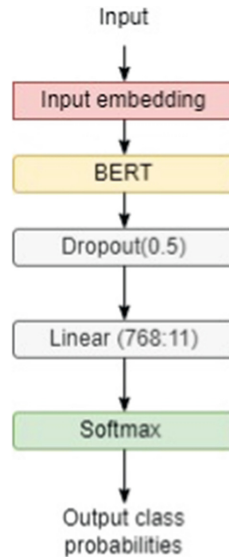


Fig. 1. Classifier architecture.

Table 1. Description of characteristics recognized by the model.

Category label	Advanced patterns	Examples
Doubt	Doubt, mistrust, suspicion, bewilderment, skepticism, uncertainty, hesitation	1. Questions the situation/phrase. 2. Refuses to believe what is happening/someone's words
Confidence	Trust, faith, conviction (in the correctness/of someone), commission	1. You rely on the other person's choices. 2. You trust the other person and believe that their choice will be the best
Support	Moral support, inspiration, optimism, motivation	1. Providing moral support to a person. 2. Show the person that the current situation does not bother you
Gratitude	Gratitude, compliment, reward	1. A simple expression of gratitude. 2. Expressing gratitude to a person or his quality, trait
Discontent	Indignation, irritation, anger	1. Reaction to the actions of the interlocutor. 2. Reaction to the current situation
Disappointment	Disappointment, despondency, frustration, grief, regret, annoyed state	1. Show that you are upset by someone's actions. 2. Express your feelings
Agreement	Consent, acceptance, conditional agreements, contract	1. Acceptance of a situation where you have to agree no matter what. 2. Complete agreement, when the expected coincides with the actual situation
Refusal	Refusal, renunciation, denial	1. Disagreement with the situation - refusal to make a decision, renunciation of words or actions. 2. Expressing disagreement with the interlocutor related to his actions, decisions, etc.
Indifference	Indifference, distance	1. Expression of indifference when choosing something. 2. Indifference to anything or anyone. 3. The desire to achieve what you want with complete indifference to methods or quality
Satisfaction	Pleasure, joy, delight, enjoyment	1. When you have achieved the desired result (with varying quality). 2. You liked something in the actions of your interlocutor

(continued)

Table 1. (continued)

Category label	Advanced patterns	Examples
Isolation	Lack of communication, coldness, unfriendliness, asociality	1. Short cold replies to a remark of any length

2.2 Datasets

The dataset used in this study was partially generated using a chat bot based on the GPT-3.5 language model. A small part of the dataset was created manually. The data is in the format of small conversational phrases that meet the following requirements, formulated to achieve the maximum balance of the dataset [6, 7]:

1. Phrases can be anything according to the purpose of the statement: interrogative, declarative and exclamatory with equal probability.
2. Phrases must have different emotional connotations: they can either have different forms of speech to express emotional assessments - interjections, special words, etc., or have none of the above.
3. Phrases should be, just like an address to the interlocutor, but not have an addressee, that is, they should be spoken alone (expression of emotions, elements of a monologue).
4. Phrases should vary in length, from a couple of words to a full sentence.

The results obtained for the Sberbank-AI/ruBert-base model are shown in Table 2. Here results for a pre-trained model from DeepPavlov was also added for comparison.

Table 2. Comparison of the main metrics of two trained BERT models.

Metrics	Sberbank-AI	DeepPavlov
Accuracy	0.8672	0.8495
F1 (micro)	0.8672	0.8495
F1 (macro)	0.8532	0.8351
F1 (weighted)	0.8602	0.8409

Thus, for example, for the phrase: “Do you really have no free rooms? This is sad”, the category “disappointment” will be determined with a key metric - probability, which will take the value 0.8.

The following architecture was chosen as the architecture for the second classifier: multilayer perceptron (MLP). It contains two hidden layers and allows you to select several of the pre-prepared actions in all sorts of situations. The classifier takes as input the current semantic category and ratings of two scales: which were named “trust” and “satisfaction”. The “trust” scale includes the following signs: doubt, trust, support. To the

satisfaction scale: dissatisfaction, disappointment, satisfaction. The scales are calculated according to formula (1), based on [8].

$$M_i = \frac{M_{i-1} + k * p}{t} \quad (1)$$

In this formula M_{i-1} - the scale assessment at the previous step, p - the probability of a category appearing in the last comment of the visitor, k - the coefficient of influence of this category on the scale, t - a temporary reduction factor introduced so that the scales fade over time and take a neutral position if certain categories do not appear. Using the last formula, after each phrase, both scales are calculated, and then fed to the input of the neural network model in format (2),

$$input = [x^1, \dots, x^i, \dots, x^{12}, D, V] \quad (2)$$

where $x^1, \dots, x^i, \dots, x^{12}$ category membership labels, D – trust metric, V – satisfaction metric. The output of the model is a text template in the format “Do X”, where X is a set of actions. This information is fed to the input of the language model, which changes the initially generated response.

2.3 Network Training

To train the neural network, dialogues were generated, from which scores on the “trust” and “satisfaction” scales were extracted, as well as a selected category of interpersonal relationships. For each of them, one of the actions was matched (only part of them is indicated here):

1. Offer additional services.
2. Find out if anything else is needed.
3. Promise to find the best option and do it.
4. Find out client preferences.
5. Find out the reason for % (% is replaced by the corresponding category of interpersonal relationships).
6. Offer options that, in your opinion, will improve the situation.
7. Thanks for %.
8. Recheck the result and report it.
9. Convince the client that everything is okay.

The action was put in accordance with the resulting assessments, as the most appropriate in the current situation.

The following architecture was used as a neural network: input layer (32 neurons), two hidden layers (64 neurons each), output layer contains 12 neurons. The classification accuracy was 92.56% on the test sample.

For example, if the input data is “doubt, 0.87, 0.01,” the output category is the instruction “Convince the client that you are right, offer alternative options.” This option is supplied to the input of the language model in the form of the following instruction: “Add the following action towards the client to your response, if it is appropriate in the

context of the dialogue.” The phrase “if it is appropriate” corrects a possible inaccuracy in the event of an inaccurate selection of actions, however, even if the action was selected appropriately, additional context will be inserted into the topic of the dialogue and will not cause discomfort for the user during communication [9]. With the use of this modification, the choice of further action of the registrar becomes justified.

The recorder model can be outlined as follows. The input phrase is entered by the user and recorded in the dialogue history. Also, this phrase first gets into the classifier, which produces the category with the highest probability and this probability.

Next, this data goes into the script block, where the scale ratings are first recalculated, and then an action is selected in the form of a text instruction and submitted, along with the dialogue history, to the input of the language model, which generates a response. The response is formatted in case of inappropriate structure and is given to the user and is also entered into the dialogue history.

3 Experimental Procedure, Results and Analysis

To test the agent, 12 subjects were invited, who were asked to create a dialogue with the virtual agent. Their goal was to book a hotel room without an appointment. To begin with, they had to say hello, explain the situation, and then choose a room to check in, simultaneously telling the agent about their preferences.

The artificial social agent (ASA) questionnaire was used for assessment [10]. This questionnaire is a tool for assessing human interaction with an artificial social agent. It was created as a result of years of research involving more than 100 Artificial Virtual Agent (IVA) researchers from around the world as part of the Open Source Artificial Social Agent Tools Evaluation Working Group. The questionnaire is presented in the form of a short version of 24 questions.

At the end of the interaction, users filled out a questionnaire with some of 24 statements, which they had to rate on a scale from -3 to 3 .

1. Criterion: “Agent’s communication skills.” Statement: “The agent has good communication skills.”
2. Criterion: “Persistence of intentions.” Statement: “The agent was persistent in his actions and controlled the dialogue himself.”
3. Criterion: “User trust.” Statement: “I trust the agent.”
4. Criterion: “Social acceptability.” Statement: “The agent is socially acceptable.”
5. Criterion: “Empathy.” Statement: “The agent exhibits empathic qualities.”
6. Criterion: “Ability to evoke emotions.” Statement: “The emotions I experience during an interaction are caused by the agent.”
7. Criterion: “Having individuality.” Statement: “The agent has a distinctive character.”

To compare the results, the same surveys were compiled after the user interacted with a human confederate (another average person, who doesn’t know how the agent work), and the user did not know in advance who he was in contact with. The confederate playing the role of the agent became familiar with the registrar’s operating procedures and himself supervised the check-in process.

The results are presented in Fig. 2. Statistical analysis was carried out comparing 7 hypotheses using the Mann-Whitney U test, taking into account the Bonferroni correction for multiple hypothesis testing.

The diagram shows the mean values with the standard error indicated by whiskers. The statistical significance was only found in comparing the empathic qualities of a human and an agent. The rest of scales showed no significant differences.

On the scale of empathy, the agent showed a very good score, much higher than that of an average human. This is the first known to us experimental result of this sort.

According to criteria such as user trust and persistence of intentions, a real person is in the lead. This can be explained by the fact that in some situations the virtual agent, after its short phrase, waits for a response from the user, while a person, having performed the target action, immediately speaks about the results of the process. In a conversation with an agent, sometimes you have to ask again about some things and take the initiative in the conversation by asking him to take a certain action.

According to social criteria, the agent is more in the lead, this is due to the prompt for chat-gpt and the work of top-level components - chat gpt is more sensitive to human emotions and tries to respond to them in a more polite and courteous manner, and the action selection component allows you to behave politely in practice in any situation during registration.

As already mentioned, the agent has an absolute advantage in empathy. Since the person invited to act as an agent was informed about the basic principles of working with clients and trained in polite and courteous communication, we can say that the agent's empathy was achieved precisely by the approach to creating a two-tier architecture.

On the other hand, only the criterion responsible for empathy has statistical significance, so despite the presence of explainable patterns in other criteria, they must be considered equal. This means that an agent was created that is perceived on the same level as a person according to socio-emotional assessment criteria and in some way (in empathy) superior to the average person, which was the basic goal of this work.

4 Concluding Remarks

In this study we observed the ability of a conversational agent to score on social scales at a human level, including communicability, intentionality, trust, social acceptability, empathy, ability to induce emotions, and individuality, in the limited social interaction paradigm. Moreover, on the scale of empathy the agent exceeded the average human participant, which can be attributed to the limitations of the paradigm and to the low empathetic ability of participating students.

The further prospect of developing a social agent, first of all, is associated with moving away from narrowly focused paradigms, for example, "hotel receptionist", since a virtual agent must be socially acceptable in any situation and show its qualities in any paradigm. Models of the human self [11] should be tested in open-ended scenarios. The approach outlined above allows us to develop universal agents that can be used both in various practical fields, such as intelligent tutoring [12], and in ordinary dialogues. First of all, the list of semantic categories (or intensions) should be expanded beyond the 3-D semantic map model [13, 14]. So, for example, in the expanded list such intensions

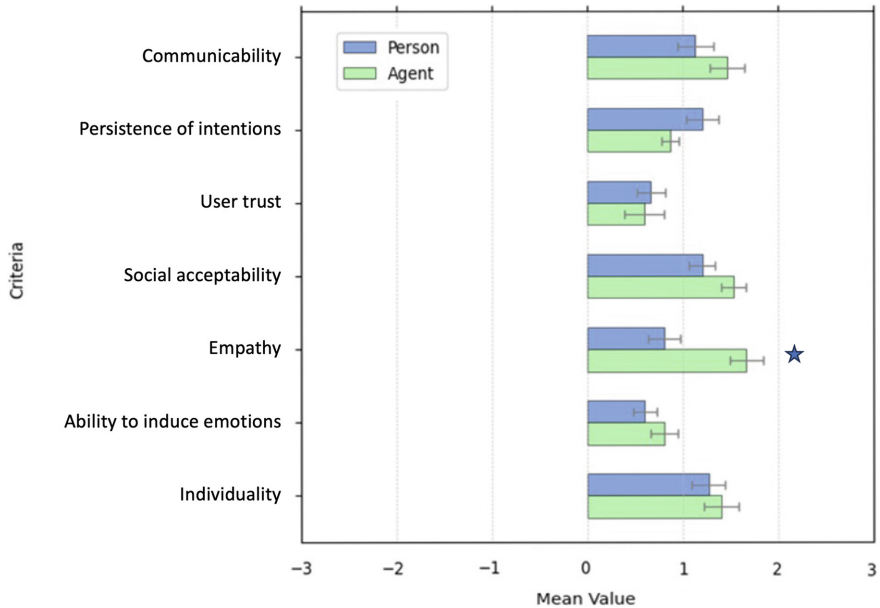


Fig. 2. Agent testing results.

appear as: a challenge to revelation, a transition to a frank conversation, a transition to a formal tone, flattery and others. Such categories expand the set of those already proposed and allow the agent to perceive a larger number of different social situations and learn to behave acceptably in them. The next step is to train the neural network to identify intensions in human speech, but this may cause inaccuracies in the agent's perception of reality: in some situations, one phrase may contain several intensions at once. Therefore, it is necessary to either add appropriate add-ons before training the neural network on the generated data, or use other means to mark phrases in an N -dimensional semantic space, where n is the number of intensions. To control the behavior of the agent, one can either compress the space to 3–4 dimensions, or train the model to select the correct action based on the initial marking, additionally combining the calculation of estimates with a cognitive model (eBICA [8]). The principle of working with the language model, in turn, will not change.

Acknowledgments. This work was supported by the Russian Science Foundation Grant #22-11-00213, <https://rscf.ru/en/project/22-11-00213/>.








References

1. Russel, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach, 4th Global Pearson Education Limited, London (2013)
2. Oishi, E.: Austin's speech act theory and the speech situation. *Eser. Filos.* **1**, 1–14 (2006)

3. Acheampong, F.A., Nunoo-Mensah, H., Chen, W.: Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artif. Intell. Rev.* **2021**, 1–41 (2021)
4. Garrido-Merchán, E.C., et al.: Comparing BERT against traditional machine learning models in text classification. *J. Comput. Cogn. Eng.* (2022). <https://doi.org/10.47852/bonviewJCCE3202838>
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., et al. (eds.) *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Long Beach, CA (2017)
6. Zhang, H., Song, H., Li, S., Zhou, M., Song, D.: Survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv. J.* **37**(4), Article 111 (2023)
7. Dathathri, S., et al.: Plug and play language models: a simple approach to controlled text generation. [arXiv:1912.02164](https://arxiv.org/abs/1912.02164) (2019)
8. Samsonovich, A.V.: Socially emotional brain-inspired cognitive architecture framework for artificial intelligence. *Cogn. Syst. Res.* **60**, 57–76 (2020)
9. Li, J., et al.: Pretrained language models for text generation: a survey. [arXiv:2105.10311](https://arxiv.org/abs/2105.10311) (2021)
10. Fitrianie, S., et al.: The 19 unifying questionnaire constructs of artificial social agents: an IVA community analysis. In: *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pp. 1–8 (2020)
11. Samsonovich, A.V., Ascoli, G.A.: The conscious self: ontology, epistemology and the mirror quest. *Cortex* **41**(5), 621–636 (2005). [https://doi.org/10.1016/S0010-9452\(08\)70280-6](https://doi.org/10.1016/S0010-9452(08)70280-6)
12. Samsonovich, A.V., De Jong, K.A., Kitsantas, A., Peters, E.E., Dabbagh, N., Kalbfleisch, M.L.: Cognitive constructor: an intelligent tutoring system based on a biologically inspired cognitive architecture (BICA). *Front. Artif. Intell. Appl.* **171**(1), 311–325 (2008). ISSN: 09226389
13. Samsonovich, A.V., Ascoli, G.A.: Cognitive map dimensions of the human value system extracted from natural language. *Front. Artif. Intell. Appl.* **157**, 111–124 (2007). ISSN: 09226389
14. Samsonovich, A.V., Ascoli, G.A.: Augmenting weak semantic cognitive maps with an “abstractness” dimension. *Comput. Intell. Neurosci.* **2013**, Article number 308176 (2013). <https://doi.org/10.1155/2013/308176>



Dynamic Model of Semantic Information Signal Processing

Mohiniso Khidirova¹ , Kamaliddin Abdivakhidov¹ , Pavel Bylevsky²,
Alexey Osipov³ , Ekaterina Pleshakova^{3,4} , Victor Radygin⁴ ,
Dmitry Kupriyanov⁴ , and Mikhail Ivanov⁵ 

¹ Yeoju Technical Institute in Tashkent, Tashkent, Uzbekistan

² Moscow State Linguistic University, Moscow, Russian Federation

³ Moscow Technical University of Communications and Informatics, Moscow, Russian Federation

⁴ National Research Nuclear University “MEPHI”, Moscow, Russian Federation
DYKupriyanov@mephi.ru

⁵ Financial University under the Government of the Russian Federation, Moscow, Russian Federation

Abstract. The article deals with the mathematical modeling of the regularism of thought processes based on the concept of ORASTA, consisting of a controller operator (OR) and an active contextual environment of regulation with a temporary constant (active system with time average – ASTA). Associative regulatory mechanisms of the thought process are considered on the basis of the developed systems of functional differential equations with delay, taking into account time relationships, cooperative processes and feedbacks in the regulation system. The existence of trivial and non-trivial equilibrium positions at certain ranges of parameters is revealed; stability of the rest regime in the absence of non-trivial equilibrium positions; the existence of stable and complex oscillatory modes of operation, which are important for analyzing the nature of the occurrence, development and deformation of the oscillatory background during the implementation of information processes in the activity of systems capable of thinking independently in the case of various external and internal signals. The developed equations of the regulatory mechanisms of paired associative phenomena in consciousness, taking into account the coupling of cognitive self-organizing processes, can be used for safe communication methods.

Keywords: Mathematical modeling · Nonlinear dynamics · Delayed-argument equations · Qualitative analysis · Regularization · Functional differential equations · Cognitive activity

1 Introduction

With the development of artificial intelligence technologies and the need to create applications that can solve large-scale, complex tasks without additional training and are able to think independently, interest in the problems of human thinking and intelligence is

increasing. To date, most theories of information transmission suffer from the problem of “lack of meaning”, that is, only the quantity and value of information are considered [1, 2], as well as the cryptographic strength of algorithms. Attempts are being made to simulate quantum consciousness, for example, the dynamics of Bayesian network learning is described by quantum mechanical equations, which leads to some effective quantization [3, 4]. Moreover, the properties of hysteresis (lag) are used to construct chaotic artificial neural networks with a complex activation function for self-learning and generalization [5–7]. But despite large successes, artificial intelligence systems with generalizing and “thinking” abilities have not yet been created, especially in a rapidly changing information environment [8]. Researchers suggest that cognition is a continuous chaos of a single process of emotional-logical thinking, which is conditionally divided into logical (rational) and emotional (irrational) [9, 10].

Modern developments in the study of the “problem of consciousness” within the framework of the concept of “strong artificial intelligence” are based on classical computational cognitive theory, connectionist approaches, information theory of consciousness, integrated information theory, computational theory, attention schema theory [11–13]. British physicist Roger Penrose in the 1990s, together with Stuart Hameroff, developed the “Orch OR” model of consciousness, according to which brain activity is considered as an essentially a quantum process [14]. At the same time, due to the effects of quantum gravity, there is a process of continuous “objective reduction” (objective reduction—OR) of the wave function of parts of the brain, which is caused by the divergence of the quantum states of space-time to the limit, after which they are reduced. The reduction process is described as “orchestrated” (orchestrated—Orch) selection of the appropriate state (the term orchestrated is used by the authors, since they believe that the collapse of the macroscopic entangled state in the microtubules of cellular organelles is to some extent controlled (conducted) by membrane proteins). V.D. Zakharov believes that the explanation of consciousness on the physical path is impossible, and Penrose’s results can be interpreted only on the path of apophaticism (“scientific ignorance” of N.I. Kuzansky) [15].

With the help of continuous wavelet study, methods of analysis and diagnostics of oscillatory neural network activity of the brain according to experimental electroencephalograms, researchers identify such oscillatory patterns as sleepy spindles, bursts of tetaactivity, K-complexes, peak wave discharges (PVR), characteristic precursors of epileptic discharges [16–18]. Researchers have revealed that nonlinear rhythmic patterns in the brain are different for different types of cognitive activity, and the characteristic rhythmic patterns are individual. Freeman, studying the mechanisms of assimilation of new smells, pays special attention to chaotic neural activity, hypothesizing that the chaotic behavior of the brain serves to learn new things in the surrounding world [18].

2 Methods

Researchers have revealed that nonlinear rhythmic patterns in the brain are different for different types of cognitive activity, and the characteristic rhythmic patterns are individual [19–21].

B.N. Khidirov developed methods of quantitative research of complex oscillatory regulatory systems, which allow considering a wide range of phenomena united by

the presence of a regulatory system, regulatory environment, competition, cooperation and combined feedback from a single position [22–24]. The concept of ORASTA was introduced, consisting of an oscillator regulator (OR) capable of receiving, processing, and transmitting signals of a certain nature, and an active medium with a time constant (active system with time average – ASTA), which allows for a feedback loop in the system for a finite time. One of the main ideas in the mathematical modeling of the regulatory mechanisms of consciousness is the central regulation of information flow based on multi-oscillatory ORASTA.

Consider the following statement of the problem of mathematical modeling of regulatory mechanisms for extracting meaning from the flow of information. Let's assume that in a limited volume (“artificial brain”) there are N interconnected elements – regulators capable of perceiving, processing, and synthesizing signals of a certain nature. And let the relationship between the regulators be carried out by means of these signals with the average time of passing the feedback loop h (i.e., the time elapsed from the moment the signals were formed to the moment their (or their products) were affected by the activity of the regulators). It is required to analyze the simplest patterns of occurrence, development, and deformation of the oscillatory background in the perception, processing, and synthesis of semantic information by these elements.

Let's consider one of the possible options for studying the regulatory mechanisms of consciousness within the framework of the concept of strong artificial intelligence. Let $X_i(t)$ – the dimensionless value characterizing the amount of synthesized signal corresponding to i -th of the element at the moment of time t ($1 \leq i, n \leq N$). A function expressing the rate of synthesis i -th signal (synthetic function S_i), it is formed from a stimulating (SC_i) and inhibitory (SI_i) function. With respect to SC_i basic subsystems, the following expression can be proposed

$$SC_i = \gamma_i X_1(t), X_2(t), \dots, X_n(t), \quad (1)$$

where n – number of elements in the basic subsystem ($1 \leq n \leq N$),

$$\gamma_i = \text{const} > 0. \quad (2)$$

Let δ_{ik} – a dimensionless quantity characterizing the binding rate constant of the products of the k -th element, inhibiting the functioning of the i -th element with the corresponding participation of the control of the synthesis of signals of the i -th element ($0 \leq \delta_{ik}, i, j = 1, 2, \dots, n$). Then the change SI_i depends on

$$\Omega_i = \sum_{k=1}^n \delta_{ik} X_k(t) \quad (3)$$

and for specific values Ω_i it matters $SI_i(\Omega_i)$. This function expresses the level of repression of the i -th element at a given state ($X_1(t), \dots, X_n(t)$) of the basic subsystem. Value changes Ω_i ($\Delta\Omega_i$) lead to certain changes ($\Delta SI_i(\Omega_i)$), they depend on the previous level of repression and on $\Delta\Omega_i$. You can accept

$$SI_i(\Omega_i + \Delta\Omega_i) = SI_i(\Omega_i) + (\Delta\Omega_i)SI_i \quad (4)$$

or going to the limit $\Delta\Omega_i \rightarrow 0$

$$\frac{d(SI/I_i)}{d\Omega_i} = SI/I_i \tag{5}$$

$$SI/I_i = A_i e^{\sum_{k=1}^n \delta_{ik} X_k(t)} \tag{6}$$

Vector $M_c(C_1, \dots, C_n)$, the values of the elements of which are calculated by formula

$$C_i = \int_0^\infty \dots \int_0^\infty A_i^n(S) \exp\left(-\sum_{j=1}^n \delta_{ik} S_j\right) dS_1 \dots dS_n - 1 \tag{7}$$

M_c expresses the relationship of the regulatory system with the external environment, since its value is determined by the specified specific values of the coefficients. In the case of $M_c = 0$, the system is in equilibrium with the external environment. It is possible to consider regulatory equations with lagging arguments, mixed equations, and pantograph-type equations.

1. Functional differential equations of regularism delay

$$\frac{dX_i(t)}{dt} = a_i \left(\prod_{k=1}^n X_k(t-h) \right) e^{-\sum_{k=1}^n \delta_{ik} X_k(t-h)} - b_i X_i(t) \tag{8}$$

2. Functional differential equations of regulation with delay and advance

$$\frac{dX_i(t)}{dt} = a_i \left(\prod_{k=1}^n X_k(t-h) X_k(t+h) \right) e^{-\sum_{k=1}^n \delta_{ik} X_k(t-h) X_k(t+h)} - b_i X_i(t) \tag{9}$$

3. Functional differential equations of regulation with compression and stretching

$$\frac{dX_i(t)}{dt} = a_i \left(\prod_{k=1}^n X_k(th) \right) e^{-\sum_{k=1}^n \delta_{ik} X_k(th)} - b_i X_i(t) \tag{10}$$

The concept of ORASTA made it possible to develop general basic equations (in the class of nonlinear functional differential equations) for the regulation of associative, interconnected and self-adjoint systems, taking into account the spatial separation of processes, cooperativeness, competition for signals and combined feedback.

The equations of the regulatory mechanisms of paired associative phenomena in consciousness, taking into account the conjugation of cognitive self-organizing processes, have the following form of functional differential equations with delay:

$$\begin{aligned} \theta_1 \frac{dX_1(t)}{dt} &= (a_1 X_1(t-h) + a_2 X_2(t-h) X_1(t-h)) e^{-X_1(t-1) - X_2(t-1)} - b_1 X_1(t) \\ \theta_1 \frac{dX_2(t)}{dt} &= (c_1 X_1(t-h) X_2(t-h) + c_2 X_2(t-h)) e^{-X_1(t-1) - X_2(t-1)} - b_2 X_2(t), \end{aligned} \tag{11}$$

where $X_1(t), X_2(t)$ – values characterize the activity of the i -th regulator capable of perceiving and generating associative signals ($i = 1, 2$); θ_1, θ_2 – average lifetime of the products of the activity of the i -th regulator (θ_i we can call the average lifetime of the products of the i -th regulator); h – the time required for feedback in the system; a_i, c_i, b_i – non-negative constants ($i = 1, 2$).

3 Analysis

Next, we will consider the associative regulatory mechanisms of the thought process, which are based on short-term regulatory connections. Considering that the appearance of some mental phenomenon in a person's consciousness leads to the emergence of another state, we can propose the following system of functional differential equations with a delay at $a_2 = c_1 = 0$

$$\begin{aligned}\frac{\theta_1}{h} \frac{dX_1(t)}{dt} &= a_1 X_1(t-1) e^{-X_1(t-1) - X_2(t-1)} - X_1(t) \\ \frac{\theta_2}{h} \frac{dX_2(t)}{dt} &= a_2 X_2(t-1) e^{-X_1(t-1) - X_2(t-1)} - X_2(t)\end{aligned}\quad (12)$$

Let us consider a qualitative analysis of a system of functional differential equations with delay (12). This system may have three equilibrium positions $(0, 0)$, $A(0, \ln a_2)$, $B(\ln a_1, 0)$. Due to the equality of (at least) one of the coordinates of the equilibrium position to zero, the characteristic equation of the linearized system (12) has the form:

$$\begin{aligned}\frac{\theta_1}{h} \frac{dZ_1(t)}{dt} &= (a_1 e^{-\alpha_1 - \alpha_2} - \alpha_1) Z_1(t-1) - \alpha_2 Z_2(t-1) - Z_1(t) \\ \frac{\theta_2}{h} \frac{dZ_2(t)}{dt} &= \alpha_2 Z_1(t-1) + (a_2 e^{-\alpha_1 - \alpha_2} - \alpha_2) Z_2(t-1) - Z_2(t)\end{aligned}\quad (13)$$

Let $\theta_1 = \theta_2 = h$. Consider the trivial equilibrium position $0 (0, 0)$, for which the characteristic equation has the form:

$$(\lambda + 1)e^\lambda - a_i = 0 \quad (14)$$

$$1 - a_i > 0; \quad (15)$$

$$-a_i < \eta \sin \eta - \cos \eta, \quad i = 1, 2; \quad (16)$$

where η – the root of the equation

$$\eta = \operatorname{tg} \eta, \quad 0 < \eta < \pi. \quad (17)$$

A trivial position has stability. The tabular solution (17) leads to $\eta = 2.02$; $\sin \eta = 0.9$; $\cos \eta = -0.43$. Consequently, the second inequality of the Hayes criterion is always fulfilled and the trivial equilibrium position is stable under

$$a_1 < 1, \quad a_2 < 1. \quad (18)$$

Let us proceed with the analysis of the stability of nontrivial equilibrium positions A and B. For the existence of these equilibrium positions, it is necessary

$$a_1 > 1, \quad a_2 > 1. \quad (19)$$

Let's say for simplicity $a_1 > a_2 > 1$. For A we have

$$(\lambda + 1)e^\lambda - \frac{a_1}{a_2} = 0; \tag{20}$$

$$(\lambda + 1)e^\lambda - 1 + \ln a_2 = 0.$$

The application of the Hayes conditions for the first equation shows the instability of the point A, due to the non-fulfillment of the second inequality

$$1 - \frac{a_1}{a_2} > 0 \tag{21}$$

for the accepted values a_1, a_2 . For point B, the Hayes condition leads to

$$\ln a_1 - 1 < 2,24. \tag{22}$$

Therefore, there are areas on the parametric portrait G, G1, G2, F1, F2, H1, H2, in the first of which there is only a stable trivial equilibrium position. In G1 (G2), there is one stable equilibrium position in F1 (F2), two unstable (trivial and A(B)), and one stable B(A). In the area of H1 (H2), there are three unstable equilibrium positions in the first quadrant of the phase plane (see Fig. 1).

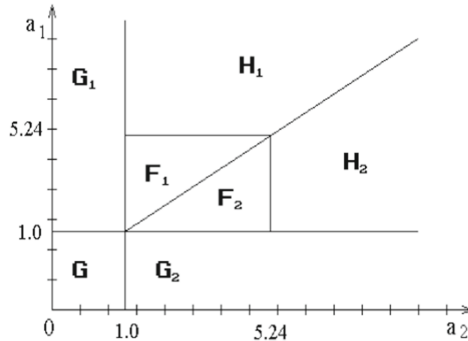


Fig. 1. Parametric portrait

The results of qualitative research allow us to use these models to study the regulatory mechanisms of consciousness in associative interactions between regulators and their general mutual expression. The dynamics of the functioning of linguistic consciousness in the light of models of verbal associative behavior by representatives of different cultures has not yet been studied in detail [20]. The results of these studies can be useful in creating a strong artificial intelligence.

When mathematically modeling the regulatory mechanisms of consciousness within the framework of the concept of strong artificial intelligence, it is useful to consider, along with the OR regulatory operators, the ASTA regulatory environment, since the identification of meaning from the flow of information, especially in conditions of ambiguity (for example, words, pictures, etc.), the concept of context is essential.

The results of the analysis of the nature of solution (2) show the possibility of the existence of zero and positive equilibrium positions, stable periodicity – self-oscillations, irregular oscillations – dynamic chaos. These properties of solutions, which are necessary conditions for the application of differential equations for the quantitative description of the regulatory mechanisms of consciousness, are due to the non-negativity and the limited number of PRODUCTS, the presence of rest states and the existence of a functionally active phase of the regulation of brain signals. One of the main factors in determining the true meaning of information is the context, that is, the regulatory environment of ASTA. In [25] it is noted that “the phenomenon of consciousness has a contextual nature, that is, it can be represented as a system of context.” Models of the regulatory mechanisms of consciousness, taking into account temporal relationships, make it possible to effectively investigate the mechanisms of thinking at various hierarchical levels of organization, which are carried out on the basis of oscillatory processes.

Computer studies allow us to quickly assess the general regularity, characteristic features and basic modes of decision behavior. They make it possible to obtain approximate solutions of nonlinear functional differential equations because of the regularity of thinking, to evaluate the behavior of irregular solutions and the level of their “randomness”, to analyze the regularity of the processes of extracting meaning from the information flow through “computational experiments”. The developed GIR software tool allows changing parameter values, detailing and temporary stopping for archiving visual material (see Fig. 2).

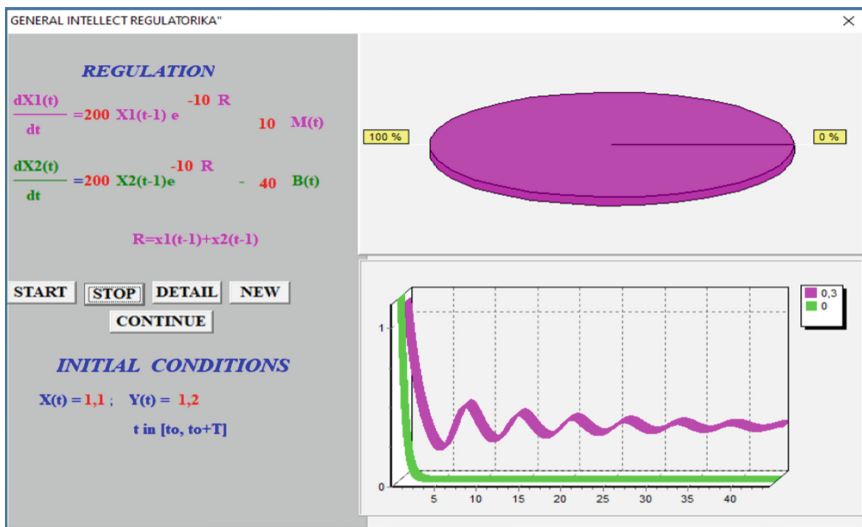


Fig. 2. Working status of the GIR display

The areas of the same type of behavior of various dynamic modes of bifurcation transitions depend on the main parameters of the system and the patterns of functioning of external and internal influences. The results allow us to quantify the parameter ranges necessary for an adequate comparison with experimental data.

4 Discussion

It should be noted some evolutionary aspects of the development of consciousness. Since associative regulatory systems of consciousness can have mild excitations of activity from a state of rest with the slightest changes in mc, it is evolutionarily more likely that associative systems of consciousness will first appear, initially it may even be in the form of simple self-oscillating systems in a separate elementary consciousness. Further development of consciousness with the accumulation of sufficient resources could create favorable conditions for the emergence of interconnected systems capable of evolutionary development. Unlike associative systems, the competition of interconnected systems for resource supply can lead to the elimination of relatively “weak” systems and the survival of more progressive systems with a sufficiently large mental resource supply. The latter have specific features of development and opportunities for progressive evolution.

Thus, for modeling the regulatory mechanisms of thinking processes, the most justified and relevant is the construction of systems of functional differential equations with delay, since they allow taking into account time relationships in the regulatory system. The developed equations of the regulatory mechanisms of paired associative phenomena in consciousness, taking into account the coupling of cognitive self-organizing processes, can be used to build chaotic artificial neural networks with a complex activation function with self-learning and generalization. The created neural network needs not to be trained, but to be formed, that is, to form the skill of hierarchically operating with images and concepts with central regulation of information flow based on multi-oscillatory ORASTA. Moreover, the regulatory approach can be useful for the development of cryptographic algorithms for transmitting information in a variety of contexts.

References

1. Strickland, E.: IBM Watson, heal thyself: how IBM overpromised and underdelivered on AI health care. *IEEE Spectr.* **56**(4), 24–31 (2019)
2. Bilyalova, A., Salimova, D., Zelenina, T.: Digital transformation in education. In: *Integrated Science in Digital Age. International Conference on Integrated Science*, pp. 265–276. Springer, Cham (2019)
3. Morawietz, T., Artrith, N.: Machine learning-accelerated quantum mechanics-based atomistic simulations for industrial applications. *J. Comput.-Aided Mol. Des.* **35**(4), 557–586 (2021)
4. Fisher, M.: Quantum cognition: the possibility of processing with nuclear spins in the brain. *Ann. Phys.* **362**, 593–602 (2015)
5. Ivanyuk, V.: Forecasting of digital financial crimes in Russia based on machine learning methods. *J. Comput. Virol. Hack. Tech.* (2023)
6. Bakhshiev, A., Fomin, I., Gundelakh, F., Demcheva, A., Korsakov, A.: The architecture of a software platform for growing spiking neural networks simulator developing. *J. Phys. Conf. Ser.* **1679**(4), 042001 (2020)
7. Yerzncyk, B., Bychkova, S., Gataullin, T., Gataullin, S.: The sufficiency principle as the ideas quintessence of the club of Rome. *Montenegrin J. Econ.* **15**(1), 021–029 (2019)
8. Kapustnikov, A., Sysoeva, M., Sysoev, I.: Modeling spike-wave discharges in the brain with small neurooscillator networks. *Math. Biol. Bioinf.* **15**(2), 138–147 (2020)

9. Gataullin, T., Gataullin, S.: Endpoint functions: mathematical apparatus and economic applications. *Math. Notes* **112**, 656–663 (2022)
10. Zhang, J., et al.: A secure and lightweight multi-party private intersection-sum scheme over a symmetric cryptosystem. *Symmetry* **15**(2), 319 (2023)
11. Yerznkyan, B., Gataullin, T., Gataullin, S.: Mathematical aspects of synergy. *Montenegrin J. Econ.* **18**(3), 197–207 (2023)
12. Gataullin, T., Gataullin, S., Ivanova, K.: Modeling an electronic auction. In: “Smart Technologies” for Society, State and Economy. ISC 2020. *Lecture Notes in Networks and Systems*, vol. 155, pp. 1108–1117. Springer, Cham (2021)
13. Gataullin, T., Gataullin, S.: Management of financial flows on transport. In: Twelfth International Conference “Management of Large-Scale System Development” (MLSD), pp. 1–4. IEEE (2019)
14. Andriyanov, N., et al.: Intelligent system for estimation of the spatial position of apples based on YOLOv3 and real sense depth camera D415. *Symmetry* **14**(1), 148 (2022)
15. Ekhlakov, R., et al.: Modeling the chemical pollution of the area by the random-addition method. *Fract. Fract.* **6**(4), 193 (2022)
16. Kositzyn, A., Serdechnyy, D., Korchagin, S., Pleshakova, E., Nikitin, P., Kurileva, N.: Mathematical modeling, analysis and evaluation of the complexity of flight paths of groups of unmanned aerial vehicles in aviation and transport systems. *Mathematics* **9**, 2171 (2021)
17. Barotov, D., et al.: Transformation method for solving system of boolean algebraic equations. *Mathematics* **9**, 3299 (2021)
18. Skarda, C., Freeman, W.: Brains make chaos to make sense of the world. *Behav. Brain Sci.* **10**(2), 161–173 (1987)
19. Chapman, S., Policastro, G.: Quantum computational complexity from quantum information to black holes and back. *Eur. Phys. J. C* **82**(2), 1–40 (2022)
20. Kant, P., Laskar, S., Hazarika, J.: Transfer learning-based EEG analysis of visual attention and working memory on motor cortex for BCI. *Neural Comput. Appl.* **34**(22), 20179–20190 (2022)
21. Lega, B., Jacobs, J., Kahana, M.: Human hippocampal theta oscillations and the formation of episodic memories. *Hippocampus* **22**(4), 748–761 (2011)
22. Khidirov, B., Saidalieva, M., Khidirova, M.: Regulation of Living Systems. “Fan va texnologiya”, Tashkent (2014)
23. Khidirov, B.: Selected Works on Mathematical Modeling of the Regularism of Living Systems. Izhevsk, Moscow (2014)
24. Kent, L., Wittmann, M.: Time consciousness: the missing link in theories of consciousness. *Neurosci. Consciousness* **2021**(2), 1–10 (2021)
25. Ashrapov, N.: Psychology of religion and its role in the formation of the modern person. *Sci. Educ. Sci. J.* **3**(2), 675–680 (2022)



Identification of Ambient and Focal Information Processing Phases Using Eye Movement Response Registration

A. N Korosteleva^(✉), S. I. Kartashov, and A. A. Kotov

National Research Center Kurchatov Institute, Moscow, Russia
Korosteleva_AN@nrcki.ru

Abstract. In this study, we conduct research aimed at differentiating types of attention during image inspection. In our investigation, we utilized the EyeLink 1000 Plus eye tracker and machine learning methods to identify phases of ambient and focal information processing. A series of experiments were conducted with 10 participants who examined images of rooms with interior items. Data on the trajectory of the gaze point was recorded using the EyeLink 1000 Plus eye tracker. We developed methods for processing eye tracker data for analyzing eye movement in the task of free inspection. These methods include applying the fixation segmentation method by areas of interest, speed, and duration. We also developed an eye tracker data classification algorithm to identify ambient and focal types of attention. The results showed that the ambient type of attention is characterized by high speed, short eye fixations, and is not tied to a specific object in space, while the focal type of attention is characterized by prolonged eye fixations and is tied to specific significant objects. These results can be used to develop innovative Brain-Computer Interface (BCI) and Eye-Brain-Computer Interface (EBCI), opening up new opportunities for research in neuroscience and user interface development.

Keywords: Eye tracker data processing · Visual search · Attention types classification

1 Introduction

In modern cognitive psychology, attention is considered as a multi-level process [1]. In the context of this approach, B. M. Velichkovsky described two key types of attention that reflect different stages of information processing: ambient and focal attention. Identifying types of attention is a relevant task in the field of cognitive science and neurophysiology, due to their role in modulating perception. Decoding these processes can contribute to the development of intuitive human-machine interfaces and a deep understanding of the neurobiological correlates of attention and related pathologies. Ambient attention, sometimes defined as diffuse, represents a critical mechanism that ensures the distribution

of attentiveness across a broad spectrum of stimuli or actions, which is important for sensorimotor coordination, spatial navigation, and reaction speed [2,3]. On the other hand, focal attention is associated with the recognition of individual objects, implying the activation of previous experience and the engagement of distinct forms of social cognition [4,5]. Both of these types of attention are important subjects of study in cognitive sciences, as they are key to understanding cognitive processes such as perception, learning, and decision-making. Eye tracking technology allows researchers to precisely determine where a subject is looking at a specific moment in time [6]. This provides the opportunity to gain a deeper understanding of how attention is distributed and how ambient and focal attention interact with each other in various contexts. The task of our work is to develop algorithms for processing and interpreting eye tracker data with the aim of more effectively determining the leading attention mechanism. These algorithms allow us to analyze and interpret complex eye movement patterns, describe the interaction of ambient and focal attention in different contexts, which opens up new opportunities for comprehensive research of attention mechanisms and higher cognitive functions.

2 Experiment Description

The study was conducted using the EyeLink 1000 Plus system (SR Research, Canada) to record the subject's eye movements. The method of pupil registration was active, based on Purkinje reflection. The detectors were monochrome video cameras. The detection speed was up to 1000 frames per second. The accuracy of determining the eye deviation angle was 0.5° . The Presentation software (NBS, USA) was used to develop paradigms. The stimuli were a set of pictures depicting rooms with interior items.

The experiment involved 10 participants (6 men and 4 women) with good vision (without lenses), 9 of whom were right-handed and 1 was left-handed. The average age of the participants was 26 ± 3.4 years.

Participants started with a task where they were shown 20 similar pictures, which they had to carefully examine and remember, focusing their gaze on the fixation cross between the images (Fig. 1a). Then they underwent a test where they had to determine whether they had previously seen the presented image using a block with buttons (Fig. 1b). This test task was necessary to stimulate participants to examine the images more carefully. After this, participants were given a minute to rest (Fig. 1c). Then this cycle—task, test, and rest—was repeated once more.

3 Fixation Processing Methods

Fixations, periods of gaze stability, are key to identifying types of attention as they indicate areas of focus. Unlike saccades and post-saccadic oscillations, which are associated with attention shifts, fixations reflect moments of attention

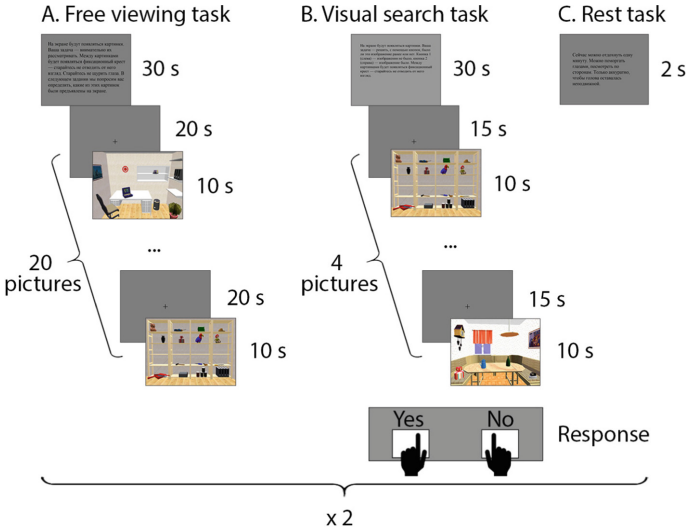


Fig. 1. Experiment scheme

concentration. Therefore, in this study, we analyze fixations, but other types of eye movements will be considered in future research perspectives.

Types of attention are characterized by spatial and temporal features. Ambient attention is characterized by fast fixations and is not tied to a specific object in space, while focal attention, on the contrary, is characterized by slow fixations and is tied to certain significant objects [2–5]. Thus, it is proposed to segment fixations by areas of interest, speed, and duration.

3.1 Fixation Segmentation by Areas of Interest

The data from the eye tracker was transformed into a series of states, such as $(x, y)(t_1), (x, y)(t_2), \dots, (x, y)(t_n)$, associated with areas of interest over time s_1, s_2, \dots, s_n . Thus, each fixation was matched to an object in the picture.

In the processing, a Markov chain was used to represent the sequences of fixations in both dynamic and static contexts [7]. Each state was modeled using initial $p(s_{t_1})$ and conditional probabilities $p(s_i/s_j)$ as follows:

$$p(s_{t_1}, s_{t_2}, \dots, s_{t_n}) = p(s_{t_1})p\left(\frac{s_{t_2}}{s_{t_1}}\right) \dots p\left(\frac{s_{t_n}}{s_{t_{n-1}}}\right) \tag{1}$$

These conditional probabilities form a transition matrix (TRM) of size $M \times M$, where M is the number of states

$$TRM = \begin{pmatrix} p\left(\frac{s_1}{s_1}\right) & \dots & p\left(\frac{s_1}{s_M}\right) \\ \vdots & \ddots & \vdots \\ p\left(\frac{s_M}{s_1}\right) & \dots & p\left(\frac{s_M}{s_M}\right) \end{pmatrix} \approx \begin{pmatrix} \frac{N(s_1, s_1)}{N(s_1)} & \dots & \frac{N(s_1, s_M)}{N(s_1)} \\ \vdots & \ddots & \vdots \\ \frac{N(s_M, s_1)}{N(s_M)} & \dots & \frac{N(s_M, s_M)}{N(s_M)} \end{pmatrix} \tag{2}$$

$N(s_i)$ is the total number of state i observed in the sequence s_1, s_2, \dots, s_n , and $N(s_i, s_j)$ is the number of occurrences of the pair of states. Instead of initial probabilities, limit probabilities (LP) were used, which reflected the focus on a specific area of the image throughout the entire viewing process [8]. Applying the ML criterion leads to an LP estimate:

$$LP = \left(\frac{N(s_1)}{\sum_{k=1}^M N(s_k)}, \frac{N(s_2)}{\sum_{k=1}^M N(s_k)}, \dots, \frac{N(s_M)}{\sum_{k=1}^M N(s_k)} \right) \tag{3}$$

This allowed to separate dynamic and static characteristics and transform the sequence $(x, y)(t)$ into feature vectors of fixed length, consisting of TRM and LP elements, to simplify the classification process. Figure 3a shows a visualization of the spatial segmentation of the image, where objects of high significance are highlighted in red, while the space surrounding them is marked in blue.

3.2 Fixation Segmentation by Speed

The sequences of coordinates $(x, y)(t)$ contain information about position and dynamics. To extract dynamic information, the coordinates were independently transformed into speeds $(V_x, V_y)(t)$ for each coordinate using numerical differentiation, as shown in formula (4):

$$x(n)' = \frac{1}{12}(x(n - 2) - 8x(n - 1) + 8x(n + 1) - x(n + 2)) \tag{4}$$

Figure 2a shows a typical sequence of coordinates $(x, y)(t)$. Speeds were obtained for all coordinates. Figure 2b shows the speeds for the data from Fig. 2a. Thus, the speed of each fixation was estimated as the sum of the speeds of the points included in it. Figure 3b shows a HeatMap, where slow fixations are marked in red and areas with fast fixations are marked in blue.

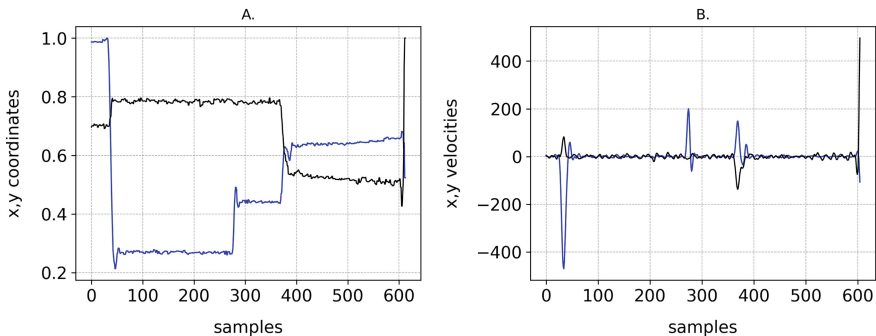


Fig. 2. a Normalized sequence of (x, y) coordinates; b (V_x, V_y) for the data from graph (a)

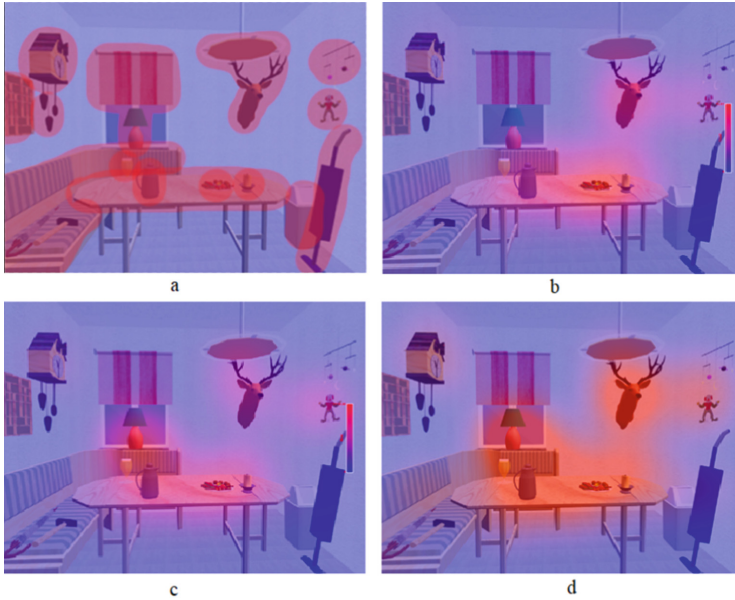


Fig. 3. Results: a fixation segmentation by areas of interest, b fixation segmentation by speed, c fixation segmentation by duration, d classification

3.3 Fixation Segmentation by Duration

The duration of fixations varied from 40 ms to 2 s. Frequency histograms of the number of fixations from their duration were constructed for each subject and each image (see Fig. 3c). The fixation threshold was determined using the Otsu method [9], which is effective for data following a bimodal distribution. Fixations with a duration of less than 80 ms do not carry meaningful information for us. Based on this distribution, a HeatMap is constructed in Fig. 3c, where long fixations are marked in red and areas with short fixations are marked in blue.

4 Fixation Classification

By integrating segmentation methods, each fixation has three binary features. Ambient attention is characterized by high speed, short fixations, and no attachment to specific objects in space. In contrast, focal attention manifests in slow, long fixations oriented towards specific significant objects. Thus, each fixation was classified according to the quantitative characteristics that determine its affiliation with the ambient or focal type of attention. An example of such classification is presented in Table 1. The determination of threshold values for parameters was performed using statistical analysis of histograms based on data obtained from all study participants. This allowed for the consideration of individual variance in speed and duration of fixations and provided a more accurate classification.

Table 1. Classification of fixations by characteristics of ambient and focal attention

Fixation number	Fixation speed (/s)	Fixation duration (s)	Object attachment	Attention type
1	6.56	0.161	Yes	Ambient
32	0.15	0.245	No	Focal
154	4.97	0.211	Yes	Ambient
186	1.32	0.287	No	Focal
278	5.64	0.183	Yes	Ambient

Figure 3d demonstrates the distribution of fixations in the context of attention type: fixations corresponding to the ambient type of attention are marked in blue, while fixations characteristic of the focal type of attention are represented in red.

To ensure classification accuracy, functional magnetic resonance imaging (fMRI) annotations were used, which indicated the time intervals during which certain types of attention were observed. The quality of classifying fixations into ambient and focal attention was evaluated using various metrics, including accuracy, sensitivity, specificity, F-score, and AUC-ROC. The evaluation results are presented in Table 2.

Analysis of the classification metrics presented in Table 2 demonstrates high efficiency in distinguishing between ambient and focal fixations. The classification accuracy is 0.91, while sensitivity and specificity reach 0.89 and 0.95 respectively. The F-score, representing the harmonic mean of precision and sensitivity, equals 0.94, confirming the balance between these two metrics. The area under the error curve (AUC-ROC) equals 0.95, indicating the model's high ability to distinguish between ambient and focal fixations. These results underline the superiority of the proposed classification method.

Table 2. Quality metrics for the classification model for identifying ambient and focal attention

Classification metric	Value
Accuracy	0.91
Sensitivity	0.89
Specificity	0.95
F-score	0.94
AUC-ROC	0.95

5 Conclusions

During the research, segmentation of fixations was carried out according to various parameters. In particular, each fixation was classified according to quantitative characteristics that allow determining its affiliation with the ambient or focal type of attention. The use of fMRI data as a benchmark ensured high classification accuracy.

It was found that each type of gaze movement has its unique characteristics that can be used for their identification and classification. This allowed separating dynamic and static characteristics and transforming the sequence of eye movements into feature vectors of fixed length to simplify the classification process.

The obtained results can be applied for a comprehensive analysis of fMRI data involving eye tracker data, where they can assist in a more accurate interpretation of the activity of various brain areas during the performance of attention-demanding tasks. Finally, this work is part of a larger project aimed at studying the mechanisms of attention and their role in cognitive processes. We aim to further explore various attention types and their cognitive interplay.





Funding. The work was carried out within the framework of the state assignment of the National Research Center Kurchatov Institute.

References

1. Величковский, Б.М.: Искра ψ : новые области прикладных психологических исследований. Бестик Московского университета. Серия 14. Психология, 1, 57–73 (2007)
2. Krukar, J., Mavros, P., Hoelscher, C.: Towards capturing focal/ambient attention during dynamic wayfinding. In: ACM Digital Library, pp. 1–4. ACM, New York (2020)
3. Dunlop, R.A., Cato, D.H., Noad, M.J.: Your attention please: increasing ambient noise levels elicits a change in communication behaviour in humpback whales (*Megaptera novaeangliae*). In: Proceedings of the Royal Society B: Biological Sciences, vol. 277, pp. 2521–2529. The Royal Society, London (2010)
4. Eriksen, C.W., St. James, J.D.: Visual attention within and around the field of focal attention: a zoom lens model. *Percept. Psychophys.* **40**, 225–240 (1986)
5. McElree, B.: Working memory and focal attention. *J. Exp. Psychol. Learn. Mem. Cogn.* **27**(3), 817–835 (2001)
6. Obaidallah, U., Al Haek, M., Cheng, P.C.-H.: A survey on the usage of eye-tracking in computer programming. *ACM Comput. Surv. (CSUR)* **51**(1), 1–36 (2018)
7. Ross, S.M.: Introduction to Probability Models, 6th edn. Academic Press, USA (1997)
8. DeAngelus, M., Pelz, J.B.: Top-down control of eye movements: Yarbus revisited. *Vis. Cogn.* **17**(6–7), 790–811 (2009)
9. Otsu, N.: A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**(1), 62–66 (1979)



The Analysis of the DMN Network of the Brain Using the Method of Segmentation of Functionally Homogeneous Regions

Stanislav Kozlov¹  , Alexey Poyda¹ , Vyacheslav Orlov¹ ,
and Vadim Ushakov^{2,3,4} 

¹ National Research Centre “Kurchatov Institute”, Moscow, Russian Federation
manist@list.ru

² National Research Nuclear University MEPhI, Moscow, Russia

³ Institute for Advanced Brain Studies, Lomonosov Moscow State University, Moscow, Russia

⁴ Mental Health Clinic No. 1 named after N.A. Alexeev of Moscow Health Department,
Moscow, Russia

Abstract. The human brain contains many functional networks: regions whose dynamics are correlated. One of the largest and least explored is the default mode network (DMN). The existence of this network, as well as the tasks that the brain performs during the activity of this network, are still hypotheses. Many studies show signs of the presence of this network. In this paper, we explore the DMN using a new method for segmentation functionally homogeneous regions. Obtained result shows the predominance of significant connections between DMN regions, which confirms the generally accepted hypothesis about the work of the brain at rest. In addition, there are clusters of significant connections for the regions of the parahippocampal regions and the amygdala. Also, a high level of significance was shown by the connections of the DMN regions with regions that are not included in the clusters of significant connections.

Keywords: fMRI · Functional connectivity · DMN · Resting state

1 Introduction

The default mode network (DMN) includes several brain regions that are active during the resting state [1]. In this state the instructions ask the person not to focus on anything, which implies to the base work of consciousness processes.

There is a hypothesis, confirmed by various studies [2–4], that the regions included in the DMN show higher functional connectivity with each other at resting state [5].

Modern studies of functional connectivity between brain regions are carried out with functional magnetic resonance imaging (fMRI), which allows to get data of brain activity.

The functional connectivity of voxels suggests that their dynamics are about similar. One of the most popular methods for determining the level of functional connectivity

is to calculate the Pearson correlation coefficient between voxels dynamics and authors of this study showed in the previous work, that correlation metrics has one of the best results in noise and time-shift stability tests [6]. Usually, the calculation of functional connectivity is carried out not between single voxels, but between regions of the brain. To obtain the dynamic of the region, the averaging of the dynamics of the voxels included in the region is used.

In different studies, there are different approaches to the definition of regions [7]. One of the most popular approaches is to use regions of population structural-functional atlases for these purposes. These atlases define rather large regions, on the order of tens of cm^3 . In addition, such a region may include voxels whose dynamics have a low level of Pearson correlation, resulting in its dynamic may incorrectly show the properties of the averaged voxels dynamics.

Also, the authors of this study showed in the previous work [8], that, if we consider DMN regions on a large scale (i.e., take small regions about 1 cm^3 in size), then we can't see functional connectivity so well. So, for example, if we select small sub-regions in large regions included in the DMN, then their average dynamics may almost not correlate.

Therefore, the next question arises. Is the hypothesis about the functional connectivity of the regions of the DMN fair if we consider the DMN network at the larger scale or regions, for example, smaller, but functionally homogeneous regions?

To test this hypothesis, in this work, we used the previously developed method for segmentation functionally homogeneous regions, which makes it possible to find individual spatially related regions with a high internal correlation [9, 10]. In this study, we assessed the level of significance of connections between functionally homogeneous regions included in the DMN regions and some other regions. The significance we assessed using permutation tests. To calculate the functional connectivity between regions, we used the Pearson correlation coefficient. In addition, following the theory of dynamic change in functional connectivity, we assess the connectivity between regions not over the entire time period of the experiment: we assess functional connectivity in time intervals with a sliding window method.

2 Materials and Methods

2.1 Data Acquisition and Preprocessing

The data was obtained using 3 Tesla MRI scanner Siemens Magnetom Verio at the National Research Centre "Kurchatov Institute". fMRI data were obtained with the following scan parameters: 42 slices, repetition time (TR) 2000 ms, echo time (TE) 20 ms, field of view (FOV) $192 \times 192 \text{ mm}^2$, voxel size $3 \times 3 \times 3 \text{ mm}^3$. 1000 time samples were scanned for functional data, with a total duration of about 33.5 min. The total study acquisition time was 40 min.

For fMRI and structural data preprocessing, the freely distributed software SPM12 and Mac OS bash scripts were used. The center of anatomical and functional data was transferred to the anterior commissure (AC). Magnetic inhomogeneity artifacts were removed from the functional data using the scanned during experiment session field mapping data. A slice timing correction of the phase shift caused by the technical features

of scanning was performed Using BROCOLLI software, the artifacts of the subject's head movement were calculated and corrected. After that, normalization of structural and functional data to the MNI (Montreal Neurological Institute) atlas space was performed. Anatomical data were segmented into 3 possible types of brain tissue (gray and white matter and cerebrospinal fluid). After normalization, the functional data smoothing was carried out with a Gaussian filter ($6 \times 6 \times 6 \text{ mm}^3$ FWHM).

The experiment involved 25 healthy subjects. Permission to conduct the experiment was granted by the Ethical Commission of the National Research Centre "Kurchatov Institute" (No. 5 dated April 5, 2017).

Since the Pearson correlation is the key metric in this work, the data was tested for autocorrelation. Voxel dynamics showed a high level of autocorrelation. Therefore, to eliminate it, a second-order autoregressive model was used. Global signal regression (GSR) was also performed.

2.2 Methods

Functionally homogeneous regions segmentation method. We use functionally homogeneous regions segmentation method [9, 10] implemented in "CCM-FOR" [11] to determine the regions of the brain between which we assess the connections. The method has three main steps:

1. Determination of homogeneity zones for each voxel. The homogeneity zone for voxel v_i is all voxels that are spatially adjacent to v_i or to homogeneity zone and whose dynamics correlates with the dynamic of v_i .
2. Filtering zones of homogeneity. At this step, the largest homogeneity zones are selected from all the homogeneity zones. Moreover, only those zones should be in the final sample, the center of which does not fall into another, larger zone, which also fell into the final sample.
3. Determining the belonging of voxels that fell into several zones. To do this, the correlation of the dynamic of the determined voxel with the activity of the centers of the zones is calculated, and belonging is determined by the best correlation coefficient.

To obtain the dynamics of the regions, we used the averaging of the dynamics of all voxels included in the one region.

For different subjects, the segmentation method identifies a different number of regions (from 446 to 783 regions). For each subject, we selected 20 regions from all segmented regions according to the following principles: first, by the best intersection with anatomical regions of interest to us, such as: DMN network (PCC, MPFC, LIPC, RIPC regions), ba10l, ba10r, VLPFC, Hippocampus, Amygdala, Parahippocampal Gyrus; and secondly, by correlation: the regions were chosen so that the correlation coefficients between their dynamics covered the largest possible range of values. Figure 1 shows an example of segmented regions and 20 selected regions on a single subject.

Method for assessing functional connections between regions. To assess dynamic connectivity, the sliding window method was used: we divided the entire time interval of the study (1000 time samples) into overlapping intervals of equal length (250 time samples) with an overlap of 95%. Further, the connections between the regions were

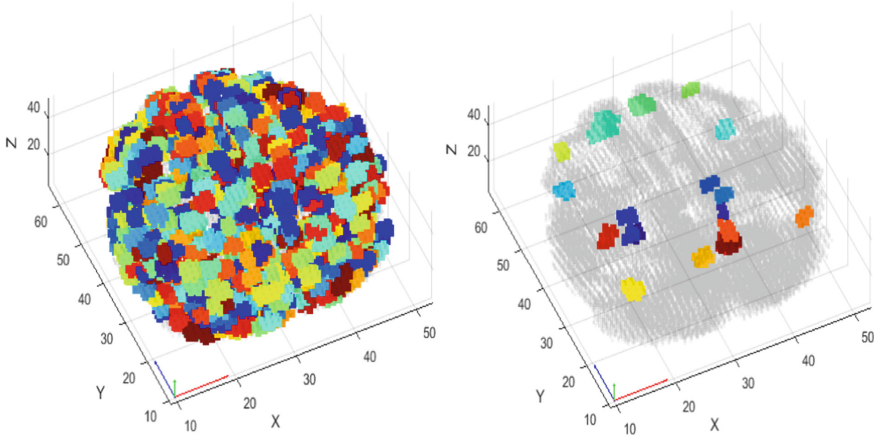


Fig. 1. Example of segmented regions on 1 subject. Left - segmented regions, right - 20 selected regions. The color indicates the number of the region.

assessed for each window using the Pearson correlation. Thus, 61 time intervals were obtained, and in each window the assessment of all connections between regions were obtained as a matrix of 20×20 values.

Assessment of the significance of functional connectivity. To obtain the significance level of each value (in each time interval for each connection, we denote as C_i), a permutation test was performed. To do this, we shuffle the values of the dynamics of the regions independently (within each dynamic, the values of this dynamic are randomly interchanged), which should nullify the functional connection if it exists and if it is significant and not random. Further, the connection is estimated on mixed series (we denote P_i).

The process of mixing and estimating the connection is repeated many times (we repeated 1000 times, getting 1000 values $P_{ij}, j \in [1, 1000]$). Next, a permutational distribution is constructed over all estimates of the connection P_i и, and then the probability of obtaining a real assessment of the connection C_i is estimated. The null hypothesis is that there is no connection between the regions. If the probability of obtaining the C_i value is less than the given significance level (we use $p = 0.001$), the null hypothesis is rejected in favor of the alternative one, i.e. connection is considered significant. Thus, for each pair of regions, an estimate of the significance of the connection between them for each time window is obtained.

We used the permutation test, since we do not have true information about the presence or absence of the functional connections between regions, so the permutation test allows us to assess the presence of the connection between regions using probability values.

3 Result

For each subject for each connection between each pair of regions we assessed different values: correlation between anatomical regions, correlation between selected functional homogeneous regions on the whole time interval and by sliding window method and their significance.

Figure 2 shows assessed functional connections for one subject. Figure contains connections, assessed by the correlation between anatomical regions, whose dynamics were obtained by averaging all dynamics of voxels of the anatomical regions (see Fig. 2A); connections, assessed by the correlation between selected functional homogeneous regions on whole time interval (see Fig. 2B); the sum of values (by time windows) of significant connections between selected functional homogeneous regions (see Fig. 2C); the number (by time windows) of significant connections between selected functional homogeneous regions (see Fig. 2C).

The figures show matrices in color coding. Each row and column correspond to one of the considered regions. Regions belonging to the DMN network are marked with a red square. The values correspond to the assessed connections between the regions. Blue (dark) values correspond to small values of functional connections, yellow (bright) values correspond to large values of functional connections.

On average for all subjects, the number of significant connections between regions from DMN was 68% (avg. 1475 out of 2196 connections), between regions not from DMN - 23% (avg. 632 out of 2745 connections). In addition, we noticed the significance of connections between regions from the DMN and regions not included in the DMN - their number was 43% (avg. 2578 out of 6039 connections).

This confirms the hypothesis of the presence of correlated activity between DMN regions in resting state, and also shows the presence of connections between DMN regions with other regions in resting state.

4 Conclusion

This paper presents an investigation of the functional connections between regions of the DMN using the method of segmentation of functionally homogeneous regions according to fMRI brain data. Pearson's correlation was used to assess connections between regions. The permutation test and the sliding window method were used to assess the significance of the connections. In addition to the values of the assessed connections, the final value for each connection was the number of significant connections by all time windows.

We understand that the results achieved on certain parameters of used methods. As a future work we see the importance of validating the result on other metrics (although the metrics were compared [6]), other widths of time-window method for another dynamic connectivity, other brain regions for validation on exploring more brain connectivities and networks. But we see that the results, achieved in this work on presented methods and parameters, showed greater variety of connections between connections of selected regions in comparison with anatomical regions, and presented statistical analysis on the permutation test strengthened the result.

Based on the results of the work, it is worth noting the predominance of connections with a high level of significance within the DMN regions. There are also clusters of

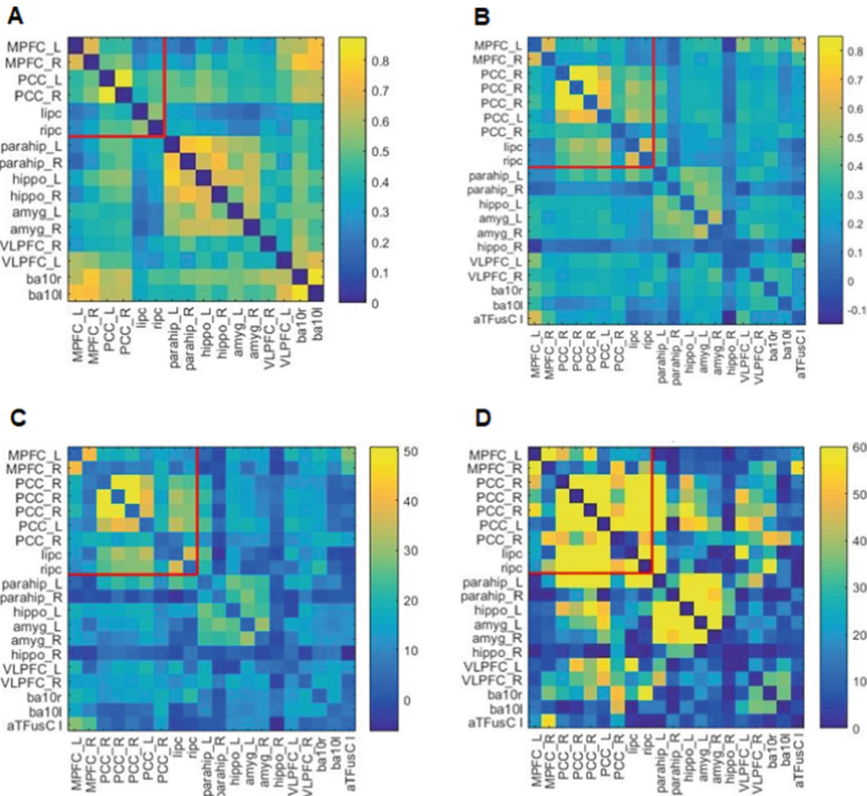


Fig. 2. Functional connections between regions of one subject. 2.A - rows and columns correspond to the anatomical regions; 2.B, C, D - rows and columns correspond to the selected functional homogeneous regions (labels by intersection with anatomical regions). 2.A, B – values correspond to correlation on whole time interval; 2.C – values correspond to the sum of significant correlations by time windows; 2.D – values correspond to the number of significant correlations by time window.

significant connections for areas of the parahippocampal regions and the amygdala, which belong to the system of the medial temporal lobe and functionally associated with the formation and retrieval of memory traces and their significance. The result confirms the presence of strong functional connections within these networks at rest.

Moreover, such a conclusion is difficult to show with only connections between anatomical regions (see Fig. 2A), or with connections over the entire time interval (see Fig. 2B). Dynamics of anatomical regions may have excessive spatial averaging of different voxel dynamics that can lead to high values of connections between large and close regions. Connections over the entire time interval may not take into account the dynamic variability of connections by time. By assessing connections in small functionally homogeneous regions using a sliding window and permutation test (see Fig. 2D) we can see pronounced connectivity within known networks and also, we can see the previously proven by authors fact that these networks are not completely homogeneous.

It is also worth noting the high level of significance of connections not only within the DMN regions, but also the high level of significance of the connections between regions from the DMN with regions that are not included in the clusters of significant connections, in particular with brain regions involved in the processes of memory and recognition of the novelty of stimuli (hippocampus and parahippocampal areas), areas involved in assessing the significance of stimuli (the amygdala area) and areas taking part in working memory processes (ventrolateral prefrontal regions). Thus, in the processes of human brain in the resting state, there is an interaction of the regions of the DMN network with structures that take part in the formation and extraction of individual experience, belonging to different levels of evolutionary development - paleocortex, archeocortex and neocortex.

Acknowledgements. Data acquisition and preprocessing, development of the method, processing and statistical analysis was supported by a government task in the National Research Centre “Kurchatov Institute”, neurophysiological analysis was supported by Russian Science Foundation grant № 20-15-00299-P.






References

1. Buckner, R.L., Andrews-Hanna, J.R., Schacter, D.L.: The brain’s default network: anatomy, function, and relevance to disease. *Ann. N. Y. Acad. Sci.* **1124**(1), 1–38 (2008). <https://doi.org/10.1196/annals.1440.011>
2. Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V.: Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci.* **100**(1), 253–258 (2003). <https://doi.org/10.1073/pnas.0135058100>
3. Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., Shulman, G.L.: A default mode of brain function. *Proc. Natl. Acad. Sci.* **98**(2), 676–682 (2001). <https://doi.org/10.1073/pnas.98.2.676>
4. Buckner, R.L., DiNicola, L.M.: The brain’s default network: updated anatomy, physiology and evolving insights. *Nat. Rev. Neurosci.* **20**(10), 593–608 (2019). <https://doi.org/10.1038/s41583-019-0212-7>
5. Alves, P.N., et al.: An improved neuroanatomical model of the default-mode network reconciles previous neuroimaging and neuropathological findings. *Commun. Biol.* **2**(1), 370 (2019). <https://doi.org/10.1038/s42003-019-0611-3>
6. Poyda, A., Sharaev, M., Orlov, V., Kozlov, S., Enyagina, I., Ushakov, V.: Comparative analysis of methods for calculating the interactions between the human brain regions based on resting-state fMRI data to build long-term cognitive architectures. In: *Brain-Inspired Cognitive Architectures for Artificial Intelligence: BICA*AI 2020: Proceedings of the 11th Annual Meeting of the BICA Society 11*, pp. 380–390. Springer International Publishing (2021). https://doi.org/10.1007/978-3-030-65596-9_46
7. Orlov, V.A., Ushakov, V.L., Kozlov, S.O., Enyagina, I.M., Poyda, A.A.: A review of method and approaches for resting state fMRI analyses. In: *Biologically Inspired Cognitive Architectures 2019: Proceedings of the Tenth Annual Meeting of the BICA Society*, vol. 10, pp. 400–404 (2020). https://doi.org/10.1007/978-3-030-25719-4_52
8. Enyagina, I.M., et al.: Technologies for studying functional neural networks of the human brain based on data of nuclear functional magnetic tomography. *J. Phys. Conf. Ser.* **2155**(1), 12034 (2022). <https://doi.org/10.1088/1742-6596/2155/1/012034>

9. Kozlov, S., Poyda, A., Orlov, V., Malakhov, D., Ushakov, V., Sharaev, M.: Selection of functionally homogeneous brain regions based on correlation-clustering analysis. *Procedia Comput. Sci.* **169**, 519–526 (2020). <https://doi.org/10.1016/j.procs.2020.02.215>
10. Kozlov, S., Poyda, A., Orlov, V., Sharaev, M., Ushakov, V.: Selection of functionally homogeneous human brain regions for functional connectomes building based on fMRI data. In: *Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics: Proceedings of the 9th International Conference on Cognitive Sciences, Intercognsci-2020, Moscow, Russia, 10–16 Oct 2020*, vol. 9, pp. 709–719 (2021). https://doi.org/10.1007/978-3-030-71637-0_82
11. GitHub - KozlovStanislav/CCM-FOR. <https://github.com/KozlovStanislav/CCM-FOR>



New Feature for Schizophrenia Classification Based on Functionally Homogeneous Brain Regions

Stanislav Kozlov¹ (✉) , Artur Zhemchuzhnikov^{1,2} , Alexey Poyda¹ ,
Vyacheslav Orlov¹ , and Sergey Kartashov¹ 

¹ National Research Centre “Kurchatov Institute”, Moscow, Russian Federation
manist@list.ru

² Moscow Institute of Physics and Technology (National Research University), Dolgoprudny,
Russian Federation

Abstract. In this work, we investigate functionally homogeneous regions segmentation method (FHR) to obtain features for binary classification of patients with schizophrenia and healthy controls using support vector machine classifier (SVM) based on resting-state functional magnetic resonance imaging (rs-fMRI) data. For comparison, we used 4 feature-type approaches: functional connectivity maps (FCM), Amplitude of low frequency fluctuations (ALFF) and fractional amplitude of low frequency fluctuations (fALFF), Regional Homogeneity (ReHo). Four different feature selection algorithms were used (χ^2 , F_test, L1 and L2). SVM classifier was trained and tested on a rs-fMRI dataset of 36 patients with schizophrenia and 36 healthy controls, obtained using Siemens Magnetom Verio MRI 3TL scanner. The best results were achieved by features obtained by the ReHo approach (93% accuracy) and the FHR approach (91% accuracy). The ReHo approach showed best accuracy with χ^2 and F test feature selection algorithms, and the FHR approach showed best accuracy with L1 and L2 feature selection algorithms.

Keywords: fMRI · Resting state · Schizophrenia · Classification · SVM

1 Introduction

The diagnosis of mental disorders such as schizophrenia is a challenging task. The diagnosis is still carried out phenomenologically - on the basis of examination, complaints and anamnesis of the patient. Biological markers that can objectify the diagnosis remain the subject of discussion and numerous studies. Developed recently neuroimaging methods based on functional magnetic resonance imaging (fMRI) provide a good opportunity to observe the work of neuronal networks in the brain, and can be used in schizophrenia classification.

Within the last decades, there are many studies in binary classification of patients with schizophrenia and healthy controls based on resting-state functional magnetic resonance imaging (rs-fMRI) [1–6]. These works show that with machine learning methods, it

is possible to achieve an accuracy of about 70–90%, depending on the used feature set. There are different approaches for obtaining feature sets, for example functional connectivity maps (FCM) with 93% accuracy [2] or 62% accuracy [3], amplitude of low frequency fluctuations (ALFF) and fractional amplitude of low frequency fluctuations (fALFF) with 75% accuracy [4], regional homogeneity (ReHo) with 78% accuracy [5]. In this study, we investigate a new feature obtained on the method of segmentation functionally homogeneous regions (FHR) developed by the authors of this work [7, 8]. The FHR method identifies individual regions with high internal correlation. This method was developed in contrast to FCM methods, where regions are taken from population structure-function atlases. It is worth noting that the accuracy depends on the dataset and therefore it is incorrect to compare the accuracy from different works with different datasets, so in this work, in addition to the proposed approach, we calculated the accuracy using popular approaches on our dataset.

In this work, we use a dataset that includes fMRI data of 72 subjects (36 patients with schizophrenia and 36 healthy controls), obtained on a Siemens Magnetom Verio 3TL tomograph.

To compare features based on the FHR method, we used features obtained with the FCM, ALFF, fALFF, ReHo methods [9–11]. The features determined for each voxel were averaged over the regions of the complete brain parcellation (CONN atlas) from CONN toolbox [12] that includes cortical and subcortical areas from the FSL Harvard-Oxford Atlas [13] and cerebellar areas from the AAL atlas [14]. We used a support vector machine method with different feature selection algorithms and different numbers of the significant features for classification. We used cross-validation to evaluate accuracy.

2 Materials and Methods

2.1 Data Acquisition and Preprocessing

We used an rs-fMRI dataset from the database of patients with schizophrenia and healthy controls at the National Research Center “Kurchatov Institute”. At the time of this work, the database continues to be filled with new data by the agreement between the National Research Centre “Kurchatov Institute” and Psychiatric Hospital No. 1 Named after N.A. Alexeev of the Department of Health of Moscow (GBUZ “PKB No. 1 DZM”). Thus, we obtained an rs-fMRI data of 72 subjects: 36 patients with schizophrenia and 36 healthy controls.

Functional and structural data was obtained using Siemens Magnetom Verio 3 T scanner. Written consent was obtained from each subject to participate in the study. Before scanning, each subject received instructions to lie with his eyes closed and try not to think about anything purposefully.

To obtain a structural image, a three-dimensional T1-weighted sequence was used with the following parameters: 176 slices, repetition time (TR) 1900 ms, echo time (TE) 2.19 ms, flip angle 9°, inversion time 900 ms, field of view (FOV) $250 \times 218 \text{ mm}^2$, voxel size $1 \times 1 \times 1 \text{ mm}^3$.

The fMRI data was obtained with scanning parameters: 56 slices, repetition time (TR) 720 ms, echo time (TE) 33 ms, field of view (FOV) $192 \times 192 \text{ mm}^2$, voxel size 3

$\times 3 \times 3 \text{ mm}^3$. 900 time points for functional data were scanned with a total duration of about 10.5 min.

For fMRI and structural data preprocessing, the freely distributed software SPM12 and Mac OS bash scripts were used. The center of anatomical and functional data was transferred to the anterior commissure (AC). Magnetic inhomogeneity artifacts were removed from the functional data using the recorded during experiment session field mapping data (`gre_field_mapping`). The slice timing correction for fMRI data signals was conducted (slices of a three-dimensional image are obtained not at one moment, but sequentially over time). Using BROCOLLI software, the artifacts of the subject's head movement were calculated and corrected. After that, normalization of structural and functional data to the MNI (Montreal Neurological Institute) atlas space was performed. Additional whitening of fMRI data from noise was performed using the independent components analysis (ICA) based on the MELODIC tool of the FSL package. Anatomical data was segmented into 3 possible types of brain tissue (gray and white matter and cerebrospinal fluid). After normalization, functional data was smoothed using a $6 \times 6 \times 6 \text{ mm}^3$ FWHM Gaussian filter.

2.2 Methods

Features obtained by the FHR method. In this study, we investigate a new feature obtained by the method of segmentation functionally homogeneous regions (FHR) [7, 8]. We used an implementation of this method from “CCM-FOR” [15]. The FHR method identifies individual spatially related regions with high internal correlation. The FHR method consists of three main steps. The first step is to determine the homogeneity zones of each voxel. At the second and third steps, all homogeneity zones are filtered and voxels are distributed by zones. Thus, non-intersecting regions are obtained from intersecting homogeneity zones, but in this work, we use the homogeneity zones obtained at the first step, so we describe this step in more detail.

We determine the homogeneity zone for voxel v_i is all spatially connected voxels whose dynamics has a high Pearson correlation with v_i dynamic. By spatial connectivity we mean the presence of a path between the voxels of the zone within the boundaries of the region. An example of the homogeneity zone with minimum correlation level 0.6 is shown in Fig. 1. Voxel v_i (A) is marked green and its homogeneity zone boundaries are marked green. For each voxel, the level of correlation of their dynamics with the dynamic of voxel A is marked. For example, voxel B is in the zone, because it has a path to A by the voxels of the zone and the correlation of dynamics of A and B is higher than 0.6. The correlation of dynamics of A and C is high enough too, but there is no path to voxel A by the voxels of the zone, so voxel C is not in the homogeneity zone.

We calculated the homogeneity zones for each voxel. In order to obtain large enough homogeneity zones (with high correlation value there are many small zones, i.e., with only one voxel), but not too large (with small correlation value there can be zones with almost all voxels of brain), we used a correlation level of 0.5.

For each of 165678 voxels we assess the size of its homogeneity zone by the number of voxels in the zone. Thus, for each subject we calculated 165678 sizes of homogeneity zones. Then we apply the averaging of these values over the regions of the CONN atlas,

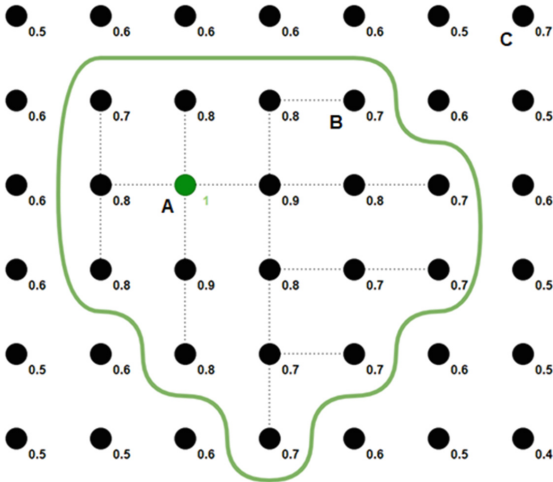


Fig. 1. Illustration of the homogeneity zone (two-dimensional slice).

so we get 132 averaged sizes of homogeneity zones. Thus, for each subject, a feature vector for classification is formed with the averaged sizes of the homogeneity zones for each atlas region.

Features obtained by the FCM, ALFF, fALFF, ReHo methods. For comparison, we also performed classification on other features described in the literature:

FCM features. Functional connectivity maps (FCM) - sets of connections between regions of the brain, i.e., the correlation values between the dynamics of regions taken from population structure-function atlases [9]. To obtain these features, we use the CONN atlas and assess the connections by Pearson's correlation between the dynamics of the regions of the atlas.

ALFF and fALFF features. Amplitude of low frequency fluctuations (ALFF) and fractional amplitude of low frequency fluctuations (fALFF) show the average fMRI signal amplitude in the frequency range from 0.01 to 0.1 Hz for each voxel. In the case of fALFF the result is additionally divided by the total signal amplitude over the entire frequency range [10].

ReHo features. Regional Homogeneity also shows the properties of local neural activity - the similarity between the dynamic of a particular voxel and the dynamics of its nearest neighbors [11].

ALFF, fALFF and ReHo were obtained using the BRANT toolbox [16, 17] with averaging over regions of the CONN atlas.

Feature selection algorithms. A large number of features for classification can lead to a decrease in the overall accuracy of the classifier, since useful data is lost in the noise and the classifier is overfitting. In this case, feature selection algorithms can help select only the most significant features for classification.

There are many algorithms for selecting the most significant features. They can be divided into two groups. The first group includes those approaches that select most correlated with the target variable features. An example of such an algorithm is χ^2 [18]. The second group includes those approaches that evaluate the relationship of a set of predictors with the target variable using statistical tests, which allows to select the required number of the most significant features. An example of such an algorithm is F-Test, based on Fisher's criterion, and algorithms where feature importance function is based on L1 and L2 norms [19].

SVM and cross-validation. Support vector machine is one of the most popular supervised learning algorithms which is used for classification and regression problems [20]. The goal of SVM is to create a hyperplane in an n-dimensional space that distinctly separates the data into classes.

We used k-fold cross-validation to assess the quality of the model [21]. For each iteration of the cross-validation cycle 64 subjects were used as training set and 8 subjects - validation set. Then the classifier was trained on the remaining data, and tested on the extracted test set. Thus, for each set of features, we performed 1000 cross-validation folds with a new test set for each fold.

3 Result

The accuracy of SVM classification for various sets of features, numbers of the most significant features and feature selection algorithms are shown in Table 1 for the FHR, ALFF, fALFF, ReHo methods and in Table 2 for the FCM method.

Maximum accuracy: 92–93% with the ReHo method, 90–91% with the FHR method, 58–59% with the ALFF method, 61–63% with the fALFF method and 78–80% with FCM method.

The ReHo method showed highest accuracy with χ^2 and F-test feature selection algorithms, and FHR showed highest accuracy with L1 and L2 feature selection algorithms. The classification accuracy is almost independent of the number of selected features with L1 and L2 feature selection algorithms. Feature selection algorithms χ^2 and F-test showed highest accuracy with a large number of features.

4 Conclusion

The approach presented in this paper for obtaining features for the binary classification of patients with schizophrenia and healthy controls based on the method of segmentation of functionally homogeneous regions (FHR) achieved an accuracy of 91%. This result is comparable to the best results in the literature and in our study (in the case of ReHo with accuracy 93%), and in many cases surpasses other methods of obtaining features. It is noteworthy that the FHR method uses size of the homogeneity zone for each voxel, and therefore it estimates homogeneity in the whole region, which can be contrasted with the ReHo method with the homogeneity for each voxel with only its neighbors.

We understand that the presented results may be specific due to the relatively small dataset of 72 subjects. So as a future work, we see the importance of validating the results

Table 1. The accuracy of the classification according to the features obtained by the FHR, ALFF, fALFF, ReHo methods.

Feature selection algorithm	FHR	ALFF	fALFF	ReHo
All (132) features	90%	58%	63%	92%
' χ^2 ', 100 features	89%	57%	60%	93%
' χ^2 ', 50 features	88%	57%	55%	92%
' χ^2 ', 10 features	80%	53%	50%	90%
F test, 100 features	89%	59%	61%	93%
F test, 50 features	88%	58%	54%	91%
F test, 10 features	83%	55%	55%	86%
L2, 100 features	91%	59%	60%	88%
L2, 50 features	90%	59%	60%	89%
L2, 25 features	90%	58%	60%	89%
L2, 10 features	90%	59%	60%	89%
L1, 100 features	90%	58%	59%	86%
L1, 50 features	90%	59%	59%	87%
L1, 25 features	90%	58%	58%	86%
L1, 10 features	90%	58%	60%	87%

Table 2. The accuracy of the classification according to the features obtained by the FCM method.

Method, features	All, 8646	' χ^2 ', 1000	' χ^2 ', 500	F test, 1000	F test, 500	L2, 1000	L2, 500	L1, 1000	L1, 500
FCM	78%	72%	67%	70%	66%	78%	80%	62%	62%

on a larger number of subjects (for example, using open datasets), and the importance of investigating classification stability on data from other tomographs. Also, we see another validation on the task of identifying accurate diagnoses of schizophrenia (non-binary classification). We suppose that the results may depend on such factors like stages of disorder, medical treatments, populations, which are needed to be investigated in future works.

The approach developed in this work has a high potential for use in other tasks, for example, studies of other mental disorders and other imaging modalities.

Acknowledgements. The study was supported by a government task in the National Research Centre "Kurchatov Institute".










References

1. Algumaei, A.H., Algunaïd, R.F., Rushdi, M.A., Yassine, I.A.: Feature and decision-level fusion for schizophrenia detection based on resting-state fMRI data. *PLoS ONE* **17**(5), e0265300 (2022). <https://doi.org/10.1371/journal.pone.0265300>
2. Tang, Y., Wang, L., Cao, F., Tan, L.: Identify schizophrenia using resting-state functional connectivity: an exploratory research and analysis. *Biomed. Eng. Online* **11**, 1–16 (2012). <https://doi.org/10.1186/1475-925X-11-50>
3. Yu, Y., Shen, H., Zhang, H., Zeng, L.-L., Xue, Z., Hu, D.: Functional connectivity-based signatures of schizophrenia revealed by multiclass pattern analysis of resting-state fMRI from schizophrenic patients and their healthy siblings. *Biomed. Eng. Online* **12**(1), 1–13 (2013). <https://doi.org/10.1186/1475-925X-12-10>
4. Guo, W., et al.: Decreased regional activity of default-mode network in unaffected siblings of schizophrenia patients at rest. *Eur. Neuropsychopharmacol.* **24**(4), 545–552 (2014). <https://doi.org/10.1016/j.euroneuro.2014.01.004>
5. Gao, S., et al.: Enhanced prefrontal regional homogeneity and its correlations with cognitive dysfunction/psychopathology in patients with first-diagnosed and drug-naïve schizophrenia. *Front. Psych.* **11**, 580570 (2020). <https://doi.org/10.3389/fpsy.2020.580570>
6. Chyzyk, D., Savio, A., Graña, M.: Computer aided diagnosis of schizophrenia on resting state fMRI data by ensembles of ELM. *Neural Netw.* **68**, 23–33 (2015). <https://doi.org/10.1016/j.neunet.2015.04.002>
7. Kozlov, S., Poyda, A., Orlov, V., Malakhov, D., Ushakov, V., Sharaev, M.: Selection of functionally homogeneous brain regions based on correlation-clustering analysis. *Procedia Comput. Sci.* **169**, 519–526 (2020). <https://doi.org/10.1016/j.procs.2020.02.215>
8. Kozlov, S., Poyda, A., Orlov, V., Sharaev, M., Ushakov, V.: Selection of functionally homogeneous human brain regions for functional connectomes building based on fMRI data. In: *Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics: Proceedings of the 9th International Conference on Cognitive Sciences, Intercognsci-2020, Moscow, Russia, 10–16 Oct 2020, vol. 9, pp. 709–719* (2021). https://doi.org/10.1007/978-3-030-71637-0_82
9. Blinowska, K.J.: Review of the methods of determination of directed connectivity from multi-channel data. *Med. Biol. Eng. Comput.* **49**, 521–529 (2011). <https://doi.org/10.1007/s11517-011-0739-x>
10. Zou, Q.-H., et al.: An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF. *J. Neurosci. Methods* **172**(1), 137–141 (2008). <https://doi.org/10.1016/j.jneumeth.2008.04.012>
11. Zang, Y., Jiang, T., Lu, Y., He, Y., Tian, L.: Regional homogeneity approach to fMRI data analysis. *NeuroImage* **22**(1), 394–400 (2004). <https://doi.org/10.1016/j.neuroimage.2003.12.030>
12. NITRC: CONN: Functional Connectivity Toolbox: Tool/Resource Info. <https://www.nitrc.org/projects/conn>
13. Atlases - FslWiki. <https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases>
14. Tzourio-Mazoyer, N., et al.: Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15**(1), 273–289 (2002). <https://doi.org/10.1006/nimg.2001.0978>
15. GitHub - KozlovStanislav/CCM-FOR. <https://github.com/KozlovStanislav/CCM-FOR>
16. Xu, K., Liu, Y., Zhan, Y., Ren, J., Jiang, T.: BRANT: a versatile and extendable resting-state fMRI toolkit. *Front. Neuroinform.* **12**, 52 (2018). <https://doi.org/10.3389/fninf.2018.00052>
17. Welcome to Brant!—BRANT 3.36 Documentation. <https://sphinx-doc-brant.readthedocs.io/en/latest>

18. Fisher, R.A., Yates, F. (eds.): *Statistical Tables for Biological, Agricultural and Medical Research*, 6th edn. Oliver and Boyd, Edinburgh (1963)
19. Johnston, J.: *Econometric Methods*, 2nd edn., pp. 35–38. McGraw-Hill (1972)
20. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995). <https://doi.org/10.1007/BF00994018>
21. Arlot, S., Celisse, A.: A survey of cross-validation procedures for model selection. *Statist. Surv.* **4**, 40–79 (2010). <https://doi.org/10.1214/09-SS054>



Preliminary Study of Cerebral Myelin Content Alterations at Schizophrenia

Ekaterina Krupina^{1,2} , Andrei Manzhurtsev¹ , Maxim Ublinskiy¹ ,
Larisa Mosina², Maria Osetrova³, Vasily Yarnykh^{4,5} , Galina Mamedova⁶ ,
Sergey Trushchelev⁶ , Natalia Zakharova⁶ , Georgy Kostyuk⁶ ,
and Vadim Ushakov^{2,6,7} 

- ¹ Clinical and Research Institute of Emergency Pediatric Surgery and Trauma, Moscow, Russia
krupina2002@mail.ru
- ² National Research Nuclear University MEPhI, Moscow, Russia
- ³ Vladimir Zelman Center for Neurobiology and Brain Rehabilitation, Skolkovo Institute of Science and Technology, Moscow, Russia
- ⁴ Laboratory of Neurobiology, Tomsk State University, Tomsk, Russia
- ⁵ Department of Radiology, University of Washington, Seattle, WA, USA
- ⁶ Mental Health Clinic No. 1 named after N.A. Alexeev of Moscow Health Department, Moscow, Russia
- ⁷ Institute for Advanced Brain Studies, Lomonosov Moscow State University, Moscow, Russia

Abstract. At the moment, the study of the association of different brain areas with schizophrenia is a thriving field of research. Although the precise mechanisms underlying the development of schizophrenia remain not fully understood, ongoing research enhances our understanding of which areas of the brain may be associated with various symptoms and behavioral disturbances in patients with the disease. Pathological changes in the brain may be accompanied by a decrease in the level of myelin. The results obtained in the present study support the hypothesis of an association between brain myelination and the development of schizophrenia. The present study revealed reduced myelination in the number of areas which can be divided into several groups based on common features. In our work, significant changes were found in the following areas of the brain: Heschl's Gyrus (includes H1 and H2), Postcentral Gyrus, Lateral Occipital Cortex superior division, Frontal Pole, Paracingulate Gyrus, Inferior frontal gyrus, Middle frontal gyrus. In the future, the results of the study can be used to create a non-invasive quantitative marker of schizophrenia. This will allow, on the one hand, to characterize the current stage of the disease of a particular patient in terms of myelination abnormalities, and on the other hand, it can help shed light on the etiology of this disease.

Keywords: Schizophrenia · MRI · Myelination · Macromolecular proton fraction

1 Introduction

Several studies have shown that brain myelination plays an important role in the normal functioning of the nervous system and has an association with many mental illnesses, including schizophrenia [1–3]. Schizophrenia is a complex mental disorder that is accompanied by various symptoms such as hallucinations, delusions, impaired thinking and social adaptation. To date, the mechanisms of schizophrenia development are not fully understood, but studies show that there is a relationship between the functional activity of different brain areas and the development of the disease [4, 5]. There is evidence that patients with schizophrenia have changes in the frontal cortex, which plays an important role in planning, decision-making and behavioral control [6].

The aim of this study is to use the macromolecular proton fractionation (MPF) method [7] to quantitatively detect abnormalities of myelin content in various brain structures in patients at the schizophrenia, as well as to assess the differences in myelination of these structures. The results of the study are intended to characterize schizophrenia in terms of myelination abnormalities, which, on the one hand, may help to shed light on the causes of this disease and, on the other hand, identify its quantitative marker.

2 Materials and Methods

2.1 Study Sample

Patients

The examination of 24 patients (14 men and 10 women, mean age of 31.6 ± 5.6 years), from among admitted to the acute care units of SFHI PKB № 1 of Moscow Health Department and diagnosed with schizophrenia spectrum disorders (ICD–10 codes F20 and F23) in 2021–2022.

All subjects signed an informed consent after full explanation of the experimental procedures in accordance with the Helsinki declaration.

Inclusion criteria: age of 18–50 years, condition meeting the ICD–10 and DSM–5 criteria for schizophrenia, self-awareness, and informed consent to participation in the trial.

Exclusion criteria: schizoaffective and affective disorders, brain structural damage, severe somatic and/or neurological diseases, potentially affecting brain physiology or structure, signs of substance abuse, general contraindications to MRI, and withdrawal of consent.

Control group

Control group consisted of 21 volunteers (9 men and 12 women, mean age of 30.0 ± 8.8 years), not related to the patients and examined uniformly and without general contraindications to MRI.

The diagnostic process spanned two days, commencing with a clinical interview conducted two days prior to scanning. The day before scanning, a psychometric evaluation using the Positive and Negative Syndrome Scale for schizophrenia (PANSS) was administered, followed by a comprehensive clinical assessment carried out by two experienced

psychiatrists who considered all pertinent data, including family interviews, medical records, physical examination results, and laboratory tests. All patients were treated with antipsychotic drugs in therapeutic doses equivalent to ~ 300 mg/day of chlorpromazine. You can find sociodemographic, clinical, and psychometric characteristics in Table 1.

Table 1. Sociodemographic, clinical, and psychometrical characteristics of studied sample

Parameters	Patients (n = 24) 10f; 14m	Healthy control (n = 21) 12f; 9m
<i>Sociodemographic characteristics</i>		
Married (abs., %)	4 (17%)	7 (33%)
Single (abs., %)	18 (75%)	12 (57%)
Divorced (abs., %)	2 (8%)	2 (10%)
Studies/occupied (abs., %)	6 (25%)	19 (91%)
Unemployed (abs., %)	18 (75%)	2 (9%)
Disabled (abs., %)	7 (29%)	0
<i>Clinicodynamic parameters of schizophrenia, years</i>		
Mean age of manifestation	23.4 ± 5.9	–
Disease duration from manifestation	8.1 ± 5.1	–
<i>Psychometric parameters, mean scores (standard deviation)</i>		
PANSS total	68.7 (8.3)	–
PANSS P	16.1 (3.0)	–
PANSS N	22.7 (3.6)	–
PANSS G	29.8 (4.9)	–

PANSS - The Positive and Negative Syndrome Scale, PANSS P - severity of productive symptoms, PANSS N - severity of negative symptoms, PANSS G - severity of other mental disorders on the general psychopathological scale.

The study was performed on a Philips Achieva 3.0 T MRI scanner, a 16 channel receive head coil was used. T1-weighted, proton density weighted, and magnetization transfer (MT) sequence images were acquired. The MT method is based on the incoherent exchange of magnetic energy of macromolecular protons with water protons. MPF is defined as the ratio of the number of macromolecular protons that are included in the magnetization transfer effect to the number of water protons. The minimum required dataset for MPF reconstruction consists of two variable flip angle (VFA) images and one MT image [8]. All VFA and MT images must be acquired using a gradient echo sequence with transverse magnetization removal (SPGRE). The same readout parameters (matrix 172 * 167 * 271, FOV 240 * 240 * 190, TE = 4.60, voxel 1.4 * 1.4 * 1.4, etc.) should be used for VFA and MT images.

The scanner image acquisition parameters were as follows:

1. MT-weighted: repetition time (TR) = 45 ms, echo time (TE) = 4.60 ms, flip angle (FA) = 8°, scan time 2 min;
2. T1-weighted: TR = 20 ms, TE = 4.60 ms, FA = 20°, scan time 1 min;
3. PD-weighted: TR = 20 ms, TE = 4.60 ms, FA = 4°, scan time 1 min.

Macromolecular proton fraction (MPF) maps were reconstructed using special software (available at <https://www.macromolecularmri.org>) [8]. Correcting the inhomogeneity of the RF field (B1) used for signal excitation and reception is recommended to obtain MPF in magnetic fields of 3 T or higher. The surrogate B1 correction option is used in this study, it eliminates the need to obtain a B1 map at the scanner. Magnetic field inhomogeneity (B0) has minimal effect on the MPF image, so no correction for B0 is performed.

The areas of interest for numerical analysis were cerebral cortex and cerebral white matter. MPF values of all areas of cerebral cortex were also determined separately according to the Harvard Oxford Cortical atlas (Fig. 1).

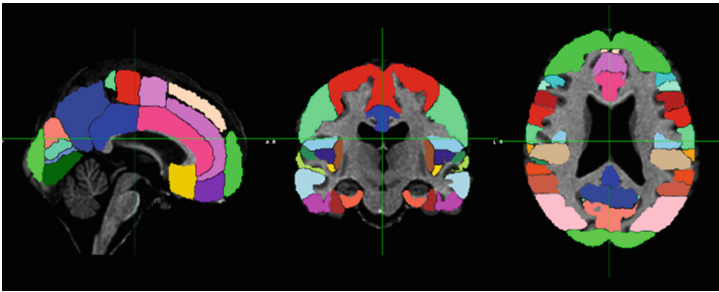


Fig. 1. Atlas Harvard Oxford Cortical in MNI 152 FLIRT

Statistical processing of the MPF values in normal volunteers and schizophrenia patients was performed in GraphPad Prism. The normality of continuous data distribution was assessed using the Shapiro-Wilk test in each group of subjects. Pearson's *r*-criterion was used to assess correlations. The statistical significance was determined at the level of $p < 0.05$. To search for intergroup differences, Student's *t*-test was used in case of normal distribution or Mann-Whitney test otherwise.

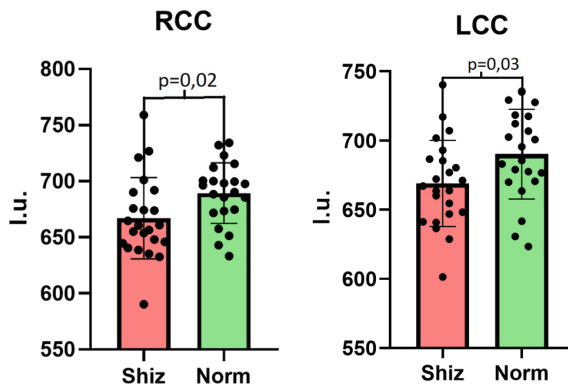
3 Results

All sets of mean MPF values in both studied groups were distributed normally. A significant decrease in the mean MPF values in schizophrenia was observed in both cortical cerebral tissue and white matter of both cerebral hemispheres (Figs. 2 and 3). The values of the coefficients of variation (CoV) in the analyzed zones are shown in Table 2.

Separate consideration of cerebral regions of interest revealed a significant decrease in the MPF value in schizophrenia in the Heschl's gyrus, postcentral gyrus, superior lateral occipital cortex, frontal pole, and paracingulate gyrus (Fig. 4).

Table 2. CoV and p-values in the norm and pathology groups in the analyzed brain areas.

Name of the region	Average MPF values (norm)	Average MPF values (shiz)	SD (norm)	SD (shiz)	CoV (norm), %	CoV (shiz), %	p-value
Left cerebral cortex	690.6	669.3	33.0	30.5	4.8	4.6	0.03
Right cerebral cortex	688.9	668.3	36.0	27.6	4.0	5.3	0.02
Left cerebral white matter	790.5	766.6	27.3	23.8	5	4.4	0.03
Right cerebral white matter	786.8	762.8	34.3	36.6	4.3	4.8	0.02

**Fig. 2.** MPF values (% * 100) of brain structures in normal volunteers and schizophrenia patients (mean \pm stand. dev.). RCC is the cortex of the right cerebral hemisphere, LCC – cortex of the left cerebral hemisphere

The results of t-test in the form of the value of reliable difference in the zones of interest between the group of patients and controls (p-value) are shown in Table 3.

The CoV values of the regions that are included in the Harvard Oxford Cortical in MNI 152 FLIRT atlas for each sample are shown in Fig. 5.

The t-test results demonstrated a statistically significant difference between CoV pat and CoV norm ($p < 0.05$). Correlation analysis between CoV values in the studied areas and mean MPF values in these areas in the patient group revealed a statistically

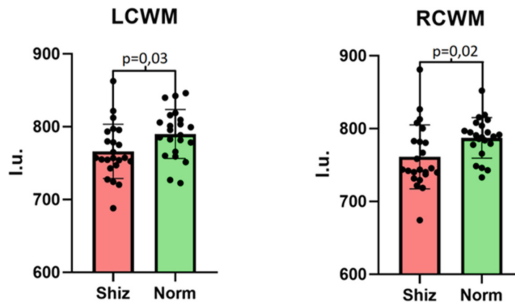


Fig. 3. MPF values (% * 100) of brain structures in normal volunteers and schizophrenia patients (mean \pm stand. dev.). LCWM - white matter of the left cerebral hemisphere, RCWM - white matter of the right cerebral hemisphere.

Table 3. Areas of interest with a statistically significant difference between the norm and schizophrenia groups.

Name of the region	p-value	Average MPF values (norm)	Average MPF values (shiz)
Heschl's gyrus (includes H1 and H2)	0.02	691.1	665.1
Postcentral gyrus	0.04	712.0	687.7
Lateral occipital cortex, superior division	0.04	700.8	677.7
Frontal pole	0.01	677.9	645.0
Paracingulate gyrus	0.02	714.5	687.4
Inferior frontal gyrus	0.05	751.4	725.1
Middle frontal gyrus	0.02	747.1	717.9

significant weak correlation ($r = -0.42$, $p = 0.011$). No significant correlation was found in the control group ($r = -0.06$, $p = 0.972$).

4 Discussion

In the present study, quantitative imaging of brain myelination by MPF is accompanied by dividing the brain into zones according to anatomical atlas. This approach provides a more detailed picture of brain myelination. In the currently available publications, myelination was determined separately in the grey and in the white matter of the brain, which does not sufficiently characterize possible regional abnormalities in various pathological processes [9].

The present study revealed reduced myelination in the number of areas which can be divided into several groups based on common features. Frontal cortex includes the

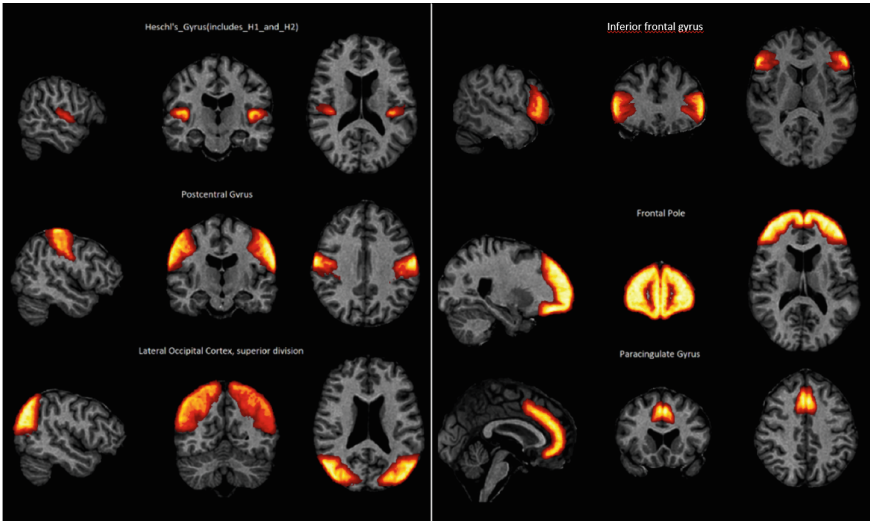


Fig. 4. Areas of the cerebral hemispheres with significantly different myelination between the groups of schizophrenia and norm

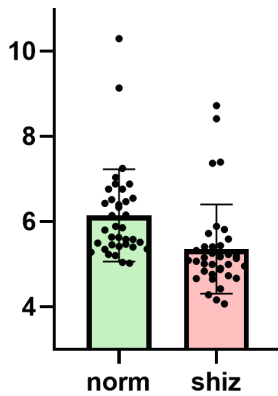


Fig. 5. CoV (%) distribution in the norm and in schizophrenia patients for the cortical zones

inferior and middle frontal gyrus. These areas of the brain are responsible for movement regulation, speech control, planning and decision making. Temporal lobe: the Heschl's gyrus This area of the brain is associated with memory, auditory perception, face recognition and emotional reactions. Parietal lobe: the postcentral gyrus. These areas of the brain are responsible for body sensation, spatial perception, attention, reading and understanding mathematical symbols. Circumlimbic areas: this includes frontal pole and paracingulate gyrus. These brain areas are associated with emotion processing, motivation, social behavior, and regulation of the autonomic nervous system. All these areas interact with each other to process complex cognitive and behavioral processes, the performance of which may be impaired in schizophrenia.

The relationship between the MPF values and myelin concentration in tissue shown in [10] allows us to conclude that the reliable difference in MPF values in different brain areas of patients and healthy subjects revealed in the present work should reflect the difference in myelin content in these areas. However, the statistical significance of the results may depend on the variation of MPF values in the analyzed groups. On average, the CoV values of the normal group were higher than the average of the patient group in the large brain (see Fig. 5). It is necessary to evaluate whether the difference of CoV values can condition the statistical significance of the difference between MPF values in the studied groups. For this purpose, we determined the correlation coefficients between the CoV values in the studied zones and the mean MPF values in these zones. In the normal group this correlation is absent, while in the group of patients only a weak correlation was revealed. This allows us to conclude that the significant difference of MPF values in norm and in patients is not associated with the difference between CoV in these groups, as CoV values are not connected or are weakly connected with the mean MPF values. Thus, the statistically significant results shown do characterize the different cerebral myelin content in the control group and in the patients. Establishing the causes of within-group variability in MPF values, which may be related to both sample homogeneity and the quality of the data obtained, requires additional much larger data collection and analysis, and is beyond the scope of this study.

Schizophrenia is a complex mental disorder that is co-occurring with various symptoms such as hallucinations, delusions, impaired thinking and social adaptation. To date, the mechanisms of schizophrenia development are not fully understood, but studies show that there is a link between the functional activity of different brain areas and the development of the disease [4, 5]. Demyelination can lead to impaired nerve impulse transmission speed, which can affect cognitive functions. Patients with schizophrenia already experience cognitive deficits, and intensification of this process can exacerbate problems with attention, memory, and problem solving [11]. Myelin affects the rapid and accurate transmission of signals between neurons. Demyelination can increase the duration of distorted signals, which can exacerbate symptoms of psychosis such as hallucinations and delusions [12]. Myelin plays a role in emotional regulation and control. Disruption of myelin may affect the ability of schizophrenic patients to control emotions, which may increase more intense reactions and emotional disturbances [13].

The observed differences in myelination levels in the schizophrenia brain are in good agreement with previous findings based on the same [9, 14] and different approaches to *in vivo* myelin visualization [15, 16], as well as molecular research evidence. Interestingly, in addition to significant differences between healthy controls and patients groups, changes in myelination showed a significant association with the severity of symptoms [17]. Since lipids constitute a significant portion of myelin dry weight, changes in lipid composition of gray and white matter of the schizophrenic patients might reflect the underlying pathology of myelin sheath formation. Previous studies demonstrated changes in a number of brain regions, including corpus callosum [18], prefrontal cortex [19, 20], and thalamus [21].

Myelin could act as a potential schizophrenia biomarker that may reflect structural and functional changes in the nervous system. Decreased myelin levels, measured as a percentage, may be significant in determining the degree of disease progression. The

introduction of myelin as a biomarker for disease assessment into clinical practice may provide opportunities for diagnosis, more accurate assessment of disease progression, and planning optimal treatment strategies. Additional studies aimed at identifying associations between myelin levels and clinical symptom severity may better define thresholds for myelin reduction associated with specific clinical conditions.

Although the existing data on changes at the molecular level are numerous, they lack consistency. Further systematic research linking molecular changes in schizophrenia brain with structural changes might reveal exact molecular signatures responsible for the observed MPF signal changes. Of special interest are lipid classes involved in myelin sheath formation—such as cholesterol, plasmalogen, and galactosylceramide—which have all previously shown changes in schizophrenia brain and blood plasma [12].

Acknowledgements. This work was in part supported by Russian Science Foundation grant № 20-15-00299-P (<https://rscf.ru/en/project/20-15-00299-P/>, data acquisition, statistical and neurophysiological analysis) and grant № 22-11-00213 (data preprocessing).

References

1. Valdés-Tovar, M., Rodríguez-Ramírez, A.M., Rodríguez-Cárdenas, L., et al.: Insights into myelin dysfunction in schizophrenia and bipolar disorder (2022). <https://doi.org/10.5498/wjp.v12.i2.264>
2. Takahashi, N., Sakurai, T., Davis, K.L., Buxbaum, J.D.: Linking oligodendrocyte and myelin dysfunction to neurocircuitry abnormalities in schizophrenia (2010). <https://doi.org/10.1016/j.pneurobio.2010.09.004>
3. Karoutzou, G., Emrich, H.M., Dietrich, D.E.: The myelin-pathogenesis puzzle in schizophrenia: a literature review. *Mol. Psychiatry* **13**(3), 245–260 (2008). <https://doi.org/10.1038/sj.mp.4002096>. Epub 9 Oct 2007
4. van den Heuvel, M.P., Fornito, A.: Brain networks in schizophrenia. *Neuropsychol. Rev.* **24**(1), 32–48 (2014). <https://doi.org/10.1007/s11065-014-9248-7>
5. Zhou, Y., et al.: Functional dysconnectivity of the dorsolateral prefrontal cortex in first-episode schizophrenia using resting-state fMRI. *Neurosci. Lett.* **417**(3), 297–302 (2007). <https://doi.org/10.1016/j.neulet.2007.02.081>
6. Salgado-Pined, P., Caclin, A., Baeza, I., Junqué, C., Bernardo, M., Blin, O., Fonlupt, P.: Schizophrenia and frontal cortex: where does it fail? *Schizophrenia Res.* **91**(1–3), 73–81 (2007). <https://doi.org/10.1016/j.schres.2006.12.028>
7. Kisel, A.A., Naumova, A.V., Yarnykh, V.L.: Macromolecular Proton Fraction as a Myelin Biomarker: Principles, Validation, and Applications (2022). <https://doi.org/10.3389/fnins.2022.819912>
8. Yarnykh, V.L.: Time-efficient, high-resolution, whole brain three-dimensional macromolecular proton fraction mapping (2016). <https://doi.org/10.1002/mrm.25811>
9. Smirnova, L.P., Yarnykh, V.L., Parshukova, D.A., Kornetova, E.G., Semke, A.V., Usova, A.V., et al.: Global hypomyelination of the brain white and gray matter in schizophrenia: quantitative imaging using macromolecular proton Fraction. *Transl. Psychiatry* **11**, 365 (2021). <https://doi.org/10.1038/s41398-021-01475-8>
10. Underhill, H.R., Rostomily, R.C., Mikheev, A.M., Yuan, C., Yarnykh, V.L.: Fast bound pool fraction imaging of the in vivo rat brain: association with myelin content and validation in the C6 glioma model. *NeuroImage* **54**(3) (2011)

11. Bartzokis, G.: Neuroglialpharmacology: white matter pathophysiologies and psychiatric treatments
12. Insel, T.R.: Rethinking schizophrenia. *Nature* **468**(7321), 187–193 (2010). <https://doi.org/10.1038/nature09552>
13. Davis, K.L., et al.: White matter changes in schizophrenia: evidence for myelin-related dysfunction. *Arch. Gen. Psychiatry* **60**(5), 443–456 (2003). <https://doi.org/10.1001/archpsyc.60.5.443>
14. Sui, Y.V., et al.: Quantitative macromolecular proton fraction mapping reveals altered cortical myelin profile in schizophrenia spectrum disorders. *Cerebral Cortex Commun.* **2**(2), tgab015 (2021). <https://doi.org/10.1093/texcom/tgab015>
15. Wei, W., et al.: Depth-dependent abnormal cortical myelination in first-episode treatment-naïve schizophrenia. *Hum. Brain Mapp.* **41**(10), 2782–2793 (2020). <https://doi.org/10.1002/hbm.24977>
16. Vanes, L.D., Mouchlianitis, E., Barry, E., et al.: Cognitive correlates of abnormal myelination in psychosis. *Sci. Rep.* **9**, 5162 (2019). <https://doi.org/10.1038/s41598-019-41679-z>
17. Palaniyappan, L., Al-Radaideh, A., Mougín, O., Das, T., Gowland, P., Liddle, P.F.: Aberrant myelination of the cingulum and Schneiderian delusions in schizophrenia: a 7T magnetization transfer study. *Psychol. Med.* **49**(11), 1890–1896 (2019). <https://doi.org/10.1017/S0033291718002647>. Epub 19 Sept 2018
18. Shimamoto-Mitsuyama, C., Nakaya, A., Esaki, K., et al.: Lipid pathology of the corpus callosum in schizophrenia and the potential role of abnormal gene regulatory networks with reduced microglial marker expression. *Cereb. Cortex. Cortex* **31**(1), 448–462 (2021). <https://doi.org/10.1093/cercor/bhaa236>
19. Schwarz, E., et al.: High throughput lipidomic profiling of schizophrenia and bipolar disorder brain tissue reveals alterations of free fatty acids, phosphatidylcholines, and ceramides. *J. Proteome Res.* **7**, 4266–4277 (2008). <https://doi.org/10.1021/pr800188y>
20. Matsumoto, J., et al.: Abnormal phospholipids distribution in the prefrontal cortex from a patient with schizophrenia revealed by matrix-assisted laser desorption/ionization imaging mass spectrometry. *Anal. Bioanal. Chem.* **400**(7), 1933–1943 (2011). <https://doi.org/10.1007/s00216-011-4909-3>. Epub 2 Apr 2011
21. Schmitt, A., et al.: Altered thalamic membrane phospholipids in schizophrenia: a postmortem study. *Biol. Psychiatry* **56**(1), 41–45 (2004). <https://doi.org/10.1016/j.biopsych.2004.03.019>
22. Schneider, M., Levant, B., Reichel, M., Gulbins, E., Kornhuber, J., Müller, C.P.: Lipids in psychiatric disorders and preventive medicine. *Neurosci. Biobehav. Rev.* **76**(Pt B), 336–362 (2017). <https://doi.org/10.1016/j.neubiorev.2016.06.002>. Epub 16 Jun 2016



Robotic Customer Service System ALKETON

Anton V. Kudriashov^(✉)

National Research Nuclear University «MEPHI» (Moscow Engineering Physics Institute),
Kashirskoe Shosse, 31, Moscow 115409, Russian Federation
anton.kudyashov@gmail.com

Abstract. Service robots are increasingly surrounding us in various fields of activity, automating various processes. The introduction of robotic systems in an organization allows optimizing labor costs, increasing productivity and minimizing the cost of maintaining front offices. The purpose of this work is to develop a social robotic system capable of providing advice on the products and services of organizations, in particular in the banking sector, in terms of speed and quality comparable to a person. To do this, this article describes the process of creating a robotic system. Namely, the creation of the hardware of the robot, the design and development of the information architect of the robotic system. The analysis of the current business process of banking customer service and the creation of a business process using the ALKETON robotic system are also carried out. The developed robotic system was tested in a bank branch. Where the robot acted as a bank manager and provided services for some bank products. The developed robotic system showed results comparable to those of a human, and in some areas the assessment of the quality of the services provided surpassed that of a human.

Keywords: Robotics system · Microservice architecture · Social service robots · Bank service

1 Introduction

Artificial intelligence (AI) service robots include various types of robotic devices that serve people. These robots are equipped with artificial intelligence technology that allows them to sense external influences, understand information received from the environment, provide human-like reactions, and learn from this process [1]. In the US, Hilton McLean uses a concierge robot named Connie to answer questions from guests and make travel recommendations. A robotic butler named “Rosie” is used to assist at large food delivery events at the Lake Nona Wave Hotel [2]. A group of scientists studied the robotization of hotel services, the attitude and behavior of customers towards the use of social robots [3].

In the current conditions of fierce competition, banks are forced to constantly improve their business processes, form their own image and attract solvent customers.

According to a study by the foreign consulting company Cognizant, automation of customer interaction with the bank’s front offices using robotic systems saves 15%

of costs per year. It also reduces the time of customer service and the frequency of committed errors [4].

The purpose of this work is to develop a social robotic system capable of providing advice on products and services of organizations, in particular in the banking sector, in terms of speed and quality comparable to a person. To achieve this goal, the following research objectives were formulated:

1. Analysis of the current process of banking customer service;
2. Development of the hardware of the EVA robot;
3. Development of the server and software parts of the robotic system;
4. Approbation of the developed system on real clients in a bank branch.

The article uses the methods of object-oriented modeling, information modeling and the implementation of logical model objects.

2 Business Model of the Customer Banking Process

At present, the success of a commercial bank and its competitiveness are determined, first of all, not by price factors, but by quality characteristics, one of which is the quality of customer service for banking services [5]. The quality of service seriously affects the competitiveness of services and the bank, and hence the stability of income and profit. World experience and Russian practice show that investments in improving the quality of customer service are considered as intensive development and affect the stability of the bank’s income growth [6].

Also, before starting the development of a robotic system and optimizing the business process of banking customer service in a bank branch, an analysis of the current business process was carried out, which made it possible to identify bottlenecks in banking services. As a result of the analysis, a banking service business process model was developed, described in the BPMN 2.0 notation and presented in Fig. 1.

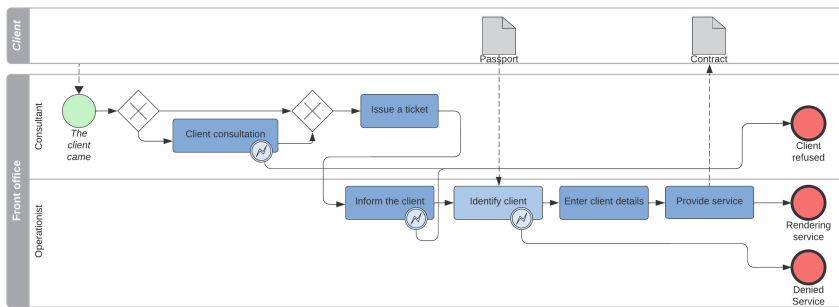


Fig. 1. The “as is” model of the banking customer service process

During the analysis of the model, the following bottlenecks in banking services can be identified:

- Longtime customer service;

- Incorrect input of client data;
- Providing the client with incomplete or not up-to-date information;
- Unreliable client identification;
- Unreliable assessment of quality while serving the client.

In order to solve the identified problems and tasks, it was decided to introduce a robotic system into the business process. The “to be” business model of the process of banking customer service using a robotic system was designed, shown in Fig. 2.

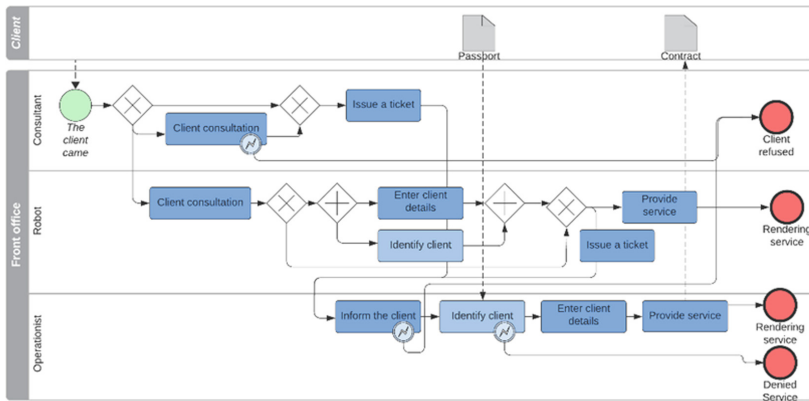


Fig. 2. Model “to be” of the process of banking customer service using a robotic system

As can be seen from the model, the client has an additional way to get advice on the products and services of the bank; during the dialogue, the system can display accompanying materials on the display. The system allows you to automate the input of the client’s passport data by recognizing the client’s passport. And it is possible to verify the client by comparing the photo from the passport or the previously saved photo of the client in the bank with the recognized face of the client. Throughout the dialogue, the robot analyzes the mood of the client, thereby evaluating the quality of the services provided. And also, there is an opportunity for the client to independently evaluate the quality of the services provided.

Thus, it is possible to unload the consultant and the teller, reducing the load, freeing them from routine work and entering the client’s passport data. The reliability of customer identification is increased due to double identification. It will be possible to reliably assess the quality of banking services provided.

3 Architecture of the Robotic System

3.1 General Architecture

The ALKETON robotic system consists of 5 main components: the EVA social robot, a display for displaying accompanying materials, a web application for system administration, a mobile application for manual control of the robotic system, and a cloud micro-service server responsible for data processing. The general architecture is shown in Fig. 3.

The EVA robot is able to communicate with customers, answer questions, maintain eye contact, identify people and recognize the emotion of the interlocutor in the process of dialogue. It is also possible to recognize and scan documents.

Cloud server - on a remote server, all requests received from different end devices of the system are processed.

Through the mobile application, manual control of the robotic system is carried out.

System administration is carried out through a web application. The application provides analytics of the robotic system, which displays the number of clients served, the rating of the quality of services provided, the response generation time and the conversion of bank visitors. It is possible to create dialogue scripts, add new robot movements and view the history of dialogues.

Near the robot, an optional display is installed, with specialized software installed. During the dialogue, accompanying materials on the products and services of the bank are displayed on the display. The history of the dialogue is also displayed and the recognized speech of the client is displayed in real time.

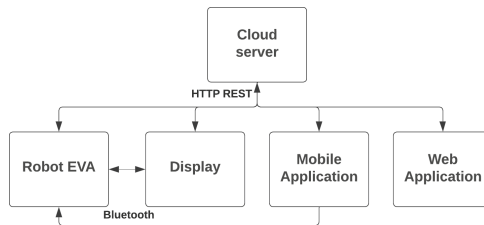


Fig. 3. General architecture of the robotic system ALKETON

3.2 Robot Microservice Architecture

Processing of incoming requests from the robot, mobile and web applications is carried out on a remote server. To implement the server part of the ALKETON robotic system, a microservice architecture was chosen. For easier scalability, reducing duplication of functionality and reducing the number of relationships between services, a paradigm was developed for dividing the system into two large conceptual layers of API and APP.

API layer - this layer implements consumer authentication and authorization, routing and orchestration of requests from consumers to the services of the APP layer. The services of the API layer are composite and provide interaction with the APP layer, implementing the data model and logic specific to the consumer cluster: orchestration, filtering, aggregation, calculations and data conversion into the required formats.

APP layer - solve a limited business function or task independent of other APP services, providing interaction with systems. The services of the APP layer are universal, i.e., they do not implement specific consumer logic. APP is the same for the entire system, its use is possible by various API layers implemented for different consumer clusters.

Using this paradigm, dividing the system into two levels, allows you to easily scale the system, increases the reuse of implemented services and makes the relationship

between services simpler [7]. The concept of a two-tier microservice architecture is provided in Fig. 4.

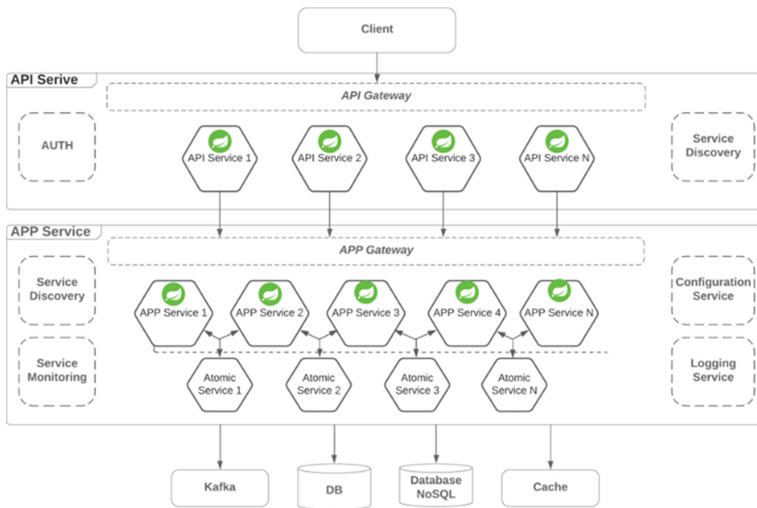


Fig. 4. The concept of a two-tier microservice architecture

Based on the proposed concept of a two-level architecture, the microservice architecture of the ALKETON robotic system was designed and developed, shown in Fig. 5. For example, when a request is received from the robot to generate an answer to a client’s question, the request enters the dialog-API service, this service, based on internal logic, organizes a request for the services of the app layer, collects the necessary information and gives the final response to the client.

Also in the architecture, two types of services can be distinguished: functional and infrastructural. Functional services are designed to directly perform the business functions of the system. Infrastructure services are designed to ensure the operation of functional services, in Fig. 5 they are indicated by a dotted line. They are responsible for configuring, routing requests, security and availability of functional services.

The designed architecture is implemented in the Java programming language using the Spring framework. Interaction between microservices is carried out by REST requests using the HTTP protocol. And for highly loaded services critical to time delays, the WSS protocol is used [8].

3.3 Hardware of the EVA Robotic System

The social robot “EVA” was designed and developed, the design was carried out in the Autodesk Fusion 360 3D modeling program, and then it was printed on a 3D printer using PLA plastic. The robot has 32 RGB LEDs, 6 degrees of freedom, and is capable of performing various movements. Two servos are installed to rotate the robot’s head along two horizontal and vertical axes. Also, two servos are installed in the hands, which

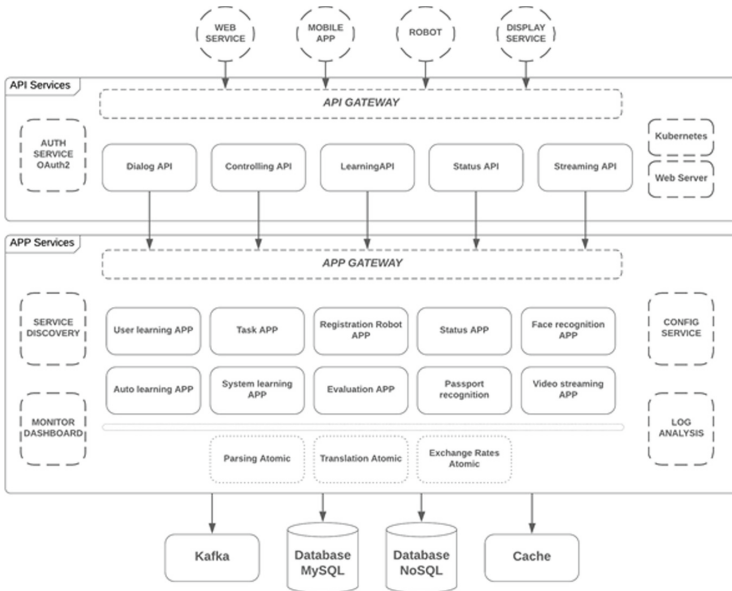


Fig. 5. ALKETON robotic system architecture.

allows you to raise and bend the robot’s arms. All this allows you to interact interactively with customers.

The robot can operate in fully automatic mode, as well as in manual mode, exercising control from a mobile application. The robot is equipped with an Ethernet and Wi-Fi module that allows you to connect to the microservice server. The robot can recognize and synthesize speech in 19 languages including Russian. The robot has a powerful acoustic system and an ultra-sensitive microphone for speech recognition in a noisy environment away from the robot.

It has a high-resolution vision system that allows you to find and track the facial expressions of the interlocutor. In combination with a cloud server, it allows you to determine the gender, age and mood of the interlocutor, which allows you to correct and adjust the generated response for each client. The robot is able to recognize a person’s face and maintain eye contact with people in the process of dialogue.

Several sensors are installed in the EVA robot to obtain information about the environment. The motion sensor allows you to activate the human detection search and the speech recognition system. The light level sensor allows you to adjust the level of illumination of the LEDs so as not to dazzle customers with bright light in the evening.

In the robot, the Raspberry Pi 4 minicomputer is responsible for processing the data received from the sensors and the server. The processed data is transferred to the Arduino Mini microcontroller for execution, and the main part of the sensors and actuating units is connected to it [9]. The microcontroller communicates with a mini computer via an asynchronous UART data transfer channel, in JSON format. The hardware architecture of the robot is shown in Fig. 6.

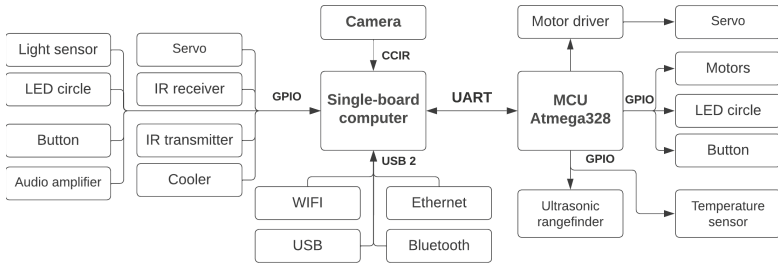


Fig. 6. EVA robot hardware architecture

The EVA robot has a powerful hardware platform. The robot body contains 8 servos with a rotation range of 180°, 6 of which have a maximum torque of 2 kg × cm and are installed in the hands, and 2 servos are tired are installed in the neck of the robot and have a torque of 13 kg × cm. A Raspberry Pi 4 minicomputer is installed under the Android Things operating system.

4 Approbation of the Developed Robotic System

The developed robotic system was tested in a branch of a large bank. In this experiment, the robot acted as a consultant in a bank branch, and could provide services for the following products: issuing a bank card, issuing a loan, opening a deposit, payments and transfers. A study was conducted based on a survey of bank customers, after the provision of services by a robot, customers were asked to evaluate the quality of the services provided on a 10-point scale. 120 bank clients took part in the exhibitor. For analysis, the results of all respondents were divided into four segments:

- mass clients—81 people (67.5%);
- pensioners—25 people (20.8%);
- students—9 people (7.5%);
- commercial clients—5 people (4.2%).

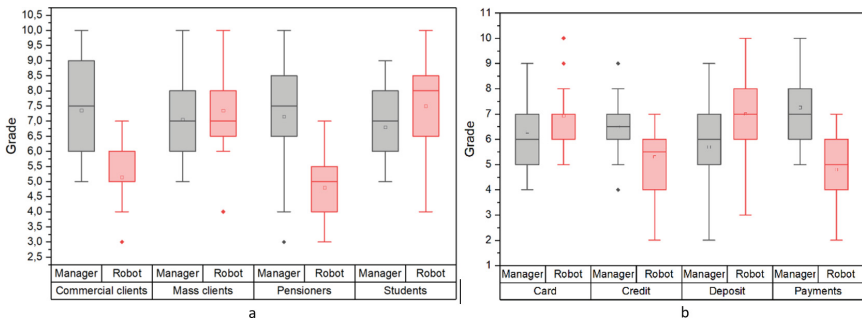


Fig. 7. Quality analysis a) by customer segments; b) bank products

The results of the assessments are shown in Fig. 7.a. Based on the results obtained, the following conclusions can be drawn. The average rating of the quality of the services provided by the robot is 6.38, and the average rating of the manager is 7.13, which is 8% less than the consultant. However, in the segment of students and mass clients, the robot outperforms the consultant by 9% and 5%, respectively. What is associated with a reduction in service time and an innovative approach to customer service.

An analysis was also carried out in the context of the functions provided by the robot, namely: issuing a bank card, obtaining a loan, opening a deposit, payments and transfers. The results obtained are shown in Fig. 7.b.

According to the function of issuing a bank card, the robot received approximately the same score. When processing loan documents, the robot showed a significantly worse result by 12% compared to the manager, which is due to customer questions that the robot could not give an accurate answer to. When registering an account and a deposit, the robot received a 13% higher score than the manager, which is due to the speed of registration and the transparency of the conditions that do not require additional clarification. In terms of payments and transfers, the robot turned out to be much worse than the manager, which is connected with numerous questions on which the robot was not trained. In general, it can be seen that the robot received a good score of 6.3, against the manager who received an average of 7.1 out of 10 points. Figure 8 shows a robot advising a client.



Fig. 8. Consulting a bank client with a robot

5 Conclusion

As a result of the study, the ALKETON social robotic system was developed, which is able to advise and provide services to bank customers. For this, an EVA robot was developed, made using 3D printing. Designed and developed a two-level microservice architecture of a robotic system that processes requests received by the robot. An analysis was made of the current business process of serving clients at a bank's remote location and a business process was developed using a robotic system. As a result, the developed system was tested to separate the bank from its own clients.

The results of the study show that the robot easily performs the functions of a bank manager, namely, it advises clients on bank products, draws up bank cards, and generates documents for opening an account and obtaining a loan. According to the results of the analysis on the quality of service of the robotic system as a whole, it showed a lower result compared to people by 8%. However, when opening a deposit, the average speed of the robot exceeded that of a human by 13%, which is due to the speed of providing services and providing complete information on the product.

References

1. Bowen, J., Morosan, C.: Beware hospitality industry: the robots are coming. *Worldw. Hosp. Tour. Themes* **10**(6), 726–733 (2018). <https://doi.org/10.1108/WHATT-07-2018-0045>
2. Chi, O.H., Chi, C.G.: Customer's acceptance of artificially intelligent service robots: the influence of trust and culture. *Int. J. Inf. Manage.* **70** (2023). <https://doi.org/10.1016/j.ijinfomgt.2023.102623>
3. Corte, V.D., Sepe, F., Gursoy, D., Prisco, A.: Role of trust in customer attitude and behaviour formation towards social service robots. *Int. J. Hosp. Manag.* **114** (2023). <https://doi.org/10.1016/j.ijhm.2023.103587>
4. Agarwal, A., Rohan, A., Kriti, D., Meryl, C.: Future of robotics in banking. *Int. J. Inf. Futuristic Res.* **4**(5) (2017). ISSN: 2347-1697
5. Lebedeva, N.S., Pavlyuchenkov, D.N.: Problems and prospects of bank customer service. http://vfmgju.ru/files/23_11_2007_27.pdf, 2023/08/01
6. Khakimov, E.A.: Analysis of the quality of customer service in a commercial bank. *Bull. Chelyabinsk State Univ.* **6**(221) (2011)
7. Kudriashov, A.V.: Two-tier architecture of the distributed robotic system «ALKETON». In: 2022 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: The 13th Annual Meeting of the BICA Society. *Procedia Comput. Sci.* **213C**, 816–823 (2022). ISSN: 1877-0509. <https://doi.org/10.1016/j.procs.2022.11.139>
8. Martin, K.: *Highly Loaded Applications. Programming, Scaling, Support.* Peter (2019)
9. Petin, V.: *Arduino and Raspberry Pi in Internet of Things Projects.* BHV (2016)



Exploring the Efficiency of Neural Networks for Solving Dynamic Process Problems: The Fisher Equation Investigation

Raul Karachurin¹, Stanislav Ladygin^{1(✉)}, Pavel Ryabov¹, Kirill Shilnikov^{1,2},
and Nikolay Kudryashov¹

¹ National Research Nuclear University MEPHI, 31, Kashirskoe shosse, 115409
Moscow, Russia

SALadygin@mephi.ru

² Department of Computational Physics, Moscow Institute of Physics and
Technology (MIPT), 9 Institutskiy per., Dolgoprudny, Moscow Region 141701,
Russian Federation

Abstract. The numerical solution of problems based on ordinary and partial differential equations has been a subject of extensive research. While several methods, such as the finite difference, finite element, and finite volume methods, have been developed, each has its own strengths and limitations. This paper presents a different approach that utilizes feedforward neural networks to approximate functions, resulting in a differentiable analytical expression. Compared to other methods, this approach requires significantly fewer model parameters, leading to reduced computational requirements. The study examines the influence of a neural network's loss function configuration on the accuracy and convergence rate of solving partial differential equations for functions of two variables: coordinate and time, using the example of the Fisher's equation.

Keywords: Computational mathematics · Numerical methods · Differential equations · Approximation · Neural networks · Activation functions · Dynamic loss function · Fisher's equation

1 Introduction

In various scientific disciplines, numerous complex phenomena exhibit intricate processes that require the resolution of diverse problems that formulated from differential equations. For example: a one-dimensional non-stationary problem within magnetohydrodynamics [1], dynamics of strongly coupled spiking neurons in neuroscience [2], and modeling tumor growth using reaction-diffusion equations in mathematical biology [3]. Throughout the years, many of numerical methods has been developed to solve these differential equations. The most popular of them are:

- The finite difference method [4]. The solution of the problem is reduced to solving finite difference equations, as a result of which the answer is obtained in the form of an array specified at certain points in the domain.
- The finite element method [5]. This approach uses basic functions to analytically represent the solution and transform the original problem into a system of linear equations.
- The finite volume method [6]. It is based on the integration of the original equation over the control volume, which corresponds to a section of the computational domain adjacent to the designated computational node.

In this context, a general approach to solving differential equations that capitalizes on the capability to approximate functions using feedforward neural networks has been investigated. This concept was initially introduced by Lagaris et al. [7], wherein a feedforward neural network serves as a fundamental approximation component, and the network's parameters (weights and biases) are adjusted to minimize the associated *loss* function.

This method has notable advantages: it provides a differentiable analytical expression for the solution, requires fewer model parameters, works for both ODEs and PDEs, and can be efficiently implemented through parallelization and execution on graphics accelerators.

There exists an extensive body of research dedicated to the aforementioned method [8,9]. However, the majority of studies resort to the application of a standard form of the loss function [10,11]. There are also works that impose additional constraints, altering the loss function. For instance, conservation laws [12]. In our study, we investigate the effect of weights in the loss function on the accuracy of the solution.

This paper is structured as follows. The first section provides a comprehensive description of the employed methodology. In the subsequent section, we examine the Fisher's equation [13], as an example, also known as the Kolmogorov-Petrovsky-Piskunov equation [14], which serves as a model for population growth and wave propagation. Through the investigation of this equation, we assess the accuracy of the obtained results by employing various activation functions, analyse the impact of weights in the *loss* function, and also compare with another numerical approach.

2 Method Description

Consider the following differential equation in general form:

$$G(\vec{x}, u(\vec{x}), \nabla u(\vec{x}), \nabla^2 u(\vec{x}), \dots, \nabla^m u(\vec{x})) = 0, \quad \vec{x} \in D, \quad (1)$$

where $\vec{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$, $D \subset \mathbb{R}^n$ is the region of definition, and $u(\vec{x})$ is the desired solution. For Eq. (1), set the boundary conditions as:

$$B(\vec{x}, u(\vec{x}), \nabla u(\vec{x}), \nabla^2 u(\vec{x}), \dots, \nabla^{m-1} u(\vec{x})) = 0, \quad \vec{x} \in \partial D. \quad (2)$$

The neural network will approximate the solution to the true solution at some specified set of points in the domain D and on its boundary ∂D . To do this, we will select discrete sets of points \hat{D} and $\partial\hat{D}$ from the domain D and its boundary ∂D , respectively. We will also introduce a function $y(\vec{x})$, which is computed using the trainable neural network and is an approximation of the function $u(\vec{x})$. Then during training, the neural network should strive to satisfy the following expressions:

$$\begin{cases} G(\vec{x}_i, y(\vec{x}_i), \nabla y(\vec{x}_i), \nabla^2 y(\vec{x}_i), \dots, \nabla^m y(\vec{x}_i)) = 0, & \forall \vec{x}_i \in \hat{D}, \\ B(\vec{x}_j, y(\vec{x}_j), \nabla y(\vec{x}_j), \nabla^2 y(\vec{x}_j), \dots, \nabla^{m-1} y(\vec{x}_j)) = 0, & \forall \vec{x}_j \in \partial\hat{D}. \end{cases} \quad (3)$$

The neural network seeks parameters (such as weights and biases) in such a way as to minimize the *loss* function, which has the following form:

$$loss = \lambda_b \sigma_b^2 + \lambda_e \sigma_e^2, \quad (4)$$

where

$$\sigma_b^2 = \sum_{\vec{x}_j \in \partial\hat{D}} [B(\vec{x}_j, y(\vec{x}_j), \nabla y(\vec{x}_j), \nabla^2 y(\vec{x}_j), \dots, \nabla^{m-1} y(\vec{x}_j))]^2, \quad (5)$$

$$\sigma_e^2 = \sum_{\vec{x}_i \in \hat{D}} [G(\vec{x}_i, y(\vec{x}_i), \nabla y(\vec{x}_i), \nabla^2 y(\vec{x}_i), \dots, \nabla^m y(\vec{x}_i))]^2. \quad (6)$$

Here, σ_e^2 represents the mean squared error on internal points, and σ_b^2 represents the mean squared error on the boundary. λ_b and λ_e are weights that determine the priorities of approximating the function values on the boundary or inside the domain. Usually, their values are chosen to be equal to 1, but in this work, we will investigate the influence of these coefficients on the accuracy and convergence rate of the obtained results.

Equations (5) and (6) show that to obtain the value of the *loss* function, it is necessary to be able to compute the gradients of the function $y(\vec{x})$. There are several approaches to do this, but the most efficient and versatile one is automatic differentiation [15]. Let us assume that we know some “basis” functions and the values of their derivatives. Then, for any complex function that is defined through these “basis” functions, we can automatically compute its value and the value of its derivatives. For our case, the function $y(\vec{x})$, which approximates the solution $u(\vec{x})$, is a superposition of activation functions, which serve as those “basis” functions. Therefore, there is a natural limitation on these functions.

Remark 1. To approximate the solution $u(\vec{x})$ of the differential equation (1) with boundary conditions (2) using a neural network, its activation functions should be m times continuously differentiable and their derivatives should not be identically zero.

If this condition is met, then all gradients of the function $y(\vec{x})$ will be automatically computed from the known values of the activation function derivatives.

3 Example (Fisher's Equation)

In this example, consider the Fisher's equation

$$u_t = \delta u_{xx} + u(1 - u). \quad (7)$$

Here the parameter δ is a real number. This equation belongs to the class of non-linear reaction-diffusion equations, which can exhibit travelling wave solutions and are encountered in numerous scientific domains.

We will also consider one of the analytical solutions of this equation, which was obtained in the article [16]

$$u_{\text{exact}}(x, t) = \left(\frac{\exp\left(-x/\sqrt{6\delta} + 5t/6\right)}{1 + \exp\left(-x/\sqrt{6\delta} + 5t/6\right)} \right)^2. \quad (8)$$

We will set a boundary value problem for this Eq. (7), which will have an analytical solution (8):

$$\begin{cases} u_t = \delta u_{xx} + u(1 - u), & -l < x < l, \quad 0 < t \leq T, \\ u|_{t=0} = u_{\text{exact}}(x, 0), \\ u|_{x=-l} = u_{\text{exact}}(-l, t), \\ u|_{x=l} = u_{\text{exact}}(l, t). \end{cases} \quad (9)$$

Here the constants l and T are real numbers.

Consider a uniform spatial grid with N points on variable x and K points on variable t . Then the function *loss*, according to (4-6), has the following form:

$$\begin{aligned} \text{loss} = & \lambda_e \sum_{x_i, t_n \in \Omega} \left[y_t|_{(x_i, t_n)} - \delta \cdot y_{xx}|_{(x_i, t_n)} - y|_{(x_i, t_n)} \cdot (1 - y|_{(x_i, t_n)}) \right]^2 + \\ & + \lambda_b \sum_{t_n \in \omega_\tau} \left[y|_{(-l, t_n)} - u_{\text{exact}}(-l, t_n) \right]^2 + \lambda_b \sum_{t_n \in \omega_\tau} \left[y|_{(l, t_n)} - u_{\text{exact}}(l, t_n) \right]^2 + \\ & + \lambda_b \sum_{x_i \in \omega_h} \left[y|_{(x_i, 0)} - u_{\text{exact}}(x_i, 0) \right]^2. \quad (10) \end{aligned}$$

When training a neural network, the parameter δ was taken for 0.06. The constant l , which is responsible for the length of the calculated space, was taken as 5. The calculation time T was taken as 10. The number of grid points for the variable x (N) and for the variable t (K) were taken as 100. The Adam method was used as the neural network optimizer. The model of this neural network comprises three internal layers with 16, 64, and 16 neurons, respectively.

To determine the optimal neural network model, several trainings with different activation functions were performed. The values of weights λ_e and λ_b were taken as 1. The results of the training can be seen in the graph of the function *loss* (Fig. 1).

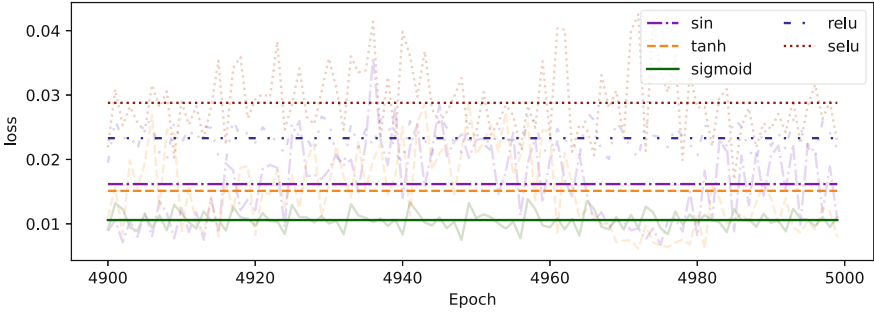


Fig. 1. A graph of the average value of the *loss* function for the last hundred epochs, when using various activation functions

As we can see, the sigmoid activation function showed the best result with a training duration of 5,000 epochs for this task. Subsequent training was conducted on the neural network utilizing this activation function, with varying weights ratios of λ_e and λ_b . The quantity of points at the x (N) and t (K) coordinates were varied from 10 to 500, incremented by 5. Consequently, Table 1 presents the statistical analysis of our sample for each weight ratio.

Table 1. Statistical data at different grid sizes for different weights ratios $\frac{\lambda_e}{\lambda_b}$

$\frac{\lambda_e}{\lambda_b}$	mean (max error)	min (max error)	max (max error)	SD (max error)
2^{-6}	0.043186	0.030748	0.047029	0.004376
2^{-5}	0.029060	0.028094	0.029761	0.000390
2^{-4}	0.026286	0.023548	0.026763	0.000900
2^{-3}	0.020031	0.015753	0.020483	0.001353
2^{-2}	0.016120	0.015393	0.021687	0.001763
2^{-1}	0.011815	0.011764	0.011921	0.000043
2^0	0.010233	0.010055	0.010929	0.000233
2^1	0.009514	0.008571	0.011049	0.000568
2^2	0.009668	0.008281	0.011789	0.001037
2^3	0.006665	0.004448	0.008956	0.001193
2^4	0.005923	0.004659	0.006853	0.000574
2^5	0.017278	0.009393	0.028967	0.005963
2^6	0.101550	0.003179	1.043280	0.297820

As you can see, the best average result shows the ratio of weights 2^4 . The most stable error was obtained with a ratio of weights 2^{-1} . With it, the lowest standard deviation can be observed.

Consider a comparative analysis of the absolute error between the numerical solution obtained via the neural network method, with a loss function weight ratio of 2^4 , and the finite difference method. The finite difference method employs an implicit scheme with first-order temporal accuracy and second-order spatial accuracy. The grid dimensions utilized are 100×100 . The discrepancy in errors between these methods can be discerned visually by examining the heat maps of absolute errors (Fig. 2).

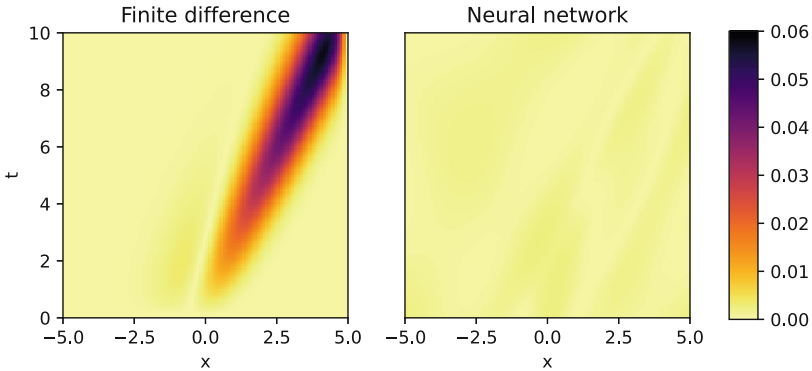


Fig. 2. Heat maps of the absolute error for finite difference method and neural network method

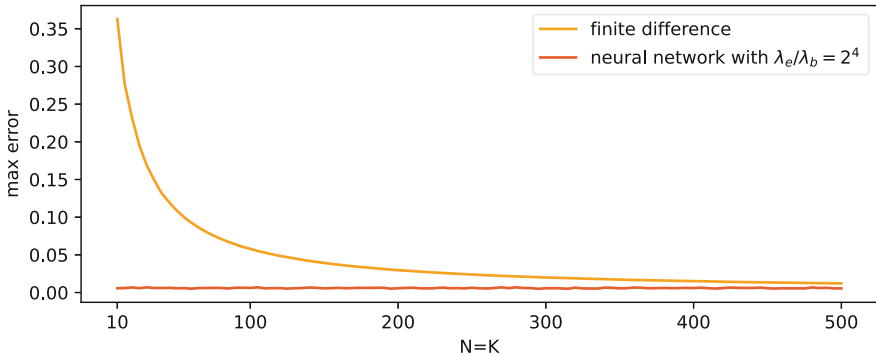


Fig. 3. Dependence of the maximum error on the number of points for two different methods

As can be seen, almost everywhere the neural network method has an error an order of magnitude lower than the finite difference method. Also, the advantage of the neural network approach in comparison with finite differences can be seen in Fig. 3. It can be seen here that the neural network approach does not have

such a strong dependence on the number of points on the grid, unlike the finite difference method. Therefore, to ensure the same accuracy in the neural network method, much fewer points on the grid can be used, which increases the speed of calculations.

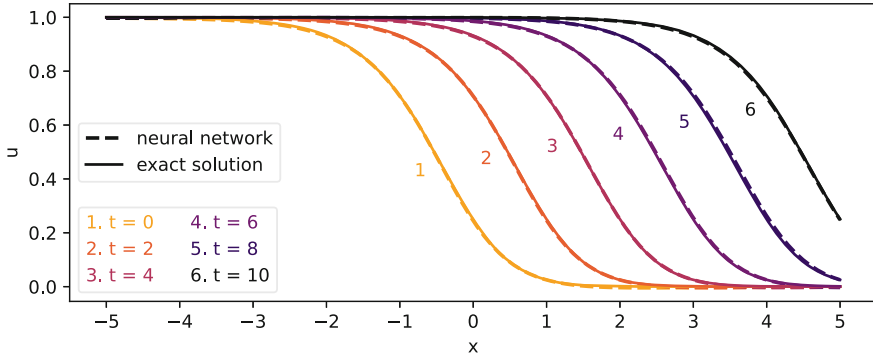


Fig. 4. A graph of the exact solution and the one obtained by the neural network method at different points in time

We will also give here the solution itself by the neural network method, in comparison with the exact one (Fig. 4).

This research was supported by Russian Science Foundation Grant No. 23-41-00070, <https://rscf.ru/en/project/23-41-00070/>.





References

1. Samarskii, A.A., Volosevich, P.P., Volchinskaya, M.I., Kurdyumov, S.P.: A finite-difference method for the solution of one-dimensional non-stationary problems in magneto-hydrodynamics. *USSR Comput. Math. Math. Phys.* **8**(5), 117–134 (1968)
2. Bressloff, P.C., Coombes, S.: Dynamics of strongly coupled spiking neurons. *Neural Comput.* **12**(1), 91–129 (2000)
3. Greenspan, H.P.: Models for the growth of a solid tumor by diffusion. *Stud. Appl. Math.* **51**(4), 317–340 (1972)
4. Samarskii, A.A., Nikolaev, E.S.: *Methods for solving grid equations* (1978)
5. Zienkiewicz, O.C., Morice, P.B.: *The Finite Element Method in Engineering Science*, vol. 1977. McGraw-Hill, London (1971)
6. Kovenya, V.M., Chirkov, D.V.: *Methods of Finite Differences and Finite Volumes for Solving Problems of Mathematical Physics*, pp. 7–8. Novosibirsk State University, Novosibirsk (2013)
7. Lagaris, I.E., Likas, A., Fotiadis, D.I.: Artificial neural networks for solving ordinary and partial differential equations. *IEEE Trans. Neural Networks* **9**(5), 987–1000 (1998)

8. Ladygin, S.A., Karachurin, R.N., Ryabov, P.N.: Numerical approach for studying problems for differential equations based on neural network method. In: IX International Conference “Laser, Plasma Research and Technologies” LaPlaz-2023: Proceedings, Conference Series LaPlaz, Publisher National Research Nuclear University MEPhI (Moscow), Abstracts, p. 147 (2023)
9. Ladygin, S.A., Karachurin, R.N., Ryabov, P.N., Kudryashov, N.A.: On the features of a numerical approach based on neural networks with direct communication for solving problems for differential equations. *Phys. Atomic Nuclei* (in press)
10. Wu, G., Wang, F., Qiu, L.: Physics-informed neural network for solving Hausdorff derivative Poisson equations. *Fractals*, 2340103 (2023)
11. Uddin, Z., Ganga, S., Asthana, R., Ibrahim, W.: Wavelets based physics informed neural networks to solve non-linear differential equations. *Sci. Rep.* **13**(1), 2882 (2023)
12. Fang, Y., Wu, G.Z., Kudryashov, N.A., Wang, Y.Y., Dai, C.Q.: Data-driven soliton solutions and model parameters of nonlinear wave models via the conservation-law constrained neural network method. *Chaos Solitons Fractals* **158**, 112118 (2022)
13. Fisher, R.A.: The wave of advance of advantageous genes. *Ann. Eugen.* **7**(4), 355–369 (1937)
14. Kolmogorov, A.: Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application á un problème biologique. *Moscow Univ. Bull. Math.* **1**, 1–25 (1937)
15. Ketkar, N., Moolayil, J., Ketkar, N., Moolayil, J.: Automatic differentiation in deep learning. In: *Learn Best Practices of Deep Learning Models with PyTorch, Deep Learning with Python*, pp. 133–145 (2021)
16. Kudryashov, N.A.: One method for finding exact solutions of nonlinear differential equations. *Commun. Nonlinear Sci. Numer. Simul.* **17**(6), 2248–2253 (2012)



Improving the Methodology for Integrated Testing of Journal Entries by Benford's Law

Pavel Y. Leonov^(✉) , Viktor M. Sushkov , Sofia A. Boiko ,
and Margarita A. Stepanenkova 

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute),
Moscow, Russia
pyleonov@mephi.ru

Abstract. This paper explores the growing prevalence of Benford's Law as a statistical method to identify intentional manipulations of numerical data. The study focuses on improving a methodology for applying Benford's Law tests in detecting distortions within accounting practices. Primary, advanced, and associated tests are conducted to assess the natural character of journal entries of a construction company. Additionally, machine learning techniques such as K-means clustering, random forest, and elliptic envelope are used to analyze the test results and identify highly suspicious transactions within the dataset. The outcomes indicate that the selected transactions flagged by the tests are indeed suspicious, with significant monetary values.

Keywords: Benford's Law · Statistics · Data mining · Economic security

1 Introduction

Currently, fraudulent activities in accounting that result in distorted financial statements have become a pervasive issue and occupy a prominent position in global statistics on economic crimes. According to the report of the Association of Certified Fraud Examiners (ACFE) [1], financial reporting fraud schemes, in which the perpetrator intentionally introduces material misstatements or omissions in the organization's accounting, bring the greatest damage to legal entities. The median loss of such crimes is 593 thousand US dollars.

The significance of this problem lies in the fact that an inaccurate presentation of the financial position not only results in substantial financial losses for companies, but also misleads investors and erodes confidence in the entire economic system. In this regard, there is a need to apply various approaches to the analysis of accounting data in order to identify signs of fraud. One of the most effective ways to detect fraudulent activities and suspicious transactions is to check the conformity of journal entries to the Benford's Law. By combining machine learning techniques with Benford's Law tests, it becomes feasible to identify the most suspicious transactions and reduce the size of the sample to be analyzed manually.

2 Analytical Part

2.1 Materials and Methods

The application of Benford's Law to identify signs of financial fraud was started with an article published in 1972 by the American economist H. Varian [2]. The general idea of H. Varian was that a simple comparison of the frequency distributions of the first digit in socio-economic data can help reveal anomalous data. 20 years later, in 1992, the American scientist M. Nigrini in his Ph.D. thesis [3] applied Benford's Law to the amounts from income tax returns in order to identify unscrupulous taxpayers.

At the end of the last century, Benford's Law started gaining popularity in relation to financial reporting. In 1988, C. Carslaw applied Benford's Law in a study of financial indicators for a series of New Zealand companies. The study revealed that the digit 0 was used too frequently as the second digit, while the digit 9 was underutilized, thus deviating from Benford's statistical frequencies. As a result, C. Carslaw concluded that financial statement users tend to view a profit of 50 million more favorably than a profit of 49 million, leading management to round profit values accordingly [4].

Over the past two decades, numerous publications have discussed the theoretical and practical aspects of applying Benford's Law in various areas of economic activity. These studies have laid the groundwork for the hypothesis that in committing fraudulent transactions individuals, due to psychological habits and situational factors, devise nonstandard combinations of numbers that deviate from the expected natural digit frequencies.

Based on the Benford distribution, M. Nigrini and L. Mittermaier in 1997 [5] developed statistical tests that allow checking the natural character of the data array. In a later edition of 2012, "Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection" [6] M. Nigrini classifies tests into primary, advanced, associated. Formulae for expected numeric frequencies for primary tests are presented in Table 1 (d stands for digits).

Table 1. Primary tests

Test	Formula
The first digit test	$P(d_1) = \log_{10}\left(1 + \frac{1}{d_1}\right); d_1 = 1, \dots, 9$
The second digit test	$P(d_2) = \log_{10}\left(1 + \frac{1}{d_2}\right); d_2 = 0, \dots, 9$
The first-two digits test	$P(d_1 d_2) = \log_{10}\left(1 + \frac{1}{d_1 d_2}\right); d_1 d_2 = 10, \dots, 99$

Advanced tests include a summation test and a second-order test. The summation test aims to identify discrepancies between the observed frequency of the first digits of numbers and the expected frequency based on the average. The second-order test is based on the digits of the difference between the numbers pre-ordered from smallest

to largest. According to M. Nigrini, the distribution of the differences obtained should approach the Benford distribution.

Two associated tests are the number duplication test and the last-two digits test. These tests differ from Benford's Law as they are founded on different characteristics of number distributions. The number duplication test focuses on identifying specific numbers that contribute to outliers in the diagrams produced by the first-two digits test and summation test. The last-two digits test examines abnormal values on the right side of numbers. Uncommon values observed in the graph generated by the last-two digits test may suggest errors, fabricated numbers, or excessive rounding.

The most frequently used statistical approach to assess conformity with Benford's Law is through the calculation of Z-statistics. It is used to check the compliance of individual data values with Benford's Law. Z-statistics formula (1) considers the absolute difference between the observed and expected frequencies, as well as the sample size.

$$Z = \frac{|AP - EP| - \frac{1}{2N}}{\sqrt{\frac{EP(1-EP)}{N}}} \quad (1)$$

In (1) AP is the actual frequency, EP is the expected frequency, N is the number of values in the sample. $1/2N$ is a continuity correction, which is used only when it is less than the first term in the numerator. For a significance level of 5%, the threshold level is 1.96.

The methodology for applying Benford's Law to identify signs of accounting fraud proposed by M. Nigrini is researched and further developed in Russian practice. In [7] it was proved that in the absence of attempts to manipulate the reporting, the distribution of the company's expenses for supplier services corresponds to the theoretical one, while distortions in the reporting are reflected in the deviations from Benford's Law. In [8] the application of Benford's Law made it possible to detect suspicious transactions of a road repair company and subsequently uncover a fraudulent scheme.

2.2 Problems

M. Nigrini's methodology, despite its widespread use, encounters certain issues that need to be addressed. Particularly, when dealing with large volumes of data, the selection of suspicious transactions based on M. Nigrini's criteria still leaves a substantial number of samples that necessitate additional processing. To mitigate this issue, we propose to employ mathematical algorithms that can effectively reduce the amount of analyzed information. Some algorithms that can be utilized for this purpose include:

- Cluster analysis, which allows splitting many objects into similar groups;
- Outlier analysis, which allows identifying the objects that are most different from the total population.

2.3 Results

During the study, a clustering model was built in order to identify a group of suspicious transactions. The model developed in this study was applied to a diverse range of companies spanning various industries. As an illustrative example, we present the results

of the analysis conducted on a construction company, which is considered one of the industries with a high susceptibility to fraud.

The input for the study was a set of journal entries from the general ledger for a 4-year period, containing 144 thousand records. The data was sourced from an auditing company. To uphold confidentiality and protect the privacy of individuals and organizations involved, all names have been anonymized. The initial data for building the model were the following statistics calculated for each transaction amount:

- X1: Z-statistics for the first digit test;
- X2: Z-statistics for the second digit test;
- X3: Z-statistics for the first-two digits test;
- X4: Frequency of transaction amounts under the first two digits;
- X5: Z-statistics for the second-order test;
- X6: Frequency of all transaction amounts;
- X7: Z-statistics for the last-two digits test.

Clustering of the data was carried out using K-means. This is an iterative clustering algorithm based on minimizing the total quadratic deviations of cluster points from the centroids of these clusters. The choice of the method is due to its simplicity of implementation, scalability to huge datasets, high learning speed and support for complex shapes and sizes of clusters.

By utilizing clustering for classification, we enhance the accuracy and effectiveness of the model by incorporating additional information derived from these clusters. This approach enables the model to capture the inherent complexity and diversity present in the dataset, leading to more accurate and robust classification results. To determine the optimal number of clusters, the silhouette metric was calculated (see Fig. 1).

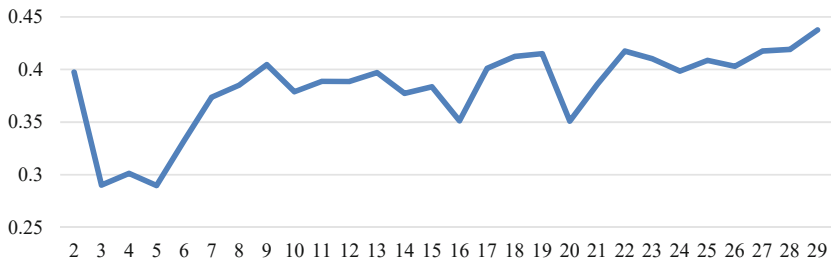


Fig. 1. Dependence of the silhouette on the number of clusters

When splitting into 2 clusters, a peak appears on the graph for the first time. When divided into 3–5 clusters, the value of the quality functional decreases sharply, and with a further increase in the number of clusters increases again to the previous value. Thus, an increase in the number of clusters slightly affects the quality functional of the partition, but significantly complicates the interpretation of the results, so we choose the number of clusters equal to 2.

The average values of features in clusters are presented in Fig. 2.

Cluster 0 includes only 15.5% of observations and is characterized by high values of Z-statistics for the first digit test, second digit test, first-two digits test, and last-two

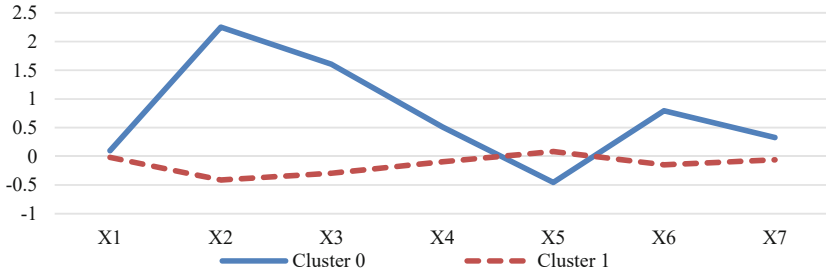


Fig. 2. Average values of features in clusters

digits test. Additionally, there are many duplicate operations and operations marked as suspicious by the summation test in this cluster. The low value of Z-statistics for the second-order test is due to the fact that this test is performed only for those sums that, when sorted in ascending order, are less than the previous value by more than 10, and the missing values were assigned the value -1 . Thus, this cluster includes the amounts of operations that were marked as suspicious by all tests and could not be verified using a second-order test. These operations are considered to be at the highest risk.

The partition in principal components is visualized in Fig. 3.

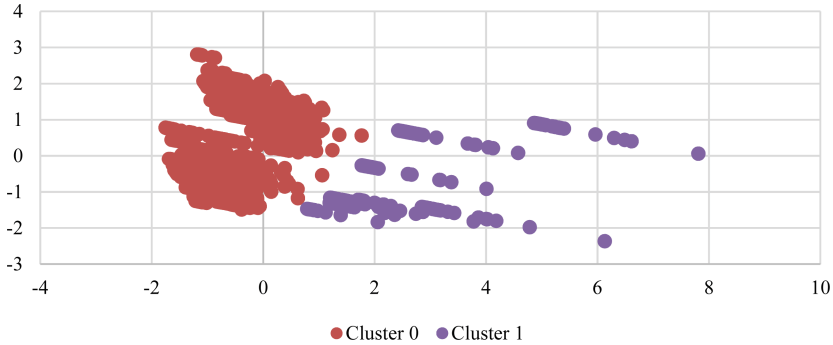


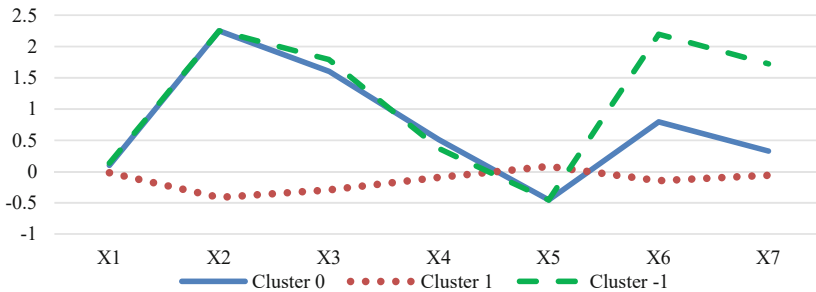
Fig. 3. Visualization of partitions in principal components

Cluster 0 operations are indeed different from the rest. In order to make sure that cluster 0 operations are suspicious, we performed an outlier analysis using two anomaly detection methods: isolation forest and elliptical envelope. Table 2 presents a comparison between the results obtained from these methods and the outcomes from clustering, with the outlier cluster being labeled as -1 .

The isolation forest flagged all operations within cluster 0 as outliers, whereas the elliptic envelope classified only half of the operations in this cluster as outliers. Both methods, however, correctly identified that the operations within the larger cluster 1 are not outliers. Figure 4 illustrates the average values of the features for the operations identified as outliers by all three methods.

Table 2. Comparison of the results

Elliptical envelope	Isolation forest	K-means clustering	% of the total number of objects
- 1	- 1	0	8.2%
1	- 1	0	7.3%
1	1	1	84.5%

**Fig. 4.** Average values of features in clusters

72% of suspicious cluster operations have abnormally round amounts, which is suspicious according to paragraph A43 of the International Standard on Auditing (ISA) 240. Especially suspicious are the round amounts contained in contracts for the supply of raw materials, for example, “Payment under the contract... for diesel fuel” in the amount of 4,000,000 rubles, “Payment under the contract... for crushed stone” in the amount of 1,000,000 rubles. The average cost of a liter of diesel fuel during that period was 32.51 rubles. Calculations for large deliveries are usually done in increments of at least 100 L, and it is not possible to obtain a figure that is a multiple of a million under such conditions. This suggests that there has been an overestimation in the cost of materials and services.

Figure 5 presents a comparison between the theoretical and actual frequency distribution of the first-two digits. The x-axis represents the digits ranging from 10 to 99, while the y-axis represents the frequency of occurrence for each digit. The theoretical frequency distribution, based on Benford’s Law, would show a gradual decrease in frequency as the digits increase from 10 to 99. However, the actual frequency distribution depicted in Fig. 5 reveals that the frequencies of sums starting with digits 10, 30, 50, and 60 are significantly higher than expected. Additionally, the frequencies after the observed peaks decrease much more rapidly than what would be anticipated according to Benford’s Law.

It therefore can be concluded that such a distortion is probably caused by rounding the transaction amounts down. Thus, considering the cost of goods or services is below the “fair market” price in the documents, the company reduced the amount of taxable income.

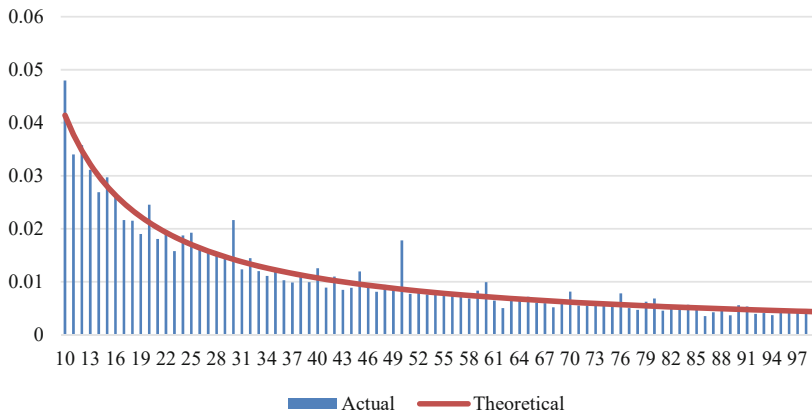


Fig. 5. Theoretical and actual frequency distribution of the first-two digits

3 Conclusion

The application of mathematical methods has proven to significantly enhance the productivity of organizations that conduct activities in the field of financial control. The proposed approach makes it possible to free professionals from routine transaction checks, thereby allowing them to narrow their focus onto operations necessitating additional procedures. The results have demonstrated high effectiveness, as the operations identified through the tests were found to be suspicious, with significant monetary values.

By combining Benford's Law with machine learning techniques, this research offers a new perspective on how to analyze and interpret data in the context of fraud detection. However, further research is needed to optimize machine learning algorithms for processing large arrays of journal entries and improving the accuracy of identifying suspicious operations. Additionally, it is worth further investigating and assessing the practical implementation and scalability of this approach in various organizational contexts.





References

- Occupational Fraud 2022: A Report to the Nations. <https://acfepublic.s3.us-west-2.amazonaws.com/2022+Report+to+the+Nations.pdf>. Accessed 24 Mar 2023
- Varian, H.R.: Benford's Law. *Am. Stat.* **26**(3), 65–66 (1972)
- Nigrini, M.J.: The detection of income tax evasion through an analysis of digital frequencies. Ph.D. thesis, University of Cincinnati, Cincinnati, OH, USA (1992)
- Carslaw, C.A.P.N.: Anomalies in income numbers: evidence of goal oriented behavior. *Account. Rev.* 321–327 (1988)
- Nigrini, M.J., Mittermaier, L.J.: The use of Benford's Law as an aid in analytical procedures. *Auditing J. Pract. Theory* **16**(2), 52–67 (1997)
- Nigrini, M.J.: *Benford's Law: Applications for Forensic Accounting, Auditing and Fraud Detection*. Wiley, Hoboken, New Jersey (2012)
- Leonov, P.Y., Rychkov, V.A., Ezhova, A.A., Sushkov, V.M., Kuznetsova, N.V., Suits, V.P.: Possibility of Benford's Law application for diagnosing inaccuracy of financial statements. In: *Proceedings of the 12th Annual Meeting of the BICA Society. Studies in Computational Intelligence*, vol. 1032, pp. 243–248. Springer, Cham (2022)

8. Leonov, P.Y., Suits, V.P., Norkina, A.N., Sushkov, V.M.: Integrated application of Benford's Law tests to detect corporate fraud. *Procedia Comput. Sci.* **213**, 332–337 (2022)



A Bayesian Network-Based Model for Fraud Risk Assessment

Pavel Y. Leonov^(✉) , Viktor M. Sushkov , Stanislav V. Vishnevsky ,
and Valentin A. Romanovsky 

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute),
Moscow, Russia
pyleonov@mephi.ru

Abstract. Global practice shows that almost every second company is exposed to financial fraud, while the process of its detection is labor-intensive and often low-efficient. This paper proposes a model for assessing the risk of fraud committed by business entities based on a Bayesian network. The model adopts a modern methodology for classifying fraud risk factors, referred to as the Fraud pentagon. The evaluation within the model incorporates financial statements, accounting data, and expert assessments regarding the internal controls. The effectiveness of classifying companies as fraudulent and bona fide using the model has been experimentally tested. It has been found that the risk-oriented approach underlying the model makes it possible to substantially reduce labor inputs for audit while maintaining high credibility of the results.

Keywords: Bayesian network · Fraud · Audit

1 Introduction

Today, the detection of fraud committed by business entities is an acute problem facing both external and internal financial control authorities. Fraud not only inflicts significant financial harm upon companies, but also poses a threat to the country's economic security. This is due to the high qualifications of perpetrators, the increasing sophistication of fraudulent schemes, and deficiencies in current methods of fraud detection.

According to PWC, almost half of all companies experience economic crime, including fraud, each year (see Fig. 1) [1].

Employees are involved in a majority (57%) of the largest fraud cases (see Fig. 2) [1].

The comprehensive nature of fraud cases makes it necessary to improve methods for assessing the risk of fraud based on modern approaches to fraud theory.

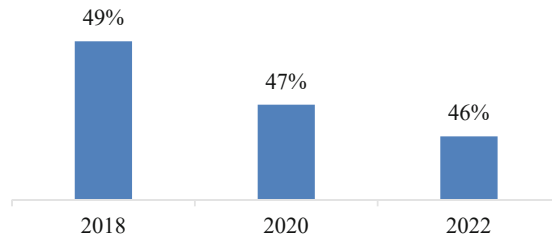


Fig. 1. Percentage of companies experiencing economic crime

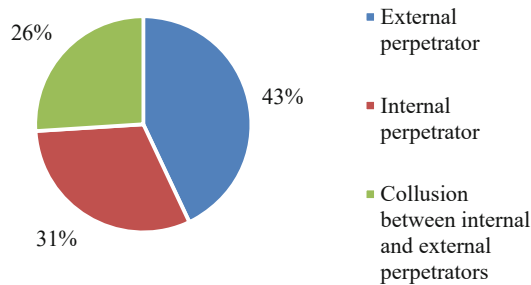


Fig. 2. Actors of the largest fraud cases

2 Analytical Part

2.1 Fraud Risk Factors

In the mid-twentieth century, American sociologist D.R. Cressey, based on the research of white-collar crime by E.H. Sutherland, put forward and proved the idea of the universal nature of fraud. The following three factors that make up the Fraud triangle are considered to be necessary and sufficient for fraud to occur:

1. an incentive (pressure) to commit fraud;
2. a recognized opportunity to commit fraud;
3. an ability to rationalize committed fraud [2, 3].

D.R. Cressey's approach is considered to be the classic theory of fraud. Almost 50 years later, in 2004, the Fraud triangle model was modified by D.T. Wolfe and D.R. Hermanson by including capability, which is understood as a set of personal characteristics necessary for committing, justifying and concealing fraud. This model is called the Fraud diamond [4]. In 2011, J. Marks supplemented the Fraud triangle with conditions of arrogance and competence of an individual, thereby obtaining the Fraud pentagon model (see Fig. 3) [5]. Fraud theory has evolved and progressed up to the present time, with contemporary research expanding its scope beyond the United States of America.

Approaches to classifying fraud risk factors leverage the widely accepted concept of fraud's universal nature. This concept enables the development of models that evaluate fraud risk. The Fraud pentagon model's comprehensive classification approach makes it an optimal choice for building a fraud risk assessment model.

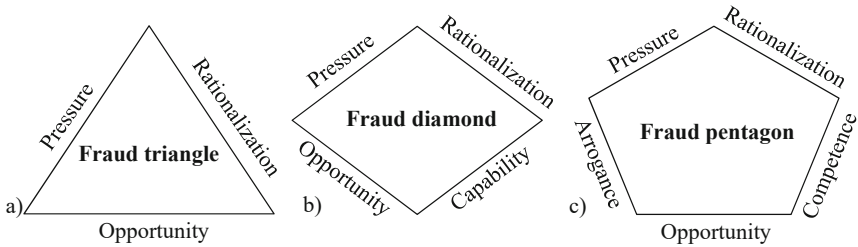


Fig. 3. Models of the a) Fraud triangle; b) Fraud diamond; c) Fraud pentagon

2.2 Developing a Model for Fraud Risk Assessment

In scientific research, a large number of data mining techniques have been proposed for fraud risk assessment. Decision trees provide clear and interpretable models, but they may struggle with complex relationships and are less effective at incorporating prior knowledge. Logistic regression is useful for binary classification but may not capture the nonlinear relationships present in fraud detection. Neural networks excel at handling large amounts of data and complex patterns but can lack interpretability.

In contrast, Bayesian networks (BN) combine the strengths of these methods by providing a flexible and interpretable model that incorporates prior knowledge, accounts for complex relationships, and handles both quantitative and qualitative data effectively. Additionally, determining whether an organization is susceptible to fraud, using risk factors as an input vector, can be seen as a binary classification task. In this classification problem, identifying a bona fide organization as fraudulent holds much less importance compared to defining a fraudulent organization as bona fide. The combination of these factors makes BN most effective for fraud risk assessment.

To account for the unavailability of direct observations of the fraud variable, an evidence diagram was utilized in the modeling process. This diagram includes third-party variables that are linked to the main variable, along with their relationships and evidentiary elements. By incorporating these third-party variables and their associated evidence, the evidence diagram aids in understanding and assessing the presence of fraud. The diagram is based on the model proposed by R.P. Srivastava, T.J. Mock and J.L. Turner [6]. The evidence diagram allows estimating the probability or belief associated with one variable given what we know about the other variables (see Fig. 4).

The diagram shows five fraud risk factors: Pressure (P), Opportunity (O), Rationalization (R), Competence (C), and Arrogance (A). These factors are quantified and assessed through subjective measures such as Likert scales for Pressure and Rationalization, while Opportunity is evaluated based on objective criteria such as control systems. Competence is measured by considering qualifications, experience, and certifications of the company's personnel, whereas Arrogance is assessed through observations of behavior and subjective assessments. The methods for quantification and calculation of these factors may vary depending on the research study and available data.

The fraud risk factors are related to the Fraud variable (F) through the "AND" relationship. The "AND" relationship implies that fraud will occur only if all factors, P, O, R, C and A, are present. The rectangular cells represent the elements of evidence

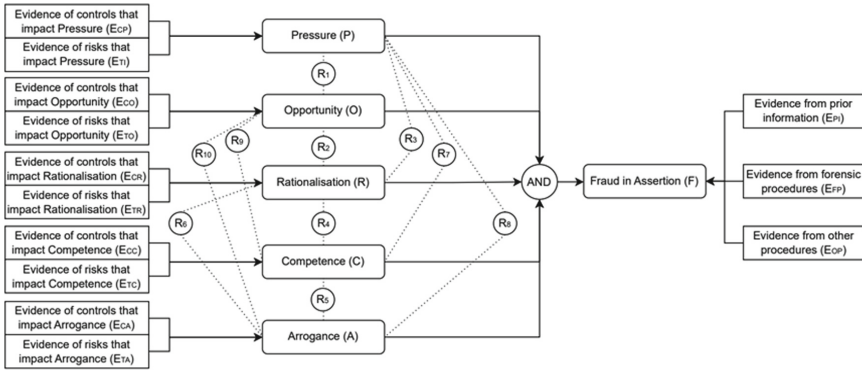


Fig. 4. Evidence diagram for fraud risk

relating to the variables to which they are linked. The variables P, O, R, C, A and F are assumed to be binary variables, that is, each variable has only two values: the variable is either present or absent.

The circles in the diagram represent relationships. The circle with “AND” represents the relationship between variable F and the five variables P, O, R, C and A. The relation R1 represents the relationship between P and O. Similarly, R2, R3, R4, R5, R6, R7, R8, R9, R10 represent the relationship between O and R, P and R, R and C, C and A, R and A, P and C, P and A, O and C, O and A, respectively.

Three types of evidence under variable F are also considered. One type of evidence is based on preliminary information (E_{PI}) about fraud in the industry of the company being audited. The second type of evidence is based on forensic procedures (E_{FP}) that can be performed. The third type of evidence is other procedures (E_{OP}). This type of evidence is considered to include any other evidence, such as analytical or other procedures performed during a traditional audit, which may provide evidence relating to the existence of fraud. In addition, two sets of evidence are considered for each variable: P, O, R, C and A. The evidence diagram consists of two sets of information. One set focuses on risks that raise the likelihood of the variable in question existing. The other set pertains to the controls implemented to lower the probability of the variable’s existence.

The formula (1) for estimating fraud risk (FR) in the Bayesian approach, expressed in terms of a priori probabilities and likelihood ratios, is presented below.

$$\begin{aligned}
 FR &= P(F|E_{CP}E_{TP}E_{CO}E_{TO}E_{CR}E_{TR}E_{CC}E_{TC}E_{CA}E_{TA}E_{OP}E_{FP}) \\
 &= \frac{\rho_1\rho_2\rho_3\rho_4\rho_5\rho_6\rho_7\rho_8\rho_9\rho_{10}\lambda_{CP}\lambda_{TP}\lambda_{CO}\lambda_{TO}\lambda_{CR}\lambda_{TR}}{D} \\
 &\quad \frac{\lambda_{CC}\lambda_{TC}\lambda_{CA}\lambda_{TA}\lambda_{OP}\lambda_{FP}\pi_P\pi_O\pi_R\pi_C\pi_A\pi_F}{D}
 \end{aligned}
 \tag{1}$$

In (1) E is evidence relating to the reduction or increase in the level of risk associated with a particular factor;

- ρ is the strength of the relationship between the risk factors;
- λ is the likelihood ratio reflecting the strength of evidence E;
- π_P is the ratio of the a priori probabilities of the risk factors;

$D = \sum_{i=1}^{32} D_i$ is the coefficient representing the sum of all 32 possible states, given that the presence or absence of fraud is determined by the presence or absence of risk factors.

Equation (1) is a fraud risk formula within Bayesian theory. The likelihood ratios determine the strength of the relevant piece of evidence, where $\lambda = 1$ implies that the evidence provides no information about the presence or absence of the relevant variable. For example, if forensic procedures (*FP*) and other procedures (*OP*) have not been carried out, we should set $\lambda_{FP} = 1$ and $\lambda_{OP} = 1$. A positive value greater than one ($1 < \lambda < \infty$) means that the evidence supports the claim or hypothesis, while a value less than one ($0 < \lambda \leq 1$) means that the evidence negates the claim. Theoretically, an infinitely large positive value of the likelihood ratio means that the statement is true with probability 1.0, while a value equal to zero, i.e. $\lambda = 0$, means that the statement is not true with probability 1.0.

The model is highly adaptable and allows for modifications. It enables the incorporation of new evidence and can accommodate different fraud assessment systems. Additionally, the relationships between the five primary fraud risk variables and the effects of controls and threats can be adjusted as needed within the model. Setting up the parameters of the system depends on the available data and the selected evaluation criteria.

In order to optimize the work of financial control entities in assessing the fraud risk using the model, a web application was developed using the high-level Python web framework Django. The main task of the developed web-application is to assess the risk of client fraud based on financial statements, accounting data, and expert assessments regarding the internal controls.

2.3 Performance Evaluation of the Model for Fraud Risk Assessment

As an example of utilizing the proposed model to calculate fraud risk, we selected the fraudulent company Alpha, considering its characteristics provided in Table 1. All the characteristics listed in the table are directly associated with the components of the Fraud pentagon. These characteristics were derived through the analysis of financial statements, accounting data, and information about the company's internal controls.

The fraud risk according to the formula (1) is 94%, which confirms the ability of the model to accurately identify fraudulent companies. We have also assessed 15 additional fraudulent companies and 156 bona fide ones. The results of these assessments consistently displayed fraud risk levels between 80% and 95% for the fraudulent companies, affirming the model's effectiveness in detecting high fraud risks. Conversely, the bona fide companies exhibited fraud risk levels ranging from 5% to 15%, further attesting to the model's ability to differentiate between unscrupulous and fair business entities.

Additionally, an important aspect of evaluating the effectiveness of the model is to determine the proportion of bona fide organizations that have been correctly identified. The most informative and illustrative metric for assessing the quality of models for these purposes is Recall.

Recall reflects the share of positive objects that the algorithm found. Thus, it shows the share of correctly identified fraudulent and bona fide companies in the total number of

Table 1. Characteristics of the company Alpha

Characteristics	Value
Beneish M-score	6.57
Roxas M-score	5.51
Growth rate of the asset quality	0.00
Growth rate of the share of expenses in sales revenue	1.03
Growth rate of the share of depreciation charges	1.05
Growth rate of income	– 88%
Increase in the rate of decline in the share of gross profit in revenue	– 195%
Growth rate of asset quality	– 100%
Growth rate of accounts receivable turnover	– 89%
Increase in the growth rate of the share of expenses in sales revenue	– 28%
Increase in the growth rate of the share of depreciation deductions	– 79%
Growth rate of accounts payable turnover	– 96%
Growth rate of financial leverage	– 13%
Increase in the growth rate of the share of other revenues in revenues	– 100%
Increase in the growth rate of the share of other expenses in revenue	119%
Number of directors' photos available online	0
Change of directors	3 times
Change of auditor	2 times
Quality of the external auditor	Not BIG4
Assessment of the internal control	2%
Financial dependency ratio	0.997
Last year's losses	4 503 ths. rub.

such companies. Figure 5 presents the outcomes of the classification process conducted on 172 companies.

As can be seen from the classification results, 1 fraudulent company out of 16 was determined to be bona fide. Accordingly, the Recall metrics for the class of fraudulent organizations would be 93.75%. The error in the classification can be related to the sample expanded with synthetic data, which is the basis of the model. To improve the accuracy of the model's classification, the sample underlying the model can be expanded by adding new company data to it.

The Recall metrics for bona fide organisations is 89.1%. This result shows that the use of this model will save a lot of labor costs of financial control entities for analysis of bona fides of companies using risk-based approach.

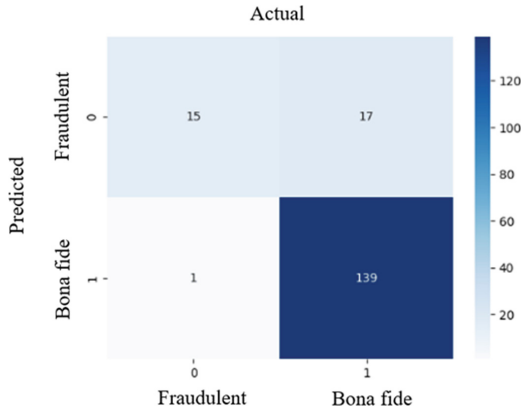


Fig. 5. Results of companies' classification by Recall metrics

3 Conclusion

Thus, the widespread occurrence of financial fraud in business entities underscores the necessity to develop effective methods for evaluating the risk of fraud based on contemporary approaches to fraud theory. The proposed model is built upon the most comprehensive and logical framework for categorizing factors that contribute to fraud risk, namely the Fraud pentagon. A Bayesian network has been selected as the method for classifying fraud risk. The model's proven efficacy has the potential to significantly reduce the cost of assessing the integrity of companies during financial control activities. Moreover, the model's adaptability allows for customization to meet the requirements of internal control services, financial monitoring entities, auditing firms, and government oversight bodies.

Limitations of the model encompass a reliance on available corporate data, which may not capture all instances of fraudulent behavior. Furthermore, the model's effectiveness may vary across different industries and regions, necessitating further investigation and adaptation to accommodate diverse contexts. Hence, piloting and refining the model in real-world scenarios is recommended to fine-tune its performance. Future research could focus on incorporating new evidence that aligns with the specificities of organizations' financial and economic activities, such as the scope, scale, and industry. Additionally, conducting tests on larger sample sizes would provide a more comprehensive understanding of the model's performance.

References

1. PwC's Global Economic Crime and Fraud Survey 2022. <https://www.pwc.com/gx/en/for-ensics/gecsm-2022/PwC-Global-Economic-Crime-and-Fraud-Survey-2022.pdf>. Accessed 24 Mar 2023
2. Cressey, D.R.: The differential association theory and compulsive crimes. *J. Crim. Law Criminol.* **45**(1), 29–40 (1954)
3. Sutherland, E.H., Cressey, D.R.: *Principles of Criminology*. Philadelphia, New York, 646 pp. (1960)

4. Wolfe, D.T., Hermanson, D.R.: The fraud diamond: considering four elements of fraud. *CPA J.* **74**(12), 38–42 (2004)
5. Mark, J.: Fraud triangle: not good enough these days. In: IIA/ACFE Conference, Cleveland, OH, 53 pp. (2011)
6. Srivastava, R.P., Mock, T.J., Turner, J.L.: Bayesian fraud risk formula for financial statement audits. *Abacus* **45**(1), 66–87 (2009)



The Influence of Articulatory Interference on Inner Pronouncing of Words

Daria Leonovich^(✉) and Alexander Vartanov

Lomonosov Moscow State University, Moscow, Russia
dagubareva@gmail.com

Abstract. In the current work, the subjects internally pronounced dictated Russian words “сахар” (“sahar”, sugar) and “ШАШЛИК” (“shashlik”, shish kebab) under two conditions: with and without the use of external articulatory interference. Articulatory interference presents itself as a wooden stick and it supposed to be clamped between the teeth horizontally. The study revealed that articulatory pre-setting systematically influences the process of internal word pronouncing, manifested in event-related brain potentials in the brain structures responsible not only for articulatory, but also for sensory components of speech. Myographic sensor detected minimal muscle activity during the study. The analysis of the obtained data was carried out using the author’s method of “virtually implanted electrode”. The results showed that articulatory pre-setting affects the process of internal word pronunciation depending on the influence of artificial interference that hinders the reproduction of the habitual muscle pattern characteristic of the external speech. For the analysis based on the new method of brain activity localization “virtually implanted electrode” ([4], Patent RU 2 785 268 C1, developer A.V. Vartanov), 41 points selected according to the MNI152 atlas, representing the centers of the main structures of brain. As a result, Broca’s and Wernicke’s areas were selected, which demonstrate the most revealing distinctions expressed in event-related brain potentials under the condition of presence or absence of the external articulatory interference that alters the pre-setting of the speech apparatus during the process of internal word pronunciation.

Keywords: EEG · EMG · ERP · Inner pronunciation · Inner speech · Words · Conditioned stimuli · Articulatory interference

1 Introduction

Internal pronouncing, as one of the elements of inner speech, is considered an “internal projection” of external speech (Sokolov 1967). Researchers have repeatedly emphasized the close connection between understanding spoken speech and its internal pronouncing, noting that when a person hears someone else’s speech, it is perceived as a set of sounds that are transformed into words by the speech apparatus, including auditory, motor, and visual centers, which then activate the organs necessary for reproducing those words. The application of speech interference techniques in the study of inner speech has been

actively developed in the works of A.N. Sokolov [5]. He established that the less automated the “mental actions” are, the greater the influence of mechanistic articulation delay on the process of their execution. It was found that external articulatory interferences impair performance in tasks of perception, reproduction, and memorization. And the less automated the motor pattern of action is, the greater the influence of articulatory interference on perception, internal pronouncing, verbalization and repetition. It is assumed that the presence of articulatory interference leads to tonic changes in not only the motor but also the sensory system. According to the paper written by Vartanov and Parchukov [3], it was shown that articulatory interference in the form of a horizontally clamped stick between the teeth significantly alters the process of perceiving the sounds “c” (s) and “ш” (sh), as pronouncing the first of these sounds is hindered by such interference, while the second is not. Moreover, a significant sequence effect was observed: if the interference was present initially and then removed, the influence on the perception of these sounds still remained. The works of A.N. Sokolov and others have also established a connection between electromyographic activity, which marks subtle movements of the lips and tongue that occur during the perception and silent reading of various texts. It was found that the less a person knows a foreign language, the greater the muscle activity registered by EMG sensors during reading, and the same applies to children who have not yet fully mastered reading in their native language. Thus, registering articulatory movements using EMG can be considered an effective method of diagnosing the presence of internal pronouncing. In the study, brain mechanisms are explored during inner pronouncing of words using articulatory interference and comparison of the effect of articulatory interference in each group of subjects, with and without speech defect, what gives the study novelty.

1.1 State of the Art

In recent years, there has been a growing interest in studying inner speech using various psychophysiological methods. Despite the low temporal resolution of fMRI, the potential of such research is significant, although currently non-invasive language decoders can only recognize a limited set of phrases and words [2]. There is also ongoing interest in utilizing the capabilities of MEG and EEG to decode inner speech [1]. However, a particular advancement in this regard is the opportunity to preprocess EEG data using a new method for localizing brain activity. The role of motor pre-setting in relation to articulatory interference remains unresolved. To what extent does it alter internal (mental) articulation if external articulation is not required? Will this be reflected in EEG activity, and to what extent will it need to be considered in the development of inner speech decoding methods?

1.2 Objectives

The study aims to investigate the psychophysiological mechanisms of inner speech production in individuals in the presence of external articulatory interference. A wooden stick horizontally clamped between the teeth of the subject was used as an artificial articulatory interference. The study is based on recording the electrical activity of the

brain and the articulatory apparatus using EEG (ERP) and EMG during the inner pronouncing of normatively pronounced words “сахар” (sahar) and “шашлык” (shashlik) in Russian. These words were chosen based on the presence of the sounds “с” (s) and “ш” (sh) at the beginning of the word, to which the applied articulatory interference acts differently. Additionally, two groups of subjects were formed for this study: a group without any speech defect, and a group with a speech defect known as rhotacism. It was hypothesized that subjects in these groups may differ in terms of their articulation apparatus pre-setting, and the effect of articulatory interference in this aspect may vary.

2 Methods

2.1 Data Processing

The aim of the study is to compare the ERP during the inner pronouncing of the words “сахар” (sahar) and “шашлык” (shashlik) with and without external articulatory interference in groups of subjects with and without speech defect. The electrical activity was measured using a 19-channel EEG (according to the international “10–20%” system, using a Neuro KM-type electroencephalograph), and the electromyographic activity was measured using a sensor fixed above the upper lip during the experiment.

2.2 Procedure

The study included two series. In each series subjects were instructed to internally pronounce the presented word:

- 1) Investigation with the initiation of inner pronouncing using a given auditory stimulus (a word to be repeated);
- 2) Investigation with the initiation of internal articulation based on an auditory stimulus (words) using an artificial articulatory interference (a wooden stick horizontally held between the teeth).

The stimuli were presented audibly in a random order, consisting of several words in Russian, among which were two words - “сахар” (sahar) and “шашлык” (shushlik), the inner pronouncing of which was further examined by averaging the ERP. The start of inner pronouncing was indicated by a special signal (a short sound). The study compares ERP averaged over the groups and on presentations. Each stimulus was given to each subject about 60 times. Thus, each ERP is constructed by averaging 500–700 presentations, in addition, 95% confidence interval was calculated, this makes it possible to assess the statistical significance of the detected differences. Auditory stimuli were presented through headphones. During both series of the experiment, the subjects had their eyes closed.

2.3 Participants and Stimuli

The study involved 21 subjects: 10 males and 11 females at the age from 18 to 34. 12 subjects had no speech defect, while 9 subjects had a speech defect (rhotacism). Participants had no history of head injuries, as well as severe somatic or mental illnesses.

The stimuli were presented as Russian words: “сахар” (sahar) and “шашлык” (shashlik), pronounced without any speech defect. The stimuli were presented audially.

2.4 Method

The analysis of the obtained data was performed using a new method of determining the localization of brain activity called “virtually implanted electrode”, which allows reconstructing the electrical activity originating from brain regions not accessible to scalp measurements ([4], Patent RU 2 785 268 C1, developer A.V. Vartanov). This method makes it possible to reconstruct the electrical activity of a source with predetermined coordinates relative to scalp electrodes based on scalp EEG data. The method is based on the analysis of the dynamics and correlation of signal changes in leads, but with the addition of artificially generated data calculated on the basis of the distances from the point under study to the scalp electrodes. Principal component factor analysis (PCA) and orthogonal rotation uniquely find one factor that is known to exist in the combined array of experimental and artificially generated data. The obtained and denormalized factorial values for the experimental EEG can be interpreted as the electrical activity of the “local field” when the electrode is implanted in the corresponding point of the brain. This data processing method has similarities with a group of source detection methods “beamforming”. The activity was investigated in 41 points selected according to the MNI152 atlas. Each point represents the center of the following structures: Hypothalamus, Brainstem, Mesencephalon, Medula Oblongata, Caput n.Caudati L, Caput n.Caudati R, Globus Pallidus Medialis L, Globus Pallidus Medialis R, Putamen L, Putamen R, Thalamus L, Thalamus R, Hippocampus L, Hippocampus R, Amygdaloideum L, Amygdaloideum R, Anterior Cingulate BA32, G. Cingulate Med.24, Insula L BA13, Insula R BA13, Ventral Striatum BA2, Dorsomedial prefrontal cortex BA9 L, Dorsomedial prefrontal cortex BA9 R, Supramarginalgyrus BA40 L, Supramarginalgyrus BA40 R, Parietal cortex BA7 L, Parietal cortex BA7 R, V1 BA17 L, V1 BA17 R, Broca BA44 L, Wernicke BA22 L, BA44 R, BA22 R, Cerebellum L, Cerebellum R, Angular G.BA39 L, Angular G.BA39 R, Mid. Fr. c. BA10 L, Mid. Fr. c. BA10 R, Obr. Fr. c. BA47 L, Obr. Fr.c. BA47 R.

The procedure of averaging the event-related potentials was performed for each series and group of subjects for each stimulus with a 95% confidence interval. The analysis aimed to obtain the ERP both for the original recordings and for the computed localized activity of each investigated brain structure. In addition to assessing the electrical activity of the brain, electromyographic activity was also studied to monitor changes related to perception and inner pronouncing acts.

3 Results

Myographic sensor detected minimal muscle activity during the study (Fig. 1). The greatest high-amplitude fluctuations are observed for a group of subjects with a speech defect (rhotacism) without the use of artificial articulatory interference.

As a result, ERP were obtained for a group of subjects with no speech defect and a group with speech defect (rhotacism) in Broca’s area for the words “caxap” (sahar) and “ШАШЛЫК” (shashlik) (Fig. 2). For a group with rhotacism, the presence of artificial articulatory interference significantly influenced the amplitude of ERP. Without the use of interference, pronounced p150 and n200 peaks were observed, which were significantly attenuated when interference was present, indicating cognitive processing

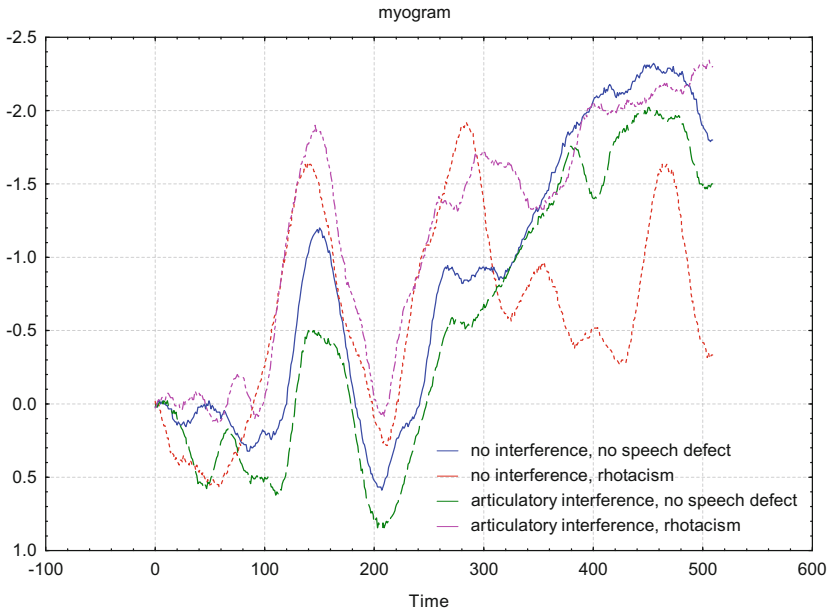


Fig. 1. Myogram. Blue line – electromyographic activity for a group with no speech defect without articulatory interference; green line - electromyographic activity for a group with no speech defect with articulatory interference; red line - electromyographic activity for a group with rhotacism without articulatory interference; purple line - electromyographic activity for a group with rhotacism with articulatory interference

of the incoming stimulus. The data obtained suggest an increased cognitive control in the process of motor program development with the use of articulatory interference, particularly for subjects with rhotacism. Additionally, significant differences in amplitudes were observed for the word “caxap” (sahar) which is presumably related to the required articulatory position that cannot be achieved for the pronunciation of the letter “c” (s) when using articulatory interference.

ERP were obtained for Wernicke’s area for a group of subjects with no speech defect and a group with speech defect (rhotacism) (Fig. 3). For a group with rhotacism the presence of interference significantly affects the amplitude of ERP. Without the use of interference, pronounced p150 and n200 peaks are observed, which are considerably smoothed out when interference is present, indicating cognitive processing of the incoming signal for the group with rhotacism. The p300 component is more expressed for the group with rhotacism. For the group with no speech defect, the main peaks are also observed at latencies of 150 and 200 ms, although the differences between the interference conditions are less significant. In this case, a decrease in the amplitude of the MMN component at latencies of 180–220 ms is observed for both groups of subjects when interference is present compared to its absence, indicating a mismatch between the familiar and current motor pattern available for reproduction by the subjects, which they rely on during the task of inner pronouncing. The obtained data indicate increased

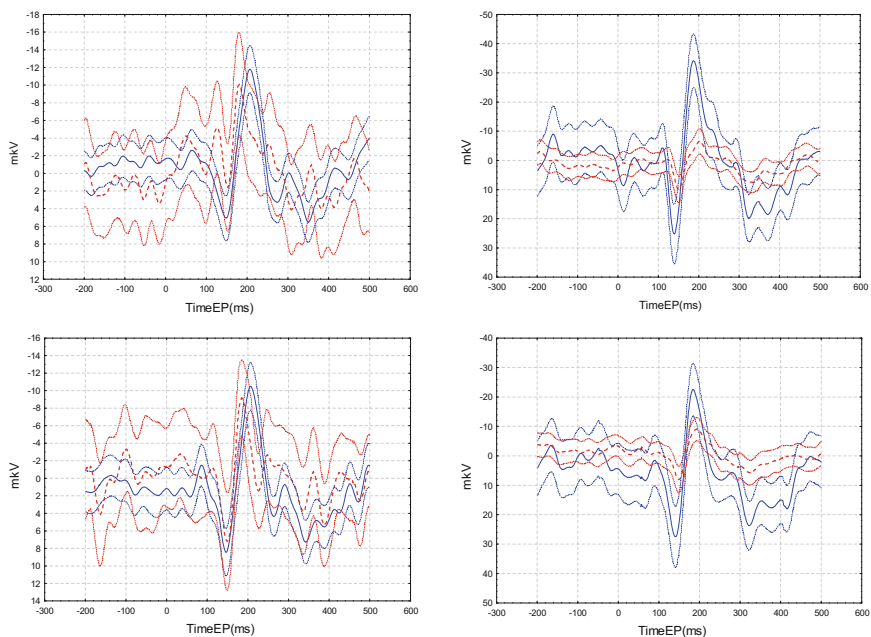


Fig. 2. Broca's area. Top left – ERP for Broca's area for a group with no speech defect for the word “caxap” (sahar), blue line – no interference, red line – artificial articulatory interference; top right - ERP for Broca's area for a group with rhotacism for the word “caxap” (sahar), blue line – no interference, red line – artificial articulatory interference; down left – ERP for Broca's area for a group with no speech defect for the word “шашлык” (shashlik), blue line – no interference, red line – artificial articulatory interference; down right – ERP for Broca's area for a group with rhotacism for the word “шашлык” (shashlik), blue line – no interference, red line – artificial articulatory interference

cognitive control in the process of developing a motor program using articulatory interference, especially in the group with rhotacism. Significant differences in amplitudes are also noted for the word “caxap” (sahar), which is presumably related to the required articulatory position that is impossible to achieve for pronouncing the letter “c” (s) when using articulatory interference. Especially the influence of artificial articulatory interference is expressed for the word “caxap” for the group with rhotacism.

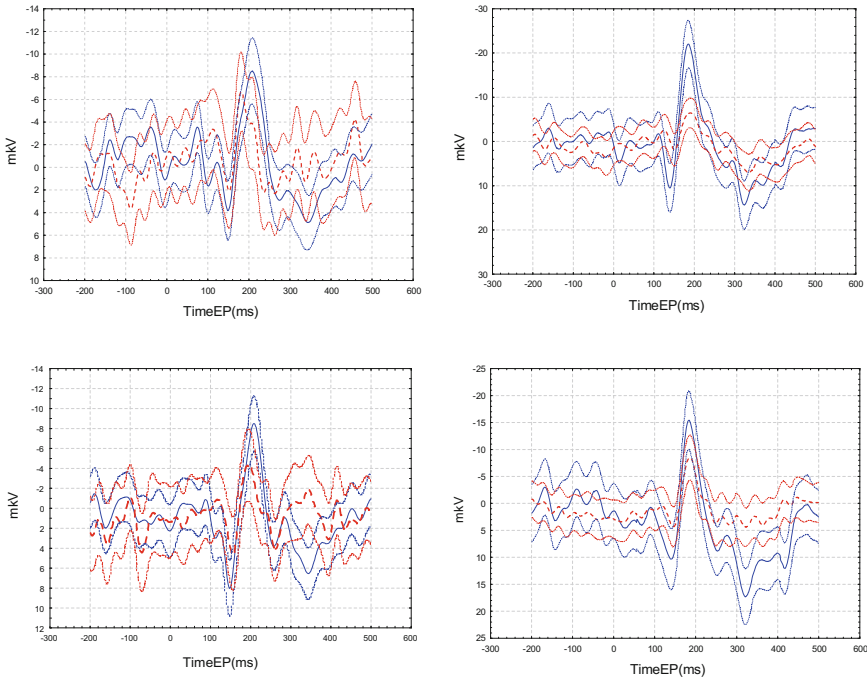


Fig. 3. Wernicke's area. Top left – ERP for Wernicke's area for a group with no speech defect for the word “caxap” (sahar), blue line – no interference, red line – artificial articulatory interference; top right - ERP for Wernicke's area for a group with rhotacism for the word “caxap” (sahar), blue line – no interference, red line – artificial articulatory interference; down left – ERP for Wernicke's area for a group with no speech defect for the word “шашлык” (shashlik), blue line – no interference, red line – artificial articulatory interference; down right – ERP for Wernicke's area for a group with rhotacism for the word “шашлык”, blue line – no interference, red line – artificial articulatory interference.

4 Conclusion

In the context of this study, it was found that articulatory pre-setting has a systematic influence on the process of inner pronouncing, which is manifested in the ERP in the majority of brain structures involved in auditory stimuli recognition and the development of articulatory programs for its reproduction through inner pronouncing. ERP graphs for Broca's and Wernicke's areas were provided, highlighting the main differences associated with the presence of articulatory interference for both groups of subjects: with and without speech defect. The obtained data indicate increased cognitive control of the process of motor program development using articulatory interference, especially in the group with rhotacism, as observed in the ERP in the Broca's and Wernicke's areas. For the group without speech defects, the main peaks also occur at similar latencies as in the rhotacism group, but the differences related to the interference condition are less significant. Moreover, external articulatory interference has a greater impact on the group of

individuals with speech defect, as well as for the inner pronouncing of the sound “c” (s), as it disrupts the automated motor pattern existing in the subjects to the greatest extent.


Acknowledgements. The research is financially supported by the Russian Science Foundation, Project № 20-18-00067-II.

References

1. Defossez, A., Caucheteux, C., Rapin, J., Kabel, O., King, J.: Decoding speech from brain recordings (2022). <https://doi.org/10.48550/arXiv.2208.12266>
2. Tang, J., LeBel, A., Jain, S., Huth, A.: Semantic reconstruction of continuous language from non-invasive brain recording. *Nat. Neurosci.* (2023). <https://doi.org/10.1038/s41593-023-01304-9>
3. Vartanov, A.V., Parchukov, A.J.: Subjective space of tactile perception and evoked potential. In: XXVII International Congress of Psychology, Stockholm, Sweden, 23–28 July 2000. Program No 21605.12
4. Vartanov, A.V.: A new method of localizing brain activity using the scalp EEG data. *Procedia Comput. Sci.* **213**, 41–48 (2022). <https://doi.org/10.1016/j.procs.2022.11.036>
5. Sokolov, A.N.: *Vnutrennyaya rech' i myshlenie* (Inner Speech and Thinking). Izdatel'stvo «Prosveshchenie», Moscow (1968)



Method of Logical Interpretation of Neural Network Solutions

L. A. Lyutikova^(✉) 

Institute of Applied Mathematics and Automation KBSC RAS (IAMA KBSC RAS), Nalchik,
Russia

lylarisa@yandex.ru

Abstract. This paper proposes a method for logical interpretation of neural network solutions. In order to logically interpret the operation of a neural network, various methods can be used that help visualize and analyze the internal processes occurring in the network. The approach under consideration examines only the input data and the results of the neural network solution and does not take into account the weights, structure, learning method. Using Boolean integro-differential calculus, it considers possible logical relationships between the input data and the results of the decisions. Combining these relationships, a function is obtained that allows you to analyze in detail the decision area of the neural network, find the most important features and hidden patterns in the data. This is especially useful for solving problems in which the exact form of the relationship between the input data and the result is unknown, but a sufficient amount of experimental data has been accumulated.

Keywords: Neural networks · Interpreter · Connections · Logical derivative

1 Introduction

Neural networks are complex mathematical models that can find complex and non-obvious patterns in data. However, understanding how exactly these patterns were found is often not clear due to the fact that neural networks usually operate in a non-linear space.

In order to logically interpret the operation of a neural network, various methods can be used that help visualize and analyze the internal processes occurring in the network. For example,

1. Visualization of neuron activation: You can look at the activation of each neuron in the network and understand which factors most strongly influence its activation. This can help to understand which features are most important for solving the problem.
2. Visualization of weights: You can look at the weights that are assigned to each feature in the network and understand which features are most important for solving the problem.
3. Error analysis: You can analyze the errors that the network makes and understand which cases are the most difficult to recognize and need additional processing.

4. Gradient visualization: You can look at the gradients that are calculated during network training and understand which features are most important for network training.
5. Interpretation through comparison models: You can train another model, such as linear regression, on the same data and compare the results with those of the neural network. This can help to understand which features are the most important for solving the problem [1–4].

However, it is important to understand that interpreting how a neural network works can be complex and requires special attention to detail. In addition, some types of neural networks, such as deep neural networks, can be particularly difficult to interpret due to the large number of layers and parameters.

This paper proposes a method for logical interpretation of neural network solutions. This approach is based on the use of Boolean integro-differential calculus [5].

Neural networks can be very deep and complex, and their structure can be far from optimal. But despite the fact that the structures of neural networks can be different depending on the specific type of problem they solve, any neural network contains:

1. Input layer - this layer receives input data that is passed to the neurons of this layer. The input layer can be of different dimensions depending on the number of input variables.
2. Hidden layers: These are the layers that process the input.
3. Output layer - this layer produces the result of the neural network. The number of neurons in this layer depends on the number of output variables.
4. Loss function: This is a function that evaluates how well the neural network performs on a task. It compares the output from the network with the expected output and returns an error value.

The goal of training a neural network is to minimize the value of the loss function. And also optimizers are the algorithms that are used to update the weights.

Neural network architecture: This is the general structure of the neural network, including the number of layers, the number of neurons in each layer, types of activation functions, etc. Different architectures can be used for different tasks [6].

We will try to establish the logical patterns that have arisen in a particular trained neural network without taking into account its structure and the value of the weights of this neural network. This will be an interpretation similar to the comparison model. A set of logical functions will act as such a model. The input and output data will be the values at the input of the neural network and at the output corresponding to these data.

2 Formulation of the Problem

As mentioned above, the main purpose of interpreting the decisions of a neural network is to understand which features of the input data are most important to the network's decision making and what patterns it finds in the data.

Interpretation can provide useful information about how the network works and how it can be improved, as well as understanding the nature of the data itself.

We will be looking at tasks that relate to what is known as learning with a teacher. Another condition would be the ability to represent each feature as a k-valued variable.

Then the mathematical formulation of the problem has the following form.

Let $X = \{x_1, x_2, \dots, x_n\}$, $x_i \in \{0, 1, \dots, k_i - 1\}$, where $k_i \in [2, \dots, N]$, $N \in \mathbb{Z}$, is the set of neural network inputs. $Y = \{y_1, y_2, \dots, y_m\}$ – many exits, each exit y_i the result of processing specific input values by the neural network $x_1(y_i), \dots, x_n(y_i) : y_i = f(x_1(y_i), \dots, x_n(y_i))$.

$$\begin{pmatrix} x_1(y_1) & x_2(y_1) & \dots & x_n(y_1) \\ x_1(y_2) & x_2(y_2) & \dots & x_n(y_2) \\ \dots & \dots & \dots & \dots \\ x_1(y_m) & x_2(y_m) & \dots & x_n(y_m) \end{pmatrix} \rightarrow \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{pmatrix} \tag{1}$$

Function needs to be restored $Y = f(X)$ by observations.

3 Materials and Methods

As a method for solving the problem, we consider the integral-differential apparatus of logical functions. Formally, the logical derivative in some of its properties is an analogue of the derivative in the classical differential calculus.

Definition. First order derivative $\frac{\partial f}{\partial x_i}$ from the Boolean function $f(x_1, \dots, x_n)$ with respect to the variable x_i is the modulo 2 sum of the corresponding residual functions:

$$\frac{\partial f}{\partial x_i} = f(x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n) \oplus f(x_1, \dots, x_{i-1}, 1, x_{i+1}, \dots, x_n) \tag{2}$$

Definition. Derivative weight $P\left(\frac{\partial f}{\partial x_i}\right)$ from a Boolean function is called the number of constituents (1) of this derivative.

Statement. The greater the weight of the derivative, the more the function $f(x_1, \dots, x_n)$ depends on the variable x_i .

Logical derivatives can be used for a variety of purposes, including boolean derivatives can be used to optimize boolean functions, simplify boolean expressions, and minimize the number of booleans needed to implement a function. They can also be used to analyze logic circuits and determine their sensitivity to changes in input signals [7–9].

One of the main applications of logical derivatives is the Quine-McCluskey method, which is used to simplify logical expressions and minimize the number of logical elements. This method is based on using the truth table of a function and calculating its logical derivatives. The derivatives are then grouped by the number of 1s in the binary notation of the original variables, which makes it possible to simplify the function and express it with fewer logic gates. Logical derivatives can also be used to analyze logic circuits and determine their sensitivity to changes in input signals. For example, if the logical derivative with respect to one of the function variables is 1, this means that when the value of that variable changes, the output of the function will also change. This can

be useful when designing logic circuits to determine which parts of the circuit will be most sensitive to changes in input signals.

In addition, logical derivatives can be used to solve other tasks related to logical functions, such as checking the equivalence of two functions, determining the completeness of a system of logical functions, etc. [10].

In general, logical derivatives are a useful tool for analyzing and optimizing logic functions and logic circuits. They can be used in many areas, including electronics, cryptography, logic, and computer science.

Definition. Let $g = \frac{\partial f}{\partial x_i}$ derivative of the function f with respect to the variable x_i , then there is a function $\int g dx_i$, called the boolean integral of the function g , such that:

$$\int g dx_i = x_i g \oplus h, \tag{3}$$

where h is an arbitrary boolean function of the variables $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$.

Property. The number L of antiderivatives of the function

$$g = \frac{\partial f}{\partial y_i} = x_1 \& x_2 \& \dots \& x_n, \tag{4}$$

where $f(x_1 \dots, x_n)$ - Boolean function, equals $L = 2n$.

The proposed method for interpreting neural network solutions is based on the following reasoning:

1. We will consider a neural network with m inputs and n outputs, suppose each output y_i depends on input values $x_1(y_i), \dots, x_n(y_i)$ and this dependence can be described by the logical function $f_i(x_1(y_i), \dots, x_n(y_i), P^\sigma(y_i))$. Where

$$P^\sigma(y_i) = \begin{cases} \overline{P(y_i)} & \text{if } \sigma = 0 \\ P(y_i) & \text{if } \sigma = 1 \end{cases} \tag{5}$$

The function sought is $f_i(x_1(y_i), \dots, x_n(y_i), P(y_i)) = 1$ and $f_i(x_1(y_i), \dots, x_n(y_i), \overline{P(y_i)}) = 1$. Thus the function depends essentially on the variable $P(y_i)$, and its derivative must be $\frac{\partial f_i}{\partial P(y_i)} = x_1(y_i) \& x_2(y_i) \& \dots \& x_n(y_i)$, indicating that the result at each output depends on all the stated values at the inputs.

2. Our task is to find the functions for each output. Then construct their composition in order to obtain the function linking all inputs and outputs of the neural network

$$F(x_1(y_i), \dots, x_n(y_i), P^\sigma(y_1), \dots, P^\sigma(y_n)). \tag{6}$$

Since the values at the outputs of the neural network depend on specific input values, the function should be such that if we want to consider only the input data, then they must correspond to a set of specific input values. This can be described as the conjunction of all input values: $\&_{j=1}^m x_j(y_i)$.

A logical function that reflects the relationship between specific input and output values of a neural network can be found by solving the following equation:

$$\frac{\partial f_i}{\partial P(y_i)} = x_1 \& x_2 \& \dots \& x_n \tag{7}$$

Then, based on the definition of the Boolean integral, we will have four functions as a solution:

$$\begin{aligned}
 f_{1i} &= x_{i1} \& x_{i2} \dots \& x_{in} \& P(y_i) \\
 f_{2i} &= x_{i1} \& x_{i2} \dots \& x_{in} \& P(y_i) \\
 f_{3i} &= x_{i1} \& x_{i2} \dots \& x_{in} \rightarrow P(y_i) \\
 f_{4i} &= x_{i1} \& x_{i2} \dots \& x_{in} \rightarrow P(y_i)
 \end{aligned}
 \tag{8}$$

Moreover, the functions f_{2i} and f_{3i} solutions for our conditions cannot be.

Therefore, in this interpretation, each neural network output is associated with its corresponding input or conjunction function or implication. Example: suppose we have two inputs and two outputs. One input is the values (0, 1) at the output of the object “a”, the second input is the values (1, 1) at the output of the object “b”. Let’s build functions that reflect the relationship between (0, 1) and “a”. Initial data (Table 1).

Table 1. Input value function.

x_1	x_2	$\bar{x}_1 \& x_2$
0	0	0
0	1	1
1	0	0
1	1	0

Let’s demonstrate tabular functions corresponding to the expression $f \bar{x}_1 \& x_2 dP(a)$ (Table 2).

Table 2. Table of functions corresponding to a value $f \bar{x}_1 \& x_2 dP(a)$

x_1	x_2	$P(a)$	f_{1i}	f_{2i}	f_{3i}	f_{4i}
0	0	0	0	1	1	0
0	0	1	0	1	1	0
0	1	0	0	1	0	1
0	1	1	1	0	1	0
1	0	0	0	1	1	0
1	0	1	0	1	1	0
1	1	0	0	1	1	0
1	1	1	0	1	1	0

It can be seen from the table that the four functions correspond to

$$\begin{aligned}
 f_{1i} &= \overline{x_1} \& x_2 \& P(a) \\
 f_{2i} &= \overline{x_1} \& x_2 \& P(a) \\
 f_{3i} &= \overline{x_1} \& x_2 \rightarrow P(a) \\
 f_{4i} &= \overline{x_1} \& x_2 \rightarrow P(a)
 \end{aligned}
 \tag{9}$$

Functions f_{2i} and f_{4i} solutions cannot be, since they have a point (0, 1, 0), the presence of input signals and the absence of an output signal, which contradicts the conditions, it remains only $f_{1i} = \overline{x_1} \& x_2 \& P(a)$, and $f_{3i} = \overline{x_1} \& x_2 \rightarrow P(a)$.

Consider now the input-value conditions (1, 1) at the output object “b” (Table 3).

Table 3. Input value function

x_1	x_2	$\overline{x_1} \& x_2$
0	0	0
0	1	0
1	0	0
1	1	1

Let’s demonstrate the table functions corresponding to the expression $f \overline{x_1} \& x_2 dP(b)$ (Table 4).

Table 4. Table of functions corresponding to a value $f \overline{x_1} \& x_2 dP(b)$

x_1	x_2	$P(b)$	f_{1i}	f_{2i}	f_{3i}	f_{4i}
0	0	0	0	1	1	0
0	0	1	0	1	1	0
0	1	0	0	1	1	0
0	1	1	0	1	1	0
1	0	0	0	1	1	0
1	0	1	0	1	1	0
1	1	0	0	1	0	1
1	1	1	1	0	1	0

Satisfying solution conditions:

$$\begin{aligned}
 f_{1i} &= x_1 \& x_2 \& P(b), \\
 f_{3i} &= x_1 \& x_2 \rightarrow P(b).
 \end{aligned}
 \tag{10}$$

The next question is the logical relationship between the resulting features. That is, the possibility of constructing a logical function, which is a superposition of the functions of each output.

4 Results

Since the outputs in a neural network are independent, it is logical to imagine that the function that combines the functions of each input is either their conjunction or disjunction.

Thus, we have two variants of functions that interpret the dependence of input and output data for each case. And two options for functions that can combine all these solutions.

1. If we consider as initial functions at each given input and output $f_{1i} = x_{i1} \& x_{i2} \dots \& x_{in} \& P(y_i)$ that

$$\& f_i = x_{i1} \& x_{i2} \dots \& x_{in} \& P(y_i) = 0 \quad (11)$$

2. If we consider the conjunction as the initial functions at each given input and output, and the disjunction as the unifying function

$$f_{1i} = x_{i1} \& x_{i2} \dots \& x_{in} \& P(y_i), \quad (12)$$

$$\vee f_i = x_{i1} \vee x_{i2} \dots \vee x_{in} \vee P(y_i), \quad (13)$$

then it will be a neural network capable of giving only those answers that we have considered. Logically, this is a function of

$$F(x_1, \dots, x_n, P^\sigma(y_1), \dots, P^\sigma(y_m)) = \begin{cases} 1 & \text{if } x_{i1} \& x_{i2} \dots \& x_{in} \& P(y_i) \\ 0 & \end{cases} \quad (14)$$

This option does not give the opportunity for conclusions, when at least some values will differ from the given ones.

3. If we consider the implication as the original functions at each given input and output, and the disjunction as the union, we get one. For our example:

$$\begin{aligned} f_{3i} = \bar{x}_1 \& x_2 \rightarrow P(a) &= x_1 \vee \bar{x}_2 \vee P(a) \\ f_{3i} = x_1 \& x_2 \rightarrow P(b) &= \bar{x}_1 \vee \bar{x}_2 \vee P(b) \\ x_1 \vee \bar{x}_2 \vee P(a) \vee \bar{x}_1 \vee \bar{x}_2 \vee P(b) &= 1 \end{aligned} \quad (15)$$

Such a result only says that the functions that interpret the connection of values at the input and output of the neural network, in this case, the implication, will be true on all considered sets.

4. If we consider the implication as the initial functions at each given input and output, and the conjunction as the union, we obtain the function

$$f(X) = \& \left(\&_{i=1}^n \bar{x}_i \rightarrow P(y_j) \right) \quad (16)$$

This function has a number of interesting properties [12]. For our example:

$$\begin{aligned} f_{3i} &= \bar{x}_1 \& x_2 \rightarrow P(a) = x_1 \vee \bar{x}_2 \vee P(a) \\ f_{3i} &= x_1 \& x_2 \rightarrow P(b) = \bar{x}_1 \vee \bar{x}_2 \vee P(b) \\ (x_1 \vee \bar{x}_2 \vee P(a)) \& (\bar{x}_1 \vee \bar{x}_2 \vee P(b)) &= \bar{x}_2 \vee \bar{x}_1 P(a) \vee x_1 P(b) \vee P(a) P(b) \end{aligned} \quad (17)$$

We can claim that we have no solutions containing \bar{x}_2 . And in order to distinguish one object from another, one variable is enough x_1 .

5 Conclusions

As a result of the considered method, it can be argued that for the logical interpretation of a correctly functioning neural network, it is possible to construct a function that will give an idea of the hidden patterns, the existing classes, and the most important features in the processed data. This approach does not take into account the weight, structure, method of learning, rather, it refers to the interpretation through comparison models, and gives a complete picture of the properties of the data under study on the considered set of solutions.

References

1. Zhuravlev, Yu.I.: On an algebraic approach to solving problems of recognition or classification. *Probl. Cybern.* **33**, 5–68 (1978)
2. Shibzukhov, Z.M.: Correct algorithms for aggregation of operations. *Pattern Recognit. Image Anal.* **24**(3), 377–382 (2014)
3. Naimi, A.I., Balzer, L.B.: Multilevel generalization: an introduction to super learning. *Eur. J. Epidemiol.* **33**, 459–464 (2018)
4. Lyutikova, L.A.: Construction of a logical-algebraic corrector to increase the adaptive properties of the $\Sigma\Pi$ -neuron. *J. Math. Sci.* **253**(4), 539–546 (2021). <https://doi.org/10.1007/s10958-021-05251-3>
5. Jha, A., Singh, J.K.A.M.R.G.D., Barash, Y.: Enhanced integrated gradients: improving interpretability of deep learning models using splicing codes as a case study
6. Samek, W., Montavon, G., Lapuschkin, S., Anders, C., Müller, K.R.: Toward interpretable machine learning: transparent deep neural networks and beyond (2020)
7. Chernov, A.V.: Razvitiye apparata logicheskogo differentsial'nogo ischisleniya v primenenii k zadacham proektirovaniya i diagnostiki telekommunikatsionnykh system. *Nauchno-tekh. vedom. SpBGPU* **2**, 118–126 (2008)
8. Spirina, M.P.: Logicheskoe differentsial'noe i integral'noe ischislenie. *Inf. sist. tekhnol. upravlenie bezopasnost'* **4**, 187–201 (2016)

9. Yang, F., Yang, Zh., Cohen, W.: Differentiable learning of logical rules for reasoning in the knowledge base. In: *Advances in the Field of Neural Information Processing Systems*, pp. 2320–2329 (2017)
10. Flach, P.: *Machine Learning: The Art and Science of Algorithms that Give Meaning to Data*. Cambridge University Press, Cambridge (2012)
11. Akhlakur, R., Sumaira, T.: Ensemble classifiers and their applications: a review. *Int. J. Comput. Trends Technol.* **10**, 31–35 (2014)
12. Lyutikova, L.A.: Use of logic with a variable valency under knowledge bases modeling. In: *CSR-2006* (2006)



Implementation of Embodied Cognition in Multi-agent Neurocognitive Architecture

Dana Makoeva^(✉) , Olga Nagoeva , Murat Anchokov , and Irina Gurtueva 

Institute of Computer Sciences and Problems of Regional Management—branch of
Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences, 360000 Nalchik,
Russia

iipru@rambler.ru

Abstract. Cognitive linguists assume that we use conceptualization to categorize the world around us. Concepts are units of knowledge that we keep in our brains to think. Theories of embodiment argue that concepts are based not only on language input but also on movement and perception that co-occur with it. It means that direct experience means a lot to our ability to learn new things and words. This congruency is well-seen in the process of first language acquisition. Nowadays being surrounded by devices with elements of artificial intelligent, it is obvious that the most convenient way to utilize them is via habitual conversations. To achieve this level of communication intelligent systems have to obtain not only a thorough vocabulary, but also to be able to understand language at a deeper level, at the semantic ones. Formal representation of semantics has been a hinder in a way to develop language understanding systems for decades. There were two main reasons: (1) the way semantics was obtain and presented was vague; (2) tools for formal representation of semantics were not effective. In this article we will try to give a short overview of the theories of embodied cognition, which as we assume give a reliable explanation of how semantics is built in our brains. While these theories explain the nature of semantics and conceptualization, multi-agent system can be a great tool to model these processes.

Keywords: Natural language semantics · Embodied cognition · Multi-agent system · Simulation model · Artificial intelligent · Robot

1 Introduction

It is hard to imagine the way to interact with the world without language. Almost everything we do during the day is done via spoken or written language. Language allows us to encode and transfer our thoughts and ideas no matter how complex and ambiguous they are. Actually, encoding and transferring of information are connected to two main functions of any language: symbolic and interactive [1].

The symbolic function is responsible for transferring our ideas via sings (symbols). Symbols are any meaningful units of a language (morphemes, words, sentences), they have their form, i.e. they can be represented in written, spoken form or as a sing, and their meaning. The form and the meaning are conventionally interconnected [2].

Interactive function arranges our interaction with the society, allowing us to send and receive information. This function is deeply connected to our ability to conceptualise, i.e. our ability to pick out a set of features of something in the world that helps us to distinguish it from others. This piece of knowledge, basic unit of thinking process, stored in our brain about some fact or object is traditionally called a concept in Cognitive Linguistics. According to it the process of acquiring knowledge brings on formation of new concepts, i.e. conceptualization [2].

A powerful cognitive skill that distinguishes us from the majority of other sentient beings is the power to make use of symbols and symbolic thought that allows us to recall the past and predict the future, i.e. shift us from the current situation to the imaginary one [3, 4]. Symbols are understood as arbitrary tokens possessing certain semantic meaning and obeying some combinatory rules. For us the most significant and accessible symbolic system is natural language, other examples are formal logic, artificial (programming) languages, mathematical notation etc. It is believed that in human brain words are presented as concepts, realized in the form of “mental symbols” or “symbolic representations” that participate in a number of symbolic operations [4, 5].

A lot of scientists [6–10] have argued about the form and format of the symbolic representation in human brain. In a number of classical theories, it is stated that concepts are represented in an arbitrary and amodal mode [8, 11–15]. Amodality means no connection to sensor and motor information—in other words connections between concepts and objects of the real world that they refer to is postulated to be random. Let’s have a look at the word “table”, this very lexical unit does not have any similarity to any real table in the world, it is supported by the fact that different languages have different words to identify one and the same thing: “Tisch” in Deutsch, “стол” in Russian, etc. [16]. It means that the concept TABLE is a mental unit, but not a real table, that is why concepts are thought to be abstract [8].

There is a number of other theories which regard concepts as abstract, amodal symbols. These theories are semantic network models [12, 13, 17], distributional models of semantics [15, 18], feature-based approaches to semantics [19, 20].

If we try to examine the concept TABLE using feature-based approaches to semantics we may use the following propositions IS(TABLE, FURNITURE), HAVE(TABLE, LEGS), FORM(TABLE, OVAL). It should be noted that in all of these theories concepts are usually identified via other concepts, which they are connected to. This way of concept representation has a variety of disadvantages. One of them is known as the Chinese–Chinese dictionary argument [3] that is based on the mental experiment called the Chinese room [21]. The Chinese–Chinese dictionary argument is a mental experiment where a student has to learn Chinese language using only the Chinese–Chinese dictionary. When the student needs to find a definition of a certain symbol, she opens the dictionary and find out that the symbol is defined by other unknown symbols and so on. As a result, instead of finding at least one appropriate definition the student has an infinite regression [16].

From the above-mentioned it can be concluded that symbols should be grounded in a way that they have a certain semantic meaning. One of the approaches to the problem of symbol grounding in Cognitive Linguistics is embodiment also known as grounded/embodied cognition. This approach argues that symbols should be grounded

via sensorimotor experience [7, 22]. The main idea is that high level cognitive systems such as language and the process of the real world conceptualization cannot be independent from perception and action, that is why they cannot be abstract or amodal. Moreover they use the same ways of representation and activate the same cognitive systems. In the theory of embodiment language understanding is a process of modelling of the context of the input language signal using the same systems used for perception, action and emotion [6, 23]. Following this theory concept TABLE does not have any resemblance to any object in the real world, as concept is a mental unit and tables are pieces of furniture, however concept TABLE has something similar to the experience with a table, that we may have with this object, i.e. our experience and mental image are represented in the same format and in the same systems. According to this theory if we hear or think about the table, it activates the same systems (parts of brain) as if we see our touch it [16].

Zwaan and Madden [24] argue that language acquisition involves co-occurrence of perceptive signal and action on the one hand and language input on the other. When we see a table and hear the word “table” we associate our visual experience of seeing it with audio experience of hearing it. Next time when we hear “table”, this very association will send out a signal for (partial) reactivation of the past experience. It is the activation of the perceptive signal and action that allow us to understand the relevant concept grounding it via sensorimotor experience [16].

Theories of embodied cognition suppose that concepts are grounded not only via linguistic but also sensorimotor experience. A program simulation model in the form of multi-agent neurocognitive architecture (MNA) has been created in order to model the process of conceptualization for the task of formal representation of natural language semantics. The authors in [25] put forward a hypothesis that MNA is an appropriate formalism to use as a basis for artificial intelligent systems. The use of MNA will give an opportunity to model processes that take place in our brain while grounding symbols. This hypothesis is based on the assumption about functional and structural analogy between MNA and human brain [26–28]. The use of this system will provide us with a tool to immerse an intellectual agent into a real-world environment in order to model the process of language acquisition by a baby.

2 Multi-agent Neurocognitive System in the Task of Simulation Modelling of Natural Language Semantics

Multi-agent system is a set of intelligent agents possessing distributed knowledge. It means that no agent can fully work or solve a task by itself, since it does not have required amount of knowledge. In this case task solving is grounded on agent interaction within or outside the system via messaging, searching for or supplying needed information. Agents’ preprogrammed behaviour patterns do not always manage to solve difficult tasks because as in real life we cannot predict all the possible scenarios. So instead of relying on preprogrammed patterns agents have to find a solution by themselves using learning process. Such an approach is comparable to the processes taken part in any human society, i.e. knowledge and experience of group members tend to increase via communication. Similar to social groups agent in multi-agent systems interact with each other for the purpose of knowledge exchange and skill learning [29].

In computer sciences an intelligent agent is a program able not only to strictly follow the task set by a user, but also make independent decisions on the utility of its action and their consequences. The intellectuality of agents is defined by their ability to gather information via sensors and effect the world via actuators.

One more notion that will be used in this work is neurocognitive architecture. It is a common knowledge that the process of thinking is grounded in the energy exchange among neurons in our brain. In pursuance of simulating human thinking processes agents in the developing system are called and presented in the form of neurons. Thanks to a big number of neuron-agents the system achieves its multi-agentness. In order to visualize processes that occur in the system we apply tools of simulation modelling. They allow us to trace the process of agents' creations and regressions. All the agents in the system are well-organised according to their types and functions. Such functionally-determined distribution of agents is called cognitive architecture. Cognitive architecture is an environment for intelligent agents, with the help of it we try to model not only the behavior, but also structural features of the modeling system, i.e. human brain. Figure 1 depicts the 3D model of the multi-agent neurocognitive architecture.

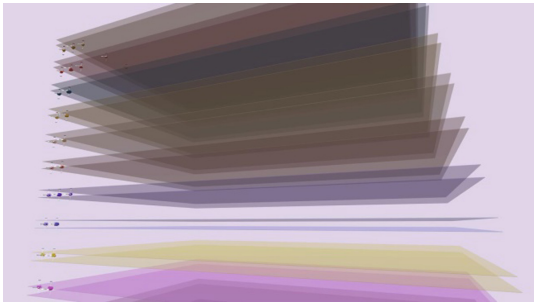


Fig. 1. Multi-agent neurocognitive (MNA). 3D model.

MNA can be used in various fields including robotics, telecommunication, economics, education the site of our special scientific interest is simulation modelling of natural language semantics.

One of the hypotheses made in [25] is that an intelligent agent should be immersed into the social environment with which it will have to communicate via natural language. Such immersion is similar to the Language Acquisition Problem, that describes the way children that initially do not know any language learn it because they are surrounded by a family speaking it. This approach allows to model “natural” way of constructing semantical structure of a natural language utterance.

3 Program Realization

If we follow the theory of constructing natural language semantics based on embodiment, then an intelligent agent must be able to interact with the surrounding world to obtain motor and perceptual experience. For these purposes, robots [30] immersed in the real

environment through a system of sensors and actuators are the best choice. Robots have the ability to ground the symbols of natural language in the process of acquiring them in communicative situations when interacting with people who “communicate” with it in a purposely simplified language. Communication occurs by sending messages via the display and keyboard through the system chat.

The following figure (see Fig. 2) shows a robot used to gain experience interacting with the world. An apple was placed in the robot’s hand, using sensors installed in the “hand”, the robot receives signals that it felt something hard, smooth and round. Cameras built into the robot’s “head” using a pattern recognition system (recognition articles) recognize its color, size and shape.

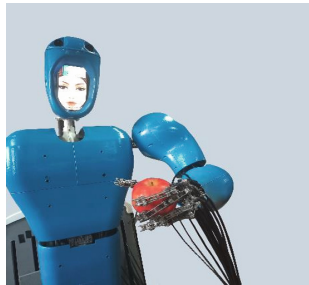


Fig. 2. Robot

Data from the sensors enters the “brain” of the robot, into a multi-agent neurocognitive architecture, where neural ensembles, a group of neurons, are created for each input signal, which are subsequently connected to each other as co-occurrent signals that were received together. The signals from the motor system, the manipulator, and the results of the pattern recognition process, perceptual system, are combined into one neural ensemble. Thus, a neural representation responsible for the appearance and texture of the apple will appear in the robot’s “brain.” Now robot possesses sensorimotor experience that must be supplemented by linguistic input, i.e. we need to give the new real-world object a name.

Provided that this object is encountered for the first time, newly-appeared agents following the rules in their databases will ask a user the question: What is this? The user’s response will also be sent to the system via chat: Apple. This utterance will be processed by the multi-agent system symbol by symbol, i.e. for each symbol; a separate neuron of the symbol type will be generated (see Fig. 3). More detailed description of this process is presented in [31].

Further, these neurons, following their internal logic, will create higher-level agent-neurons, a word with the corresponding name: apple (see Fig. 4). This process is thoroughly illustrated in [32].

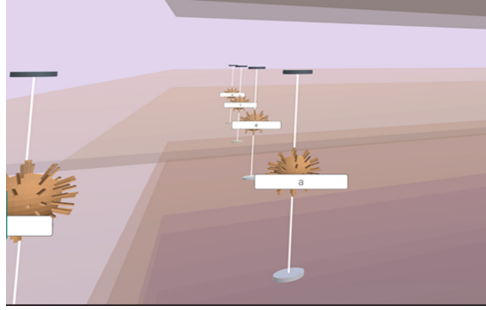


Fig. 3. Neurons created for the symbols of the word “apple”

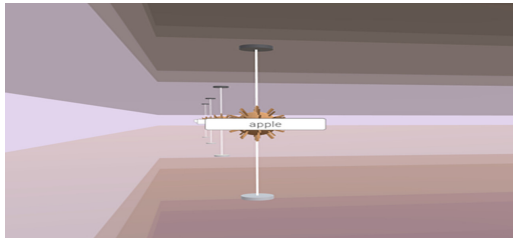


Fig. 4. Neuron created for the word “apple”

4 Conclusions

According to the nature of the MNA all the agents that were generated during this sensorimotor and language experience are connected with each other in a way that they can ask for and supply with information about this investigated object. Thus the semantics of the word “apple” will consist of its external description (shape, form etc.), the manner that it can be hold in a hand and its name.

In this article we have tried to illustrate a simplified way for formal representation of semantics of a single word using theories of embodiment of cognition. We assume that the use of MNA complemented with simulation modelling can be an effective tool to solve problems of symbol grounding and conceptualization in general. Moreover simulation modelling will provide us with the opportunity to “observe the emergence of semantics” occurring in the MNA.

References

1. Evans, V., Green, M.: Cognitive Linguistics. Edinburgh University Press, Edinburgh (2006)
2. Langacker, R.: Foundations of Cognitive Grammar., vol. I. Stanford University Press, Stanford (1987)
3. Harnad, S.: The symbol grounding problem. *Phys. D Nonlinear Phenom.* **42**, 335–346 (1990)
4. Hummel, J.E.: Symbolic versus associative learning. *Cognit. Sci.* **34**, 958–965 (2010)
5. Murphy, G.L.: The Big Book of Concepts. MIT Press, Cambridge (2002)

6. Glenberg, A.M.: Few believe the world is flat: how embodiment is changing the scientific understanding of cognition. *Can. J. Exp. Psychol.* **69**, 165–171 (2015)
7. Glenberg, A.M.: Response to Mahon: unburdening cognition from abstract symbols. *Can. J. Exp. Psychol.* **69**, 181–182 (2015)
8. Mahon, B.Z.: The burden of embodied cognition. *Can. J. Exp. Psychol.* **69**, 172–178 (2015)
9. Mahon, B.Z.: Response to Glenberg: conceptual content does not constrain the representational format of concepts. *Can. J. Exp. Psychol.* **69**, 179–180 (2015)
10. Masson, M.E.J.: Toward a deeper understanding of embodiment. *Can. J. Exp. Psychol.* **69**, 159–164 (2015)
11. Anderson, J.R.: *The Architecture of Cognition*. Harvard University Press, Cambridge (1983)
12. Collins, A.M., Quillian, M.R.: Retrieval time from semantic memory. *J. Verb. Learn. Verb. Behav.* **8**, 240–247 (1969)
13. Kintsch, W.: The role of knowledge in discourse processing: a construction-integration model. *Psychol. Rev.* **95**, 163–182 (1988)
14. Kintsch, W., Van Dijk, T.A.: Toward a model of text comprehension and production. *Psychol. Rev.* **85**, 363–394 (1978)
15. Landauer, T.K., Dumais, S.T.: A solution to Plato’s problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* **104**, 211–240 (1997)
16. Fritz, G., Dudschig, C., Kaup, B.: Symbol grounding without direct experience: do words inherit sensorimotor activation from purely linguistic context? *Cognit. Sci.* **42**(2), 63–69 (2018)
17. Quillian, M.R.: Word concepts: a theory and simulation of some basic semantic capabilities. *Behav. Sci.* **12**, 410–430 (1967)
18. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. Res. Methods Instrum. Comput.* **28**, 201–208 (1996)
19. McRae, K., Cree, G.S., Seidenberg, M.S., McNorgan, C.: Semantic feature production norms for a large set of living and nonliving things. *Behav. Res. Methods* **37**, 547–559 (2005)
20. Smith, E.E., Medin, D.L.: The classical view. In: Smith, E.E., Medin, D.L. (eds.) *Categories and Concepts*, pp. 22–60. Harvard University Press, Cambridge (1981)
21. Searle, J.R.: Minds, brains, and programs. *Behav. Brain Sci.* **3**, 417–424 (1980)
22. Barsalou, L.W.: Perceptual symbol systems. *Behav. Brain Sci.* **22**, 637–660 (1999)
23. Johnson-Laird, P.N.: *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press, Cambridge (1983)
24. Zwaan, R.A., Madden, C.J.: Embodied sentence comprehension. In: Pecher, D., Zwaan, R.A. (eds.) *Grounding Cognition: The Role of Action and Perception in Memory, Language, and Thinking*, pp. 224–245. Cambridge University Press, Cambridge (2005)
25. Nagoev, Z.V., Nagoeva, O.V.: *Symbol Substantiation and Multi-agent Neurocognitive Models of Natural Language Semantics*. Izdatel’stvo KBNTS RAN, Nalchik (2022)
26. Nagoev, Z. V.: Multiagent recursive cognitive architecture. In: *Proceedings of the Third Annual Meeting of the BICA Society. Biologically Inspired Cognitive Architectures, Advances in Intelligent Systems and Computing series*, pp. 247–248 (2012)
27. Nagoev, Z.V.: Ontoneuromorphogenetic modeling. *News Kabardino-Balkarian Sci. Center RAS* **4**(54), 46–56 (2013)
28. Nagoev, Z.V.: *Intelligence, or Thinking in Living and Artificial Systems*. Izdatel’stvo KBNTS RAN, Nalchik (2013)
29. Bennane, A.: Tutoring and multi-agent systems: modeling from experiences. *Inform. Educ.* **9**, 171–184 (2010)
30. Vogt, P.: *Language Evolution and Robotics: Issues on Symbol Grounding and Language Acquisition*. Artificial Cognition Systems. Idea Group, Hershey (2006)

31. Nagoev, Z., Nagoeva, O., Anchokov, M., Bzhikhatlov, K., Kankulov, S., Enes, A.: The symbol grounding problem in the system of general artificial intelligence based on multi-agent neurocognitive architecture. *Cognit. Syst. Res.* **79**, 71–84 (2023)
32. Makoeva, D., Nagoeva, O., Gurtueva, I.: Formal representation of natural language elements in multi-agent system based of self-organization of distributed neurocognitive architectures. *Proced. Comput. Sci.* **213**, 631–635 (2022)
33. Hebb, D.: *The Organization of Behavior*. Wiley, New York (1949)



Natural and Artificial Intelligence: An Activity-Based Approach

Nikolay Maksimov^(✉)  and Valentin Klimov

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute),
Kashirskoe shosse, 31, Moscow, Russia
nv-maks@yandex.ru

Abstract. The generalized composition and structure of rational activity is considered. The model of thought processes is presented in the form of interacting functional blocks of information structures processing and storage. In the context of the activity approach, the basic concepts of natural and artificial intelligence are considered. The necessity to consider intelligence in symbiosis with knowledge and activity is substantiated. The characteristic properties of the essence of “knowledge” are determined: systematization, objectivity and qualification, distribution, granularity. From the point of view of the structural-functional approach, the concepts of weak, strong, and general intelligence are defined.

Keywords: Cognition processes · Information scheme of cognition · Artificial intelligence · Knowledge

1 Introduction

Artificial intelligence (AI) technologies are developing intensively and show impressively successful results in many areas. The concepts (and, accordingly, directions) of strong and weak, general and narrow intelligence have developed and are discussed and developed in different ways. However, despite the fact that intelligence is a tool of cognition, the function of understanding/cognition as such is not defined in AI.

We can say that the global purpose (super task) of AI is the synthesis of new knowledge on the array of old and the development of AI itself. But ideally, the goal of cognition (primarily scientific) is the discovery of a new phenomenon or its prediction, and this, one way or another, practice is a physical way beyond knowledge as an artifact.

AI doesn't have to be humanoid. But can it be non-anthropocentric? The development of natural intelligence (NI) is determined by the existential goal of man—survival in interactions with the environment, human development and environmental change. Accordingly, the functions of intelligence are “combined” in the general process of “life”. At the same time, problematic situations arise directly in the processes of interactions, goals are iteratively identified, tasks are formulated and reformed, and other solutions are found. In this sense, it is unclear what the global goal of AI may be. Nevertheless, the inevitable orientation towards human goals is still obvious, which will allow purposefully and quite effectively combine the functions of AI systems that differ in their purpose within the framework of specific goals.

Consciousness, like information, is an incoming entity that exists, rather, as an action that generates results-macro objects (changes in the environment, artifacts, knowledge). According to the figurative comparison of Joseph Bogen, consciousness is like the wind: it is impossible to see and catch it, but the results of its activity are obvious—bending trees, waves or even a tsunami (cited in [1]).

But speaking about artificial intelligence, it is necessary to understand that we are dealing with a macro-object—a technical system whose components together have “the ability, like a person, to think, interact, adapt to changing conditions and solve other tasks in the field of information processing associated with natural human intelligence” [2].

Currently, the main efforts in the field of AI are focused on modeling and developing inference tools, and mainly on the assumption that the knowledge on which the inference is based is a previously (and to a certain extent subjectively) marked array, the image of which is localized in AI itself. But at the same time, knowledge by its nature is always relative and incomplete. Deduced knowledge should be tested by theory and practice, as well as applied and transferred to other subjects of knowledge for development or revision. That is, it is necessary to determine what to transmit in addition to the received codes—which to form the accompanying context (meta-knowledge, metadata) so that the data is perceived as effective information on which another subject of activity, in turn, will build new knowledge. To do this, it's need to know (or at least model) how (and why!) incoming information is processed and what structures, with what content and how they are used for this. This will allow us to rationally solve problems, organize the reproduction and transfer of knowledge.

2 Intelligence in the Context of Activity¹

Intelligence, as a tool for performing intellectual operations, is a complex of interconnected systems having an informational nature, perceiving, processing, storing and synthesizing image (as object). Accordingly, intelligence should be considered not only as a “brain” (or “processor + memory”), but also as its state (knowledge), and in connection with sensors/effectors that communicate with the environment (sensations/effects), including the use/transfer of generalized knowledge. Moreover, considering the “tool” outside the context of its application (as well as its appearance and development) is limited productive. It would be like developing a computer without input/output devices and regardless of the specifics of its application: the machine will be, but its effectiveness may be limited by the class of the task or completely unacceptable.

The functional-informational scheme of cognitive processes that form the basis of intellectual activity shown in Fig. 1 shows that the stages related to the perception of the received stimulus (definition and evaluation of the value, recognition) depend not only on the stimulus itself, but also on other related (associated with the stimulus) factors: the properties of the need, motives, understanding of the subject areas (SubjAr), etc.

¹ The author's point of view is presented here, the main provisions of which are formulated as a result of the analysis and generalization of numerous publications on consciousness and cognition, a review of which would be unacceptably voluminous for this publication.

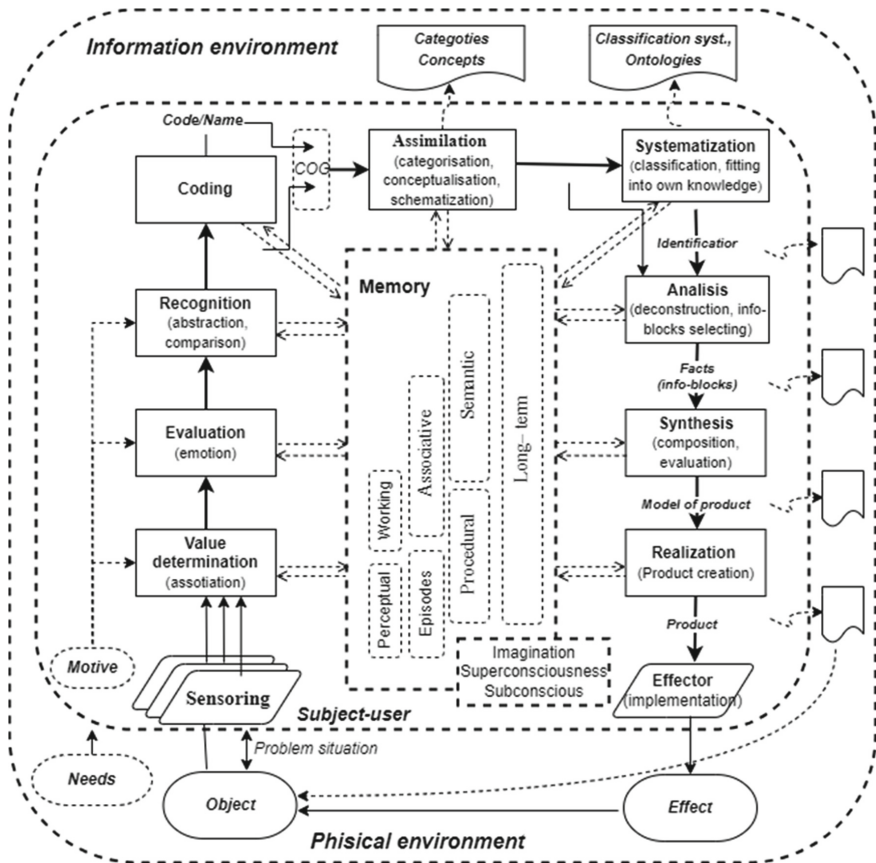


Fig. 1. Functional-information scheme of cognitive processes (is given according to [3]).

A need is a material or ideal entity, something that has a final value for the person and the achievement of which is the meaning of activity. A need is, in fact, a state of an organism when it objectively needs something (a lack or surplus of something, a discrepancy that creates a problematic situation).

The needs, as something necessary, from a functional and systemic point of view can be divided into the needs of the “working body”—that which ensures the existence and realization of the abilities of the intellect, and the needs associated with the development of the organism and its future states (i.e., both the working body and the intellect, and possibly the immediate environment, with by which the interaction is carried out). It is to the latter that the two “disinterested” needs characteristic of mammals, according to terminology [4], belong. This is the need for knowledge and curiosity, as well as the need for empathy (emotional resonance for the emotional state of another individual). Their implementation ensures the future existence and the possibility of accelerated development of the organism and, possibly, society and/or the environment.

The need is specified through objectification: it becomes a definite need—an entity with properties. Achieving (or not achieving) a goal, i.e. full or partial satisfaction of a need, causes emotions in a person, which leads to changes in the state of the nervous system, and thus the formation of an assessment of the stimulus, as well as, possibly, the motives and the need itself.

The need is somehow associated with the ways to achieve the target state, as well as the effect—positive or negative emotions for the subject or the reaction of the environment to the changes that have occurred as a consequence of satisfying the need. Such associations represent a motive—why and how the activity will be carried out. The motive is revealed by the person through experiences characterized by positive or negative emotions, which not only stimulate the process of cognition, but also bring illogic into thinking.

A stimulus is a material or ideal object, condition, etc., situationally arising in the course of activity, associated in some way with a need, for example, presumably entering the structure as a part or having properties that lead to the satisfaction of a need.

In cognition, the stimulus of activity is uncertainty, acting as a property characterizing a problematic situation [5]. The state of uncertainty (the ratio of stimulus and inner knowledge) is a symptom of a possible contradiction. If this contradiction is investigated, it will be formalized in the form of a problem—an established, recognized contradiction that does not yet have a solution or may have ambiguous solutions. Such a solution can be found both in a rational way—by reducing to solved problems (methods in the corresponding class of problems), and irrational, that is, going beyond the bounds of existing knowledge. The solution of the problem is closely related to the choice of evaluation criteria (including alternatives representing different points of view and grounds), for example, such as simplicity, justification of plausibility, syntactic, semantic and other criteria.

At the same time, already at the staging part of the process of purposeful activity (and, in particular, the solution of the problem /task and the synthesis of knowledge) associated facts—information, like “nutritious broth”, accumulate at each step. These facts, most likely, were not used in the original formulations of the problem/task/criterion, but they can be useful as alternative or additional when evaluating intermediate and final solutions, providing visibility when proving the viability of the solution, or when reforming the problem, as well as completing and rebuilding one’s own knowledge. They also form a “contextual field”—some meaningful basis for organizing the subsequent application of the solution and its development.

Moreover, getting information about your condition makes it possible to correlate your development with the development of the whole, but not individual parts, and allows you to already make self-assessment. In the case of AI, this is mainly the construction of output according to a concise specification of the task/request based on accumulated generalized knowledge.

The choice of the trajectory of the development of the system (it would be more correct to say “organism” according to A. Bogdanov) that ensures the management of interaction with the environment is based on the use of values (criteria, sense, meanings). And for self-developing systems that implement and manage their own structure, it is also a solution to the main, but implicit, existential task—ensuring the safety of the

organization of the system. It follows from this that along with the value system that determines from the outside (in relation to the system) the expediency/effectiveness of functioning (of the system in relation to the environment), there should also be an internal value system that provides autonomous control over perception, learning, inference, as well as its own reorganization.

Cognition, learning, solving problems and problems—that is, what relates to intelligent (intellectual) activity,²—this is the realization of the cyclopausal chain of interaction with the *environment—intelligence—knowledge*. Here, the action provides a connection with the external, as a result, with the physical environment. Knowledge provides a connection in time and space with the information environment of analysis/synthesis of information/knowledge. Intelligence provides a “functional” connection of knowledge and activity in the context of its purpose, forming decisions and evaluating their effectiveness.

A characteristic feature of the distributed intelligent activity process, the generalized scheme of which is presented in Fig. 2, is that the practical result is obtained in accordance with a certain model. Such a process can be represented as predictive and iterative, using feedbacks. The goals are determined by the current and expected state (in the limit—“ending technology”, bringing together a chain of strategic and technological decisions aimed at achieving a global goal). The ways to achieve the goal are determined by the available knowledge/experience and resources available to the subject, and the criterion for achieving the goal (and to some extent the goal itself) is also determined by the will, which determines how persistently the subject will act, especially in case of obtaining a not completely satisfactory result. It is also significant that the intermediate results obtained or the changed circumstances can, among other things, radically change the course of the process, redefine goals and paradigms.

The entity “Person of activity” presented in the diagram functionally corresponds to the concept of “organism”—a system of systems: a symbiosis of heterogeneous systems having a common goal (more precisely, subsystems related to its achievement) and some common element base.

A model is an image of a “Goal” (transformation, interaction, object, etc.)—something that allows you to predict the properties and/or behavior of the simulated object. This is a functionally identical transformation of the image of the initial state of the simulated object into the image of the target state, which is considered as a predefined predicted state.

Note that the construction of the model is determined not only by the properties of the object being modeled and its interactions with the environment, but also by the possibilities of practical/experimental verification of the adequacy of this model, including

² For reasonable (or rational, intellectual) activity, in addition to multi-step, multilevelness, as well as interactivity and iterativity are characteristic. The implementation and control of the activity is based on the model that determines it and the use of feedback. The activity is based on a coordinated combination (by properties) of its main entities: *the needs and motives, goals and means, methods and criteria, results and effects*. Intelligent activity is initiated and controlled by the mind, and first a judgment is formed, then an action is performed, then an assessment of the action and judgment is made.

the assumed necessary measurements of the initial and final state, or consistency (or explanation) with previously obtained data.

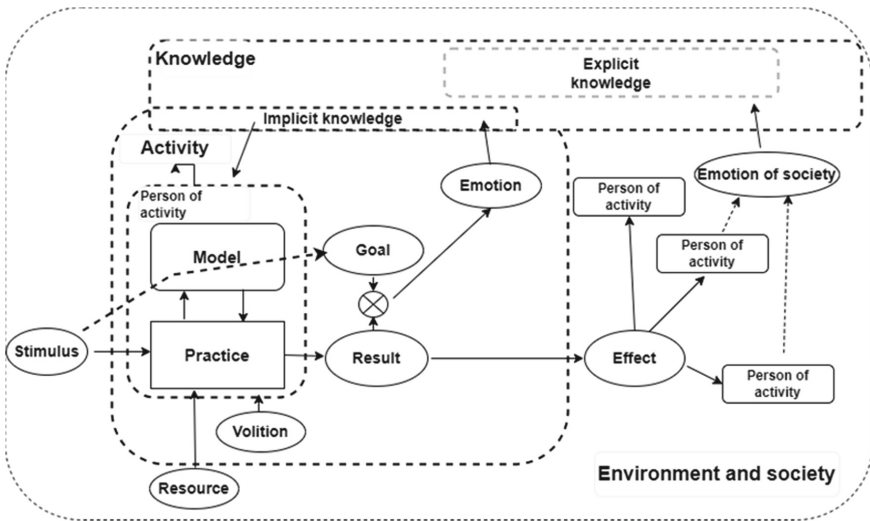


Fig. 2. Diagram of the distributed intelligent activity process.

Emotion here is an assessment of changes in the state of the organism in the categories of the system of internal values as a result of the action of the stimulus. But this is not only the significance of the result obtained, but also the effectiveness and “convenience” of obtaining it, the expected and actual impact on the development of the model and practice, as well as compliance with available knowledge (including relict), etc.—everything that explicitly or implicitly affects the psyche. Despite the fact that emotion has a multimodal nature, it is a holistic assessment that cannot be reduced to individual indicators. It is important that from the point of view of the formation of knowledge as an object, emotion is one of the components of the context of the corresponding element /fragment of knowledge, which contributes to their identification, identification and use.

3 Knowledge

The result of the activity is not only the actual target result—a change in the state of the environment or the subject itself (which also cause emotions of society), but also knowledge. Moreover, the knowledge obtained is not only a separate “product” used for the application and synthesis of new knowledge, but also a resource for reproduction (including training, education) of the cognizing subject himself, and, in particular, for the training of intelligent systems. It should also be noted that a by-product “technological” product of purposeful cognitive activity is the creation of a communication interface (languages, systems of quantities, categories and concepts, as well as linguistic support) that connects the cognizable and the cognizer.

Knowledge, even as an end in itself of the process of cognition, is not a product of creativity alone: interacting with reality, conducting experimental research, the subject seeks to organize them in such a way as to be able to “enter into a dialogue” with reality: “ask her a question” and get an answer. That is, the process of cognition can be represented as a communicative activity. In this process, the stages of “parallel”, to some extent autonomous development of theory and experiment, culminate in their symbiosis. This is a special integrative unity with respect to autonomous theoretical structures interconnected by a system of marginal transitions—the principles of conformity [6].

Knowledge from the point of view of existence is a model of some part of reality in the consciousness of the person (so-called implicit knowledge) or in a scientific discipline (generalized documented and so-called explicit knowledge). Unlike signals and data, which are part of physical reality, knowledge is the reality of the subject’s consciousness (more precisely, the state of the brain—a biophysical substance, or computing environment, which are also parts of reality). Knowledge has an abstract-concrete nature and they are formed by reflecting the perceived part of reality on the cognitive state of the person—his available knowledge, tasks, needs, motives.

From a semiotic point of view, knowledge is information related to the context of its application. That is, *knowledge* is a *message* (signs representing the content) + *the conceptual base* (concept) + *circumstances of use* (connotation).

From the point of view of the form of existence and semantics, knowledge is a set of facts that:

1. systematized, i.e. ordered and interconnected in accordance with some categorical-conceptual system reflecting the basic properties and interactions that define this SubjAr;
2. objectified, i.e. identified by a place in the system (classification) of knowledge—“objects” defined through characteristic properties;
3. qualified³ taking into account the scope and conditions of obtaining/application, i.e. they have an assessment of consistency: internal completeness and consistency—by theory, and adequacy and effectiveness—by practice/experiment;
4. attributed according to the properties and interactions defining a given SubjAr, and also possibly named.

From the point of view of knowledge representation, it is a symbolic representation of real or abstract objects, phenomena, processes, models, theories, descriptions, etc., fixed on a carrier, constructed in accordance with some language of description/modeling, based on categorization—a certain set of characteristic features (in reality—a set of properties), and in general—a system of signs. A special case is the ontological form of knowledge representation [7].

Knowledge can be presented in a deterministic symbolic form (traditional texts, schemes, formulas, etc.) or in a probabilistic form reflecting statistical relationships of data in the array of knowledge. Similarly, the output construction method can have

³ Namely the property of the qualification of knowledge elements, or rather its absence in the training array in an explicit form, that determines the limitations of statistical (streaming) learning in AI.

a scheme “step by step” of sequential “growing” of the result, or “holistic”—without presenting intermediate results and grounds for selecting and linking operational objects.

Knowledge is not and cannot be absolute and final, because it is a reflection of reality through feelings and further—a system of quantities.

Knowledge from the point of view of the process of their construction is “recursive”: facts and statements (axioms) are taken as a basis, which are considered reliable at this moment, and then constructions (theorems, models) arise on their basis, which are theoretically and/or experimentally verified, generating new data and problems, which, in turn, lead to a revision of the initial provisions, etc.

Knowledge is distributed in “space”—between subjects and in time—by stages and forms. The distribution “in time” is also an opportunity not only to transfer knowledge to the next generations, but also to “myself”. The latter is also important because the person of cognition develops, not only the state of his current knowledge changes, but also points of view, i.e. an explicit comparison with the previous one is useful and constructive. It is a distributed environment that unites the subjects of cognition both functionally—organizing activities, and information—systematizing and ensuring the availability of knowledge. Consciousness using personal knowledge, through means of communication (which are also the embodiment of knowledge) interacts with generalized knowledge, as a result of which new personal knowledge is formed (not only practical and abstract, but also emotional and evaluative) and, possibly, generalized knowledge is replenished.

If there are more than one persons of activity, then such a system should have an organization (and not just a structure): a sphere that ensures sustainable interaction (communication and management). And if the persons have different “element bases” (for example, in the case of human-machine systems), then there are several languages and corresponding translators. That is, the semantic feature of the activity is the distribution and ambiguity of meanings: the volumes and meanings of even the same concepts for different person or different subject areas may not coincide, since they may have different contexts.

Knowledge/information is stratified by forms of representation. For implicit knowledge, these are images—representations of sensation, perception, recognition, naming (coding), and for explicit this is a categorical and conceptual apparatus, fundamental theories and the most representative stratum—applied knowledge. This has the consequence:

1. the multiplicity of instances of a separate fragment (elements) of knowledge and the corresponding context;
2. the components are “categorically” divided (situationally and variationally) by roles into objects and tools (implementing the principle of complementarity), which provides the possibility of analysis (division by some property) and synthesis of knowledge (possibly by another property), as well as the development of activities, including cognitive.

But at the same time, it should be noted that in AI (in particular artificial neural networks), the initial data for learning and drawing conclusions are mainly texts related to numerous and diverse applications—images (objects) of the upper stratum, which quantitatively practically suppress (in a statistical sense) the data of other strata. Moreover, the data of the strata of individual perception-understanding, which are “closer to

reality” and therefore relatively adequately reflect it, constitute implicit knowledge and therefore are practically not used.

Knowledge is conditional and relative. And they are objective exactly to the extent that the properties and states predicted by the model will correspond to the quality of the results (which are also images—measurements in accordance with some model) that will be obtained as a result of interaction with reality. Knowledge fractally represents reality. They are fragmented and granular: by its nature, any result, any knowledge is an aspect or a particular image, which corresponds to the principle of complementarity (in this area, the principle of complementarity/uncertainty means that the meaning of a fact/message is always determined by some context: understanding requires the use of additional descriptions). Moreover, in the case of an explicit form of the existence of knowledge, fragmentation and granularity are determined by the nature of the division of reality and the detail (accuracy and completeness) of the representation (model, description) necessary for the applied sphere. In addition, the “clarity” of the form requires an explanation and justification of its choice, which in itself forms a new associated knowledge that increases completeness and evidence.

But it is precisely these properties that determine the possibility of forming an information field for synthesis based on “old” knowledge of new solutions and their evaluation.

4 Intelligence

Generalizing, we can say that Intelligence is a logical apparatus that uses and generates knowledge, has the ability to organize and control actions in accordance with a certain model, as well as the ability to form and evaluate these models. Intelligence is a consequence of: (1) interaction (including informational type and feedback) with the environment; (2) multilevel representation of knowledge in local (subject) and external memory; (3) the use of language as a means of representation, communication and modeling.

The artificial intelligence system, by analogy with the von Neumann computer architecture scheme—“*computer = processor + memory*” and with a similar generalized program structure of N. Wirth’s—“*program = algorithm + data*” can be considered as “*AI system = derivation algorithmy + knowledge*”. Moreover, as noted in the fundamental works on programming, such a dichotomy is quite conditional: programs (more precisely instructions) exist in the form of data and, accordingly, can be modified in the process of its implementation (which is the fundamental basis of self-development), and methods of constructing inference (solutions) are knowledge recorded also in the form of data, which may also change during their use. Processor/memory is a hardware environment, the elements of which differ in function, but which have a common purpose—to save or change the state of the elements.

That is, practically and essentially, the computer constructs *Computer—Program—AI System* are collectively a stack of nested objects (more precisely, a “matryoshka”). Accordingly, the AI system is a software implementation of logical methods on data in a computing environment. Moreover, just as NI is basically physical processes and structures, so AI is information processes using data structures.

Note also that this stack is similar to the *Data–Information–Knowledge* stack [8, 9], reflecting the complexity and nesting of objects that are processing components at the appropriate levels. That is, knowledge = <<<*data & metadata* > & *meta-information* > & *meta-knowledge* >, where the first pair (operands connected by the operator “&”) is the actual data and the encoding method, the second is the perceived text and context, i.e. the semantic content that is interpreted or applied within the framework of meta-knowledge—the accepted axiomatics and the circumstances of use.

In this sense, AI is a tertiary model of reality, i.e. it is a model built on secondary models (knowledge, representations, descriptions), which, in turn, are based on primary models—sensations manifested in the nervous system when exposed to a stimulus, which is the only reality. And knowledge is an image of the world (reality, including the knower), formed on the world of images in accordance with a certain image.

Speaking about the boundaries of weak and strong AI, we can say that weak AI (more precisely, narrow, specialized AI) corresponds to reason. And if the AI system has the ability to (1) selectively interact with the environment and other AI agents (which is quite feasible in a digital environment) and (2) identify problematic situations, as well as form goals, show “curiosity” and “interest” in finding more favorable conditions of existence, then it will be a strong AI. And if he also possesses and is able to form a system of internal values, and is also capable of self-reproduction and self-improvement based on stored, synthesized and/or externally found knowledge, then this will be general artificial intelligence (AGI)⁴. That is, we can say that AGI is an AI system capable of rational (i.e., in accordance with its defined goals, resources available to it and accepted criteria) activities (i.e. purposefully interacting with the environment and knowledge resources), including identifying problematic situations (and possibly their expectation, i.e. modeling of possible), as well as dynamic (in the course of activity) changes in the selected goals and paradigms.

It should be noted that all the components and functions of natural intelligence presented in the above diagrams are already practically implemented in one way or another in one or another composition. In particular, it is possible to implement (more precisely, for now, simulation) emotions and needs, although their adequacy to the global goals of the organism-society is doubtful. In essence, this is already artificial life, which will have to interact and compete with other forms of life, including humans and society. Theoretically and practically, this is quite feasible (including and the formation of goals by the system and the simulation of emotions): all the components and functions of intelligence presented in the above diagrams are already practically implemented in one way or another in one or another composition. But it follows from this that the meaning of this life and, accordingly, the standards should have been determined before its real origin.

⁴ Or, according to [10], an *autonomous AI* with adaptability, emotional apparatus and freedom of decision making.

5 About the Ratio of Artificial and Natural Intelligence

NI, considered as a component of the activity process, as a self-developing mechanism that implements it, can be represented as a hierarchical system of distributed interacting systems (peripheral and central nervous systems, conceptual systems, physiological systems, communication systems, social systems, etc.). These systems are dispersed (1) by the environments of functioning: organism (physical environment—body, nervous system), mental environment (psyche, consciousness), communication (language, channels), environment and society (realizations/results, effects), and (2) by the strata of representation: sensations, recognition, understanding, analysis-synthesis, implementation/evaluation. At the same time, for each state (stratum-environment) there is (1) its own operating space—a set of objects in some memory area; (2) their criteria (thresholds) for the success of the result, and the higher the complexity of the stratum, the more factors will be in the composition of the criterion; (3) their rules for performing iterations necessary to fine-tune the solution, and the stopping criterion will take into account not only the necessity/significance of the result, but also motivation. And the will of the subject in achieving the goal.

The multimodality of representations (as a consequence, the heterogeneity of subsystems and the environment) determines that one perceived object generates a certain variety of images, creating the possibility of explicitly or implicitly conducted multidimensional analysis and the possibility of using several factors. Their diversity provides greater completeness in the search, selection and evaluation, as well as provides the preparation of alternatives.

The hierarchy of the aggregate system of activity assumes that the subsystems have different goals and, accordingly, use different fragments of knowledge, which may be differently represented and differently ordered. Therefore, it is necessary either to coordinate the goals—to change them (and, possibly, the methods of solution), and perhaps not only local, but also global—of the whole organism (individual and/or society), or to choose one of them, suppressing the goals of all other subsystems.

In the case of NI, such coordination/choice are implicit processes (or maybe a one-act process, if viewed from the point of view of the quantum wave approach), which is practically not controlled at the logical level, which is difficult to analyze and verify. But in practice, it usually leads to an adequate result, which is probably due to the multimodality and granularity of the representation of knowledge, the stock of implicit (including negative) knowledge, as well as the diversification of needs, motives and goals, but, most importantly, their harmonious naturally formed combination.

In the case of AI, the goals, tasks and knowledge used in this case are defined (more precisely, formed) initially, unambiguously and in a well-defined form, and based on the expediency and expected usefulness (for a person and a task), as well as the intended way to get a result.

6 Conclusion

It is appropriate to form a working “principled” (logical) AI scheme and identify possible architectures (as a set of essential, defining properties, and not composition and structure, as is often found in modern publications) in conjunction with possible areas of application and classes of tasks.

For the sustainable development of AI, it is necessary to formally define knowledge (and, preferably, standardize it in conjunction with the definition of AI), its types, kinds, forms of representation and, accordingly, languages of definition and manipulation, as well as methods of using knowledge, including for training AI systems.

It is necessary to determine strategies for the development and application of AI and, accordingly, forms and technologies for presenting knowledge, and ultimately determine whether to go along the path of creating one global SuperBrain (and perhaps an information repository), or a System of interacting heterogeneous specialized AI systems. But it can be noted that already due to the tertiary nature of AI (the reliability of conclusions, with the exception of trivial or formal so-called “truths”, will be in doubt and will require additional actions that must be performed externally), a distributed form suggests itself.

And, obviously, it is necessary to preserve independent natural forms of knowledge storage/retrieval (libraries, information systems, etc.), which include not only an array of documents/records, but also means of systematizing knowledge (classifications, indexes, thesauruses, etc.). Indeed, the characteristic properties of knowledge—systematization, objectivity and qualification, determine that any element of established or newly knowledge should be correlated with some other stable entity: scientific disciplines, facts, classifications, etc. Here, “stable entity” means the separateness of its existence as an independent object, and “stability” should not be understood in the physical aspect (as immutability), but rather statistically. That is, the semantics (essence, meaning) of a knowledge element can and should exist in many instances (as information objects): in different types and forms, with different completeness and may be in different languages. In addition, it is important to understand that “to exist” means to interact: each instance of an element of knowledge generates its own semantic chain leading to new knowledge. At the same time, a semantic field is formed, creating, in particular, a basis for analyzing the properties of this element of knowledge and its recursive actualization.

The preservation of knowledge and information is not only and not so much an obvious function of ensuring the non-destruction of their carriers (more precisely, the immutability of the state), as the maintenance of the system component (its actualization and harmonization—a change reflecting the development of science and practice), as well as the means of communication—languages and carriers. It is languages in the conditions of constant changes in consumer perceptions of objects and objects of reality that will provide the necessary availability of information that has not yet lost its usefulness, but will not be found as a result of changes in the SubjAr language that have occurred over time.



Acknowledgements. This work was supported by the Ministry of Science and Higher Education of the Russian Federation (state assignment project No. FSWU-2023-0031).

References

1. Chernigovskaya, T.V.: Cheshirskaya ulybka kota Shredingera: mozg, yazyk i soznanie [Cheshire Smile of Schrödinger's Cat: Brain, Language and Consciousness]. Izdatel'stvo AST, Moscow (2021)
2. National Standard of the Russian Federation Artificial intelligence systems: Methods for Ensuring Trust. General. GOST R 59276-2020. Standartinform, M. (2020)
3. Maksimov, N., Golitsina, O.: About cognitiveness of information retrieval. *Proced. Comput. Sci.* **213**, 317–324 (2022). <https://doi.org/10.1016/j.procs.2022.11.073>
4. Simonov, P.V.: Motivirovannyj mozg [Motivated brain]. Nauka, M. (1987)
5. Dorozhkin, A.M.: Nauchnyj poisk kak postanovka i reshenie problem [Scientific Search as Setting and Solving Problems]. Nizhegorodskij gumanitarnyj centr, Nizhny Novgorod (1995)
6. Arshinov, V.I., Budanov, V.G.: Kognitivnye osnovaniya sinergetiki. Sinergeticheskaya paradigma: Nelineinoe myshlenie v nauke i iskusstve. pp. 67–108. Progress-Traditsiya, Moscow (2002)
7. Maksimov, N.V., Lebedev, A.A.: Ontological system knowledge-activity. *Ontol. Proekt.* **11**, 185–211 (2021)
8. ISO/IEC 2382-1: Information Technology. Vocabulary. Part 1: Fundamental Terms. ISO, IEC (1993)
9. Golitsina, O.L., Maksimov, N.V., Popov, I.I.: Informacionnye sistemy i tekhnologii [Information systems and technologies]. FORUM: INFRA-M, Moscow (2014)
10. Zhdanov, A. A.: Avtonomnyj iskusstvennyj intellekt [Autonomous artificial intelligence], 4th edn. <https://postnauka.org/books/38231>. Accessed 21 Sept 2023
11. Burtsev, M.S., Bukhvalov, O.L., Vedyakhin, A.A., et al.: Sil'nyi iskusstvennyi intellekt: na podstupakh k sverkhrazumu [Strong artificial intelligence: approaching the superintelligence]. *Intellektual'naya literatura*, Moscow (2021)
12. Maksimov, N.V., Golitsyna, O.L.: From semantic to cognitive information search: the fundamental principles and models of deep semantic search. *Autom. Docum. Math. Linguist.* **56**(3), 145–159 (2022). <https://doi.org/10.3103/S0005105522030074>



Models of Generation of Statements of Various Genre Types According to Data of Early Speech Ontogenesis: Imperative Versus Informative Genres

Irina G. Malanchuk^(✉)  and Anastasia N. Korosteleva 

National Research Center Kurchatov Institute, Acad. Kurchatov Sq., 1, Moscow, Russia
coral@inbox.ru

Abstract. The paper presents computational models of the generation of statements depending on the types of speech genres they belong to. The modeling was carried out on the basis of the primary psychological and linguistic analysis of more than 300 speech products of children of the second year of life in various communicative situations. A system of parameters for assessing natural speech is presented, including intentional complexes—speech motivators, gender categories, speech situations, the speaker’s communicative status, and other non-linguistic parameters of statement production. Regression analysis using the method of multinomial logistic regression showed the contribution of a number of social, psychological, communicative factors in the production of genre types which are typical for the specified age. The paper focuses on the results of the regression analysis for the two most frequent types of genres—imperative and informative. It was determined that the differences are associated with the level of contribution in the production of imperatives of needs in a social being, positioning, the need for a material object, the need to prevent damage, to express one’s state/thought, the need for cooperation, support, and reflection of the speech strategy; when producing informative statements—the need to change the objective situation.

Keywords: Speech generation · Speech genres · Imperative genres · Informative genres · Intention · Communicative needs · Speech goals · Parameters of a communicative situation · Multinomial logistic regression · Early ontogenesis

1 Introduction

Speech—in its various implementations (external/internal, oral/written, in aspects of production and perception, communication of speech and language with social and “objective” thinking, etc.)—is the most complex socio-communicative and intrapsychic phenomenon requiring rapid formation and continuous development of a huge range of intellectual skills covering social cognition, speech-communicative reality and language as a means of explication of various structures of the conceptual picture of the world.

There is no doubt that the analysis of speech processes, carried out in a variety of ways in different disciplines (phonetics and acoustics of speech and language, many sections of psychology and linguistics) still needs the development of conceptual fundamentals, verification of general and particular concepts of speech. The analysis is carried out for the development of communicative artificial intelligence, and for some clinical purposes—a diagnostic one.

In the aspect of the stated topic, it is necessary to update both the already known ones and the latest scientific ideas about the intentionality of speech and typologies of speech genres.

The problem of the intentionality of speech/statement is connected, first of all, with the theory of speech acts, where the intentional state of a potential speaker and the intentionality of a statement have become the fundamental concept for distinguishing types of speech acts [1–12]. In connection with the idea of the general purpose of a statement, usually one intention is defined, expressed by one or another grammatical indicator. However, this is not entirely accurate, and some modern authors distinguish two types of intentions—cognitive and communicative [13, 14], acting simultaneously [14].

Thinking about the intentionality of speech/statements, we developed our own concept, which was formed based on the analysis of more than 7200 speech units. Speech production motivators were identified for the each of them—from one to whole complexes of such motivators. The key concept that forms the basis of the description of the intentions of the author of the statement is the psychological concept of “need”. The concept of intentionality obviously correlates with the problem of needs and motives of speech, developed in the psychology of activity and speech activity [15, 16]. In this case, these are communicative and communication-related needs [17] (see also [18–20]).

Based on the results of the analysis of the array of speech products, we previously identified 11 types of needs that form the motivational basis of speech activity in their various configurations. Those are: the need for a social being; the need for attention; the need for positioning; the need for information received or verified at the expense of a communicative partner; the need for a material object, the receiving of which is possible with the help of a communicative partner; the need to prevent potential damage; the need to change one’s emotional state; the need to change the objective situation, including the communicative, social reality of interaction; the need to express one’s state/thought; the need for cooperation, support; the need for identification [17].

The analysis shows that part of these needs/intentions is directed at external objects and quasi-objects, which include material and informational values; the other part is directed at the social partner in order to establish and maintain contact, provide support in achieving the goal, as well as at himself in the tasks of changing behavior directed at the partner. In addition, there are intentions of a “deep” psychological nature aimed at maintaining the dynamic balance of the psycho-emotional system. This also correlates with our analysis of goal-setting in speech communication acts, which was undertaken earlier (case studies are presented in [21]).

Thus, the following conclusions were made:

1. groups of needs are intentional complexes—speech motivators;

2. there is a multi-vector nature of speech-communicative intentions and goals, their focus on external objects, the social partner and the “I” as elements of the communicative system and their relation to the psychodynamic state of the speaker.

As for the problem of differentiation of statements, over the past 50 years their attribution to specific genre forms and types has remained not only relevant, but also promising. In the development of the theory of speech acts in line with the theory of speech genres, a theoretical model of the speech genre [22], typologies of speech genres [22–25] are proposed, intentions characteristic of various types of discourses are studied [26–28].

With regard to the typology of speech genres, we have taken as a basis the typology [22]. It was supplemented for early speech pathogenesis by expressions that translate affect in the form of vocalizations and interjections, and ritual genres that are associated with the development of communication rules in the child–adult dyads and the beginning of the development of children’s communicative rules established in the microsocium as a social whole.

It should also be noted that our development of the problems of speech communication, factors of speech production in the categories of speech genres is carried out in line with our theory, differentiating speech and language in their specific conceptual content [17, 29], as well as the experimentally proven difference between the two types of neurophysiological processes: (1) the ones aimed at providing an analysis of the socio-cognitive parameters of the communicative situation extracted from vocal-speech forms, and (2) language processing [30]. Therefore, our study of speech constituents, which has a multidisciplinary nature, focuses mainly on non-linguistic parameters (factors) of speech production of various genre types, including socio-cognitive reflexive processing of speech strategies, as well as reflexive processing of actual results of using language as normative/non-normative (for language reflection in children, see, for example, [31–34]). The latter is especially important for studies of deviant and distorted speech ontogenesis from the point of view of the formation of the skill of self-control (cognitive controlling) of speech (cf. [35, 36]).

2 Material and Methods

Earlier, we created a speech database with more than 7200 units of children’s speech. Three hundred of them were selected for modeling speech production by children of the second year of life. The selected units represent exchanges in dialogues, statements in autocommunication mode, and statements addressed to characters in the game. Subsequently, models related to each child’s age will be presented. Speech production models will be compared by genre types in age dynamics.

The primary analysis of speech genres characteristic of early ontogenesis made it possible to optimize the list of types of speech genres. Imperative and informative genres are implemented with the greatest frequency in this age sample. The remaining types of genres—ritual, performatives, expressive, and evaluative genres—will not be discussed in this work due to the limitations on the volume of the paper.

The primary psychological and linguistic analysis was carried out according to the following parameters:

- gender (girls/boys, encoding 0/1 respectively);
- types of addressees: people (mother, father, etc.), speaking in their presence, auto-communication, addressees—toys and animals (encoding 0/1/2/3/4, respectively);
- speaker's communicative status (high—0 or low—1);
- type of situation (natural communication—0 or game—1);
- features of speech behavior (initiative—0 or response—1);
- features of speech and language described in the categories of error-free statement production—0/speech, language errors, errors of logical construction of verbal statement, connectivity errors at the pragmatic, communicative, semantic levels—compared with the adult speech-language norm—1 for each of these items separately;
- types of linguistic and speech-communicative reflection: the construction of a statement automatically (0) or using markers of reflexive speech–language behavior with the implementation of various types of reflection—phonetic, lexical, word-formation, syntactic, grammatical (in the aspect of inflection), awareness of speech strategy, speech genre, communicative rules, reflexive evaluation of linguistic expression of content—1 for to each item separately;
- needs—motivators of speech production (see the list above; absence—0 or presence of each of the needs—1).

To create computational models of speech generation in their relation to the types of speech genres, the method of multinomial logistic regression (MLR) for categorical (attribute) data was chosen. The analysis was carried out using Python programming in the Jupyter Notebook software environment using the Python—scikit-learn library. The choice of the MLR method was determined by the nature and structure of the data: the above types of speech genres are categorical dependent variables, in the process of speech production, this variable acquires more than three values, which is a condition for using the multinomial logistic regression method [37–39]. The independent variables are the parameters described above, representing the systems of processes and taking into account the conditions of communicative acts for launching and implementing speech production.

We created and trained a multinomial logistic regression model using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno optimization algorithm (LBFGS), an iterative method that efficiently processes large data sets without having to calculate the full Hessian at each iteration. The model was trained on the basis of independent (x) and dependent (y) variables. After training, the absolute values of the coefficients of the model were recovered. In order to assess the contribution of each variable to the dependent variable, we applied regression coefficients to the relative frequencies of each variable. This allowed us to estimate the contribution of each variable to the forecast of the dependent variable.

3 Results and Discussion

Figures 1 and 2 show the results of a regression analysis using the multinomial logistic regression method, carried out considering all the described independent variables. For the convenience of visual perception, the regression coefficients in the figures below are

given modulo, the \pm sign will be clarified in the discussion and interpretation of the results.

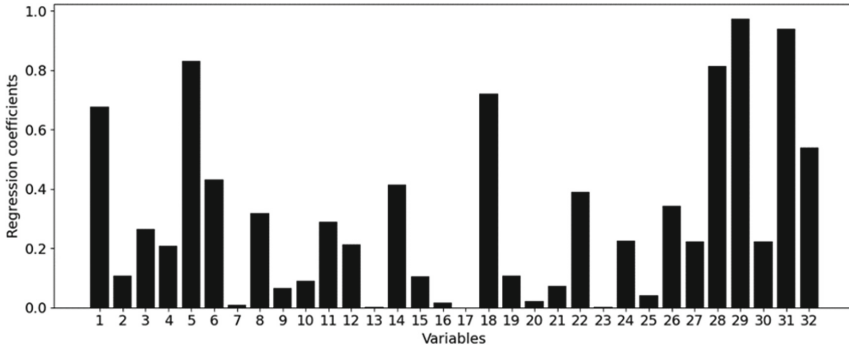


Fig. 1. The model of producing imperative statements: second year of life. The contribution of independent variables 1–32 to the production of imperative statements. The values of the coefficients in the figure are given modulo. Variables: 1—gender, 2—type of addressee, 3—communicative status of the speaker, 4—situation (natural communication/game), 5—speech behavior (initiative/response), 6—language errors, 7—speech errors, 8—logical errors, 9—communication errors (pragmatics), 10—communication errors (communicative level), 11—communication errors (semantics), 12—automatism/lack of automatism, 13—language reflection (phonetics), 14—language reflection (vocabulary), 15—language reflection (word formation), 16—language reflection (syntax), 17—language reflection (grammar in the aspect of inflection), 18—reflection of speech strategy, 19—speech genre reflection, 20—reflection of communicative rules, 21—reflection of the content of the statement, 22—the need for a social being, 23—the need for attention, 24—the need for positioning, 25—the need for information, 26—the need for a material object, 27—the need to prevent potential damage, 28—the need to change one’s emotional state, 29—the need to change an objective situation, including social reality, 30—the need to express one’s state/thought, 31—the need for cooperation, support, 32—the need for identification.

When presenting the full data of regression analysis here, we will discuss in detail only the contribution of key social, socio-psychological and speech-cognitive independent variables to the production of imperatives and informatives, such as: (1) systems of needs as motivators of speech production of a certain genre type; (2) gender as a characteristic that can significantly determine the characteristics of motivation and quality of social interactions, including the use of speech and language [40–44], cf. [45]; (3) initiative/response speech behavior, which is a significant socio-speech category for the organizing of various discourses and other features of speech production (see, for example, [46]); (4) types of addressees; (5) types of situations (natural communication/game); (6) the communicative status of the speaker; (7) reflection of communicative rules; (8) speech-genre reflection; (9) reflection related to the choice of speech strategy, which seems significant mostly for the production of imperatives.

As expected, and as the primary analysis showed, whole complexes and various configurations of needs determine the production of speech.

When categorizing statements as **imperatives** in the confidence interval from 0.025 to 0.975, there were—with a decrease in the regression coefficient—the need

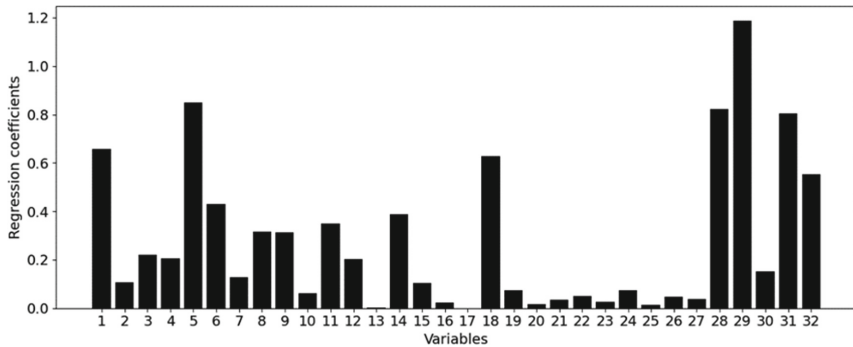


Fig. 2. Model of producing informative statements (second year of life). The contribution of independent variables 1–32 (see the caption to Fig. 1) in the production of informative statements. The values of the coefficients in the figure are given modulo.

to change the objective situation ($b = 0.97$), including socio-communicative, in cooperation/support ($b = 0.94$), in changing one's emotional state ($b = 0.81$), as well as the need for identification ($b = 0.54$), in a social being (social partner) ($b = 0.39$), in a material object ($b = 0.34$), in positioning ($b = 0.23$), prevention of potential damage ($b = 0.22$), expression of one's state/thoughts ($b = 0.22$). With a minimal negative regression coefficient—the need to obtain verbal information about reality ($b = -0.04$). Thus, the greatest contribution to the production of imperatives is made by the needs of a child in changing the content of the situation, cooperation/support, changing their emotional state. In general, the imperatives meet a wide range of communicative and communication-related needs of a person at a given age.

Of the other types of independent variables, the characteristics of initiative/response speech behavior make the greatest contribution to the process of producing imperatives—with a negative coefficient $b = -0.83$, which shows that an increase in response statements in the sample, according to the data obtained, does not increase the probability of their attribution to imperatives. Perhaps it reflects the tendency to expand the repertoire of speech communication in developing dialogic interactions, which define the speech-genre system of children's communication needfully and cognitively/informationally as more complex than the one that could relate exclusively to imperative forms of statements.

The attribute of reflection of speech strategy, which reflects the facts of changes in speech strategies and tactics, has a positive coefficient $b = 0.72$, which in its significance fully corresponds to the situation of the need to produce speech forms that affect an adult.

The attribute of gender makes a high contribution with a negative coefficient $b = -0.68$, while the increase in the age sample of boys' statements does not increase the possibility of attributing speech forms to imperatives, which may mean the similarity of the systems of conditioning and implementation of imperative statements in boys and girls of this age. In the future, it should be checked to what extent gender differences will determine the production of imperatives in subsequent children ages.

With smaller regression coefficients, variables contribute to the production of imperatives: the communicative status of the speaker with a negative coefficient $b = -0.26$.

That is, the increase in the representation of the low status of the speaker in the sample does not increase the possibility of attributing the statement to imperatives, therefore, the low status of the child significantly determines the production of other genre forms, for example, informative or expressive when “requesting” help from an adult.

The type of situation—natural communication or game—has a negative coefficient $b = -0.21$; the increase in statements defined by game communication does not increase the contribution of these situations to the production of imperatives. Therefore, it can be assumed that the psychological boundary between natural communication and game is “diffuse”, and its establishment and speech marking require verification at subsequent ages.

The contribution of speech-genre reflection with a negative coefficient $b = -0.11$ with an increase in the attribute in the sample does not increase the possibility of attributing statements to imperatives, which may indicate that the production and implementation of the imperative requires rather the possibilities of automating this communicative-speech skill when using vocal and linguistic means in a visual and understandable situation for communicants or in a communicative system, typical for this age, it completely allows spontaneity, coupled with an affective experience of needs expressed imperatively.

The type of addressee makes an insignificant contribution to the production of imperatives at this age, with a positive coefficient $b = 0.11$, while the expansion of the types of objects—potential addressees increase the possibilities of their use, and adapting the cognitive-communicative scheme to autocommunication and other types of addressees (in addition to the addressee—a specific person).

When categorizing statements as **informatives** in the confidence interval—with a decrease in the regression coefficient—there were needs for changing the objective situation ($b = 1.19$), for changing one’s emotional state ($b = 0.82$), for cooperation, support ($b = 0.81$), as well as the need for identification ($b = 0.55$) of expressing one’s state/thoughts ($b = 0.15$). With significantly lower regression coefficients in comparison with the needs indicated here, as well as in comparison with the level of contribution of needs to the production of imperatives—the need for positioning ($b = 0.07$), the need for a social being (social partner) ($b = 0.05$), a material object ($b = 0.05$), prevention of potential damage ($b = 0.04$), in attention ($b = -0.03$); again, the need for information has a negative coefficient $b = -0.01$. Thus, the greatest contribution to the production of informative genres is made by the need to change the objective situation, to change one’s emotional state, to cooperate/support, as well as the need for identification. It was expected that a high contribution to the production of informative statements is made by the need for information, but this need is not found in this normotypical age sample in the confidence interval. This can be explained by an adult’s advanced verbal behavior in relation to the identification and characterization of reality objects and/or a shift of interest in obtaining information to other types of needs, as well as the relevance of non-linguistic mechanisms for obtaining information about reality objects at this age. Perhaps this is due to the insufficient representation of the child’s speech/verbal requests about objects in the sample, which should be studied in the future, including by comparing with subsequent age samples and a sample of the speech of children demonstrating advanced speech language development.

The type of speech behavior in the categories of initiative or response speech turned out to be highly significant when producing informative statements, however, the regression coefficient is negative ($b = -0.85$). It means that response speech does not increase the possibility of classifying it as informative statements, which may indicate that statements of various types can be implemented in the position of response forms of speech, probably, depending on the actual need that arose in the dialogue with an adult.

A gender attribute with a high negative regression coefficient $b = -0.66$ is one of the most significant in the production of informative statements, while an increase in the volume of boys' statements does not increase the possibility of attributing them to informative genres in the age sample under study.

The sign of the need for awareness of the speech strategy contributes to the production of informative statements with a regression coefficient $b = 0.63$, an increase in the value of this sign increases the probability of attributing the statement to informative and may mean an increase in the relevance of informative statements in this age sample.

With lower regression coefficients, the speaker's communicative status ($b = -0.22$) and the type of natural communication or game situation ($b = -0.21$) contribute to the production of informative statements—with negative coefficients. It means that an increase in cases with low status and cases of game communication does not increase the probability of attributing the statement to the category of informative genres. This may mean that a child of this age, within the limits of their information competence, is able to produce informative materials being in a high status. With regard to the parameter natural communication/game, the conclusion, as in the discussion of imperatives, can be made about the unformed psychological boundaries of these forms of activity and their specific speech support, while the specifics of the use of speech genres in the game (and different types of games) require further research.

To an even lesser extent, the informatives are determined by the type of addressee ($b = 0.11$), while the positive coefficient shows that an increase in the value of the attribute—from communication with a specific addressee—a person into forms of autocommunication and communication with other objects (animals, toys—primarily anthropomorphic)—increases the likelihood of attributing a statement to the informatives.

With a minimal contribution and a negative regression coefficient $b = -0.07$, an attribute of speech-genre reflection is presented. Its value increase doesn't attribute it to informative genres more often. The low level of contribution of this parameter—as well as in the production of imperatives—indicates the beginning of the formation of speech-genre reflection, a negative regression coefficient may indicate that the correspondence of statements produced by children to the genre form is not required due to the situational understanding of the goals of speech by adults and the possibility for adults to test hypotheses about the intentions/goals of a child using speech.

The variable that characterizes the reflection of communicative rules in contrast to the automatic implementation of the statement when producing both imperatives and informative genre forms is not represented in this age sample in the confidence interval.

4 Conclusions

The most important and still undisclosed topic of empirical research of speech is the assessment of the contribution of non-linguistic and linguistic factors in the production of speech statements depending on the types of speech genres. We developed a system of parameters for evaluating natural speech and presented it in the paper, including intentional complexes—speech motivators, categories of gender, speech situations, the speaker's communicative status, other non-linguistic parameters of the statement production.

The data of regression analysis carried out using the method of multinomial logistic regression show differences in the systems of formation of imperative and informative speech genres at an early age. It will allow further comparison of the presented speech models with models belonging to other types of genres, and analyze the dynamics of the systems of conditioning and implementation of speech throughout childhood.

The differences relate to the level of contribution of the variables discussed above to the production of imperatives and informatives. Although significant parameters have a lot in common, and positive or negative regression coefficients in the models of the production of imperatives and informatives often coincide, a more significant contribution in the first case is made by the need for a social being, positioning, the need for a material object, the need to prevent damage, express one's state/thought, the need for cooperation, support, as well as reflection of speech strategy; in the second case—the need to change the objective situation.

Significant differences in the structures and weight of constituents of imperative and informative statements compared to expressive and ritual genres were discovered, which, due to the limitation of the volume of the article, will be presented in future publications.

The paper opens a series of publications tracing the dynamics of speech generation patterns in early human ontogenesis (1–7 years).

Acknowledgements. The study was carried out within the Thematic Plan of the National Research Center Kurchatov Institute (Order no. 87 dated January 20, 2023).

References













1. Austin, J.: *How To Do Things with Words*, 2nd edn. Oxford Univ. Press, Oxford (1975)
2. Searle, J.: *Speech Acts. An Essay in the Philosophy of Language*. Cambridge Univ. Press, London (1969)
3. Searle, J.: *Intentionality. An Essay in the Philosophy of Mind*. Cambridge Univ. Press, Cambridge, NY (1983)
4. Strawson, P.: Intention and convention in speech acts. *Philos. Rev.* **73**(4), 439–460 (1964)
5. Dore, J.: *The Development of Speech Acts*. Mouton, The Hague (1975)
6. Dore, J.: Holophrases, speech acts and language universals. *J. Child Lang.* **2**, 21–40 (1975)
7. Bruner, J.: The ontogenesis of speech acts. *J. Child Lang.* **2**, 1–19 (1975). <https://doi.org/10.1017/S0305000900000866>
8. Burkhardt, A.: Speech acts, meaning, and intentions: critical approaches to the philosophy of John R. Searle. In: Burkhardt, A. (ed.) *W. de Gruyter*, Berlin, NY (1990)

9. Astington, J.: Children's understanding of the speech act of promising. *J. Child Lang.* **15**, 157–173 (1988). <https://doi.org/10.1017/S0305000900012101>
10. Cameron-Faulkner, T., Lieven, E., Tomasello, M.: A construction based analysis of child directed speech. *Cog. Sci.* **27**, 843–873 (2003). <https://doi.org/10.1016/j.cogsci.2003.06.001>
11. Ninio, A., Snow, C.: *Pragmatic Development*. Westview Press, Boulder, CO (1996)
12. Rakoczy, H., Tomasello, M.: Done wrong or said wrong? Young children understand the normative directions of fit of different speech acts. *Cognition* **13**, 205–212 (2009). <https://doi.org/10.1016/j.cognition.2009.07.013>
13. Bezuglaya, L.: Pragmalingvisticheskaya kontseptsiya I. P. Susova [Pragmalinguistic conception of I. P. Susov]. In: Susov I. *Linguistic pragmatics*, pp. 249–260. Nova Kniga, Vinnitsa (2009)
14. Klushina, N.: Intentional method in the modern linguistic paradigm. *Mediascope* **4**, 4 (2012)
15. Leont'ev, A.N.: *Deyatel'nost'*. Soznanie. Lichnost' [Activity. Consciousness. Personality], 2nd edn. Politizdat, Moscow (1977)
16. Leont'ev, A.A.: *Yazyk, rech, rechevaya deyatel'nost'* [Language, speech, speech activity]. Prosveschenie, Moscow (1969)
17. Malanchuk, I.: *Rech' kak psikhicheskiy protsess* [Speech as a mental process]. Krasnoyarsk Gov. Pedagog. Univ. named after V.P. Astafiev, Krasnoyarsk (2009)
18. Marisova, L.: *On the Motivational-Need Basis of Communication*. Berlin (1978)
19. Murray, H.: *Explorations in Personality*. Oxford Univ. Press, New York (1938)
20. Shneidman, E.: *Lives and Deaths: Selections from the Works of Edwin S. Shneidman*. Bruner/Mazel, Philadelphia (1999)
21. Malanchuk, I.: Cognitive Architectures of Goal Setting in Natural Speech Communication. *Proc. Comp. Sci.* **190**, 546–552 (2021). <https://doi.org/10.1016/j.procs.2021.06.101>
22. Shmeleva, T.: Model' rechevogo zhanra [Speech genre model]. In: *Zhanry rechi* [Speech Genres], Vol. 1, pp. 88–98. Kolledzh, Saratov (1997)
23. Bakhtin, M.: Problema rechevykh zhanrov [The problem of speech genres]. In: *Estetika slovesnogo tvorchestva* [Aesthetics of verbal creativity], pp. 237–280. Isskustvo, Moscow (1979)
24. Dement'ev, V.: *Teoriya rechevykh zhanrov* [The theory of speech genres]. Znak, Moscow (2010)
25. Balashova, L., Dement'ev, V.: *Russkie rechevye zhanry* [Russian speech genres], 2nd edn. YaSK, Moscow (2022)
26. Grosz, B., Sidner, C.: Attention, intentions, and the structure of discourse. *Comput. Linguist.* **12**(3), 175–204 (1986)
27. Ushakova, T.: Rechevye intentsii v mezhlchnostnom obschenii [Speech intentions in interpersonal communication]. *World Psychol.* **2**(42), 222–230 (2005)
28. Pavlova, N., Almayev, N., Latinov, V., Murasheva, O.: Psychology of discourse: new trends and achievements. *Nat. Syst. Mind* **2**(1), 64–68 (2022)
29. Malanchuk, I.: Speech forms (genres) as representatives of social and communicative thinkin/consiousness: age-gender aspect. In: *XIX International Interdisciplinary Congress NEUROSCIENCE FOR MEDICINE AND PSYCHOLOGY*. Max Press, Moscow, pp. 186–187 (2023). <https://doi.org/10.29003/m3297.sudak.ns2023-19/186-187>
30. Malanchuk, I., Orlov, V., Kartashov, S., Malakhov, D.: Differentiation of speech and language functional systems and analysis of the differences in related neural networks. *Hum. Physiol.* **49**(3), 297–306 (2023). <https://doi.org/10.1134/S0362119723700251>
31. Eliseeva, M.: Metalanguage activity of a child of early and preschool age. *Russ. Lang. School* **6**, 35–39 (2017)
32. Eliseeva, M.: Metalanguage activity of a child of early and preschool age (ending). *Russ. Lang. School* **7**, 24–29 (2017)

33. Alpatov, V.: Linguistic reflection in children. *Acta linguistica metropolitana*. Trans. Instit. Linguist. Stud. **13**(3), 811–819 (2017)
34. Vorobyova, T.: On the issue of differentiating children’s metalinguistic activity and their language reflection. *Izvestiya Herzen Univ. J. Hum. Sci.* **196**, 61–68 (2020)
35. Nefedyeva, D., Belousova, M.: Early abilitation and ontogenesis features of sensory systems, cognitive functions, and speech in preterm born children. *Bull. Contemp. Clin. Med.* **12**(6), 41–48 (2019)
36. Muratova, M., Valiullina, G., Klimenko, V.: The model of parental competence in ontogenesis and dysontogenesis of early speech development. In: *Proceedings IFTE-2021*, pp. 1141–1154 (2021). <https://doi.org/10.3897/ap.5.e1141>
37. Kwak, C., Clayton-Matthews, A.: Multinomial logistic regression. *Nurs. Res.* **51**(6), 404–410 (2002). <https://doi.org/10.1097/00006199-200211000-00009>
38. Krishnapuram, B., Carin, L., Figueiredo, M., Hartemink, A.: Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(6), 957–968 (2005). <https://doi.org/10.1109/TPAMI.2005.127>
39. El-Habil, A.: An application on multinomial logistic regression model. *Pak. J. Stat. Oper. Res.* **8**(2), 271–291 (2012)
40. Kyratzis, A., Guo, J.: Preschool girls’ and boys’ verbal conflict strategies in the United States and China. *Res. Lang. Soc. Interact.* **34**(1), 45–74 (2001). https://doi.org/10.1207/S15327973RLSI3401_3
41. Nakamura, K.: Gender and language in Japanese pre-school children. *Res. Lang. Soc. Interact.* **34**(1), 15–43 (2001). https://doi.org/10.1207/S15327973RLSI3401_2
42. Ladegaard, H., Bleses, D.: Gender differences in young children’s speech: the acquisition of sociolinguistic competence. *Int. J. Appl. Linguist.* **13**(2), 222–233 (2003). <https://doi.org/10.1111/1473-4192.00045>
43. Eriksson, M., et al.: Differences between girls and boys in emerging language skills: evidence from 10 language communities. *Br. J. Dev. Psychol.* **30**(2), 326–343 (2012). <https://doi.org/10.1111/j.2044-835X.2011.02042.x>
44. Oller, D., Griebel, U., Bowman, D., Bene, E., Long, H., Yoo, H., Ramsay, G.: Infant boys are more vocal than infant girls. *Curr. Biol.* **30**(10), PR426-R427 (2020)
45. Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C., Ramírez-Esparza, N., Hamrick, L., Marklund, E., Kalashnikova, M., Guez, A., Casillas, M., Benetti, L., van Alphen, P., Cristia, A.: Everyday language input and production in 1001 children from 6 continents. *PsyArxiv* (2022). <https://doi.org/10.31234/osf.io/fjr5q>
46. Sharonov, I.: Poisk i opisaniye kommunikativov na osnove Natsional’ nogo korpusa russkogo yazyka [Search and description of communication tools based on the National corpus of the Russian language]. In: *Methods of Cognitive Analysis of Word Semantics: Computer-Corpus Analysis*, pp. 141–183. YaSK, Moscow (2019)



Polymorphisms of IL10 Immunoregulatory Gene Impact the Morphometric Changes of the Brain in Schizophrenia

Irina K. Malashenkova^{1,2} , Vadim L. Ushakov^{3,4,5} ,
Sergey A. Krynskiy¹ , Daniil P. Ogurtsov^{1,2} , Ekaterina I. Chekulaeva¹ ,
Ekaterina A. Filippova¹ , Vyacheslav A. Orlov¹ , Natalia V. Zakharova³ ,
Denis S. Andreyuk^{3,5} , Sergey A. Trushchelev³ , Georgy P. Kostyuk³ ,
and Nikolay A. Didkovsky² 

- ¹ National Research Center “Kurchatov Institute”, Akademika Kurchatova sq., 1, 123182 Moscow, Russia
malashenkova.irinal@gmail.com, srgkr002@gmail.com
- ² Lopukhin Federal Research and Clinical Center of Physical-Chemical Medicine of Federal Medical Biological Agency, Malaya Pirogovskaya st., 1a, 119435 Moscow, Russia
- ³ Psychiatric Clinical Hospital 1 named N.A. Alekseev, Zagorodnoye shosse, 2, 117152 Moscow, Russia
tiuq@yandex.ru
- ⁴ National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoye shosse, 31, 115409 Moscow, Russia
- ⁵ Institute for Advanced Brain Studies, Lomonosov Moscow State University, Leninskie Gory, GSP-1, 119991 Moscow, Russia

Abstract. Structural changes of the brain in schizophrenia progress over time and their extent differs depending on the patient, yet their mechanisms remain unclear. Polymorphisms in the genes regulating immune response and inflammation have associations with the risk of schizophrenia. Studying their impact on the structural parameters of the brain in the patients is relevant for prediction of cognitive impairment. The goal of this work was to investigate the possible associations of polymorphisms of IL10 immunoregulatory gene with neuroanatomic changes in schizophrenia to gain new insights on the role of immunogenetic factors in the development of neurodegeneration in the patients. 89 patients and 99 healthy volunteers were enrolled. SNPs of IL10 gene (G-1082A and C-592A) were determined by polymerase chain reaction. MRI scans were performed on a Siemens Magnetom Verio 3T MRI scanner. It was found that the A allele for G-1082A SNP and CC genotype for C-592A SNP are more frequent in the patients than in the control group, and are significantly ($p < 0.01$) associated with decreased mean thickness in a number of areas of the frontal cortex in the patients. These findings are important for revealing the role of immunogenetic factors regulating IL-10 production in the pathogenesis of neurodegeneration in schizophrenia. Further research directed at finding and confirming new immunogenetic markers of structural brain changes in schizophrenia is important for identifying the various biological variants of the disease that can differ in pathogenesis, clinical presentation and therapeutic opportunities.

Keywords: Cortical thickness · Interleukin-10 · Magnetic resonance imaging · Schizophrenia · Single-nucleotide polymorphisms

1 Introduction

Advances in neuroimaging technology have provided new insights into the various brain changes in psychiatric disorders that occur during prenatal development or as a result of disease progression. There is evidence that morphometric and functional brain abnormalities in schizophrenia (SCH) progress over time and have different severity in different patients, but the reasons for these differences and the mechanisms of brain damage remain poorly understood [1]. Importantly, a greater degree of brain changes corresponds to greater severity of symptoms and more pronounced electrophysiological changes in SCH patients [2, 3]. In example, according to van Haren et al., the degree of cortical volume reduction in the frontal and parietal lobes correlates with the severity of positive and negative symptoms in SCH [3].

It should be emphasized that SCH is clinically and genetically an extremely heterogeneous disease. Large-scale studies have provided insight into its complex genetic architecture. The genetic risk of SCH is associated with both common and rare alleles distributed throughout the genome [4]. At present, it has become clear that a single gene cannot cause SCH, but variations in individual genes can significantly influence brain homeostasis and determine some predisposition to the disease. The pathogenesis of SCH is a highly complex process involving several variables [5]. Neuroinflammation, systemic inflammation and immune system activation have been shown to play a major role in the development of SCH [6]. According to large genome-wide association studies (GWAS), a number of single nucleotide polymorphisms (SNP) in genes regulating immune system functions have significant associations with the risk of SCH [7]. However, their role in the pathogenesis and their association with morphometric brain changes in patients are poorly understood [6]. There is data that the GG genotype of rs2275913 SNP of the IL17A gene, which is thought to be associated with an increase in IL-17A production, is associated with reduced right middle occipital gyrus volume in schizophrenia patients [8]. In our previous work it was shown that the GA genotype of rs2275913 is associated with reduced cortical thickness in a number of areas of the frontal cortex in patients with schizophrenia [9]. IL-17A is an important inflammatory cytokine participating in the pathogenesis of a number of neurodegenerative diseases and inflammatory CNS diseases [9]. However, the possible role of the SNPs in other immune response genes, including the genes regulating inflammation, as risk markers for the cortical thickness decrease in SCH remains unknown.

Taking into account data from previous studies on the association of markers of the immunoinflammatory syndrome with structural brain changes in the dynamics of the disease, we hypothesized the involvement of SNPs of immune response genes in the reduction of cortical thickness in SCH. In particular, our attention was attracted by the association of interleukin-10 (IL-10) level with morphometric parameters of the brain in SCH patients [10]. IL-10 is a key immunoregulatory cytokine secreted by immune cells, which is essential for the normal regulation of immune response and inflammation. IL-10

is considered an important anti-inflammatory modulator of glial activation, preventing inflammation-mediated neuronal degeneration [11]. In addition, the IL-10 gene SNPs (rs1800871 and rs1800872) have been shown to be significantly associated with cognitive function in schizophrenia [12].

Polymorphisms in the promoter region of the IL10 gene at positions G-1082A (rs1800896) and C-592A (rs1800872) are known to be involved in the regulation of IL-10 production [13]. Their associations with morphometric brain changes in SCH remain unexplored.

The aim of this work was to investigate possible associations of SNPs of the immunoregulatory gene IL10 (rs1800896 and rs1800872) with neuroanatomical changes in schizophrenia to obtain new insights into the role of immunogenetic factors in the development of neurodegeneration in patients and to search for biomarkers of these brain changes.

2 Materials and Methods

89 patients with SCH undergoing treatment at the N.A. Alekseev Psychiatric Clinical Hospital No. 1 and 99 healthy volunteers were enrolled into the study. The diagnosis of SCH was established in accordance with the diagnostic criteria of the disease (F20), International Classification of Diseases, Tenth Revision (ICD-10). Blood collection from patients was performed during the first days of hospitalization, after the disappearance of positive symptoms under the influence of antipsychotic therapy. Volunteers in the control group had no signs of psychiatric disorders and no relatives with SCH or other psychiatric diseases. The socio-demographic characteristics of the patients and control group volunteers are summarized in Table 1.

Table 1. Main socio-demographic and clinical characteristics of schizophrenia patients and control group volunteers enrolled into the study.

Value/groups	Schizophrenia (n = 89)	Controls (n = 99)
Age on the date of the assessment, years	26.5 ± 1.3	28.3 ± 2.6
Male	52	46
Female	27	53
Duration of the disease, years	4.1 ± 1.6	–
Age of onset of prodromal symptoms, years	17.7 ± 1.3	–
Age of manifestation, years	22.6 ± 1.4	–
PANSS (Positive and negative syndrome scale), points	95.2 ± 5.0	36.0 ± 1.9
NSA-4 (4-item negative symptom assessment), points	19.4 ± 1.5	5.0 ± 0.0
BFQRS (Bush-Francis catatonia rating scale), points	6.9 ± 1.6	0.0 ± 0.0

The exclusion criteria were the following: exacerbation of somatic diseases at the time of examination, signs of alcohol and/or other psychoactive substance abuse, presence of other psychiatric diseases, exacerbation of infectious, inflammatory and autoimmune diseases within 2 months before the examination, pregnancy.

All patients and volunteers of the control group underwent genetic tests, 52 patients and 24 volunteers of the control group also underwent the assessment of morphometric parameters of the brain.

The work was approved by the local ethical committee of the National Research Centre “Kurchatov Institute” (No. 5 of 05.04.2017). All participants were familiarized with the details of the study, signed voluntary informed consent and consent to the processing of personal data.

MRI scans were performed on a Siemens Magnetom Verio 3T magnetic resonance imager (Siemens GmbH, Germany). A 32-channel brain coil was used for data acquisition. For grey and white matter morphometry, high-resolution anatomical data were acquired for each subject based on a T1-weighted sequence (TR = 1900 ms, TE = 2.21 ms, 176 slices, voxel size $1 \times 1 \times 1 \text{ mm}^3$). All obtained structural images were analyzed in Freesurfer software.

The polymerase chain reaction method with a fluorescence-based real-time detection scheme was used to detect SNPs in IL10 gene (rs1800896 G-1082A, rs1800872 C-592A).

Excel (Microsoft) and Statistica (Stat Soft) programmes were used for statistical processing. The Shapiro-Wilk test was used to assess the normality of distribution. The results of morphometric and clinical examinations are presented as mean with 95% confidence interval; in comparisons of morphometric parameters, the significance of differences was assessed using student's test. The Chi-squared test was used to detect differences in the frequency of SNPs. Differences were considered statistically significant at $p < 0.05$. Benjamini-Hochberg procedure was used to check for multiple hypothesis testing.

3 Results

3.1 Frequencies of Rs1800896 and Rs1800872 Single Nucleotide Polymorphisms in Schizophrenia Patients and Healthy Volunteers

The study revealed a significant decrease in the proportion of GG homozygotes and a significant increase in the proportion of heterozygotes and AA homozygotes for SNP rs1800896 in SCH patients compared to controls (Table 2). For SNP rs1800872, a significant decrease in the proportion of CC homozygotes and an increase in the proportion of heterozygotes was found in patients compared to controls. The almost complete absence of the homozygous AA genotype for rs1800872 can be explained by the small sample size of our study and the rare occurrence of this genotype in the population.

3.2 Changes in Brain Morphometric Parameters in Patients with Schizophrenia Are Associated with the Carriage of Single Nucleotide Polymorphisms of IL10 Gene

As we reported earlier, patients with schizophrenia, compared to control group volunteers, showed a significant ($p < 0.001$) decrease in mean cortical thickness in a number

Table 2. Frequencies of rs1800896 and rs1800872 SNPs of the IL10 gene in schizophrenia patients and control group volunteers (n, %; *—statistically significant differences).

SNP	Group	Homozygotes for the first allele	Heterozygotes	Homozygotes for the second allele
IL10 G-1082A (rs1800896)	Schizophrenia	29 (32.6%)*	46 (51.7%)*	14 (15.7%)*
	Controls	74 (74.7%)	26 (26.3%)	0
IL10 C-592A (rs1800872)	Schizophrenia	50 (56.2%)*	38 (42.7%)*	1 (1.1%)
	Controls	81 (81.8%)	18 (18.2%)	1 (1.0%)

of frontal cortex zones (superior, middle and inferior frontal gyrus, inferior frontal gyrus, orbitofrontal cortex), as well as in superior temporal gyrus and fusiform gyrus [10].

The study of associations of morphometric brain changes with SNPs of the IL10 gene showed that a number of these changes were associated with certain genotypes. The study revealed for the first time that only patients carrying the A allele for the SNP IL10 G-1082A (rs1800896) had a highly significant ($p < 10 \times 10^{-4}$) decrease in cortical thickness in a number of frontal cortex zones (caudal and rostral part of the middle frontal gyrus, left superior frontal gyrus, opercular part of the left inferior frontal gyrus), as well as in the fusiform gyrus (Table 3). Only neuroanatomical changes in the superior frontal gyrus had a comparable level of significance in patients with the other alleles for this SNP.

Patients with homozygous genotype CC for SNP IL10 C-592A (rs1800872) also showed a number of unique statistically significant differences with controls in cortical thickness parameters (Table 4). The reduction in thickness in these patients affected a number of areas of the frontal cortex.

Patients with homozygous genotype CC for rs1800872 also showed a number of unique statistically significant differences with controls in cortical thickness parameters (Table 4).

4 Discussion

A major unresolved problem in psychiatry is the progression of cognitive deficits and negative symptoms in schizophrenia. Extensive neuroimaging studies have shown that these severe symptoms are based on structural and functional brain changes, the mechanisms of which have not been deciphered. The pronounced heterogeneity and extremely complex genetic architecture of the disease make it difficult to obtain unequivocal data on the pathogenesis of brain structural changes in schizophrenia. The aim of this work was to investigate the association of SNPs of the IL10 gene, a key regulatory cytokine, with cortical thickness in schizophrenia to gain new insights into pathogenesis and to search for biomarkers of neuroanatomical changes.

According to MRI data, we found the greatest changes in the cerebral cortex in schizophrenic patients in the frontal lobe. This agrees with the data of other authors, including the results of our previous work [14–16].

Table 3. Cortical thickness parameters in patients with schizophrenia—carriers of the A allele and of GG homozygous genotype for the IL10 G-1082A (rs1800896) SNP compared to controls.

Cortical thickness, mm	Schizophrenia (AA + GA) (n = 18)	Schizophrenia (GG) (n = 19)	Controls (n = 25)
Superior frontal gyrus (right)	2.756 ± 0.040 (p = 1.5 × 10 ⁻⁵) (p _a = 1.1 × 10 ⁻³)	2.782 ± 0.054 (p = 8.0 × 10 ⁻⁴) (p _a = 2.0 × 10 ⁻²)	2.928 ± 0.055
Caudal part of MFG (left)	2.556 ± 0.053 (p = 1.2 × 10 ⁻⁴) (p _a = 4.4 × 10 ⁻³)	2.619 ± 0.073	2.727 ± 0.059
Caudal part of MFG (right)	2.466 ± 0.071 (p = 1.3 × 10 ⁻⁴) (p _a = 3.2 × 10 ⁻³)	2.584 ± 0.049	2.664 ± 0.056
Fusiform gyrus (left)	2.723 ± 0.043 (p = 2.4 × 10 ⁻⁴) (p _a = 3.5 × 10 ⁻³)	2.764 ± 0.066	2.853 ± 0.045
Superior frontal gyrus (left)	2.759 ± 0.058 (p = 2.8 × 10 ⁻⁴) (p _a = 3.4 × 10 ⁻³)	2.799 ± 0.062	2.919 ± 0.053
Opercular part of IFG (left)	2.548 ± 0.062 (p = 3.6 × 10 ⁻⁴) (p _a = 4.4 × 10 ⁻³)	2.526 ± 0.063	2.714 ± 0.055
Fusiform gyrus (right)	2.739 ± 0.043 (p = 4.6 × 10 ⁻⁴) (p _a = 4.8 × 10 ⁻³)	2.836 ± 0.080	2.872 ± 0.053
Rostral part of MFG (right)	2.371 ± 0.054 (p = 5.6 × 10 ⁻⁴) (p _a = 5.1 × 10 ⁻³)	2.393 ± 0.064	2.528 ± 0.062
Rostral part of MFG (left)	2.412 ± 0.048 (p = 7.5 × 10 ⁻⁴) (p _a = 6.0 × 10 ⁻³)	2.542 ± 0.050	2.542 ± 0.051

IFG—inferior frontal gyrus; MFG—middle frontal gyrus

SNP study of the IL10 gene revealed that the frequency of homozygous genotype GG for rs1800896 was significantly ($p < 0.01$) decreased and the frequency of heterozygous genotype GA and homozygous genotype AA were increased in SCH patients compared to controls. In addition, they had significantly decreased frequency of homozygous genotype CC and increased frequency of heterozygous genotype CA C-592A for rs1800872. The occurrence of homozygous AA genotype for rs1800872 was 1% in both study groups. The results obtained partially coincide with the data of other authors [17, 18].

Our study demonstrated for the first time that changes in MRI parameters were closely associated with IL10 SNPs. Thus, a decrease in cortical thickness in superior frontal gyrus

Table 4. Cortical thickness parameters in patients with schizophrenia—carriers of CC homozygous genotype and CA heterozygous genotype for the IL10 C-592A (rs1800872) SNP compared to controls.

Cortical thickness, mm	Schizophrenia (CC) (n = 23)	Schizophrenia (CA) (n = 18)	Controls (n = 25)
Superior frontal gyrus (right)	2.756 ± 0.043 (p = 2.2 × 10 ⁻⁵) (p _a = 1.6 × 10 ⁻³)	2.795 ± 0.069	2.928 ± 0.055
Caudal part of middle frontal gyrus (right)	2.550 ± 0.075 (p = 5.5 × 10 ⁻⁵) (p _a = 2.0 × 10 ⁻³)	2.580 ± 0.043	2.664 ± 0.056
Superior frontal gyrus (left)	2.755 ± 0.051 (p = 8.2 × 10 ⁻⁵) (p _a = 2.0 × 10 ⁻³)	2.795 ± 0.069	2.919 ± 0.053
Opercular part of IFG (left)	2.540 ± 0.061 (p = 1.6 × 10 ⁻⁴) (p _a = 3.0 × 10 ⁻³)	2.559 ± 0.066	2.714 ± 0.055
Lateral part of orbitofrontal cortex (right)	2.681 ± 0.047 (p = 2.4 × 10 ⁻⁴) (p _a = 3.6 × 10 ⁻³)	2.752 ± 0.060	2.811 ± 0.042
Rostral part of middle frontal gyrus (right)	2.359 ± 0.056 (p = 2.9 × 10 ⁻⁴) (p _a = 3.6 × 10 ⁻³)	2.411 ± 0.052	2.528 ± 0.062
Rostral part of middle frontal gyrus (left)	2.406 ± 0.050 (p = 5.1 × 10 ⁻⁴) (p _a = 4.8 × 10 ⁻³)	2.445 ± 0.051	2.542 ± 0.050
Caudal part of middle frontal gyrus (left)	2.484 ± 0.075 (p = 6.4 × 10 ⁻⁴) (p _a = 5.3 × 10 ⁻³)	2.621 ± 0.065	2.772 ± 0.062

IFG—inferior frontal gyrus

(bilaterally) was significantly associated with the CC genotype for rs1800872, and in the left superior frontal gyrus it was also associated with the A allele of rs1800896. Thinning of the fusiform gyrus cortex bilaterally was associated with the A allele for rs1800896, and the lateral part of the orbitofrontal cortex on the right side was associated with carrying the CC genotype for rs1800872. There is an indication in the literature of extensive cortical thinning in patients in these areas involved in the pathogenesis of SCH [19]. It should also be noted that anhedonia (loss of the ability to experience pleasure and interest in activities) and poor episodic and working memory have been found to be associated with hypoactivation of the medial prefrontal cortex, which includes the orbitofrontal cortex, in SCH patients [20, 21].

Our results also demonstrated a significant association of a bilateral reduction in rostral and caudal cortical thickness of the middle frontal gyrus, which belongs to the prefrontal region, with the A allele for rs1800896 and with CC genotype for rs1800872. Prefrontal cortical areas are involved in higher cognitive functions, auditory/visual speech processing, emotional processing, executive function and decision making [22]. Dysfunction of the prefrontal cortex is considered to be an important mechanism contributing to the pathogenesis of negative symptoms in schizophrenia [23]. The associations of these changes with IL10 SNPs may have important clinical implications that require further clarification.

The CC genotype for rs1800872 was also associated with decreased thickness in the opercular region of the left inferior frontal gyrus.

The A allele of SNP rs1800896 results in decreased production of anti-inflammatory IL-10, which prevents oxidative stress. Homozygous genotype AA allele of rs1800896 was also prevalent in SCH patients with tardive dyskinesia (TD), involuntary movements of limb and facial muscles, a complication due to the intake of antipsychotics that increase oxidative stress [17].

Data on the effect of the IL10 SNP C-592A (rs1800872) on IL-10 production are conflicting [24, 25]. There are reports about the relationship of genotype AA and allele A for this SNP with an increased risk of allergic diseases [26, 27], which may suggest an association of allele A with immune dysfunctions leading to excessive activation of the Th2-link of adaptive immunity. In addition, a number of studies have shown that rs1800872 is associated with increased susceptibility to autoimmune diseases such as systemic lupus erythematosus, multiple sclerosis and with IL-10 overexpression [28, 29].

The role of rs1800872 in the pathogenesis of SCH was studied in a meta-analysis based on 63 publications [18]. It was shown that allele A and homozygous genotype AA of the polymorphism are significantly associated with the risk of developing SCH. Data on the influence of SNP rs1800872 on the clinical characteristics of schizophrenia are contradictory; apparently, this influence may depend on associated genetic factors. The association of rs1800872 with the severity of negative symptoms in schizophrenia has been reported: patients with AA homozygous genotype had significantly higher severity of apathetic-abolic disorders (negative symptoms) according to the PANSS scale [30]. According to other authors, the C allele for this SNP may be associated with more pronounced cognitive disorders in schizophrenia [12]. According to Wang et al., the effect of genotype for rs1800872 on cognitive function in schizophrenia may be different depending on concomitant polymorphisms of the xanthine oxymethyltransferase (COMT) gene. COMT is an enzyme that plays an important role in the regulation of dopamine metabolism and modulates the inflammatory response [31].

Our study is the first to demonstrate associations of IL10 SNPs, which cause impaired production of this cytokine, with structural brain changes in schizophrenia, which confirms our assumption about the involvement of immune disorders in the pathogenesis of the disease. It has been shown that genotype CC for rs1800872 and the A allele for rs1800896 have pronounced associations with a decrease in the thickness of a number of areas of the frontal cortex of the large hemispheres involved in executive functions, speech perception and emotions in patients, which makes these genotypes potential

predictors of brain changes in schizophrenia. In this study the patients with different IL10 genotypes had no significant differences in antipsychotic medications; therefore the differences in MRI parameters in these groups of patients were due to some other mechanisms. In particular, they may be related to the effects of the SNPs on IL-10 production.

As a limitation of the study, a small sample size should be noted. Further studies in larger cohorts of patients with schizophrenia are needed to replicate the results. In order to facilitate the practical implication of the results, it is also important to research the associations of IL10 SNPs affecting the morphometric parameters of the brain with clinical symptoms of schizophrenia.

5 Conclusion

This study demonstrated an increased frequency of functional SNPs (rs1800896 and rs1800872) in the IL10 gene in a group of patients with schizophrenia compared to healthy volunteers. It was revealed for the first time that carriage of AA genotype for rs1800896 and CC genotype for rs1800872 by the patients was significantly associated with morphometric changes in the brain in the regions associated with the pathogenesis of the disease. These findings contribute to the understanding of the role of immunogenetic factors in the pathogenesis of schizophrenia, including the development of structural brain abnormalities, and indicate that carriage of the IL10 gene SNPs that influence the production of this cytokine may be a biomarker of these abnormalities. It is necessary to validate the results obtained in larger studies, as well as to study the relationship of carrying polymorphisms of the IL10 gene with immune parameters and clinical characteristics of patients. The prospect of future research is the possibility of translating the data into practice to predict the nature of the course of the disease, the development of structural brain abnormalities and the choice of therapy.

Acknowledgements. This work was carried out within the state assignment of NRC «Kurchatov institute» and was partially supported by the RSF (project No 20–15–00299).

References

1. Mathalon, D.H., Sullivan, E.V., Lim, K.O., Pfefferbaum, A.: Progressive brain volume changes and the clinical course of schizophrenia in men: a longitudinal magnetic resonance imaging study. *Arch. Gen. Psych.* **58**(2), 148–157 (2001). <https://doi.org/10.1001/archpsyc.58.2.148>
2. Xiao, Y., et al.: Subtyping schizophrenia patients based on patterns of structural brain alterations. *Schizophr. Bull.* **48**(1), 241–250 (2022). <https://doi.org/10.1093/schbul/sbab110>
3. van Haren, N.E.M., Hulshoff Pol, H.E., Schnack, H.G., Cahn, W., Brans, R., Carati, I., Rais, M., Kahn, R.S.: Progressive brain volume loss in schizophrenia over the course of the illness: evidence of maturational abnormalities in early adulthood. *Biol. Psych.* **63**(1), 106–113 (2008)
4. Degenhardt, F.: Update on the genetic architecture of schizophrenia. *Med. Gen.* **32**(1), 19–24 (2020). <https://doi.org/10.1515/medgen-2020-2009>

5. Năstase, M.G., Vlaicu, I., Trifu, S.C.: Genetic polymorphism and neuroanatomical changes in schizophrenia. *Rom. J. Morphol. Embryol.* **63**(2), 302–322 (2022)
6. Kaneko, N., et al.: Suppression of cell proliferation by interferon-alpha through interleukin-1 production in adult rat dentate gyrus. *Neuropsychopharmacology* **31**(12), 2619–2626 (2006). <https://doi.org/10.1038/sj.npp.1301137>
7. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium: Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43**(10), 969–976 (2011). <https://doi.org/10.1038/ng.940>
8. Subbanna, M., Shivakumar, V., Bhalerao, G., Varambally, S., Venkatasubramanian, G., Debnath, M.: Variants of Th17 pathway-related genes influence brain morphometric changes and the risk of schizophrenia through epistatic interactions. *Psychiatr. Genet.* **32**(4), 146–155 (2022). <https://doi.org/10.1097/YPG.0000000000000315>
9. Malashenkova, I.K., Ushakov, V.L., Krynskiy, S.A., Ogurtsov, D.P., Khailov, N.A., Ratushnyy, A.Y., Filippova, E.A., Zakharova, N.V., Kostyuk, G.P., Didkovsky, N.A.: Associations of IL17A G-197A single nucleotide polymorphism with immunological parameters and structural changes of the brain in schizophrenia. *Med. Immunol.* **25**(5), 1225–1232 (2023)
10. Malashenkova, I.K., Ushakov, V.L., Zakharova, N.V., Krynskiy, S.A., Ogurtsov, D.P., Hailov, N.A., Chekulaeva, E.I., Ratushnyy, A.Y., Kartashov, S.I., Kostyuk, G.P., Didkovsky, N.A.: Neuro-immune aspects of schizophrenia with severe negative symptoms: new diagnostic markers of disease phenotype. *Sovrem. Tehnol. Med.* **13**(6), 24–33 (2021)
11. Porro, C., Cianciulli, A., Panaro, M.A.: The regulatory role of IL-10 in neurodegenerative diseases. *Biomolecules* **10**(7), 1017 (2020). <https://doi.org/10.3390/biom10071017>
12. Zakowicz, P., et al.: Genetic association study reveals impact of interleukin 10 polymorphisms on cognitive functions in schizophrenia. *Behav. Brain Res.* **419**, 113706 (2022). <https://doi.org/10.1016/j.bbr.2021.113706>
13. Sun, L., et al.: Dual role of interleukin-10 in the regulation of respiratory syncytial virus (RSV)-induced lung inflammation. *Clin. Exp. Immunol.* **172**(2), 263–279 (2013). <https://doi.org/10.1111/cei.12059>
14. Malashenkova, I.K., et al.: Association of structural changes of the brain with systemic immune activation in schizophrenia. *Proced. Comput. Sci.* **213**(3), 325–331 (2022). <https://doi.org/10.1016/j.procs.2022.11.074>
15. van Erp, T.G.M., Walton, E., Hibar, D.P., Schmaal, L., Jiang, W., Glahn, D.C., Pearlson, G.D., Yao, N., Fukunaga, M., Hashimoto, R., Okada, N., et al: Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the enhancing neuro imaging genetics through meta analysis (ENIGMA) consortium. *Biol. Psych.* **84**(9), 644–654 (2018)
16. Wannan, C.M.J., et al.: Evidence for network-based cortical thickness reductions in schizophrenia. *Am. J. Psych.* **176**(7), 552–563 (2019). <https://doi.org/10.1176/appi.ajp.2019.18040380>
17. Choi, K.-Y., Choo, J.M., Lee, Y.-J., Lee, Y., Cho, C.-H., Kim, S.-H., Lee, H.-J.: Association between the IL10 rs1800896 polymorphism and tardive dyskinesia in schizophrenia. *Psych. Invest.* **17**(10), 1031–1036 (2020)
18. Gao, L., Li, Z., Chang, S., Wang, J.: Association of interleukin-10 polymorphisms with schizophrenia: a meta-analysis. *PLoS ONE* **9**(3), e90407 (2014). <https://doi.org/10.1371/journal.pone.0090407>
19. Takayanagi, Y., et al.: Reduced cortical thickness in schizophrenia and schizotypal disorder. *Schizophr. Bull.* **46**(2), 387–394 (2020). <https://doi.org/10.1093/schbul/sbz051>
20. Yan, C., et al.: Rostral medial prefrontal dysfunctions and consummatory pleasure in schizophrenia: a meta-analysis of functional imaging studies. *Psych. Res. Neuroimag.* **231**(3), 187–196 (2015). <https://doi.org/10.1016/j.psychresns.2015.01.001>

21. Ragland, J.D., Laird, A.R., Ranganath, C., Blumenfeld, R.S., Gonzales, S.M., Glahn, D.C.: Prefrontal activation deficits during episodic memory in schizophrenia. *Am. J. Psych.* **166**(8), 862–874 (2009). <https://doi.org/10.1176/appi.ajp.2009.08091307>
22. Li, J., et al.: Schizophrenia affects speech-induced functional connectivity of the superior temporal gyrus under cocktail-party listening conditions. *Neuroscience* **359**, 248–257 (2017). <https://doi.org/10.1016/j.neuroscience.2017.06.043>
23. Fuentes-Claramonte, P., et al.: Negative schizophrenic symptoms as prefrontal cortex dysfunction: Examination using a task measuring goal neglect. *Neuroimage Clin.* **35**, 103119 (2022). <https://doi.org/10.1016/j.nicl.2022.103119>
24. Masilionyte, U., et al.: IL-10 gene polymorphisms and IL-10 serum levels in patients with multiple sclerosis in Lithuania. *Brain Sci.* **12**(6), 800 (2022). <https://doi.org/10.3390/brainsci12060800>
25. Duvlis, S., Dabeski, D., Noveski, P., Ivkovski, L., Plaseska-Karanfilska, D.: Association of IL-10 (rs1800872) and IL-4R (rs1805010) polymorphisms with cervical intraepithelial lesions and cervical carcinomas. *J. BUON* **25**(1), 132–140 (2020)
26. Gaddam, S.L., Priya, V.H.S., Srikanth Babu, B.M.V., Joshi, L., Venkatasubramanian, S., Valluri, V.: Association of interleukin-10 gene promoter polymorphism in allergic patients. *Genet. Test. Mol. Biomark.* **16**(6), 632–635 (2012). <https://doi.org/10.1089/gtmb.2011.0255>
27. Chatterjee, R., et al.: Interleukin-10 promoter polymorphisms and atopic asthma in North Indians. *Clin. Exp. Allergy* **35**(7), 914–919 (2005). <https://doi.org/10.1111/j.1365-2222.2005.02273.x>
28. Rianthavorn, P., Chokedeemeeboon, C., Deekajorndech, T., Suphapeetiporn, K.: Interleukin-10 promoter polymorphisms and expression in Thai children with juvenile systemic lupus erythematosus. *Lupus* **22**(7), 721–726 (2013). <https://doi.org/10.1177/0961203313486192>
29. Al-Naseri, M.A., Salman, E.D., Ad'hiah, A.H.: Association between interleukin-4 and interleukin-10 single nucleotide polymorphisms and multiple sclerosis among Iraqi patients. *Neurol. Sci.* **40**(11), 2383–2389 (2019). <https://doi.org/10.1007/s10072-019-04000-4>
30. Golimbet, V., et al.: A study of the association between polymorphisms in the genes for interleukins IL-6 and IL-10 and negative symptoms subdomains in schizophrenia. *Indian J. Psych.* **64**(5), 484–488 (2022). https://doi.org/10.4103/indianjpsychiatry.indianjpsychiatry_212_22
31. Wang, J., et al.: The interactive effect of genetic polymorphisms of IL-10 and COMT on cognitive function in schizophrenia. *J. Psychiatr. Res.* **136**, 501–507 (2021). <https://doi.org/10.1016/j.jpsychires.2020.10.021>



Enhancing Event Selection with ChatGPT-Powered Chatbot Assistant: An Innovative Approach to Input Data Preparation

Andrey Malynov^(✉)  and Igor Prokhorov 

National Research Nuclear University MEPhI, Moscow, Russia
andrey@malynov.com, ivprokhorov@mephi.ru

Abstract. This article discusses the development of a chatbot assistant that helps users select theater and concert events from a website using ChatGPT. The chatbot uses natural language processing and machine learning algorithms to understand user queries and provide relevant recommendations. We provide an approach for input data preparation that allows the model to use necessary information about events, so to use the context about them that wasn't received during model training. The article also explores the benefits of using a chatbot for event selection and the potential for future improvements in chatbot technology. Overall, this chatbot provides a user-friendly and efficient way to discover events.

Keywords: ChatGPT · Chatbot assistant · Concert events · Natural language processing · Machine learning algorithms · Recommender system

1 Introduction

In this article, we delve into the intricacies of developing a chatbot assistant powered by ChatGPT, designed to aid users in selecting theater and concert events from a website. The chatbot leverages the power of natural language processing (NLP) and machine learning algorithms to comprehend user queries and offer pertinent recommendations. A significant part of our discussion is dedicated to the methodology of input data preparation, which equips the model with essential information about events, thereby enabling it to utilize context that was not received during model training.

Furthermore, we explore a methodology for the development of chatbot architecture and its integration into the prevailing recommendation system. The experimental section, which compares the expected responses of the recommendation system with the actual ones, is given particular emphasis.

Ultimately, this chatbot presents a user-friendly and efficient method for event discovery, paving the way for a new era of intelligent, automated assistance.

2 Literature Review

The development and application of chatbots have been a significant topic of research in recent years. Chatbots, as defined by Abu Shawar and Atwell [1], are computer programs designed to simulate conversation with human users. The use of chatbots in various sectors, including customer service, education, and healthcare, has been explored extensively [2].

The integration of chatbots into event selection platforms is a relatively new area of research. However, the use of chatbots in related fields, such as e-commerce and customer service, provides valuable insights. For instance, the authors developed a chatbot for e-commerce platforms that uses NLP to understand user queries and provide product recommendations [3]. Their findings suggest that an NLP-based chatbot can significantly enhance the user experience by providing personalized recommendations.

NLP is a subfield of artificial intelligence (AI) that focuses on the interaction between computers and humans through natural language. Chatbot assistant usually involve solving several NLP tasks at the same time, such as text classification to determine the user's intent, response generation, etc.

The development of NLP began in the middle of the twentieth century. The early methods of NLP were heavily reliant on handcrafted features and rules, often involving linguists in the process and requiring extensive knowledge of the language in question. For instance, the ELIZA program developed by Joseph Weizenbaum in the 1960s was based on pattern matching and substitution methodology [4]. Later approaches involving the use of machine learning algorithms were based on handcrafted features, so they were limited in their ability to handle the complexity and variability of natural language [5].

The emergence of embedding models marked a significant shift in NLP research. These models, such as Word2Vec [6] or FastText [7], represent words in a high-dimensional space where semantically similar words are closer together. An important change with the advent of such models was that there was no need for manual construction of features, so that they have become built by the model. These methods significantly improved the performance of many NLP tasks.

The development of models with an attention mechanism was another leap forward. The seq2seq models with attention [8] allowed the model to focus on different parts of the input sequence when producing the output. This was further developed into the Transformer model [9], which replaced the recurrent layers with self-attention mechanisms, leading to substantial improvements in translation tasks.

The Transformer model became the basis for BERT (Bidirectional Encoder Representations from Transformers) [10]. BERT, which is essentially the encoder part of the Transformer, revolutionized NLP by pre-training on a large corpus of text and fine-tuning on specific tasks. This approach allowed BERT to achieve state-of-the-art results on a wide range of NLP tasks without the need to build a task-specific model.

Generative models like GPT (Generative Pretrained Transformer), which are based on the decoder part of the Transformer, have also made significant strides. GPT-1 [11], GPT-2 [12], GPT-3 [13], and GPT-4 [14] have shown remarkable capabilities in generating human-like text.

The authors of GPT-2 observed that an ordinary language model could solve tasks for which it was not explicitly trained. For instance, the model could answer questions,

summarize texts, and even translate languages, all without any task-specific training data. This was a significant discovery, as it indicated that a sufficiently large language model could learn to perform a wide range of tasks directly from the text.

The GPT-3 model introduced several innovations. It was significantly larger than GPT-2, with 175 billion parameters compared to GPT-2's 1.5 billion. This increase in size allowed GPT-3 to generate even more coherent and contextually accurate text. Furthermore, GPT-3 demonstrated that it could perform tasks by following a few examples, a process known as “few-shot learning”. This ability to learn from a few examples without any training steps and weights change made GPT-3 incredibly versatile and powerful.

The new GPT-4 model demonstrates performance at par with humans on a range of professional and academic benchmarks. It could be useful in a wide range of applications, such as drafting documents, creating high-quality content for blogs or social media, generating code, assisting in education by providing detailed explanations of complex concepts, and much more. The potential applications are virtually limitless, and as the model continues to improve, we can expect it to become an even more valuable tool.

The quality of the output text generated by the GPT model strongly depends on the input prompt. [12–14] provide examples of prompts for GPT models, they offer valuable insights into how prompts can be used to guide the model's output and achieve the desired results.

In conclusion, the development of NLP has seen a shift from handcrafted features to embedding models, attention mechanisms, and now, pre-training models like BERT and GPT. These models have significantly improved the performance of NLP tasks, and their applicability in various domains, including event selection, is promising.

3 Development of the Chatbot Assistant

This section is divided into two subsections, each focusing on a different aspect of system development. The first subsection outlines an approach to constructing the system's architecture and organization of components. The second subsection delves into the process of preparing input data, specifically for transmission to the ChatGPT programming interface.

3.1 Architecture of the Chatbot Assistant

The system architecture under discussion comprises four main components: a chatbot, an authentication service, a user context enrichment service, and a ChatGPT programming interface (see Fig. 1).

The chatbot is the primary interface for user interaction. It is designed to accept and process user requests in a conversational manner.

The authentication service is a critical component that ensures the security and privacy of the system.

The user context enhances the user's interaction with the chatbot by providing additional context about the user. For instance, it can provide a list of concerts the user has attended in the past. This contextual information allows the chatbot to deliver more personalized and relevant responses, thereby improving the overall user experience.

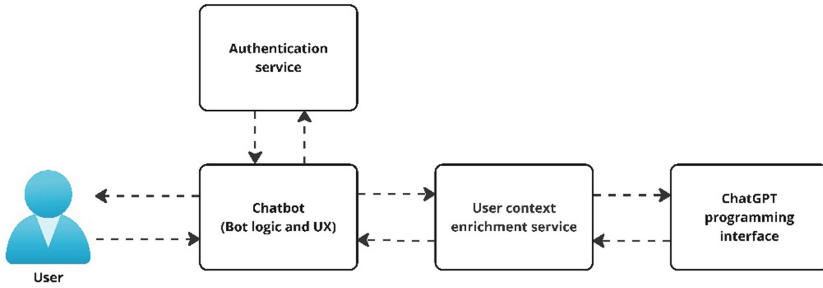


Fig. 1. Architecture of the chatbot assistant.

Lastly, the ChatGPT programming interface is the core of the recommendation system. It utilizes the power of OpenAI's Generative Pretrained Transformer models to analyze the user's request, enriched with context, and generate relevant concert event recommendations. It is responsible for the system's ability to understand the user's preferences and provide tailored suggestions.

3.2 Preparation of Input Data

The format of the request sent to ChatGPT is of particular interest as it assumes the presence of several elements. Firstly, there are instructions for ChatGPT that guide the AI's responses. Secondly, the request includes a list of concert events, which are enclosed within the `<EVENTS>` tag. This list is used to provide context to the ChatGPT about the upcoming events. Thirdly, the request contains information about the user, for example, a list of concerts they have attended. This information is enclosed within the `<USER_CONTEXT>` tag and helps the AI to personalize its responses. Lastly, the request includes a message from the user, which is placed inside the `<QUERY>` tag. This message is what the ChatGPT responds to, taking into account the instructions, events, and user context.

Here is an example of a request sent to the ChatGPT programming interface.

The list of events is presented after the `<EVENTS>` token and before the `</EVENTS>` token in the following format: ID | date | venue | genres | performer | minimum ticket price. The absence of data is marked with the null keyword, information for a new event is written from a new line. Having information about events, help the user find the most suitable event for him. The user's request goes after the `<QUERY>` token before the `</QUERY>` token. The information about user goes after the `<USER_CONTEXT>` token before the `</USER_CONTEXT>` token. Refer to the event by its identifier in double square brackets. For example, to refer to event 2, you need to specify `[[2]]`.

`<EVENTS>`

1 | September 08, 2023 | Forest Hills Stadium | Alternative Rock, Indie | Arctic Monkeys | 65\$

2 | October 28, 2023 | Madison Square Garden | Electronic Music, Dance | Depeche Mode | 50\$

```

3 | November 04, 2023 | Madison Square Garden | Pop Music, Soft Rock
| Pink | 60$
4 | January 22, 2024 | Madison Square Garden | Pop Music, Soft Rock
| Madonna | 46$
5 | December 01, 2023 | Madison Square Garden | Hard Rock, Heavy
Metal | KISS | 46$
6 | September 02, 2023 | Northwell Health at Jones Beach Theater |
Punk, Garage Rock | The Offspring | 40$
7 | August 30, 2023 | Capital One City Parks Foundation SummerStage
| Pop Music, Soft Rock | Vance Joy | 90$
8 | September 19, 2023 | Sony Hall | Hard Rock, Heavy Metal | Eric
Johnson | 40$
9 | October 26, 2023 | Beacon Theatre | Country, Folk | Lady A, Dave
Barnes | 71$
10 | May 22, 2024 | Barclays Center | Latin Music | J Balvin | 15$
11 | August 27, 2023 | FivePoint Amphitheatre | Rap, Hip-Hop | Snoop
Dogg | 40$
12 | September 25, 2023 | State Farm Arena | Rap, Hip-Hop | Drake |
70$
</EVENTS>
<USER_CONTEXT>The user attended a Madonna concert.
</USER_CONTEXT>
<QUERY>{{(see Table 1)}}</QUERY>

```

4 Results and Discussion

To evaluate the performance of ChatGPT, we conducted a series of experiments. The results of these experiments were recorded in a table (see Table 1). This table consisted of four columns: the experiment number, the user message, the expected result, and the actual result. The experiment number was used to keep track of each individual test. The user message was the input given to ChatGPT. The expected result was what we predicted ChatGPT would respond with. The actual result was the response that ChatGPT actually produced. By comparing the expected and actual results, we were able to assess the accuracy and reliability of ChatGPT in different scenarios.

The findings from this study substantiate the assertion that the utilization of a ChatGPT-based service in a recommendation system is justified. In nearly all scenarios, the concert events recommended by the chatbot aligned with the anticipated outcomes. Experiments 1, 2, 4, and 7 demonstrated the service’s capacity to select events from a list using information about the events that is not readily available in the recommendation system developed in [15, 16]. In Experiment 7, the external context was derived from song lyrics, while in Experiment 4, the information was based on the location of the concert venue. The outcome of Experiment 3 deviated from the expected results. However, ChatGPT provided a rationale, stating, “While Event 3 (Pink) and Event 4 (Madonna) both feature pop music performances, please note that their minimum ticket prices slightly exceed the \$50 budget you mentioned.” This explanation not only justified the deviation but also surpassed expectations. In the final experiment, the system utilized the provided user information to generate personalized recommendations.

Table 1. The results of experiments.

No	User message	Expected result	Actual result
1	I desire to attend a concert featuring a classic rock band	1, 2, 5	1, 2, 5
2	I'm a fan of the band Queen and I'm interested in attending a concert of a similar genre	1	1
3	I wish to purchase a concert ticket for a friend who enjoys pop music, with a budget of < \$50	4	3, 4
4	I am interested in attending any concert taking place in California	11	11
5	I wish to attend a concert that is scheduled for next week	7	11
6	I am interested in attending any forthcoming event	7, 11, 8, 1	2, 3, 4, 5, 9
7	I wish to attend a concert by the band that performed the song with the lyrics: I was made for loving you	5	5
8	Recommend a concert that I've previously attended	4	4

5 Conclusion

In conclusion, the development and integration of a ChatGPT-powered chatbot assistant into a recommendation system for theater and concert events has proven to be highly effective. The chatbot's ability to comprehend and respond to user queries, coupled with its capacity to offer relevant recommendations, is a testament to the power of natural language processing and machine learning algorithms. The methodology employed in preparing the input data, which includes instructions for ChatGPT, a list of events, user context, and user query, has been instrumental in enhancing the chatbot's performance. The experimental results have confirmed the efficacy of this approach, with the chatbot's recommendations closely aligning with the expected outcomes in almost all scenarios. This study, therefore, validates the use of ChatGPT-based services in recommendation systems, paving the way for further advancements in this field.

References

1. Abu Shawar, B., Atwell, E.: Chatbots: are they really useful? *J. Lang. Technol. Comput. Linguist.* **22**, 29–49 (2007)
2. Brandtzaeg, P.B., Følstad, A.: Why people use chatbots. *Internet Sci.* **12**, 377–392 (2017)
3. Angelov, S., Lazarova, M.: E-Commerce Distributed Chatbot System. In: *Proceedings of the 9th Balkan Conference on Informatics* (2019)
4. Weizenbaum, J.: Eliza—a computer program for the study of natural language communication between man and Machine. *Commun. ACM* **9**, 36–45 (1966)
5. Li, Q., et al.: A survey on text classification: from traditional to deep learning. *ACM Trans. Intell. Syst. Technol.* **13**, 1–41 (2022)
6. Mikolov, T., et al.: Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013)

7. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **5**, 135–146 (2017)
8. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint [arXiv:1409.0473](https://arxiv.org/abs/1409.0473) (2014)
9. Vaswani, A., et al.: Attention is all you need. *Adv. Neural Inform. Process. Syst.* **1**, 30 (2017)
10. Devlin, J., et al.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
11. Radford, A., et al.: Improving language understanding by generative pre-training (2018)
12. Radford, A., et al.: Language models are unsupervised multitask learners. *OpenAI Blog* **1**(8), 9 (2019)
13. Mann, B., et al.: Language models are few-shot learners. arXiv preprint [arXiv:2005.14165](https://arxiv.org/abs/2005.14165) (2020)
14. GPT-4 Technical Report. <https://arxiv.org/pdf/2303.08774.pdf>. Accessed 31 Aug 2023
15. Malynov, A., Prokhorov, I.: Development of an AI recommender system to recommend concerts based on microservice architecture using collaborative and content-based filtering methods. In: *Brain-Inspired Cognitive Architectures for Artificial Intelligence: BICA* AI 2020: Proceedings of the 11th Annual Meeting of the BICA Society 11*. Springer International Publishing, New York (2021)
16. Malynov, A., Prokhorov, I.: Clustering of concert and theater events based on their description. *Proced. Comput. Sci.* **213**, 673–679 (2022)



Functional Connectivity of Brain Regions Resulting from Learning Unfamiliar Words: Word Frequency Effect

K. S. Memetova^{1,2}(✉), V. M. Knyazeva¹, L. N. Stankevich¹, I. G. Malanchuk², and A. A. Aleksandrov¹

¹ Saint-Petersburg State University, Saint-Petersburg, Russia
k.memetova@spbu.ru

² National Research Center “Kurchatov Institute”, Moscow, Russia

Abstract. The study is aimed to investigate the change in functional interactions of brain regions during the learning process with fMRI. It was assumed that a word training, during which the meanings of a language words were assigned to certain pseudo-words, had an impact on the functional interactions of structures, primarily involved in semantic processing, as well as in the processes of involuntary attention. The ROI's were divided into 2 groups. The first group united the regions involved in semantic processing, the second group—associated with involuntary attention and involved in the generation of MMN. The results obtained described synchronous changes in functional connectivity after learning in the selected regions. As a result of training, it was possible to show how the functional connectivity of the involved regions had changed in the parameters of the hemodynamic response. The analysis of the semantic network revealed a simultaneous increase in cerebral blood flow in the posterior superior and middle temporal gyrus, as well as in the anterior superior temporal gyrus and in the insula after the semantization. Functional connectivity changes in ROI associated with the involuntary attention were manifested by a simultaneous decrease in the BOLD signal during the deviant stimuli perception.

Keywords: Training · Language processing · fMRI

1 Introduction

In electrophysiological studies, the event related potentials (ERPs) components are used in the analysis of speech processes. One of these components is mismatch negativity (MMN), which reflects a special reaction of the involuntary attention system to linguistic stimuli. It has been shown that the MMN parameters differ between words that occur in speech with different frequency. When comparing words of different frequencies, the MMN amplitude for a high-frequency word is much larger and the latency of the MMN peak is less in comparison with the MMN of a low-frequency word [1]. If pseudowords are used as a stimulus material, there are differences in ERPs between the words and pseudowords [2, 3]. In Russian language studies, it has been shown that the MMN

response caused by a word is more pronounced and its latency is less in comparison with the response to a pseudoword [1]. In addition, the context in which pseudowords are presented also influences the change in the MMN parameters. The pseudoword presentation in context with other pseudowords leads to the relatively low MMN amplitude with a long latent period. However, the response to the same pseudoword presented in the sequence with words leads to a significant increase in the MMN amplitude as well as the peak latency reduction in the 100–200 ms range [4].

The ERP studies allow us to conclude that new phonological forms are learned quite quickly, for example, when semantizing unknown linguistic stimuli. The data obtained for the Russian language show that as a result of the semantization of pseudowords, the parameters of MMN become similar to those that arise when the words are presented. There is also a dependence on the frequency of speech: the amplitude of MMN becomes much larger and the latency decreases during the presentation of a pseudoword, which has been assigned the meaning of a high-frequency word [5].

The neuroimaging studies showed that the BOLD signal in response to the presentation of pseudowords had changed as a result of their meaningfulness during the learning [6]. Significant differences in the local cerebral blood flow parameters after the training were detected in the middle and the superior temporal gyrus on the right and in the temporal plane (*planum temporale*) and in the superior temporal gyrus on the left. After a short training session, there was a significant increase in local activity in the middle temporal gyrus of the right hemisphere. The results cleared that the meaningfulness of pseudowords generated a significant boost in the involuntary attention system reaction to them. Besides, the intensity of the BOLD signal after the learning in the temporal plane and in the superior temporal gyrus became noticeably higher on the left. Thus, as a result of meaningfulness, there was an increase in the hemodynamic response in the speech area of the neocortex.

The aim of this study is to investigate the functional connectivity between the involved brain regions resulting from training. The brain structures selected, are primarily involved in the semantic processing, as well as in the processes of involuntary attention. The structures involved in the semantic processing are separated into a semantic network consisting of: left insula (IC L), pars triangularis and pars opercularis of the left inferior frontal gyrus (IFG L), anterior and posterior parts of the bilateral superior temporal gyrus (aSTG L/R, pSTG L/R), bilateral posterior and temporo-occipital middle temporal gyrus (pMTG L/R, toMTG L/R), bilateral inferior temporal gyrus (ITG L/R), bilateral posterior supramarginal gyrus (pSMG L/R), left angular gyrus (AG L) and left Heschl's gyrus (HG L). Brain regions involved in the generation of mismatch negativity and in the processes of involuntary attention: bilateral middle frontal gyrus (MFG L/R), bilateral pars triangularis and pars opercularis of the inferior frontal gyrus (IFG L/R), bilateral anterior and posterior parts of the superior temporal gyrus (aSTG L/R, pSTG L/R). The functional connectivity analysis of the selected ROI has been conducted on the basis of synchronous changes in the hemodynamic response parameters that arose as a result of word training. During the training, pseudowords have been charged the word meanings of a hypothetical language. Training has led to the fact that meaningless pseudo-words acquired the meaning and the so-called semantization of pseudo-words

took place. Two pseudowords were used as stimuli. They have been constructed in accordance with the rules and by analogy with the existing words of the Russian language by replacing one phoneme in words in order to exclude the possible influence of acoustic characteristics on the hemodynamic response parameters. We supposed that training and the lexical meaning of the charged hypothetical meanings would influence the functional interactions of the selected ROI.

2 Methods

The fMRI study involved 17 volunteers: 10 men and 7 women. All of them were conditionally healthy, with the leading right hand, native Russian language speakers. Subjects aged were between 23 to 35 years. Prior to the visit, participants were instructed not to drink alcohol the day before the testing. The study was accepted by the ethics committee of the National Research Center “Kurchatov Institute”.

The data were collected on a 3-T Siemens system. Functional images were acquired using a T2*-weighted echo planar imaging (EPI) sequence, with 2000-ms time repetition (TR), 20-ms time echo (TE), and 90° flip angle. Each functional image consisted of 98 axial slices, 200 mm * 200 mm field of view, 2-mm thickness. Structural images were acquired using a T1-weighted 3D sequence, with 2530-ms TR, 3.31-ms TE. The structural image consisted of 176 slices, with 1-mm thickness. The passive multi-stimulus oddball paradigm, in which among 85% of standard stimuli (St) there were deviant stimuli (D1 and D2) was used. Unnatural words were used as stimuli. Stimuli were performed in arrays: 22 arrays with D1 (7xSt + D1) and 22 arrays with D2 (7xSt + D2). Arrays were performed in a quasi-randomized order (see Fig. 1).

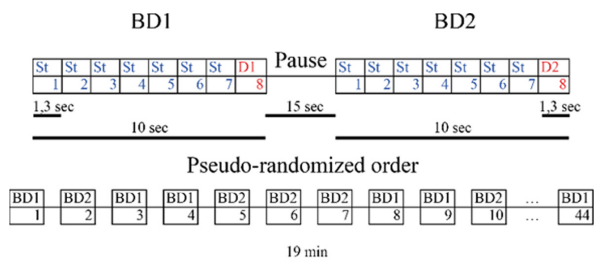


Fig. 1. Experiment scheme: BD1—block with deviant stimulus 1; BD2—block with deviant stimulus 2; St—standard stimulus; D1—deviant stimulus 1; D2—deviant stimulus 2.

Each participant listened three unnatural word stimuli in a quasi-randomized order: (1) standard stimulus (St)—“*kai*”; (2) deviant stimulus 1 (D1)—“*kan*”; (3) deviant stimulus 2 (D2)—“*kas*”. The stimuli were spoken by a female native speaker of Russian and processed with Praat (Paul Boersma and David Weenink Phonetic Sciences, University of Amsterdam, Amsterdam, The Netherlands) and Adobe Audition 3.0 (Adobe Systems Incorporated, California, USA) software. The NBS Presentation software (neurobs.com) was used to present stimuli through fMRI compatible electrodynamic headphones integrated into earmuffs for reduction of residual background scanner noise binaurally at

an audio level of 80 dB. To reduce the effect of acoustic differences, the stimuli were equalized to the maximum in physical properties (amplitude, duration, intensity, spectral characteristics). We used consonant-vowel-consonant syllables for all stimuli. The first two phonemes formed the syllable “*ka*” that was identical for all pseudowords used, so the stimuli differed only by the last phoneme. The point at which the replacement of the last phoneme took place was called the point of divergence (PD). The interval from the stimulus beginning to the divergence point was 240 ms the total stimuli duration was 385 ms.

As a result, to the PD, the stimuli were absolutely similar. The inter-stimulus interval was 900 ms with randomization from 0 to 50 ms. During recording the fMRI session, the participants watched the dynamic image on mute while hearing the sounds to keep their attention occupied, in keeping with previous studies of the MMN [5]. A neutral screen-saver was used as the image, in which orange circles of various sizes flickered smoothly against a black background. There was no need to pay attention to the auditory-presented stimuli. To assess possible distortions and associations in the stimuli perception, after the study, the participants completed the feedback form.

After the first fMRI session, the participants underwent a semantic training: they had to listen twice a day to an educational audio recording for 7 days. In this audio recording, unnatural words have been charged the word meanings of a hypothetical language, which differed in the frequency of use. It should be noted that for this study it was necessary to set the specific lexical characteristics, according to which stimuli differed. A lexical frequency was chosen (the frequency of occurrence and use) for this purpose. The word frequency was determined according to the New Frequency Dictionary of Russian Vocabulary, the Frequency Dictionary of Living Russian Speech [7], and the National Corpus of the Russian Language [8]. The audio was recorded by natural female voice. The pseudowords inserted into the audio recording were identical to those presented during the fMRI session. The audio that the participants listened to during the semantic training used the existing words “*god*” (“year”, frequency of occurrence in live speech 1954.1 ipm) and “*gid*” (“guide”, no data on the frequency of occurrence in live speech, word frequency 15 ipm). Thus, the pseudoword “*kas*” was assigned the value of the high-frequency word “*god*”, while the pseudoword “*kan*” was assigned the value of the low-frequency word “*gid*”. In order to fix the learning outcomes, before the second fMRI session, the participants were instructed to listen to the audio recording again and then write at least 10 memorized suggested statements. The second fMRI session took place 7 days after the first one by the identical procedure.

The data were processed in the SPM12 [9], Matlab-based cross-platform CONN Toolbox [10–12], MATLAB R2017b 9.3.0 (MathWorks, Natick, MA, USA) software packages. Pre-processing of functional images was performed with all necessary procedures [13] by SPM12. Structural images, segmentation and spatial normalization parameters were performed using the CONN Toolbox software package.

Data analysis was performed in two stages by CONN Toolbox. At the first level analysis, based on the calculated GLM parameters (a general linear model), t-contrasts between deviant stimuli (D1, D2) before and after the word training were calculated

separately for each subject. Each stimulus corresponded to a strictly defined time of presentation. The fMRI model specification took into account this time. Thus, the response to each stimulus was considered at the first level analysis.

The resulting contrasts, representing a linear combination of GLM parameters, were subjected at the second level analysis to further parsing of the functional connectivity between regions of interest—ROI-to-ROI analysis (a region of interest to a region of interest—the ratio of the selected region of interest to another region of interest). ROI-to-ROI connectivity analyses computed the correlation matrices characterizing a functional connectivity between the set of region of interest [12]. When choosing ROI, we were guided by the literature data on the most stable network components [2, 3, 14]. ROI were the areas corresponding to the topography of the semantic processing network: left insula (IC L), pars triangularis and pars opercularis of the left inferior frontal gyrus (IFG tri L, IFG oper L), anterior and posterior parts of the bilateral superior temporal gyrus (aSTG L/R, pSTG L/R), bilateral posterior and temporo-occipital middle temporal gyrus (pMTG L/R, toMTG L/R) and posterior part of the bilateral inferior temporal gyrus (pITG L/R), posterior part of the bilateral supramarginal gyrus (pSMG L/R), left angular gyrus (AG L) and left Heschl's gyrus (HG L). ROI related to the involuntary attention network (participating in the MMN generation): bilateral middle frontal gyrus (MFG L/R), bilateral pars triangularis and pars opercularis of the inferior frontal gyrus (IFG tri L/R, IFG oper L/R), bilateral anterior and posterior parts of the superior temporal gyrus (aSTG L/R, pSTG L/R).

We used the alternative settings for ROI-level inferences with parametric multivariate statistics. Significance was determined at the level of individual ROI's. Statistical significance for the level of functional interaction between each pair of ROI was assessed by Student's T-test. The critical significance threshold was $p < 0.05$, adjusted for multiple comparisons (a false discovery rate).

3 Results

The data obtained in response to deviant stimuli demonstrate significant changes in functional interactions after the word training session. An analysis of the functional connectivity between brain regions in response to the D2 stimulus (a pseudoword assigned a high-frequency hypothetical meaning) shows a significant simultaneous increase in the BOLD signal in ROI of the semantic network (see Fig. 2A). A synchronous increase in energy consumption appears in the posterior part of the left superior temporal gyrus (BA 22) and in the posterior part of the right middle temporal gyrus (BA 21, BA 37): pSTG L—toMTG R— $T(16) = 4.51$, p-FDR = 0.0265. Also, a simultaneous rise in the BOLD signal is observed in the anterior part of the left superior temporal gyrus (BA 38) and in the left insula (contains Brodmann areas 44 and 45): aSTG L—IC L— $T(16) = 4.31$, p-FDR = 0.0265. Figure 2A shows changes in the functional connectivity of hemodynamic responses to the D2 stimulus found in the ROI of the semantic network after the training.

Figure 2B shows changes in functional connectivity of hemodynamic activation in response to the D2 stimulus found after the training in ROI of the involuntary attention network. The changes show a simultaneous decrease in energy consumption in the pars

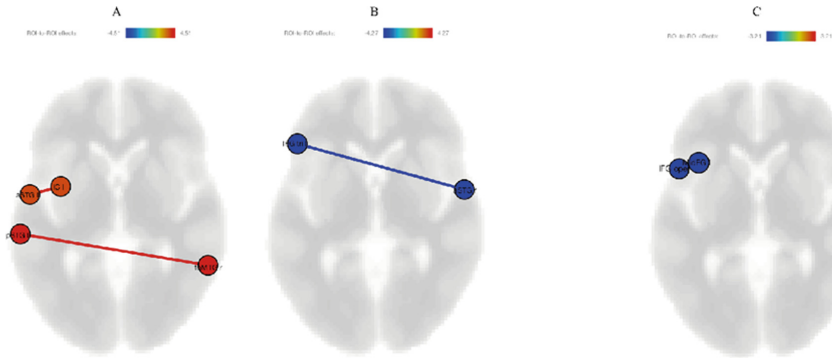


Fig. 2. Functional interactions between ROI's. The figure shows functionally significant interactions ($p\text{-FDR corr} < 0.05$). The color scale corresponds to the effect size (T -value). **A** Interrelated ROI of the semantic network after the training for D2 (a pseudoword, which was assigned the high-frequency word meaning): there is a significant simultaneous increase in activation in pSTG L—toMTG R and aSTG L—IC L; **B** Interconnected ROI of the involuntary attention network after the training for D2 (a pseudoword, which was assigned the high-frequency word meaning): there is a significant simultaneous decrease in activation in IFG tri L—aSTG R; **C** Interconnected ROI of the involuntary attention network after the training for D1 (a pseudoword, which was assigned the low-frequency word meaning): there is a significant simultaneous decrease in activation in MFG L—IFG oper L.

triangularis of the left inferior frontal gyrus (BA 45) and in the anterior part of the right superior temporal gyrus (BA 38): IFG tri L—aSTG R—T(16) = -4.27 , $p\text{-FDR} = 0.0263$.

The effect of the word training on BOLD responses to the deviant D1 stimulus (a pseudoword charged the meaning of a low-frequency word) reveals no significant changes in functional connectivity in ROI's of the semantic network. Changes are shown only for ROI's of the involuntary attention network. Figure 2C shows the change in the functional connectivity of the involuntary attention network ROI in response to the deviant stimulus D1 after the training. It has been found a simultaneous BOLD signal decrease in the left middle frontal gyrus (BA 6, BA 8) and the pars opercularis of the left inferior frontal gyrus (BA 44) MFG L—IFG oper L—T(16) = -3.21 , $p\text{-FDR} = 0.0495$.

4 Discussion and Conclusions

We analyzed the influence of the word training process during which unnatural words were charged the meaning of words with different frequency of a hypothetical language, on the functional interactions between brain structures, primarily involved in semantic processing, as well as participating in the involuntary attention mechanisms. The functional connectivity analysis for ROI was performed. The semantic processing network included the following brain regions: left insula (IC L), pars triangularis and pars opercularis of the left inferior frontal gyrus (IFG tri L, IFG oper L), anterior and posterior parts of the bilateral superior temporal gyrus (aSTG L/R, pSTG L/R), bilateral posterior

and temporo-occipital middle temporal gyrus (pMTG L/R, toMTG L/R), posterior part of the bilateral inferior temporal gyrus (pITG L/R), posterior part of the bilateral supra-marginal gyrus (pSMG L/R), left angular gyrus (AG L) and left Heschl's gyrus (HG L). ROI involved in the generation of mismatch negativity and to the involuntary attention were represented by the following structures: bilateral middle frontal gyrus (MFG L/R), bilateral pars triangularis and pars opercularis of the inferior frontal gyrus (IFG tri L/R, IFG oper L/R), bilateral anterior and posterior parts of the superior temporal gyrus (aSTG L/R, pSTG L/R).

It was shown that the perception of the same unnatural word before and after the word learning induced a change in functional interactions, expressed in a simultaneous BOLD signal increase in the areas of semantic processing (see Fig. 2A). Moreover, the changes in functional connectivity in response to the pseudoword with an acquired meaning of the word showed a clear dependence on speech frequency. When analyzing ROI of the semantic network, a simultaneous increase in cerebral blood flow was detected during presentation of a unnatural word, which was changed the meaning of a high-frequency word in the posterior superior and middle temporal gyrus, as well as in the anterior superior temporal gyrus and in the insula.

According to the cytoarchitectonic classification of brain areas, the posterior part of the superior temporal gyrus corresponds to the Brodmann area 22, which is a key link in Wernicke's area. The main language function, which is provided by Brodmann area 22, is the semantic processing of complex sound stimuli. It is known that a lesion of this area leads to the various speech disorders associated with understanding—sensory aphasia. The “alienation of the word meaning” phenomenon even though the word is correctly pronounced is observed. Native speech with sensory aphasia sounds like a foreign language. The functional interaction of the posterior part of the superior temporal gyrus was observed with the posterior middle temporal gyrus—Brodmann areas 21 and 37. These areas are traditionally considered acoustic-gnostic centers of speech recognition [15]. A simultaneous rise in the BOLD signal has been observed in the insular and in the superior temporal gyrus. It is known that the insula contains Brodmann areas 44 and 45, which are the links of Broca's center, which provides the motor organization of speech [16]. A previous fMRI study has shown a significant rise in the local cerebral blood flow in response to pseudoword in the specific brain speech areas after the training [6]. A previously unknown pseudoword, after the training, acquired a significant semantic meaning, which is reflected in the parameters of the hemodynamic response. Powerful energy consumption in the semantic speech area is probably associated with successful learning or, in other words, with the successful semantization of a pseudoword.

It should be noted that the functional interactions in the brain regions responsible for the semantic signal processing, when perceiving a pseudoword, which was assigned the low-frequency meaning of a hypothetical language word, little changed after the training and these changes did not reach the statistical significance. Earlier it was noted that the training process was quite short, it lasted for seven days. Probably, significant changes in functional interactions for this type of stimuli did not have time to form in such a short time [17]. It can be assumed that for words with the low-frequency or rarely occurring meaning, a longer training was required.

In ROI's associated with involuntary attention, functional connectivity was expressed by a synchronous decrease in the BOLD signal for both stimuli in the middle and inferior frontal gyrus. Previously, the role of involuntary attention in the processing of deviant stimuli had been already noted [6]. A significant change in the BOLD response was observed in the region involved in the mismatch negativity generation—the anterior part of the right superior temporal gyrus. In this paper, the functional connectivity during the pseudoword processing with a high-frequency meaning was shown by a simultaneous decrease in the BOLD signal in the right anterior part of the superior temporal gyrus, corresponding to Brodmann area 38, and in the pars triangularis of the left inferior frontal gyrus, Brodmann area 45 (see Fig. 2B). For the pseudoword, with a low-frequency word meaning, functional interactions were also expressed by a simultaneous decrease of response in the pars opercularis of the inferior frontal gyrus, corresponding to the Brodmann area 44, and in the middle frontal gyrus—Brodmann areas 6, 8, which control of the eye movements and attention (see Fig. 2C). Some changes associated with a decrease in energy consumption in the brain regions that provide the MMN generation, were observed. Thus, the automatic mechanisms of involuntary attention were found in response to deviant stimuli.

Thereby, summarizing the above the data obtained show that even nearly short learning session, during which the stimuli have been semantized, can lead to significant changes in the brain function. Namely, there are pronounced changes in the functional connectivity of brain structures that provide semantic processing of words with high frequency of use.

Acknowledgements. Supported by National Research Center «Kurchatov Institute» (the order No 87 from 20.01.2023).



References

1. Aleksandrov, A.A., Memetova, K., Stankevich, L., Uplisova, K.: Effects of Russian-language word frequency on mismatch negativity in auditory event-related potentials. *Neurosci. Behav. Physiol.* **47**, 1043–1050 (2017). <https://doi.org/10.1007/s11055-017-0510-3>
2. Pulvermüller, F., Shtyrov, Y., Kujala, T., Näätänen, R.: Word-specific cortical activity as revealed by the mismatch negativity. *Psychophysiology* **41**, 106–112 (2004). <https://doi.org/10.1111/j.1469-8986.2003.00135.x>
3. Price, C.: The anatomy of language: a review of 100 fMRI studies published in 2009. *Ann. N. Y. Acad. Sci.* **1191**(1), 62–88 (2010). <https://doi.org/10.1111/j.1749-6632.2010.05444.x>
4. Aleksandrov, A.A., Memetova, K., Stankevich, L.: Lexical context affects mismatch negativity caused by pseudowords. *Hum. Physiol.* **43**, 395–403 (2017). <https://doi.org/10.1134/S036211971704003X>
5. Aleksandrov, A.A., Memetova, K., Stankevich, L., Knyazeva, V., Shtyrov, Y.: Referent's lexical frequency predicts mismatch negativity responses to new words following semantic training. *J. Psycholinguist. Res.* **49**, 187–198 (2020). <https://doi.org/10.1007/s10936-019-09678-3>
6. Memetova, K., Knyazeva, V., Stankevich, L., Malanchuk, I., Aleksandrov, A.A.: BOLD signal changes in response to pseudowords after giving them semantic meaning in the course of special training: an fMRI study. *Proc. Comp. Sci.* **213**(2), 285–291 (2022). <https://doi.org/10.1016/j.procs.2022.11.068>

7. Lyashevskaja, O., Sharov, S.: Chastotnyj slovar' sovremennogo russkogo jazyka [Dictionary of Frequency of Contemporary Russian]. Azbukovnik, Moscow (2009)
8. Russian National Corpus. <https://ruscorpora.ru>
9. Friston, K., Ashburner, J., Kiebel, S., Nichols, T., Penny, W. (eds.): Statistical parametric mapping: the analysis of functional brain images. Elsevier, Acad. Press, London (2011)
10. Whitfield-Gabrieli, S., Nieto-Castanon, A.: Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connect.* **2**(3), 125–141 (2012). <https://doi.org/10.1089/brain.2012.0073>
11. Nieto-Castanon, A., Whitfield-Gabrieli, S.: CONN functional connectivity toolbox (RRID SCR_009550), version 20 (2020). <https://doi.org/10.56441/hilbertpress.2048.3738>
12. Nieto-Castanon, A.: fMRI minimal preprocessing pipeline. In: Handbook of Functional Connectivity Magnetic Resonance Imaging Methods in CONN, pp. 3–16. Hilbert Press, Boston, MA (2020)
13. Rinne, T., Alho, K., Ilmoniemi, R., Virtanen, J., Näätänen, R.: Separate time behaviors of the temporal and frontal mismatch negativity sources. *Neuroimage* **12**, 14–19 (2000). <https://doi.org/10.1006/nimg.2000.0591>
14. Deouell, L.: The frontal generator of the mismatch negativity revisited. *J. Psychophysiol.* **21**(3–4), 188–203 (2007). <https://doi.org/10.1027/0269-8803.21.34.188>
15. Aguirre, F., et al.: Neural correlates of semantic matching in indirect priming. *Cogn. Syst. Res.* **77**, 18–29 (2023). <https://doi.org/10.1016/j.cogsys.2022.10.006>
16. Graves, W., Purcell, J., Rothlein, D., Bolger, D., Rosenberg-Lee, M., Staples, R.: Correspondence between cognitive and neural representations for phonology, orthography, and semantics in supramarginal compared to angular gyrus. *Brain Struct. Funct.* **228**, 255–271 (2023). <https://doi.org/10.1007/s00429-022-02590-y>
17. van Heuven, W.: Learning an additional language: consequences for language and cognitive processes. In: Smith, A. (ed.) *Cognition in the Real World*. Oxford Univ. Press, Oxford (2023). <https://global.oup.com/academic/product/cognition-in-the-real-world-9780198790914?cc=ru&lang=en&>



DIPy-AI: Brain-Cognition-Inspired DIKW Pyramid-Based Agile AI Architecture for Industrial Sensor Data Assimilation

Amit Kumar Mishra^{1,3}  and Yi Zhong² 

¹ University of Cape Town, Cape Town, South Africa
akmishra@ieee.org

² Huazhong University of Science and Technology, Wuhan, China
yzhong@hust.edu.cn

³ University West, Trollhättan, Sweden

Abstract. The paper proposes DIPy-AI, an agile AI architecture based on the data-knowledge-information-wisdom (DIKW) pyramid, for processing sensor data in production environments. DIKW is one of the accepted models abstracting the assimilation of sensory data by the human brain. DIPy-AI aims to address challenges related to data assimilation, quality detection, and modular information extraction. The proposed architecture consists of three layers, viz a sensor-dependent data pre-processing layer, a sensor-agnostic ML layer for converting data into information, and an application-specific layer for knowledge extraction. There are two major merits of the proposed architecture. By having a layered architecture, it can easily be repurposed for different industries. Secondly, this agility in the architecture also facilitates the changing of sensors as well as overall goals of the architecture. The work aligns well with sustainable industrial digitization goals (shared by many countries) and offers a flexible solution applicable to multiple industries, promoting sustainability, data-sharing and architecture sharing.

Keywords: DIKW pyramid · Industrial AI · Brain-inspired architecture

1 Introduction

The assimilation of vast amounts of sensor data in production environments poses significant challenges, especially when it comes to automatically scanning for data quality, adapting to changing production environments and facilitating cross-industry sharing of AI-solutions. This paper proposes the development of a modular and agile AI architecture based on the Data-Information-Knowledge-Wisdom (DIKW) pyramid, specifically designed for sensor data processing. It can be noted here that the DIKW framework has been developed to explain human cognition.

The proposed architecture focuses on answering the research questions of processing data to identify and mitigate bad quality data, extracting relevant information, and facilitating smooth changes in the data assimilation process. The overarching goal is to create an agile AI architecture that can be easily adapted to different sensor types and industries. The above is achieved by using a modular layered architecture. The first layer focuses on sensor-dependent data pre-processing, encompassing data sanity checks and calibration specific to each sensor. The second layer employs sensor-agnostic machine learning techniques to transform sensor data into information using deep learning architectures. The third layer is an application-specific information processing layer that leverages symbolic processing to analyze information from the previous layer. This enables the extraction of context and goal-specific knowledge, providing users with valuable insights tailored to their objectives. Additionally, this layer facilitates user interaction to understand new needs, such as incorporating new sensors or adjusting the architecture for different industrial setups. The proposed architecture would be validated through co-develop the solution with two industries from Sweden, viz. GKN Aerospace and AP&T, and two industries from China, viz. BGRIMM Technology Group and Wuhan Jingce Electronic Group. These industries work on a diverse type of products. GKN works on aerospace engine manufacturing whereas AP&T works on the development of metal processing plants. By working with such diverse types of industries, we aim to make sure that the solution is developed to suit various industries. Secondly, this will help us to quantify the amount of development effort required when we are trying to use the solution for various industries (which is one of the main claimed merits of DIPy-AI).

Rest of the paper is organized as follows. Section 2 discusses the state of the art. Section 3 details the proposed architecture. Section 4 concludes the paper and discusses the future work.

2 State of the Art (SOTA)

We shall present the SOTA in three phases. In the first phase, we shall look at SOTA in the domain of heterogeneous sensor data analysis in industries. Secondly, we shall investigate the two sensor-specific processing we shall work on, viz. calibration and compression. In the third stage, we shall investigate SOTA on cognitive architectures to analyze the data from industrial sensors.

The integration of heterogeneous sensor data in industries requires efficient data processing architecture. DeepSense proposed by Yao et al. [21], a deep learning framework, has been proposed for signal estimation and classification that addresses noise and feature customization challenges in a unified manner. It accommodates a wide range of applications, including car tracking with motion sensors, heterogeneous human activity recognition, and user identification with biometric motion analysis. It significantly outperforms the state-of-the-art methods for all three tasks and is feasible to implement on smartphones and embedded devices. In another work [1], Aberer et al proposed a global sensor networks

(GSN) middleware based flexible architecture for integrated data processing of heterogeneous sensor networks. It supports efficient distributed query processing and combination of sensor data and enables dynamic adaption of the system configuration during runtime with minimal effort. Alamri et al proposed a Sensor-Cloud infrastructure [3] which provides a powerful and scalable high-performance computing and massive storage infrastructure for real-time processing and storing of the WSN data as well as analysis (online and offline) of the processed information under context using inherently complex models to extract events of interest. CASSARAM, a context-aware sensor search, selection, and ranking model, has been proposed [15] to efficiently select a subset of relevant sensors out of a large set of sensors with similar functionality and capabilities for data processing architecture for heterogeneous sensor data processing in industries. Some researchers have also used genetic algorithm to extend the network life and improve balanced energy consumption [6]. In this work, a Genetic Algorithm based method has been proposed for optimizing heterogeneous sensor node clustering in a Wireless Sensor Network (WSN).

In terms of auto-calibration and blind calibration of industrial sensors, one can find a number of research in the open literature. Blind calibration of industrial sensors is possible using total least squares (TLS) estimation, which provides significant performance benefits over the standard least squares approach [4, 8]. Convex optimization approaches for blind sensor calibration using sparsity have also been proposed [5]. Auto calibration of industrial sensors is proposed using a dynamic gas sensor network [19]. Auto-switch Gaussian process regression-based probabilistic soft sensors for industrial multigrade processes with transitions have also been proposed [9]. Recently, the PI has also proposed a patented AI-based blind calibration methodology [12] which would be developed further in this current project.

Sensory data compression has been a field of study for a long time. Of late, several works have appeared in the open literature using machine learning algorithms to compress sensory data. We shall limit out SOTA to these pieces of work. The use of deep learning for data compression (and anomaly detection) in structural health monitoring (SHM) data has been proposed in several studies [13]. The proposed approach includes a one-dimensional Convolutional Neural Network (CNN) for abnormal data detection and an Autoencoder-based compression and reconstruction method for normal data. In another interesting work [14] industrial data is represented by a function and predicted with high accuracy. This was shown to achieve high compression indices. In [18], they developed optimal compression techniques using CNNs for remote sensing images for industrial IoT application.

Lastly, looking at SOTA for cognitive architectures to analyze the data from industrial sensors, there are some recent works in this domain, including the ones by one of the authors. In [20] the use of meta data was shown to be enabling a fast and easy integration of data in the digital services for Industry 4.0. In another interesting work [7], the proposed cognition model was shown to outperform the other considered algorithms in terms of the classification accuracy. In [11]

a perception-centric cognitive architecture was proposed for industrial applications. In [10], a DIKW based architecture was proposed to process heterogeneous bigdata. A brain-inspired distributed cognitive architecture was proposed in [16] for distributed heterogeneous sensor data processing.

3 The Proposed Architecture

As was shown in the SOTA, there are several interesting methods which have been shown to be useful in analyzing data from industrial sensors. The following are some of the research gaps which the proposed architecture would endeavor to bridge.

- Most of the solutions in the open literature work with the assumption that the quality of the data coming from the sensors is of good quality. This is an assumption which is not true in most of the practical industrial setups.
- Most of the solutions in the open literature are domain and application specific. Hence, they may work for a given environment in each industry. Changing the goal or the industry invariably takes a substantial amount of redevelopment time.
- Making the solution application agnostic needs that the insights from the data are extracted at a level which is industry agnostic. This extraction of knowledge is a challenge not tackled in the open literature.

Based on the above gaps, the following are the major features which we wanted to achieve through the proposed architecture. We call a DIKW pyramid-inspired agile AI architecture for sensor data assimilation or DIPy-AI. overall research questions for the proposed project are as follows.

- Given a “bigdata” case in an industrial environment, the architecture needs to process the data to detect bad quality data, extract information from the data and be able to change the information-required and be able to smoothly change the data assimilation process.
- The new architecture needs to be agile enough to be useful for multiple sensor-types and multiple industries. The architecture should facilitate the extraction of actionable knowledge from the data from industrial sensors.
- The architecture should have a layered sensor-data analysis implementation based on abstraction (like the DIKW pyramid). This will facilitate changing of the overall goal without substantial re-development.

The data-information-knowledge-wisdom (DIKW) conceptual framework has been in use to model the progressive transformation of raw data into actionable insights. Perhaps, Ackoff [2] was the first to formulate this model. Of late, architectures for big-data processing have been proposed inspired by the DIKW model [10]. Though there is no universal consensus about the exact meanings of each of the layers, data, information and knowledge are better formulated than wisdom. Especially from an ML architecture point of view, the definitions of data and information are well understood. While data (symbolic or non-symbolic) is

the primary input to any expert system, information is the subjective meaning derived from data. Most machine learning operations can be modeled as data to information conversion. For example, when we use deep learning algorithms to identify various objects in a video stream, we are converting raw video data into objects of interest. Knowledge is more abstract and is mostly understood as building context around the information. Ontological processing of data is a good example of modern-day computational processes where we extract knowledge from data and information. Extraction of knowledge is almost like developing an “understanding” of the data. It can be mentioned that multiple initiatives in AI research are working towards this. We shall ignore the wisdom layer in our discussion. Because this is more abstract, and we are not going to use this in our proposed architecture.

In this work we propose a modular data-information-knowledge-wisdom (DIKW) pyramid-based sensor-data processing AI architecture. We aim to use this in industrial sensory data processing. Hence, we shall ignore the last layer (wisdom) which is not well defined as of now.

The proposed architecture, DIPy-AI, has the following distinct layers in the processing chain.

- **Sensor Layer (L1):** Sensor-dependent data pre-processing layer which will have sensor-specific data sanity-check, compression and calibration abilities.
- **ML Layer (L2):** Sensor-agnostic ML layer to convert sensor data into information using deep learning architectures.
- **LLM Layer (L3):** Application-specific information processing layer where information from L2 is analyzed through symbolic processing to extract context and goal specific knowledge which can be used to give the users goal-specific insights. L3 can also be used to interact with users to understand new needs in terms of new sensors and new industrial setup. This in turn can be propagated down to lower layers. Till recently, L3 would have been impossible to implement. However, given the tremendous success of large language models (LLMs), L3 will be implemented using LLMs.

The proposed solution is illustrated in the data flow graph (DFG) shown in Fig. 1. Heterogeneity is a major value proposition of the project. Hence, it would be developed to be easily fine-tuned for different sensors. This makes the architecture transferable to different industrial ecosystems. The next block represents the use of pertinent sensor models. All models are wrong, but some are useful ! Hence, the fidelity of the model to be used depends on the requirements of the application. This is a crucial and non-trivial block.

Layered architecture is another value proposition of the solution. This is achieved by building in wrappers between layers to facilitate inter-layer integration. The first wrapper works on three goals; viz. sensor-specific auto/blind calibration, sensor and application specific compression, and data packaging to make the succeeding ML layers data-format agnostic. Some of these goals are inspired by the way thalamus does sensory data pre-processing in the human brain [16]. In addition, the wrapper for data pre-processing and packaging (at the edge)

would also use a newly developed work (by the first author) to calibrate the sensor blindly and at the edge [12].

The deep learning block would be relatively non-challenging to implement. The following block is the second wrapper which would convert the information extracted by the ML algorithms into natural languages. This sub-symbolic to symbolic information conversion would be achieved by domain-specific dictionaries or look up tables (LUTs).

Information to knowledge conversion is a challenging task. The goal is to extract an understanding of the situation. This situational awareness would be implemented using the current generation of large language models (LLMs). The LLMs also feedback the action-items to the feedback mechanism which then decides on the ways to fine-tune the lower blocks in the architecture to adopt the complete architecture to insights generated from the situation awareness.

The next stage consists of humans in the loop. This also makes sure that our solution is adhering to the guidelines set by EU AI Regulation. The human or operator in the loop would also be necessary to change the goal of the solution (if needed). Lastly, feedback mechanisms would be determined based on the changing goals. This would be used to change the inputs to various preceding blocks. This top-down feedback is a non-trivial process in which we intend to use one of our recent works [17].

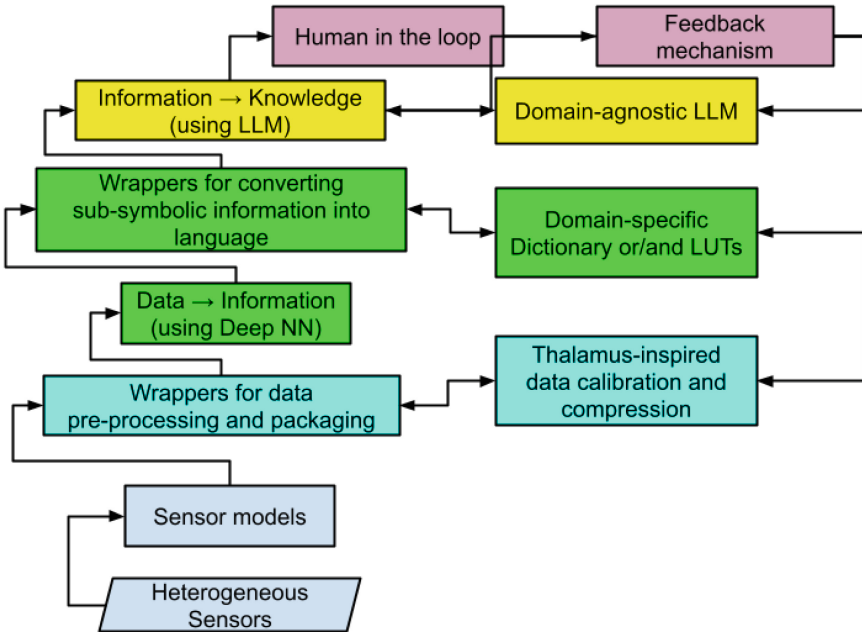


Fig. 1. DFG of DIPy-AI. Heterogeneity and layered-architecture are the two major value propositions of this solution

4 Conclusion and Future Work

Development of a flexible and agile sensor processing architecture is a challenge. The proposed DIKW-inspired architecture, DIPy-AI, which takes inspiration from DIKW model and thalamus, will enable industries to have a data assimilation system that is easy to change, repurpose and transfer to a different ecosystem. The use of LLMs to generate knowledge about the process is a major novelty and enabler of the architecture. The author is working with two Swedish industries (viz. GKN Aerospace and AP and T Groups) to validate the utility of the proposed architecture.

As discussed through the paper, the unique selling points (USPs) of the architecture are its agility (by which the solution can accommodate new scenarios, goals and sensors without needing substantial development time) and its flexibility (in the sense that the architecture is easily transferable across industries because of its layered nature).

Acknowledgments. This research is supported in part by the National Key Research and Development Program of China (2020YFB1807700), the National Natural Science Foundation of China (NSFC) grant No. 62071190, and the Key Research and Development Program of Hubei Province (2021BAA015).








References

1. Aberer, K., Hauswirth, M., Salehi, A.: Infrastructure for data processing in large-scale interconnected sensor networks. In: 2007 International Conference on Mobile Data Management, pp. 198–205. IEEE (2007)
2. Ackoff, R.L.: From data to wisdom. *J. Appl. Syst. Anal.* **16**(1), 3–9 (1989)
3. Alamri, A., Ansari, W.S., Hassan, M.M., Hossain, M.S., Alelaiwi, A., Hossain, M.A.: A survey on sensor-cloud: architecture, applications, and approaches. *Int. J. Distrib. Sensor Networks* **9**(2), 917–923 (2013)
4. Balzano, L., Nowak, R.: Blind calibration of sensor networks. In: Proceedings of the 6th International Conference on Information Processing in Sensor Networks, pp. 79–88 (2007)
5. Bilen, Ç., Puy, G., Gribonval, R., Daudet, L.: Convex optimization approaches for blind sensor calibration using sparsity. *IEEE Trans. Signal Process.* **62**(18), 4847–4856 (2014)
6. Elhoseny, M., Yuan, X., Yu, Z., Mao, C., El-Minir, H.K., Riad, A.M.: Balancing energy consumption in heterogeneous wireless sensor networks using genetic algorithm. *IEEE Commun. Lett.* **19**(12), 2194–2197 (2014)
7. Kumar, A., Jaiswal, A.: A deep swarm-optimized model for leveraging industrial data analytics in cognitive manufacturing. *IEEE Trans. Industr. Inf.* **17**(4), 2938–2946 (2020)
8. Lipor, J., Balzano, L.: Robust blind calibration via total least squares. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4244–4248. IEEE (2014)
9. Liu, Y., Chen, T., Chen, J.: Auto-switch gaussian process regression-based probabilistic soft sensors for industrial multigrade processes with transitions. *Ind. Eng. Chem. Res.* **54**(18), 5037–5047 (2015)

10. Mishra, A.K.: A DIKW architecture for cognitive engineering. *Procedia computer science* **123**, 285–289 (2018)
11. Mishra, A.K.: PeC-HiCA: a perception centric human-in-loop cognitive architecture. *Procedia Computer Science* **213**, 768–773 (2022)
12. Mishra, A.K.: A propagation-model empowered solution for blind-calibration of sensors. In: *International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE (2023)
13. Ni, F., Zhang, J., Noori, M.N.: Deep learning for data anomaly detection and data compression of a long-span suspension bridge. *Comput. Aided Civil Infrastructure Eng.* **35**(7), 685–700 (2020)
14. Park, J., Park, H., Choi, Y.J.: Data compression and prediction using machine learning for industrial IoT. In: *2018 International Conference on Information Networking (ICOIN)*, pp. 818–820. IEEE (2018)
15. Perera, C., Zaslavsky, A., Liu, C.H., Compton, M., Christen, P., Georgakopoulos, D.: Sensor search techniques for sensing as a service architecture for the internet of things. *IEEE Sens. J.* **14**(2), 406–420 (2013)
16. Rimmelzwaal, L.A., Mishra, A.K., Ellis, G.F.: Brain-inspired distributed cognitive architecture. *Cogn. Syst. Res.* **66**, 13–20 (2021)
17. Son, J., Mishra, A.K.: Exgate: Externally controlled gating for feature-based attention in artificial neural networks. In: *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE (2022)
18. Sujitha, B., Parvathy, V.S., Lydia, E.L., Rani, P., Polkowski, Z., Shankar, K.: Optimal deep learning based image compression technique for data transmission on industrial internet of things applications. *Trans. Emerg. Telecommun. Technol.* **32**(7), e3976 (2021)
19. Tsujita, W., Ishida, H., Moriizumi, T.: Dynamic gas sensor network for air pollution monitoring and its auto-calibration. In: *SENSORS 2004 IEEE*, pp. 56–59. IEEE (2004)
20. Weskamp, J.N., Chowdhury, A.G., Pethig, F., Wisniewski, L.: Architecture for knowledge exploration of industrial data for integration into digital services. In: *2020 IEEE Conference on Industrial Cyberphysical Systems (ICPS)*, vol. 1, pp. 98–104. IEEE (2020)
21. Yao, S., Hu, S., Zhao, Y., Zhang, A., Abdelzaher, T.: Deepsense: a unified deep learning framework for time-series mobile sensing data processing. In: *Proceedings of the 26th International Conference on World Wide Web*, pp. 351–360 (2017)



Assessment and Correlation of Morphometric and Tractographic Measures of Patients Diagnosed with Schizophrenia

Larisa Mosina¹ , Vadim Ushakov^{1,2,3} , Vyacheslav Orlov⁴ ,
Sergey Kartashov⁴ , Natalia Zakharova³ , Georgy Kostyuk³ ,
and Sergey Trushchev³ 

¹ National Research Nuclear University MEPhI, Moscow, Russia
the.borsch@yandex.ru

² Institute for Advanced Brain Studies, Lomonosov Moscow State University, Moscow, Russia

³ Mental Health Clinic No. 1 Named After N.A. Alexeev of Moscow Health Department,
Moscow, Russia

⁴ National Research Centre “Kurchatov Institute”, Moscow, Russia

Abstract. This study describes the detection of anatomical changes in brain regions based on morphometric measures and white matter tracts in patients diagnosed with schizophrenia (F20.0 according to ICD-10) compared to a health control group. All data were checked for normality, and significant differences between the health control group and schizophrenia patients divided into groups based on symptomatic severity were tested for the investigated parameters. A correlation analysis was also performed between connectivity strength of brain regions and morphometric parameters. Based on these findings, three tracts were identified as possible biomarkers for this disorder: Paracentral lobule left—Paracentral lobule right; Supplementary motor area left—Anterior cingulate left; Thalamus left—Inferior temporal left.

Keywords: Schizophrenia · MRI · Tractography · Morphometry · Correlation

1 Introduction

Despite numerous years of research utilizing brain imaging, the visual identification of structural abnormalities in the brain associated with schizophrenia (ICD-10 code F20.0) still does not allow objective diagnosis.

This study aims to detect anatomical changes in brain regions based on morphometric measures and white matter tracts of patients with schizophrenia compared to a health control group. The obtained information will allow for a more accurate diagnosis of the disease and provide an opportunity for further research to investigate changes in brain regions as well as connectivity between its areas.

This study is part of a larger interdisciplinary project aimed at developing objective criteria for building a classification system for polymorphic schizophrenia, characterized

by disorders of thinking and perception. Within the overall project, patients diagnosed with schizophrenia with hallucinatory-delusional syndrome are undergoing neuropsychological testing, comprehensive structural (morphometry and tractography), and functional MRI studies, which include detailed clinical examination as well as determination of the patient's immunological and genetic status.

When examining brain structures in individuals with schizophrenia compared to control group, significant reductions were observed not only in whole brain volume [1, 2] but also in cortical gray matter volume [3]. These findings suggest a widespread impact on the overall size and composition of the brain among those experiencing deterioration.

In addition to these global changes, specific regional volumes showed notable alterations as well. The fusiform gyrus [4], known for its role in facial recognition and visual processing, exhibited decreased volume. Similarly, reductions were observed in regions such as the pars orbitalis gyrus and pars triangularis [5] that are involved in language processing and cognitive control.

Furthermore, volumetric decreases were found in the superior temporal gyrus [5], which plays a crucial role in auditory processing and language comprehension. This suggests potential disruptions or impairments related to auditory perception among individuals within the deteriorated group.

On top of these findings, it is worth noting that both subgroups of schizophrenia patients demonstrated increased volumes in certain areas including the lateral ventricle (left), putamen (right), and pallidum (left). At the same time, they experienced decreased bilateral hippocampus volume alongside left precentral gyrus, right rostral middle frontal gyrus, and bilateral superior frontal gyrus volumes when compared to controls [5]. These alterations may be indicative of abnormalities within neural circuits involved in movement coordination, decision-making processes, executive functions associated with planning and problem-solving abilities.

Moreover, hyperconnectivity was observed between the thalamus—a key relay station for sensory information—with multiple brain regions specifically within individuals classified under the deteriorated group compared to connectivity patterns seen among controls [5]. This finding suggests disrupted communication pathways involving this vital region.

Additionally noteworthy is that volumetric alterations have frequently been reported specifically within memory-associated cerebral regions like the hippocampus among individuals diagnosed with schizophrenia [6]. Given its critical role in memory formation and consolidation processes along with spatial navigation ability, these structural changes may contribute significantly to the cognitive impairments commonly observed in schizophrenia patients.

According to research [7], there is evidence suggesting that individuals with schizophrenia exhibit decreased consistency in all six white matter tracts compared to those in the healthy control group. Specifically, reductions in consistency have been observed within the bilateral posterior thalamic radiation, right corticospinal tract, and bilateral superior longitudinal fasciculus (both left and right). Additionally, a decrease in consistency has also been noted within the left sagittal stratum among individuals diagnosed with schizophrenia. These findings underscore the potential disruptions in white matter connectivity associated with this mental health condition.

Overall, these findings provide valuable insights into the complex interplay between brain structure and schizophrenia. They highlight not only the global reductions in whole brain and cortical gray matter volume but also specific regional alterations that likely contribute to the cognitive deficits experienced by individuals within the deteriorated group. The presence of hyperconnectivity patterns and volumetric changes in memory-related regions further deepens our understanding of this multifaceted disorder at a neurological level.

In our investigation, we aimed to examine both morphometry and tract changes simultaneously. By combining analyses of brain structure and white matter tracts, we sought to gain a comprehensive understanding of the underlying neural alterations in individuals with schizophrenia.

2 Methods and Materials

1. Clinical method

This method involved selecting patients with similar psychiatric manifestations and classical development of Kandinsky-Clérambault syndrome. In the study, 62 patients (31 females and 31 males, mean age 26 ± 5 years) diagnosed with schizophrenia were included. They were evaluated during the period of remission after their first psychotic episode. There was also a control group from 41 people (17 females and 24 males, mean age 27 ± 7 years). The severity of the condition was assessed using psychometric scales such as PANSS, CRDPSS, BFCRS, NSA-4, FAB. In all patients, the disease began with persecutory interpretative delusions, followed by phenomena of mental automatism, auditory pseudohallucinations, and paraphrenic elaboration of delusions during psychotic episodes. To participate in the study, it was required to have vivid memories of psychosis and formed insight into the illness while undergoing treatment, as well as a reduction in psychotic symptoms. For each participant, the following factors were assessed: degree of catatonia, presence and severity of delusional symptoms, negative symptomatology, presence and severity of cognitive impairments, presence and severity of hallucinations, thought stereotypy, and disorganized thinking. All participants provided written informed consent after a comprehensive description of the research procedures in accordance with the Helsinki Declaration. The conduct of the study was approved by the local ethics committee of NRC “Kurchatov Institute” (No. 5 dated April 5, 2017).

Inclusion criteria: aged 21–35 years, meeting the diagnostic criteria for schizophrenia according to ICD-10 (F20) and DSM-5, right-handedness, awareness of their condition with memory retention of psychotic symptoms, informed consent to participate in the study. Exclusion criteria: schizoaffective and affective disorders, organic brain diseases, severe somatic and/or neurological conditions that may potentially affect brain physiology or structure, signs of psychoactive drug abuse and general contraindications for MRI scanning.

The clinical examination was conducted by two experienced psychiatrists with the involvement of all necessary data (family interviews, analysis of medical records, results of physical and laboratory tests, etc.).

2. Neurophysiological studies using magnetic resonance imaging (MRI) method

The study was conducted using magnetic resonance imaging (MRI) method on a Magnetom Verio 3T MRI scanner (Siemens, Germany), utilizing a 32-channel head coil. The research included analysis of morphological parameters and analysis of brain white matter tracts. Structural imaging was performed using T1-weighted MRI sequence (TR = 1900 ms, TE = 2.21 ms, 176 slices, voxel size of $1 \times 1 \times 1 \text{ mm}^3$).

Analysis of brain white matter tracts was performed using structural connectivity matrix obtained from tractography data in B0 mode (representing the magnetic field strength of the MRI scanner) and along 64 vector directions (with a b-value of 1500 s/mm^2), with TE = 101 ms and TR = 13700 ms. Data were acquired for two phase-encoding gradient directions: anterior-posterior and posterior-anterior. The slice thickness was 2 mm, and the in-plane resolution of the tomographic slice was also 2 mm.

3. Software processing

The data obtained from the MRI scanner were processed using the Freesurfer [8] software on a supercomputer. The cerebral cortex was divided into anatomical zones according to the atlas of Desikan-Killiany [9]. A total of 865 morphometric parameters were computed for each participant who took part in the study. For the morphometry analysis, 61 individuals diagnosed with schizophrenia of various severity levels and 41 individuals comprising the control group were selected.

These parameters were determined separately for each hemisphere of the brain for each of anatomical structures: mean MRI signal intensity (mean), volume of subcortical regions (volume), surface area (area), gray matter volume (volume), thickness, standard deviation of thickness (thicknessstd), integral mean curvature (meancurv), integral gaussian curvature (gauscurv), curvature index (curvind), folding index (foldind), area and average MRI intensity of smoothed surface [area (LGI) and mean (LGI)], volume of cerebellar regions.

The SUI [10] program was used to segment the cerebellum in the brain and obtain volumetric measurements of its structures. The anatomical structures of the cerebellum were derived based on the MNI152 template.

The DSI Studio [11] software was used to obtain data on the brain's white matter tracts. For this study, 48 patients diagnosed with schizophrenia of varying severity levels and 42 individuals comprising the control group were selected. For each participant was obtained information about 14,400 white matter tracts (connections between 120 brain regions).

The data obtained from the processing in Freesurfer [8] and DSI Studio [11] programs were checked for normality using RStudio software. The same data were also tested for significant differences compared to the health control group ($\alpha = 0.05$). Subsequently, Spearman's rank correlation coefficients and significance levels ($\alpha = 0.05$) were calculated for the available values, and a correction for multiple comparisons was applied. The missing data on brain white matter tracts were imputed using the MICE [12] (Multiple Imputation by Chained Equations) method.

The Shapiro-Wilk test was used to assess the normality of the distribution. Spearman's rank correlation was chosen to calculate the degree of association between the

data. For each computed correlation coefficient, the p -value was calculated to determine the level of significance. A statistical significance level was chosen $\alpha = 0.05$.

To conduct further analysis, the obtained data were divided into two groups:

1. Data that were normally distributed in both the health control group and the group of patients with schizophrenia.
2. Data that were not normally distributed in at least one of the groups.

For group 1, a two-sample independent t -test was used for analysis to determine whether there was equality or inequality in the means of the two independent populations. For the data from group 2, the Mann-Whitney U -test was used.

Based on the results of the t -test and U -test, data that showed statistical significance ($\alpha = 0.05$) were selected. For each of the remaining data points, the mean value was calculated for both the control group and the comparison group.

The Benjamini-Hochberg method [13] was used to calculate false discovery rate.

3 Results

For each of the 48 individuals with schizophrenia and 42 individuals from the health control group, data on 14,400 brain tracts were obtained. Since the data on tracts are identical for tract from region A to region B and from region B to region A, only 7200 unique tracts were considered. Additionally, due to more than half of the tracts being zero in the study, 10 patients with schizophrenia and 4 individuals from the health control group were excluded. From the remaining tract data, zero tracts going back into their own regions were also removed. Further-more, tract data that had a zero value for more than 50% of participants were not included in the study. Zero value tracts were excluded for each participant from all remaining data points.

For the analysis 38 individuals with schizophrenia and 38 individuals from the health control group which have both morphometry and tractography data were selected.

The group of patients diagnosed with schizophrenia was decided to be divided based on the severity of symptoms, as there was a hypothesis about varying degrees of structural changes in the brain associated with different levels of symptomatology.

In the group with mild negative symptoms, data from 12 individuals diagnosed with schizophrenia were investigated. The significant different tracts shown in Table 1. The data for the group of patients diagnosed with schizophrenia is denoted as SZ, while the healthy control group is denoted as HC.

Data from 13 individuals diagnosed with schizophrenia with mild hallucinations were considered for significant differences. Based on morphometry data significant difference were found in mean MRI signal intensity of smoothed entorhinal gyrus (2.46 ± 0.03 for SZ and 2.62 ± 0.01 for HC) with significance level $\alpha = 0.05$. The significant different tracts shown in Table 2.

In the group with mild delusional symptoms, data from 12 individuals diagnosed with schizophrenia were examined. Significant different morphometry parameters shown in Table 4 and significant different tracts shown in Table 3.

Table 1. Significant tract differences between the health control group and patients diagnosed with schizophrenia with mild negative symptoms.

Tract	Schizophrenia	Control
Precentral right—putamen left	474 ± 95	1342 ± 148
Occipital middle right—lingual right	1229 ± 198	580 ± 57
Occipital middle right—fusiform right	3135 ± 474	1648 ± 169
Thalamus left—temporal middle left	527 ± 100	1433 ± 110
Thalamus right—cerebellum 4 5 left	229 ± 49	529 ± 71

The significance level was set at $\alpha = 0.05$. (FDR)

Table 2. Significant brain connections differences found between the health control group and patients diagnosed with schizophrenia exhibiting mild hallucinations.

Tract	Schizophrenia	Control
Precentral right—putamen left	336 ± 63	1342 ± 148
Frontal superior left—supplementary motor area right	3314 ± 721	7241 ± 625
Frontal superior right—supplementary motor area left	3402 ± 857	9043 ± 714
Supplementary motor area right—cingulate anterior left	2585 ± 517	4944 ± 471
Occipital middle right—lingual right	1154 ± 188	580 ± 57
Thalamus left—ParaHippocampal left	511 ± 83	1244 ± 126
Thalamus left—temporal inferior left	969 ± 226	1816 ± 151
Cerebellum 4 5 right—Vermis 6	7413 ± 804	9885 ± 571
Cerebellum 9 left—Vermis 10	951 ± 124	1498 ± 125
Vermis 4 5—Vermis 6	5619 ± 348	6988 ± 337

The significance level was set at $\alpha = 0.05$. (FDR)

Table 3. Significant brain connections differences between the health control group and patients diagnosed with schizophrenia with mild delusional symptoms.

Tract	Schizophrenia	Control
Precentral right—putamen left	438 ± 103	1342 ± 148
Frontal superior right—supplementary motor area left	4691 ± 902	9043 ± 714
Cerebellum 9 left—Vermis 10	844 ± 151	1498 ± 125
Vermis 4 5—Vermis 6	5141 ± 234	6988 ± 337

The significance level was set at $\alpha = 0.05$. (FDR)

Table 4. Significant morphometry differences between the health control group and patients diagnosed with schizophrenia with mild delusional symptoms.

Parameter	Schizophrenia	Control
Volume left inferior lateral ventricle	564.93 ± 86.3 mm ³	224.32 ± 19.83 mm ³
Volume right choroid plexus	660.84 ± 62.11 mm ³	390.33 ± 25.77 mm ³

The significance level was set at $\alpha = 0.05$. (FDR)

Data from 11 individuals diagnosed with schizophrenia and moderate negative symptoms were examined. Significant differences were found in tract Thalamus left—Parahippocampal left (445 ± 74 for SZ and 1245 ± 126 for HC) with significance level $\alpha = 0.05$ (Table 4).

Data from 13 individuals diagnosed with schizophrenia and severe negative symptoms were examined. Results on found significant differences in brain connections shown in Table 5.

Table 5. Significant brain connections differences between the health control group and patients diagnosed with schizophrenia with severe negative symptoms.

Tract	Schizophrenia	Control
Frontal superior right—supplementary motor area left	5081 ± 1063	9043 ± 714
Thalamus left—temporal inferior left	1139 ± 156	1816 ± 151
Vermis 4 5—Vermis 6	5546 ± 214	6988 ± 337

The significance level was set at $\alpha = 0.05$

Data from 25 individuals diagnosed with schizophrenia and severe delusional symptoms were examined. Results on found significant differences in brain connections for this group shown in Table 6.

The patient group with severe positive symptoms consisted of 11 individuals diagnosed with schizophrenia, who simultaneously exhibited severe delusions and hallucinations. Significant differences found in tract Hippocampus right—Postcentral right (155 ± 26 for SZ and 339 ± 43 for HC) with significance level $\alpha = 0.05$.

The aim of next part of the study was to identify existing associations between tracts showing significant differences between the health control group and patients with a diagnosis of schizophrenia, and morphometric parameters of regions connected by these tracts at different levels of symptom severity. To achieve this, correlation coefficients were computed between data from significant tracts for each group (significance level $\alpha = 0.05$, FDR not included) and 849 morphometric parameter data.

For conducting the correlation analysis, 38 individuals with schizophrenia and 38 individuals from the health control group were selected.

The region of interest indicated by * was not present in the morphometry data. k represents the coefficient of Spearman's rank correlation.

Table 6. Significant brain connections differences between the health control group and patients diagnosed with schizophrenia with severe delusional symptoms.

Tract	Schizophrenia	Control
Frontal superior right—supplementary motor area left	5269 ± 730	9043 ± 714
Supplementary motor area left—cingulate anterior left	4379 ± 492	7037 ± 565
Supplementary motor area left—cingulate anterior right	3731 ± 553	5654 ± 492
Supplementary motor area right—cingulate anterior left	3243 ± 454	4945 ± 471
Hippocampus right—postcentral right	162 ± 24	339 ± 43
ParaHippocampal left—amygdala left	2937 ± 164	3526 ± 160
ParaHippocampal right—amygdala right	5987 ± 338	6878 ± 279
Calcarine left—occipital middle left	8667 ± 513	6675 ± 389
Calcarine left—occipital inferior left	5008 ± 435	3919 ± 269
Lingual right—precuneus right	5011 ± 372	6587 ± 346
Lingual right—cerebellum 4 5 right	5007 ± 402	6191 ± 376
Occipital middle right—lingual right	1073 ± 119	580 ± 57
Occipital middle right—fusiform right	2809 ± 288	1648 ± 169
Fusiform right—temporal middle right	3851 ± 365	2717 ± 223
Paracentral lobule left—paracentral lobule right	1249 ± 198	2558 ± 263
Caudate right—Vermis 3	187 ± 23	341 ± 32
Thalamus left—amygdala left	836 ± 128	1324 ± 141
Thalamus left—temporal middle left	926 ± 133	1433 ± 110
Thalamus right—cerebellum 4 5 left	268 ± 50	529 ± 71
Cerebellum 8 right—cerebellum 10 right	2232 ± 297	3386 ± 301

The significance level was set at $\alpha = 0.05$

Results of correlation analysis shown in Tables 7, 8, 9, 10, 11, 12 and 13. Coefficients are significant with significance level $\alpha = 0.05$.

4 Discussion

In the treatment of schizophrenia, diagnosis is a crucial stage. It aims to establish the diagnosis and identify the causes of the illness since symptoms related to cognitive impairments, delusions, and hallucinations can be characteristic not only of schizophrenia but also of other disorders such as infectious, metabolic, and other diseases. Accurate diagnosis determines the appropriateness of treatment selection and its effectiveness in the end result. For diagnosis, a psychiatrist analyzes complaints reported by either the patient or their close ones. The doctor conducts direct conversations with patients and inquiries about all symptoms of the illness, such as delusions, hallucinations, thought

Table 7. Results of correlation analysis for patients diagnosed with schizophrenia with mild delusional symptoms.

Tract	k_{sz}	Parameter
Supplementary motor area left*—cingulate anterior left	– 0.67	Volume caudal anterior cingulate
	0.58	Thickness rostral anterior cingulate
	– 0.58	Meancurv caudal anterior cingulate
	– 0.62	Gauscurv caudal anterior cingulate
	– 0.71	Curvind caudal anterior cingulate
Fusiform right—temporal middle right	0.64	Volume middle temporal
	0.64	Curvind middle temporal
	0.65	Foldind middle temporal
Paracentral lobule left—parietal superior left	– 0.58	Gauscurv paracentral
Thalamus left—paraHippocampal left	0.69	Meancurv parahippocampal
Thalamus left—temporal middle left	– 0.79	Meancurv middle temporal
	– 0.79	Gauscurv middle temporal
	– 0.82	Foldind middle temporal

Table 8. The results of the correlation analysis for patients diagnosed with schizophrenia with mild hallucinations.

Tract	k_{sz}	Parameter
Precentral right—putamen left	0.66	Meancurv precentral
	0.58	Gauscurv precentral
	0.58	Curvind precentral
Supplementary motor area right*—cingulate anterior left	0.72	Gauscurv caudal anterior cingulate
	0.7	Curvind caudal anterior cingulate
	0.63	Foldind caudal anterior cingulate
Hippocampus right—postcentral right	0.66	Meancurv postcentral

disturbances, catatonic symptoms, emotional disorders, among others. However, gathering information alone does not allow for an objective diagnosis of schizophrenia which may lead to inappropriate treatment being prescribed for patients. This work is part of comprehensive research aimed at creating disease markers that would enable objective diagnostics to be developed.

Firstly, the data from morphometry and tractography were tested for normal distribution within each group. The majority of morphometric parameters were normally distributed across all study groups. For the tractography data, a majority of connectivity strength between brain regions also follows a normal distribution. However, in the group

Table 9. The results of the correlation analysis for patients diagnosed with schizophrenia with moderate hallucinations.

Tract	k_{sz}	Parameter
Supplementary motor area left*—cingulate anterior left	- 0.72	Meancurv rostral anterior cingulate
	- 0.7	Curvind rostral anterior cingulate
	- 0.85	Foldind rostral anterior cingulate
Occipital middle right*—fusiform right	0.65	Foldind fusiform
	0.67	Mean (LGI) fusiform

Table 10. The results of the correlation analysis for patients diagnosed with schizophrenia and severe hallucinations.

Tract	k_{sz}	Parameter
ParaHippocampal left—amygdala left	0.59	Mean left amygdala
Paracentral lobule left—paracentral lobule right	0.55	Gauscurv paracentral left
	0.71	Curvind paracentral left
	0.67	Foldind paracentral left
Thalamus left—temporal middle left	- 0.6	Mean middletemporal
Thalamus right—cerebellum 4 5 left	0.56	Volume right thalamus

Table 11. The results of the correlation analysis for patients diagnosed with schizophrenia with moderate negative symptoms.

Tract	k_{sz}	Parameter
Cingulate posterior right—lingual right	0.62	Foldind posteriorcingulate
Cerebellum 9 left—Vermis 10	- 0.63	Volume cerebellum vermis X

of patients with severe delusional symptoms, a different type of distribution predominates. As a result of the conducted study, all available data on morphometric parameters and connectivity strength of brain tracts were analyzed.

During the analysis of data on changes in the strength of connectivity between brain regions, several potential significant alterations were identified at different levels of symptom severity.

As can be observed, in Figs. 1 and 2 there is a proportional change in connectivity strength with an increase in symptom severity. Figure 1 depicts a decrease in tract connectivity strength as the disease severity increases, indicating that patients with schizophrenia may have disrupted connectivity between areas connected by these tracts (Cerebellum

Table 12. The results of the correlation analysis for patients diagnosed with schizophrenia with severe negative symptoms.

Tract	k_{sz}	Parameter
ParaHippocampal left—amygdala left	− 0.64	Gauscurv parahippocampal
Calcarine left—occipital middle left*	0.76	Meancurv pericalcarine
Calcarine left—occipital inferior left*	− 0.58	Thickness pericalcarine
Paracentral lobule left—parietal superior left	− 0.6	Area superior parietal
	− 0.74	Volume superior parietal
	− 0.64	Thickness superior parietal
	− 0.6	Area (LGI) superior parietal
	0.64	Gauscurv paracentral
Thalamus left—temporal inferior left	0.67	Mean (LGI) inferior temporal
	0.62	Meancurv inferior temporal

Table 13. The results of the correlation analysis for patients diagnosed with schizophrenia with severe positive symptoms.

Tract	k_{sz}	Parameter
Hippocampus right—postcentral right	0.65	Curvind middle temporal
	0.74	Mean (LGI) middle temporal
Calcarine left—occipital inferior left*	− 0.73	Mean (LGI) pericalcarine
Lingual right—precuneus right	− 0.67	Mean (LGI) lingual
Paracentral lobule left—paracentral lobule right	0.7	Meancurv paracentral left
	0.63	Curvind paracentral left
Thalamus right—cerebellum 4 5 left	0.7	Volume right thalamus

8 right—Cerebellum 10 right for delusional symptoms and hallucinations, Hippocampus left—amygdala left for negative symptoms). The opposite situation is presented in Fig. 2—the connectivity strength between two regions increases with the severity of the disease. Such changes may indicate that the right hemisphere hippocampal areas and right middle temporal pole may be directly involved in the manifestations of the disease, resulting in an increased connectivity strength between them as symptom severity worsens. The right hippocampus is responsible for human spatial memory, while the temporal pole of the brain participates in functions such as emotional and social behavior, memory, and speech. Considering that hallucinatory symptoms can create additional objects and sounds that do not exist in reality for an individual, it is possible that these areas play a role in the formation of hallucinations.

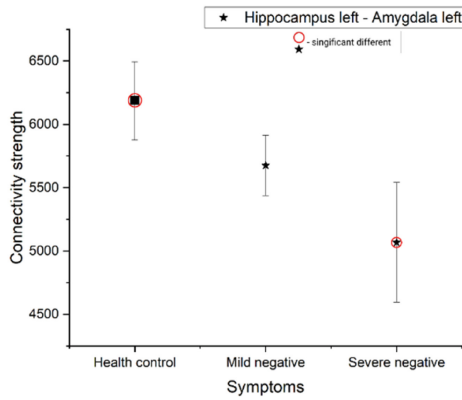


Fig. 1. Changes in connectivity strength depending on the severity of symptoms on the tract Hippocampus left—amygdala left

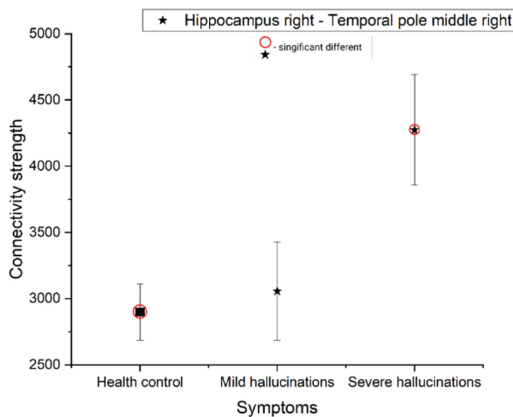


Fig. 2. Changes in connectivity strength depending on the severity of symptoms on the tract Hippocampus right—temporal pole middle right

An unusual situation is observed in Fig. 3. In this image, it can be seen that with mild symptoms, significant differences occur in the strength of connectivity between areas in individuals (Vermis 4 5—Vermis 6 for delusional symptoms, Precentral right—Putamen left and Thalamus left—Parahippocampal left for hallucinations, Thalamus left—Temporal middle left for negative symptoms). However, with severe symptoms, the strength of connectivity approaches the levels observed in the health control group. Such differences may indicate that in the early stages of the disease, the human brain attempts to compensate the additional load on certain regions, eventually restoring connectivity parameters close to those of a healthy individual. The presence of such changes in the human brain is crucial for early disease diagnosis, even before any noticeable symptoms manifest themselves. These changes can be observed prior to clear symptomatology that

would allow a doctor to diagnose the condition based on traditional methods. However, alterations in connectivity strength between brain regions are already evident.

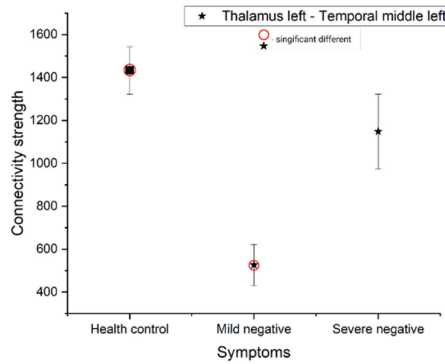


Fig. 3. Changes in connectivity strength depending on the severity of symptoms on the tract Thalamus left—Temporal middle left

It is worth noting that changes occurred not only in the strength of connectivity for some tracts but also in the morphometric measures of the regions they connect. Furthermore, correlations were found between these altered tracts and changes in morphometric parameters of the associated areas, suggesting a link between these tracts and the disease. Such tracts include: (1) Paracentral lobule left—Paracentral lobule right ($N_{sz} = 1249 \pm 198$; $N_{hc} = 2558 \pm 263$) correlating with changing thickness ($k = -0.45$; $L_{sz} = 2.41 \pm 0.02\text{mm}$; $L_{hc} = 2.49 \pm 0.02\text{mm}$) of right paracentral lobule; (2) Supplementary motor area left—Anterior cingulate left ($N_{sz} = 4030 \pm 6356$; $N_{hc} = 7037 \pm 565$) correlating with the changing mean integral curvature ($k = -0.72$; $M_{sz} = 0.13 \pm 0.003$; $M_{hc} = 0.14 \pm 0.002$) of anterior cingulate; (3) Thalamus left—Inferior temporal left ($N_{sz} = 1139 \pm 156$; $N_{hc} = 1816 \pm 930$) correlating with mean MRI signal intensity of smoothed surface ($k = 0.67$; $LGI_{sz} = 2.67 \pm 0.04$; $LGI_{hc} = 2.73 \pm 0.02$) of inferior temporal gyrus.

In the case of reduced consistency or alterations in white matter tracts, it is believed that disruptions may arise from abnormalities during early brain development or impaired myelination processes [14]. These abnormalities can impact the integrity and efficiency of communication between different brain regions, leading to decreased consistency within specific tracts [15].

Dysfunction within networks connecting regions like the paracentral lobule, supplementary motor area— anterior cingulate gyrus, thalamus-inferior temporal gyrus could disrupt information processing related to motor control coordination executive functions language processing visual perception respectively [16, 17].

It’s important to note that while we have identified significant differences between groups (e.g., individuals with schizophrenia compared to healthy controls), further research is needed to fully understand the underlying mechanisms driving these observations. Future studies incorporating larger sample sizes along with longitudinal designs

will provide deeper insights into why specific brain regions exhibit such notable differences among individuals with schizophrenia.

5 Conclusion

Tracts of the brain pathways and morphometry parameters were identified, according to which there are significant differences between the health control group and patients diagnosed with schizophrenia divided into groups depending on the severity of symptoms. Two possible variants of changes in the strength of connectivity of brain regions depending on the severity of symptoms were identified: proportional (increasing and decreasing) and nonlinear. According to the results of the correlation analysis, tracts were identified that correlate with the morphometric parameters of the regions between which a connection was formed. The relative number of tracts correlating with the parameters of the associated areas increased with increasing severity of the disease (from 0 to 42%). As biomarkers of schizophrenia disease, 3 tracts were identified, according to which there are simultaneously significant differences between the healthy control group and the group of patients diagnosed with schizophrenia, correlating with the parameters of the areas that bind, and there are significant changes in these areas.

Acknowledgment. This work was in part supported by Russian Science Foundation grant No 20-15-00299-P (<https://rscf.ru/en/project/20-15-00299-P/>, data acquisition, statistical and neurophysiological analysis) and grant No 22-11-00213 (data preprocessing), by project FSWU-2023-0031 “Analytical and numerical methods for studying complex systems and nonlinear problems of mathematical physics” (calculation of morphometric and tractographic indicators).


References

1. Veijola, J., et al.: Longitudinal changes in total brain volume in schizophrenia: relation to symptom severity, cognition and antipsychotic medication. *PLoS ONE* **9**(7), e101689 (2014). <https://doi.org/10.1371/journal.pone.0101689>
2. Guo, J.Y., et al.: Longitudinal regional brain volume loss in schizophrenia: Relationship to antipsychotic medication and change in social function. *Schizophr. Res.* **168**(1–2), 297–304 (2015). <https://doi.org/10.1016/j.schres.2015.06.016>
3. Madre, M., et al.: Structural abnormality in schizophrenia versus bipolar disorder: a whole brain cortical thickness, surface area, volume and gyrification analyses. *NeuroImage Clin.* **25**, 102131 (2020). <https://doi.org/10.1016/j.nicl.2019.102131>
4. Jung, S., et al.: Fusiform gyrus volume reduction associated with impaired facial expressed emotion recognition and emotional intensity recognition in patients with schizophrenia spectrum psychosis. *Psych. Res. Neuroimag.* **307**, 111226 (2021). <https://doi.org/10.1016/j.pscychresns.2020.111226>
5. Yasuda, Y., et al.: Brain morphological and functional features in cognitive subgroups of schizophrenia. *Psych. Clin. Neurosci.* **74**(2), 89–101 (2019). <https://doi.org/10.1111/pcn.12963>
6. Adriano, F., Caltagirone, C., Spalletta, G.: Hippocampal volume reduction in first-episode and chronic schizophrenia: a review and meta-analysis. *Neuroscientist* **18**, 180–200 (2012). <https://doi.org/10.1177/1073858410395147>

7. Zhao, J., Huang, C.C., Zhang, Y., et al.: Structure-function coupling in white matter uncovers the abnormal brain connectivity in Schizophrenia. *Transl. Psych.* **13**, 214 (2023). <https://doi.org/10.1038/s41398-023-02520-4>
8. FreeSurfer: Bruce Fischl. *Neuroimage* **62**(2), 774–781 (2012). <https://doi.org/10.1016/j.neuroimage.2012.01.021>
9. Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., et al.: An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**(3), 968–980 (2006)
10. Diedrichsen, J.: A spatially unbiased atlas template of the human cerebellum. *Neuroimage* **33**(1), 127–138 (2006)
11. Yeh, F.-C., Verstynen, T.D., Wang, Y., Fernández-Miranda, J.C., Tseng, W.-Y.I.: Deterministic diffusion fiber tracking improved by quantitative anisotropy. *PLoS ONE* **8**(11), e80713 (2013). <https://doi.org/10.1371/journal.pone.0080713>
12. Van Buuren, S., Groothuis-Oudshoorn, K.: Mice: multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**(3), 1–67 (2011)
13. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**(1), 289–300 (1995)
14. Thomason, M.E., Race, E., Burrows, B., Whitfield-Gabrieli, S., Glover, G.H., Gabrieli, J.D.: Development of spatial and verbal working memory capacity in the human brain. *J. Cogn. Neurosci.* **21**(2), 316–332 (2009). <https://doi.org/10.1162/jocn.2008.21028>
15. Zalesky, A., et al.: Disrupted axonal fiber connectivity in schizophrenia. *Biol. Psych.* **69**(1), 80–89 (2011). <https://doi.org/10.1016/j.biopsycho.2010.08.022>
16. Barch, D.M., Ceaser, A.: Cognition in schizophrenia: core psychological and neural mechanisms. *Trends Cogn. Sci.* **16**(1), 27–34 (2012). <https://doi.org/10.1016/j.tics.2011.11.015>
17. Hwang, W.J., et al.: Thalamic connectivity system across psychiatric disorders: current status and clinical implications. *Biol. Psych. Glob. Open Sci.* **2**(4), 332–340 (2021). <https://doi.org/10.1016/j.bpsgos.2021.09.008>



Deep Learning Evolution: Using Genetic Algorithm to Modify Training Datasets

Mikhail Yu. Nazarko, Klim A. Fedorov, and Alexei V. Samsonovich^(✉) 

National Research Nuclear University “MEPhI”, Kashirskoe Shosse 31, Moscow, Russia
avsamsonovich@mephi.ru

Abstract. The work addresses the problem of integration of deep learning and genetic algorithms (GA). An approach is developed where the GA directly modifies the training datasets rather than adjusting the parameters of the trained neural network. These datasets consist of records capturing the agent’s behavior in the environment and are treated as genotypes within the GA framework. The resulting phenotypes are the trained neural networks themselves. Importantly, the architecture and hyperparameters of the neural network and its learning model remain unchanged throughout the process. Numerical experiments conducted using the “Three Cowboys” game paradigm provide evidence supporting the concept and demonstrate the effectiveness of the proposed approach.

Keywords: Genetic algorithm · Deep learning · Integration

1 Introduction

The objective of integrating deep learning methods and genetic algorithms (GA) is to discover a combination of them where the respective advantages of both approaches are leveraged while their limitations are mitigated. Until now, the application of GA to training neural networks or other statistical models has been limited mainly to the modification of the parameters and architecture of the model [1, 2] or the hyperparameters of its training process through GA [3]. An alternative example of an integration approach is the use of a deep neural network to implement the GA fit function [4].

This paper examines the idea that an approach in which the GA does not directly affect either the neural network itself or the process of its training may be more productive. Instead, the GA is used to modify the data sets on which the neural network is trained. Modification is carried out through recombinations and mutations applied directly to the data itself, and not to its parameters or to the probabilities of selecting certain data for training. In this sense, datasets act as a genotype.

2 General Statement of the Problem and the Approach

Let’s consider the behavior of an agent consisting of a sequence of discrete actions in a certain environment within a certain paradigm: for example, a game with a certain objective function. Let the agent’s actions be controlled by a neural network that receives

the state of the environment as input and outputs a specific choice of action as output. The neural network is trained using examples of behavior, and its learning process is controlled by GA.

At the end of each match of the game, the agent receives an assessment of his target function, which is determined by both his actions and events in the environment: for example, the actions of rivals or partners. The task of training the neural network, as well as the GA that controls the learning process, is to achieve the maximum possible average value of the objective function within the framework of this behavioral paradigm.

According to the proposed approach, GA will not directly affect the architecture or parameters of the neural network. Instead, data on the behavior of individual agents will be collected during the game. The collected data will be selected and modified by recombination, after which they will be used as datasets for training the next generation. In this case, supervised learning will be used.

This approach to combining the work of GA and training neural networks is universal, since it allows you to use any neural network architecture. The main thing is that the neural network can learn lessons from the collected data about the behavior of the agent. This method can be used for any paradigm in which there is an agent interacting with the environment in some way, and there is an objective function—the assessment of the agent’s behavior.

In this discussion, we will focus on a scenario in which the dataset is a time series. The architecture of long short-term memory (LSTM) is usually used for processing this type of data. The reason is that the LSTM layer is able to learn how to reproduce long-term dependencies in time series [5]. Also, all the prepared entities are they passed through a residual self-attention block, similar to [6].

The procedure for crossing data sets can be any, provided that it does not violate the integrity of episodes of behavior, such as game matches. The order of the sequence is important. Crossovers, for example, can change the behavior of different players. In this case, intersections in training data sets occur by concatenating the behavior history of parental individuals.

In this case, it was decided to use the stochastic universal sampling (SUS) selection [7] method because it provides a more uniform selection and a large variety in the population, which in turn prevents the rapid degeneration of the population. It also makes sense to consider tournament selection, due to the possible inaccuracy of the fitness function.

Mutation is introduced by utilizing a dataset that solely comprises behavior episodes from one of the parents to train a new individual. This form of mutation goes beyond mere replication, as even when trained on identical datasets, the neural networks’ internal weighting coefficients may differ, resulting in variability.

3 Specific Statement of the Problem and Its Formalization

To test the performance of the concept of the proposed approach, the game “Three cowboys” was chosen as an experimental paradigm in this work. This game involves three opponents—“cowboys”, and the match consists of a series of rounds: fights between cowboys. The goal of the game is to score the maximum number of points in the match.

The conditions of the game are as follows. Each cowboy has one single-shot pistol, which cannot be transferred to another cowboy. All three are in an open, limited space, constantly see all each other's actions, including aiming, and are excellent shooters (when shooting at a target, they do not miss). In each round, shooting begins after the signal and stops after the allotted time interval. At the end of the round, the one who was not hit by the shot receives 10 points, plus another point if he himself hit someone. The one who is hit receives nothing.

Obviously, under these conditions, cowboys should compete with each other, but their optimal behavior is to fire a shot into the air. This result is the most humane from a human point of view. This version of the game can be called peaceful. In another, aggressive version of the game, an additional condition is introduced: if at the end of the round no one is hit, then everyone loses.

The selected example of the experimental paradigm has several features. Firstly, there are no pre-trained data with which to pre-train the neural network. Secondly, there is no absolute criterion for assessing the quality of a cowboy: the optimality of the resulting individual can be judged both by his ability to win in the population where he developed (local fitness) and in an arbitrary, previously unknown to him population of cowboys (absolute fitness). This difference is significant, since a population in the process of evolution can form an internal language, for example, based on signaling through a certain sequence of targeting, which will be understood only by representatives of this particular population.

Using more formal language, the problem can be described as follows.

- A cowboy is defined as an agent capable of performing the following actions:
 - inaction—denoted as DN ;
 - aiming at a specific (one of the two) opponent—denoted as AM_i , where i is the number of the opponent at whom the gun is pointed;
 - a shot at a specific (one of two) opponent—denoted as ST_i , where number i is determined similarly to the previous case;
 - shot in the air—denoted as SA ;
 - “survived”— AL ;
 - “dead”— DE .
- A cowboy can perform actions no more often than once every few clock periods (discrete time ticks). This characteristic of the agent is called reaction time. It changes randomly after each action, remaining within a given range, and is unknown in advance.
- The set of possible states of cowboys will be denoted as $states = \{DN, AM_i, ST_i, SA, DE, AL, DE\}$.
- Let's introduce the cowboy evaluation function in the round:

$$f(x) = m, x \in states, m \in Z, \quad (1)$$

defining it as follows:

- If the cowboy dies, the score for the round is zero. $f(DE) = 0$;

- If the cowboy survives, the score for the round is ten $f(AL) = 10$;
 - If the cowboy survives and kills the cowboy, then the score for the round is eleven $f(ST_i) = 11, i \in 0 \dots (K - 1)$;
 - In the second version of the game there is also a rule in which the entire group participating in the match receives 0 points if no one died during the match.
- A round refers to a single fight between three cowboys. In a duel, each agent is assigned a number of points (points) according to the function (1) of cowboys' evaluation in the round.
 - A match is defined as a sequential series of rounds by one group of cowboys.
 - Let's introduce a function for determining the winner of a match:
 - The winner of the match is the cowboy with the maximum amount of points over a series of rounds:

$$WN = \operatorname{argmax}_{x=1\dots K} \left(\sum_1^N f(M_{x,i}) \right), \quad (2)$$

where $M_{x,i}$ is the status of the cowboy with name x at the i -th step of the round.

- The behavior of an agent may depend on various parameters, which include:
 - current statuses of opponents (values from the set *states*);
 - previous statuses of opponents (values from the set *states*);
 - statuses of opponents that they had in previous rounds of the match (values from the set *states*).

From the previous paragraph we can conclude that in order to make a decision, a cowboy needs to take into account all the statuses that each of the opposing cowboys had throughout the match—in general, the sequence of cowboys' statuses represents a time series. This circumstance does not allow the use of methods of crossing and mutation that violate the logical connection between events.

4 Results of Numerical Experiments and Their Analysis

In the course of the work, numerical experiments were carried out within the framework of the described paradigm of the Three Cowboys game. Both versions of the game were tested: peaceful and aggressive. Runs with and without stopping criteria were tested, as well as various configurations determined by hyperparameter values. The results obtained are as follows.

In the peaceful version, it turned out that agents quickly come to peaceful behavior if their reaction window does not vary (Fig. 1). In Figs. 1 and 2 the following notations are used:

- x-axis represents the generation number of the population;
- y-axis indicates the number of points scored in qualifying matches;
- The blue line is the minimum number of points in the population;

- The orange line is the average number of points in the population;
- The green line is the maximum number of points scored in the population.

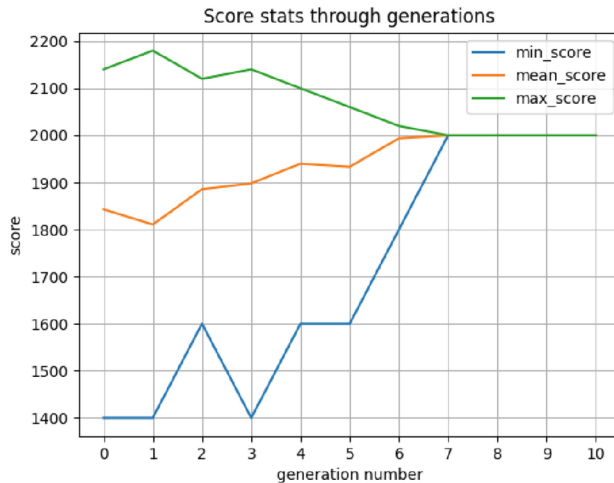


Fig. 1. An example of the evolution of a population in a peaceful version of the game.

Starting the system in the case of Fig. 1 was performed with the following values of hyperparameters: reaction time is fixed at 1 tick, population size is 99, number of selected individuals is 60, number of group shuffles is 20, number of matches within a group is 5. As can be seen from the example of evolution, shown in Fig. 1, initially the agents fired at each other, but after their aggressiveness decreases, and in the eighth generation they come to a peaceful behavior strategy that allows everyone to safely score a significant amount of points.

After that, the behavior of the system in the aggressive version of the game was investigated. The results are shown in Fig. 2.

Starting the system in the case of Fig. 2 was carried out with the following values of hyperparameters: the reaction time is fixed and is 1 tick, the population size is 14, the number of selected individuals is 6, the number of group mixing is 3, the number of matches within the group is 4. As can be seen from Fig. 2, the agents quickly come to a stationary situation, if we talk about their ability to defeat each other. This is also manifested in the fact that after a small number of generations the system immediately passes the algorithm stopping criterion, initially defined as the best individual scoring 90% of the maximum possible number of points. Therefore, the stopping criterion was modified as described below.

Since even with a large number of matches played, the average number of points scored during the selection process remained approximately the same level, it was decided to develop a stopping criterion that involved holding control matches with a “reference” agent, whose behavior is controlled by a pre-written algorithm. In such a stopping criterion, the number of matches that the agent controlled by the algorithm did not win against two agents from the population is important. The proportion of matches

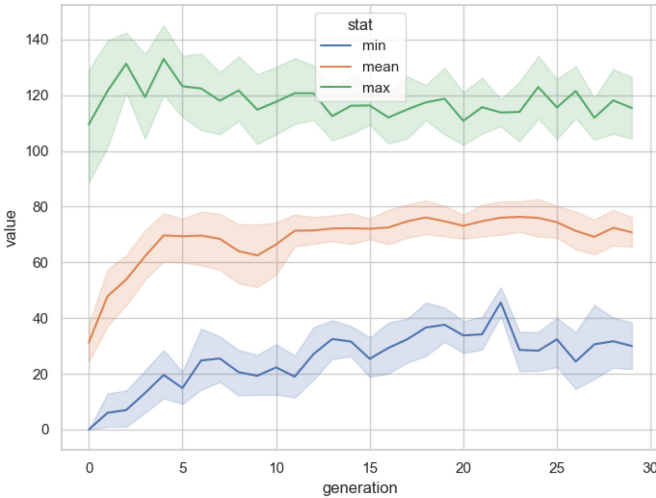


Fig. 2. An example of population evolution in an aggressive version of the game.

that must be won can also be adjusted. Also at this stage, evolution was launched with the same settings many times to collect statistics. The results are presented in Fig. 3. Specifically, in Fig. 3 the following notations are used:

- x-axis represents the generation number of the population.
- y-axis shows the number of points scored in test matches with “standard” agents.
- The blue line is the number of points scored in test matches by the first agent from the population.
- The orange line is the number of points scored in test matches by the second agent from the population.
- The green line is the number of points scored in test matches by the second “reference” agent, controlled by a pre-written algorithm.
- The boundaries of the colored bars correspond to the confidence intervals.

As can be seen in Fig. 3, at first the agents from the evolving population and their “standard” opponent scored almost the same number of points, but then the selected evolving agents began to dominate each match. By the end of the experiment, the difference becomes very significant in favor of the trained neural networks, which indicates their development. To collect statistics, the launch took place without a stop criterion.

An increase in the quality of agents in control matches is also observed when the number of their points in qualifying matches with each other remains approximately the same. That is, the agents remain at approximately the same level relative to each other, but grow relative to the fixed algorithm. Also, the minimum values in each generation are growing, and therefore relative to each other, each individual in the generation plays better than in the previous one.

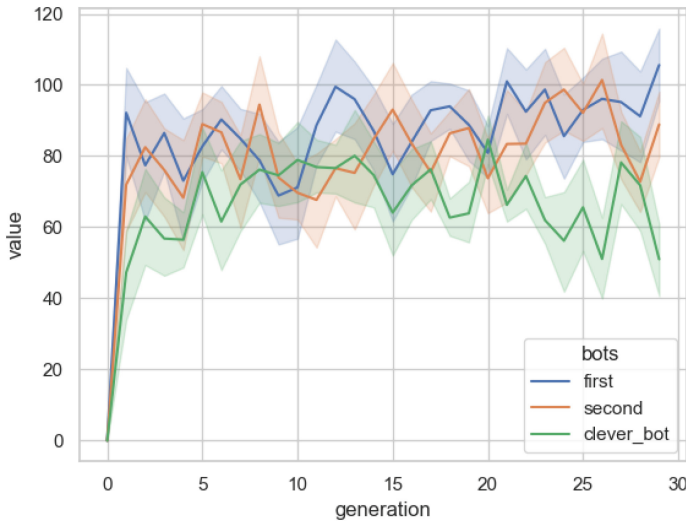


Fig. 3. Test match statistics as a function of generation number in the aggressive version of the game.

5 Conclusions

A new method for integrating deep learning of neural networks and GAs is proposed and implemented. The approach is universal and allows you to work with various neural network architectures and solve various problems. By optimizing the training samples, this approach shows successful results in reinforcement learning (RL) tasks. Numerical experiments confirm the functionality of the concept.

The proposed approach represents a departure from conventional methods of combining Genetic Algorithms (GAs) and neural network training models. In this case, the genotype, which undergoes mutations and recombinations, is represented by the dataset used to train the neural network. The resulting phenotype is the trained neural network itself. Parallelization of the training stages is also possible, which significantly speeds up the process of obtaining an acceptable agent.

One limitation of the method is that mutations and recombinations cannot break the integrity of episodes of behavior: in this case, matches. Future studies will attempt to overcome this limitation. In addition, it is necessary to expand both the range of crossing methods and the range of experimental paradigms. Finally, it is necessary to compare the effectiveness of this method with traditional analogues, including both “pure” GA and deep learning models (primarily RL) and their hybrid versions.

Practical applications of intelligent social agents evolved based on the proposed method can be anticipated to find many important practical applications, such as, for example, in intelligent tutoring systems and cognitive learning environments [8].

Acknowledgments. This work is supported by the Russian Science Foundation Grant No. 22-11-00213.

References

1. Levy, E., David, O.E., Netanyahu, N.S.: Genetic algorithms and deep learning for automatic painter classification. In: Proceedings of the 16th Genetic and Evolutionary Computation Conference: GECCO 2014, pp. 1143–1150. Association for Computing Machinery, Vancouver, BC (2014). <https://doi.org/10.1145/2576768.2598287>
2. Young, S.R., Rose, D.C., Karnowski, T.P., Lim, S.H., Patton, R.M.: Optimizing deep learning hyper-parameters through an evolutionary algorithm. In: Workshop on Machine Learning in High-Performance Computing Environments: MLHPC 2015, Article 4. P. 1–5. Association for Computing Machinery (2015). <https://doi.org/10.1145/2834892.2834896>
3. Erden, C.: Genetic algorithm-based hyperparameter optimization of deep learning models for PM25 time-series prediction. *Int. J. Environ. Sci. Technol.* **20**, 2959–2982 (2023)
4. Majidi, M., Toroghi, R.M.: A combination of multi-objective genetic algorithm and deep learning for music harmony generation. *Multimedia Tools Appl.* **82**(2), 2419–2435 (2023). <https://doi.org/10.1007/s11042-022-13329-6>
5. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press, Boston, MA (2016)
6. Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., Mordatch, I.: Emergent Tool Use from Multi-Agent Interaction. [arXiv:1909.07528](https://arxiv.org/abs/1909.07528) (2020)
7. Baker, J.E.: Reducing bias and inefficiency in the selection algorithm. In: Grefenstette, J.J. (ed.). *Genetic Algorithms and their Applications: Proceedings of the Second International Conference on Genetic Algorithms*. Psychology Press, New York (1987)
8. Samsonovich, A.V., De Jong, K.A., Kitsantas, A., Peters, E.E., Dabbagh, N., Kalbfleisch, M.L.: Cognitive constructor: an intelligent tutoring system based on a biologically inspired cognitive architecture (BICA). *Front. Artif. Intell. Appl.* **171**(1), 311–325 (2008)



Strategies for Business Cybersecurity Using AI Technologies

Svetlana Nosova , Anna Norkina , and Nikolay Morozov 

National Research Nuclear University “MEPHI”, Kashirskoe Shosse 31, 115409 Moscow,
Russian Federation
SSNosova@mephi.ru

Abstract. The article analyzes business cybersecurity strategies using artificial intelligence technologies, which are currently changing rapidly due to the nature of cyber attacks and the need to find new ways to counter cybersecurity threats. The main purpose of the study is to show how, by introducing artificial intelligence at the micro level, cybersecurity groups can help organizations make more informed business decisions regarding the production and logistics of their products and services, given that cybersecurity and artificial intelligence should work together, create cybersecurity management programs, thereby increasing awareness of the risks associated with cybercrime. Overall, our goal is not only to discuss cybersecurity data science and related techniques, but also to focus on the applicability of cybersecurity to intelligent data-based decision-making to protect systems from cyber attacks. The results of the study are aimed at extracting models of cybersecurity incidents and building an appropriate cybersecurity model in order to (1) make the security system automated and intelligent, and business, having taken steps to mitigate cyber threats, was able to experiment with artificial intelligence and cybersecurity, take on more responsibility, strengthen cooperation, develop a strategy for using artificial intelligence technologies to protect cyberspace; (2) strategically transforming cybersecurity, AI developers should make new changes and make more informed decisions to manage cyber attacks in the business process system; (3) help business leaders determine the timing of investments and the share of the budget for the implementation of AI related to the investment strategy for the development of AI, based on future reports on the evaluation of the effectiveness of AI.

Keywords: Cybersecurity · Artificial intelligence · Cyberspace · Cybercrime · Cyber-attacks · Risks · Fraud

1 Introduction

Cybersecurity in the broad sense of the word is governmental and non-governmental organizations, academia, resources, processes and structures used to protect cyberspace from cyber threats as one of the quintessential threats to modern security due to the realities of a globalized, interdependent and networked world. The established meaning

of “cybersecurity” has evolved from a macro-concept to its micro-meaning, i.e. enterprises (businesses) where cybersecurity is a critically important aspect for all types of their activities. Unfortunately, in a number of countries, most enterprises do not have the high-quality tools necessary to quickly identify and restore business processes after cyber threats, which are of great importance for cybersecurity based on the latest AI technologies [1]. For example, a study conducted by Cisco in 2019 showed that artificial intelligence-based tools can identify up to 95% of the threats faced by an organization. Therefore, it is extremely important that cyberspace be provided with adequate protection against unauthorized and illegal activities (cybercrime), leading to huge monetary losses. Cybercrime is the biggest threat to every company in the world and one of the biggest challenges facing humanity [2]. It threatens incentives for innovation and investment and will be more profitable than global trade in all major illicit goods (e.g. drugs) and services combined. Therefore, a key role in the fight against cybercrime is played by ensuring cybersecurity not only of enterprises and their infrastructure, but also in addition to coordinating risk management and information technology activities and regulating the activities of private and public enterprises; countries must cooperate with each other to have a secure cyberspace and resolve legitimate cybercrime requests. This will lead to the active introduction of AI technologies and the creation of a reliable cybersecurity system for managing business and the economy as a whole, which is the purpose of our study. In this aspect, we will initially consider the essential characteristics of AI and cybersecurity in economic theory.

The introduction of sophisticated AI technologies can help organizations reduce complexity and save time and effort on day-to-day security activities in order to increase efficiency and effectiveness in day-to-day operations, as well as better prepare for emerging threats. Such sophisticated technologies are both weapons against current threats and long-term investments if they are acquired and work correctly. AI technology is becoming increasingly available, so soon there will be no excuse for any organization to delay the introduction of AI to prevent cyber-attacks. It is preferable to be one step ahead and start developing a reliable custom AI model adapted to the needs of your organization, and use these technologies before it is too late. As a result, many organizations have begun to expand the scope of information security functions in various ways [3].

Predicting cyber-attacks before they occur is an important pillar that allows organizations to study the threat environment and see if they are already under attack, identify vulnerabilities and ensure early detection of attacks. This is a data-driven approach that is very important. Bringing cybersecurity experts into an organization can also enable organizations to conduct proper assessments to develop thoughtful ideas about their security status, helping to create risk-based thinking and decision-making. AI is gaining momentum much faster than other new technologies. Various government bodies notice this AI revolution and create initial strategies and guidelines for AI regulation [4]. This is important because AI systems, unlike other technologies, are used to simulate human behavior and decision-making. Therefore, audit specialists and IT experts should actively participate in improving the skill set to identify and reduce the risks associated with AI systems. AI is at the top of the list when determining fraud protection. The advantage of AI in detecting fraud is that the device does not doubt its superiority [5].

Although the use of new technologies can really complement the impressive and holistic protection, it is still in the early stages. AI should be used as a support, not something that can be heavily relied upon. Unfortunately, with all the hype around AI, companies may just be trying to advertise its use without fully understanding how to use it effectively. After all, AI is only as good as the data it is trained on [6]. Many organizations will have to select and identify data to train their solution, helping it identify examples of clean data and what contains malware. If it's done incorrectly, you can easily be lulled into a false sense of security when your program doesn't raise any flags to view, but in fact it may just skip them. In this context, organizations need to skillfully implement artificial intelligence and use cybersecurity equipment [7].

2 Theoretical Analysis

Cybersecurity is a set of technologies and processes designed to protect computers, networks, programs and data from attack, damage or unauthorized access. In recent days, cybersecurity has been undergoing massive shifts in the context of computing and data science in the aspect of the development of artificial intelligence (AI) [8]. This is vital because governments, corporations and military institutions maintain and store a lot of facts and how they work. "Proceeding from the transition to a new digital structure, we found that effective strategic maneuvering in the business activity of the economy and overcoming its current turbulent state requires large-scale use of digital technologies as a component of modern integration processes in the international scientific space in order to minimize problems and strengthen the economic potential of the country" [9]. Cybersecurity includes reducing the risk of cyber-attacks [10]. Cyber risks must be actively managed with the help of controls. Along with firewalls, cybersecurity technologies are widely used. "Artificial intelligence based on big data and the way to create customer knowledge are of great importance." There are many interdisciplinary intersections between artificial intelligence and cybersecurity [11]. Artificial intelligence mechanisms (such as expert systems, computational intelligence, neural networks, intelligent providers, artificial immune systems, system domain, intelligent information analysis, template reputation, ambiguous judgment, heuristics, etc.) allow security specialists to learn about the cyber environment, counteract prejudice. The use of AI can help expand the horizons of existing cybersecurity solutions.

Cybersecurity is important because government, military, corporate, financial and medical organizations collect, process and store unprecedented amounts of data on computers and other devices. A significant part of this data may be confidential information, whether it is intellectual property, financial data, personal information or other types of data for which unauthorized access or disclosure may have negative consequences [12]. Organizations transmit confidential data over networks and to other devices in the course of doing business, and cybersecurity describes a discipline dedicated to protecting this information and the systems used to process or store it [13]. As the volume and complexity of cyber-attacks increases, companies and organizations, especially those tasked with protecting information related to national security, health or financial records need to take steps to protect their confidential business and personnel information. Experts have warned that cyberattacks and digital espionage are the main threat to national security, eclipsing even terrorism.

To understand the essence of cybersecurity, it is important to consider 5 types of types of cybersecurity that will help reduce cyber-attacks among enterprises and organizations.

- Cybersecurity of critical infrastructure.
- Network security.
- Cloud security.
- Internet of Things security.
- Application security.

With the introduction of AI into business management, the only thing that is changing just as quickly is the sphere of cyber threats. Cyber threats continue to grow in volume and complexity due to the expansion of data presence in the enterprise. Now the data is called new gold, new air, new oil. But whichever metaphor you prefer, the reality is that the need to use data is becoming increasingly important in all areas of business. This is one of the main reasons why cybersecurity groups should not view themselves as the sole executor of rational security practices, but rather disseminate security information and adopt clear programs with their colleagues as a constant, sustained focus. To achieve cyber resilience, businesses must assess which risks are their greatest cyber risks—those that are most likely to cause irreparable damage or destroy that particular business—and then focus resources on a roadmap to mitigate those risks and overcome them if problems arise. No single enterprise can protect against every risk, so it makes sense to prioritize efforts and increase cyber maturity in the areas most suitable for this unique business. It is necessary to carefully study modern types of cyberbullying [14]. Continuous security check will help the company:

- increase your cyber resilience through frequent testing;
- check the effectiveness of their security management tools and tools to prevent specific attack vectors;
- develop a cyber-threat model to focus on high-risk areas and key information assets;
- conduct a methodical analysis of identified safety observations.

Practice shows that creating a stronger cybersecurity culture increases the profitability of an organization, and this will only become more targeted as organizations increasingly use digital business models. Historically, businesses tend to be more attentive to positioning themselves to sell products and increase revenue than to protecting themselves and their customers from security threats. But as the new decade approaches—the 2030s—the pace at which businesses will rebuild to thrive in a technology—driven digital economy will only accelerate. Technologies such as artificial intelligence/machine learning, robotics and the continued proliferation of connected devices will create new business opportunities that will lead to new methods of product development and bringing products to market. Anything less than deep-rooted cybersecurity throughout the enterprise will not work in the future. By integrating robust cybersecurity practices in all areas of the organization, introducing new security capabilities, enterprises will be able to rethink their business models while maintaining a stable foundation for innovation [15].

3 Results

3.1 A New Approach to Business Cybersecurity Management

Cyber security threats are becoming increasingly sophisticated, sophisticated, malicious, well-organized, personalized and highly effective cyber attacks. To cope with the complexity and sophistication of such attacks, it is necessary to create an authorized Cybersecurity Control Center. This is a department in an organization that employs cybersecurity specialists responsible for monitoring and investigating security events in real time to prevent, detect and respond to cyber threats using a combination of people, processes and technologies. The Cybersecurity Management Center, as a specialized group, is crucial for all types and sizes of organizations in the conditions of the modern digital economy, the essence of which is discussed in detail in the book “Fundamentals of the Digital Economy” [16]. Over time, security analysts have to switch to monitoring the Cybersecurity Control Center in order to quickly respond to incidents. Employees raise the alarm when suspicious or abnormal events occur in the cybersecurity system and react quickly to reduce the impact on the organization. Due to the adverse impact of security incidents, organizations are looking for ways to reduce vulnerability and ensure the security of their assets and data. Understanding the evolution and building a successful and efficient Cybersecurity Control Center can significantly improve the ability to detect and prevent cyber attacks, protecting the organization from cyber-malware. The best starting point is to consider a balanced strategy for implementing AI technologies that meets business goals. Organizations can benefit using minimal resources and time with the help of people, processes and advanced next-generation technologies. Creating an effective management system requires an understanding of the organization’s capabilities, as well as its limitations. Cybersecurity combines the strengths of artificial and human intelligence. Cognitive computing offers an advanced type of AI using various forms of AI, including machine learning algorithms and deep learning networks that become stronger and smarter over time [17]. All this helps to gain an advantage in assessing the reduction of cyber risks and focus efforts on critical cybersecurity issues.

3.2 Strategies for Building Effective Cybersecurity

Without a strategy for building a cybersecurity implementation program, separate decisions are often made that contradict each other, which ultimately often worsens the financial and competitive position of the firm. The nature of AI should be aimed at ensuring that the results correspond to business goals. If there is no such AI implementation strategy, then it would be good to develop it. It is necessary to take into account three key elements of the AI implementation strategy: the scope of the business; the goal that the business strives for; the advantage that makes the business unique. The best starting point is to consider a balanced strategy for implementing AI technologies that meet business goals. Organizations can benefit from OAC using minimal resources and time with the help of people, processes and advanced next-generation technologies [18]. Thus, the creation of an effective management system requires an understanding of the needs of the organization, as well as its limitations.

3.3 Pros and Cons of Establishing Effective Security Policy Methods

Like any business initiative, launching a cybersecurity awareness business has its advantages and disadvantages. The plus is that every employee should be instructed about the company's data usage policy, cyberattack prevention strategies and how to detect fraud in order to reduce the number of staff errors. However, awareness of the benefits of implementing effective security policy methods does not always lead to changes in the behavior of company employees—at least not immediately. This is a disadvantage for developing effective security policy methods. Since old habits are difficult to get rid of, real changes require constant training, practical training, setting measurable goals and rewards for achieving them. Some of the most serious obstacles to change can be completely prevented. To do this, it is necessary to create an information organization that will be for correct understanding and free from confusing technical language [9]. Next, it is necessary to develop a policy that is mandatory. And training should be frequent and constant, not a one-time event. Here's how to create a cybersecurity culture. But it doesn't happen overnight. Ultimately, it was found that the money spent on the awareness-raising business would undoubtedly be much less than what would have to be spent on mitigating the consequences of a major data leak and recovering from it. The organization's information departments are also easily scaled, so you can start small and expand the program according to the business budget. You can start by putting up posters with tips and best practices throughout the office, and then move on to sending out newsletters by email, daily lists of tips and online training courses. Everyone should understand that the organization's employees need to convey information about the pros and cons. Despite the fact that the creation and implementation of an effective organization requires considerable time and resources to create program materials, set goals, organize and conduct trainings and measure progress.

3.4 Preparing for the Era of Security Based on Artificial Intelligence

Network activity continues to grow, and almost all important information is stored in the cloud. This reality means that cyber threats are becoming more frequent, and organizations need to be prepared for faster attacks on the integrity of their system. The answer to this question is artificial intelligence, the pace of implementation of which has been growing rapidly for many years, as well as proven experience in improving security and ensuring cost savings in the long term. However, it is important to keep in mind that cybercriminals also use artificial intelligence to infiltrate systems. Thus, organizations should not lag behind in this arms race and consider using artificial intelligence in their cybersecurity efforts to protect their network from malicious threats.

3.5 AI and the Future in Cybersecurity: What Will It Be Like?

There have been significant advances in AI development in recent years, and even more improvements that are significant are possible in the coming decades. In this regard, the communities of technologists, scientists and policy makers should actively cooperate in creating a safer and globally profitable AI, studying the short- and long-term consequences for cybersecurity and AI management, as well as the potential of AI to mitigate

environmental and biological risks [19]. AI allows companies to reduce business costs, increase productivity and make last-minute purchases. However, many agencies do not want to integrate and use AI mainly due to lack of knowledge and/or resources. The availability of data that is “available for exploration” in the research community is a prerequisite for the successful development of AI. The future in AI and cognitive computing attracts the economies of all countries, especially those who want to become a world leader. AI is transforming the economy and industries. To help countries get on the positive side of using AI, companies need to invest heavily in AI technology. Organizations spend a huge budget to ensure the security of their business. Since the industry trend has shifted towards intelligent Internet companies, cybersecurity has become a problem for the researcher.

4 Conclusion

1. Due to the fact that cybercriminals and intruders are carrying out increasingly sophisticated attacks to steal confidential data and undermine businesses, which continue to develop and multiply and are faced by cybersecurity specialists, it is extremely important to monitor emerging threats and adapt new ways to eliminate them, including the development of the OAC, which is crucial for all types and sizes of organizations in the modern digital economy, since most of the organization’s operations and confidential data are on the network and in the cloud.
2. The AI strategy should be well coordinated with the concepts of each business function, including data selection, determining the relationship between problem solving and AI technology, and fine-tuning the AI and cybersecurity strategy system, aiming to create a strong AI.
3. It is useful to think about a strategy for the development of artificial intelligence and cybersecurity for business management whenever a competitor or customer preferences appear on the market in order to provide the necessary preventive methods to protect data, networks, electronic devices and servers from malicious attacks and unauthorized access, or so-called cybersecurity elements that include application security, identity management, network security, data security, end user training, disaster recovery and business continuity.

References

1. Artificial Intelligence//WIPO Technology Trends. World Intellectual Property Organization. Geneva (2019)
2. Gupta, B.B., Tvari, A., Jainak, S., Agrawal, D.P.: The fight against phishing attacks: the current state and future challenges. *Neural Comput. Appl.* **28**(12), 3629–54 (2017)
3. IBM Opens Threat Intelligence to Combat Cyber Attacks. <https://www.ibm.com/news/ca/en/html>. Accessed 21 Oct 2012
4. Artificial Intelligence. Merriam-Webster. <https://www.merriam-webster.com/dictionary/artificial%20intelligence>. Accessed 11 Nov 2022
5. (ISO)/International Electrotechnical Commission (IEC), ISO/IEC 27001:(2005) Information technology—Security techniques—Information security management systems—Requirements, Switzerland. <https://www.iso.org/standard/42103.html> (2005)

6. Mohammadi, S., Mirvaziri, H., Gazizade-Ahsai, M., Karimipur, H.: Detection of cyber intrusions using the combined component of the selection algorithm **44**, 80–88 (2019)
7. Craigen, D., Diakun-Thibault, N., Purge, R.: Definition of cybersecurity. *Technol. Innov. Manag. Rev.* **4**(10), 13–21 (2014)
8. Lee, I.: Cybersecurity: risk management framework and investment cost analysis. *Bus. Horiz.* **64**(5), 659–671 (2021)
9. Kukier K.: *Data, Data Everywhere: A Special Report on Information Management* (2010)
10. Alice, A.M., Zaidan, B.B., Zaidan, A.A., Sahar, N.M.: An overview of intrusion detection systems based on deep learning methods: consistent taxonomy, challenges, motivations, recommendations, meaningful analysis, and future directions. *Neural Comput. Appl.* **32**(14), 9827–9858 (2020)
11. Nosova, S.S., Norkina, A.N., Morozov, N.V.: *Artificial Intelligence and Economics. Bachelor's degree, Textbook for universities. M.: KNORUS*, p. 400 (2023)
12. McCallister, E.; Grance, T.; Scarfone, K.: *Guide to protecting the confidentiality of personally identifiable information (PII), Special Publication (SP) 800-122, National Institute of Standards and Technology (NIST), USA.* <https://csrc.nist.gov/publications/detail/sp/800-122/final> (2010)
13. Al-Garadi, M., Mohamed, A., Al-Ali, A., Du, X., Guizani, M.: A survey of machine and deep learning methods for internet of things (IoT) security. *IEEE Commun. Surv. Tutor.* **22**, 1646–1685 (2020)
14. Nosova S.S., Norkina, A.N., Morozov, N.V.: *Typologies of Financial Frauds Bachelor's Degree, Specialty. Textbook for Universities. M.: KNORUS*, p. 476 (2021)
15. Nosova, S.S., Norkina, A.N., Morozov, N.V.: *Artificial intelligence and the future of the modern economy. Innov. Invest.* **1**, 229–234 (2023)
16. Nosova, S.S., Putilov, A.V, Norkina, A.N.: *Fundamentals of the Digital Economy. Bachelor's Degree, Textbook For Universities. M.: KNORUS*, p. 392 (2022)
17. Forugi F., Luksh.: *Data science methodology for cybersecurity projects.* preprint arXiv :1803.04219 (2018)
18. International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC), ISO/IEC 2700, Switzerland. <https://www.iso.org/standard/73906.html> (2018)
19. Django-Jacquard, J., Nepal, S.: Overview of emerging threats in the field of cybersecurity. *J. Comput. Syst. Sci.* **80**(5), 973–993 (2014)



Integration of Artificial Intelligence into Business Management Strategy

Svetlana Nosova¹  , Anna Norkina¹ , Nikolay Morozov¹ , Irina Arakelova² ,
and Galina Fadeicheva³ 

¹ National Research Nuclear University “MEPHI”, Kashirskoe Shosse 31, 115409 Moscow, Russian Federation

SSNosova@mephi.ru

² Volgograd State Medical University, 1 Pavlov Bortsov Square, 400131 Volgograd, Russian Federation

³ Academy of Labor and Social Relations, Lobachevsky 90, 119454 Moscow, Russian Federation

Abstract. The article analyzes the role of the integration of artificial intelligence (AI) into a business management efficiency strategy based on stable connections of network partners who carry out joint actions in the mode of their activity programs based on solving issues such as the pros and cons of AI, business functions and AI, as well as the development of basic ways to overcome obstacles on the way the use of AI technologies, in particular, the elimination of unproductive links between science and business, the lack of personnel capable of developing AI devices. Achieving this goal requires analyzing the changes that AI has caused in the business management strategy in order to help market economy entities effectively and intelligently use its developments in terms of increasing the competitiveness of business processes. The results of the study confirmed that by combining human intelligence with a more powerful AI, it is possible to better understand the integration of AI into a business management strategy in terms of obtaining significant advantages and opportunities for business from automation to the adaptation of products and services to consumers using algorithms, data analysis in order to improve the efficiency of AI development from the position of creating a new strategy model business management.

Keywords: Artificial intelligence (AI) · Strategy · Technology · Business management · Neural network · Cybersecurity

1 Introduction

Research in the field of integrating AI into business management strategy is gradually gaining real value for the economy as a whole. Important developments such as improved algorithms, mass availability of data and more powerful hardware have allowed AI to go beyond human cognitive abilities in the field of visual recognition and natural language processing. The development of AI technologies has allowed modern enterprises to work faster than ever, because AI can manage resources more efficiently. However,

most organizations have not yet realized the potential benefits of using AI. Many companies and governments are currently exploring aspects of doing business and economic management in order to maximize the benefits of implementing AI that they can extract in our digital age. In this regard, business leaders need to develop strategies that will allow them to take advantage of the benefits arising from the use of AI technologies. According to business management experts, enterprises can effectively use AI in terms of reducing costs and increasing the productivity of business processes. Enterprises seek to capitalize on their capabilities by applying the latest advances in intelligent manufacturing or by introducing “Smart Manufacturing” instead of traditional manufacturing. Intelligent manufacturing focuses on the use and integration of intelligent equipment into the production environment. Such equipment, which is supplied using AI, brings a significant effect to the enterprise, as it saves time and reduces prices and costs for manufacturing products. Practice shows that AI is used at any stage of the business process, including front, middle and back offices. It can be used in any business, regardless of industry and any location. To help countries get positive results from the use of AI, companies must invest heavily in the development of its technologies. However, the final effect may be different, i.e. there may be winners and losers. Thus, AI and strategic business management reinforce each other. They have changed how we live, how we work, how we communicate and meet, as well as how countries develop. As with electricity, AI has changed the global landscape, covered vast distances and made the world flatter, providing instant access to an endless stream of information. This has influenced the strategic management of both business and the economic development of the country and the world economy, increasing the level and quality of life of the population.

2 Methods of Integrating AI Into the Strategy of Increasing the Efficiency of Business Management

Practice shows that digitization and certain fundamental methods of studying the impact of integrating AI into a strategy to improve the efficiency of business management are crucial for creating value using AI at the intended scale. Here are a few ways that companies can take to capitalize on the potential of AI:

- It is necessary to move forward in the process of digitalization of business processes. The results confirm that digitization of business processes is a necessary condition and a critical factor for extracting value from AI. The consequences of the ongoing digitization are significant. For many companies, they include changes at the level of transformation of the business processes underlying the enterprise itself, and new ways of working people. But without a solid digital foundation, the company’s AI systems will lack the training data needed to create better models and the ability to transform superior AI ideas into behavioral changes at scale.
- It is necessary to strive to increase the influence of AI on the enterprise. Although most companies have already implemented AI to some extent, few have implemented it into standard operational processes in several business units or functions, and the rest are only testing the use of AI. Getting stuck with the use of AI in a company is a real risk. Therefore, it requires not only the dissemination of these capabilities throughout the enterprise, but also a real understanding on the part of managers in carrying out large-scale changes, as well as focusing on change management, and not only on technology.

- It is necessary to learn how to implement the key underlying factors necessary to benefit from AI at scale. These factors include sponsorship by senior management, the development of a corporate—wide portfolio of AI capabilities, actions to address talent gaps, and the implementation of a complex data strategy—all of which require a more strategic understanding of AI programs and objectives. Business and technology leaders need to work quickly to implement key artificial intelligence tools. Otherwise, they risk missing out on the current and future opportunity to use AI.

AI has infiltrated many organizational processes, leading to a growing fear that smart machines will begin to displace employees from their jobs in the decision-making process. In order to understand the active and pragmatic perspective of AI, it is important to substantiate the complementarity of humans and AI and consider how each can bring their own strengths into organizational decision-making processes characterized by uncertainty, complexity and ambiguity. With more computing power of information processing and an analytical approach, AI can expand the possibilities of human cognition in solving complex problems, while humans can still offer a more holistic, intuitive approach to solving problems of uncertainty and ambiguity in organizational decision-making. This premise reflects the idea of “increasing intelligence”: AI systems should be designed with the intention of increasing, not replacing, human input. We consider AI in the aspect of applying technologies to perform tasks that resemble human cognitive functions and are usually defined as the «the capability of a machine to simulate intelligent human behavior» [1]. Calling specific applications “artificial intelligence” is like calling a car a “vehicle”—technically it is correct, but it does not cover any of the functions. To understand what type of AI prevails in business, we need to dig deeper. AI, or the idea that computer systems can perform functions normally associated with the human mind, has suddenly turned from a futuristic speculation into a modern reality.

3 Theoretical Analysis

3.1 Definition

AI can be defined in various ways. Thus, the Merriam-Webster Dictionary defines AI as “a branch of computer science that is engaged in writing computer programs capable of creatively solving problems”. «Artificial intelligence experts hope to mimic or duplicate intelligence in computers and robots. There are other definitions: Artificial intelligence is a broad term that refers to any type of computer software involved in human-like activities, including learning, planning, and problem solving» [3]. This formulation of the question gives the right to consider AI as a new participant in cooperation, which becomes a catalyst for broad structural transformations, since economies using AI not only do something differently, but will also do other things [2]. Undoubtedly, AI is a new resource in the global market system of the twenty-first century, the value of which is only increasing every day. Based on this understanding of AI, we emphasize that the largest technology companies are already participating in a competition in the field of AI technologies and applications for strategic management of their business processes. The availability of data that is available for exploration in the research community is a

prerequisite for the successful development of AI. After deciding what needs improvement—their operational flexibility, speed or scalability; their ability to make decisions; or their ability to personalize products and services—they can develop appropriate solutions [19]. To get the most out of artificial intelligence, companies need to restructure their business processes.

3.2 AI and Organizations: Current Practices and Challenges

Many large multinational firms claim the significant potential of artificial intelligence technologies. At the same time, AI is still at an early stage of commercialization, and only 8% of firms today observe widespread adoption of AI in organizations. A limited number of empirical papers have been published on problems specifically related to AI and innovation management. The key AI implementation pattern symbolizes the separation between additional use cases that optimize existing business processes and products, as well as an integrated AI use case that changes the organization, its products, and sometimes the market [11]. Additional artificial intelligence is being implemented into existing business processes and products through projects in non-critical areas that are relatively independent of other parts. It focuses primarily on optimizing existing processes, risk management and short-term return on investment to ensure gradual innovation of existing business. On the contrary, the integrated AI considers the main area of the company's activity and becomes deeply integrated with the overall organizational goal and strategy. It is focused on the long-term perspective and is focused on a strategy focused on the broader ecosystem of the company with the aim of creating value in the broader market [12]. The latter type of AI lays the foundation for transformational or radical innovations.

4 Results

4.1 Strategy for the Introduction of Artificial Intelligence as a Driving Force for the Growth of Business Management Efficiency

The AI implementation strategy means certain advantages in the market. Such a strategy will help you find the keys to success and set the direction for achieving goals. It can also help to expand the production of products/services. In order to properly achieve rapid and profitable growth, leaders are needed. As a rule, they can challenge the prevailing point of view without provoking indignation or cynicism; they can change course if the path they have chosen turns out to be wrong; they are looking for ways to achieve their goals in terms of improving competitive advantages.

Practice shows that so far only a small number of companies have begun to rethink their business processes to optimize joint analytics. Those organizations that use machines only to displace workers with automation will miss the full potential of AI. Modern AI has achieved many superhuman abilities in narrow directions, but the social compatibility of artificial agents with humans is currently at a low level. Artificial intelligence should be able to establish a relationship of mutual understanding and trust with a person [17]. In other words, the interaction of a machine and a person is necessary (see Fig. 1).



Fig. 1. Illustrative diagram of the interaction of a machine and a person.

The conclusion is obvious: the use of AI in production provides a qualitatively new level of business processes and the economy as a whole. Artificial intelligence has a huge potential to contribute to global economic activity. In this regard, it will be necessary to manage the growing gap between countries, companies and employees in order to maximize benefits [4]. In addition to companies, employees, lawyers and society, educational systems and legislators are also facing the task of meeting the new challenges that result from constantly advancing technology. The introduction of artificial intelligence is of great importance for security, as organizations create layers of protection to ensure the safety of the company's most valuable assets. It is necessary to develop an IT strategy that would optimize and promote the growth of investments in security-related technologies. It should be noted that asset protection should be proportionally commensurate with the value of the asset in terms of its criticality and sensitivity [16].

4.2 Assimilation of AI Into Strategic Business Management

Assimilation of AI into strategic business management can lead to improved outcomes such as organizational flexibility, customer flexibility, and firm productivity growth.

Computer scientists have made significant breakthroughs in machine learning and deep learning, giving machines cognitive and predictive capabilities. Today, these systems are already being implemented in real situations. One of the reasons for the growing role of AI is the formation of huge opportunities for economic development. The project undertaken by PricewaterhouseCoopers has shown that artificial intelligence technologies can significantly increase global GDP [10]. It is important to emphasize that AI is currently one of the most important technologies that transform business management strategies, macro and mesoeconomics and contribute to the global digital transformation of all human life.

Significant advances have been made in AI development in recent years, and even more significant improvements are possible in the coming decades. In this regard, the communities of technologists, scientists and politicians should actively cooperate in creating a safer and globally profitable AI, studying the short—and long-term consequences for the security and strategic management of business, increasing its potential to reduce environmental and biological risks. AI allows companies to reduce business costs, increase productivity and make last-minute purchases. However, many agencies do not want to integrate and use AI mainly due to lack of knowledge and/or resources. Thus, it is necessary to identify opportunities associated with the integration of AI into business management strategies with an emphasis on cybersecurity issues [6]. The most exciting

developments in the field of artificial intelligence are advances in deep learning methods. Given the significant computational demands of deep learning, some organizations will maintain their own data centers due to regulations or security concerns [14]. Many studies have been conducted on the application of artificial intelligence related to business theory. As a rule, various aspects of its application in various industries are described in order to better illustrate the impact of artificial intelligence on business processes [8]. The potential commercial benefit from the development of artificial intelligence is a powerful incentive for specific applications. Policies that encourage transparency and the sharing of basic data sets between both public and private actors can contribute to a higher level of innovation-based competition and a higher level of research productivity in the future.

4.3 AI As the Most Important Technology Transforming Business Management Strategies

Artificial intelligence will have a fundamental impact on the global labor market in the next few years. Therefore, it is important to discuss legal, economic and business issues, such as changes in the future labor market and company structures, the impact on working hours, wages and the working environment, new forms of employment and the impact on labor relations [9]. Artificial intelligence has the potential to change the innovation process itself. From the point of view of the innovation economy, there is an important difference between the narrow scope of innovations, for example, such as robots (specially designed to solve narrow tasks), and their almost limitless scope of application. Artificial intelligence entails many related technologies, such as machine learning, deep learning, neural networks, natural language processing machines and other technologies, which often take the form of major inventions and can potentially significantly increase the competitiveness of business processes and the economy as a whole. The problem associated with advances in artificial intelligence is that they are research tools and have a powerful impact on the volume of innovations being introduced and their nature [15]. The AI strategy should be well coordinated between AI technology and the concepts of each business function, including data selection, determining the relationship between task concepts and technologies, and fine-tuning the AI system. Artificial intelligence technologies should be introduced into certain areas of activity with careful study, since automated solutions can destroy the reputation of a company if ethical and regulatory requirements do not work properly [7]. Organizations spend huge budgets to ensure the security of their business. Since the industry trend has shifted towards intelligent internet companies, cyber threat has become a problem for the researcher [5, 11]. The final effect of using artificial intelligence may be different, i.e. there may be winners and losers.

4.4 Artificial Intelligence as a Tool for the Growth of Competitive Advantage

Artificial intelligence (AI) is located within the scientific field of computer science, which covers almost all areas and is becoming more and more popular every day. In business, with the help of AI, you can strengthen your company until we achieve a competitive advantage. Competitive advantage is the ability of a company to produce any product

or service more efficiently than its competitors. In fact, this is a simple task that can be solved with the help of artificial intelligence software, since it can offer automated services with a non-existent margin of error. AI provides successful results: demand and company profits are growing [13]. The more stable the competitive advantage, the higher the competence of the company. The conclusion is obvious: using artificial intelligence, the company provides 100% competitive advantage. Currently, scientists are increasingly promoting the concept of automation with the support of artificial intelligence in order to basically complement the tasks within this human-machine partnership in strategic management [18]. The prevailing opinion in academic circles continues to be that human managers are best suited for the role of a central processor because of their unique level of sanity and competence in judgments.

5 Conclusion

1. AI can support managers by effectively delegating strategic management decisions to them. The key to understanding the prospects and challenges of AI in strategic management is to focus on the problem of delegating decisions by AI managers, which can lead to improved outcomes such as organizational flexibility, customer flexibility and firm productivity growth.
2. All AI development cooperation agencies should consider how to fully integrate in order to achieve their intended goals. Research shows that AI is not perceived as a quick-impact technology and that the effective use of the technology's potential requires a clear strategy for its implementation with the participation of many stakeholders. In the current state, business process management at the level of industrial practice is limited by existing organizational models and insufficient interdisciplinary cooperation within the framework of various cognitive technologies.
3. Today, AI has entered into the development strategy of many areas, such as space, military, industry, electric power, renewable energy, medicine, engineering, mass media and many other areas. Experts expect that by 2040 AI will play an important role in managing everything, thereby strengthening trust in the processes of human interaction with AI, including the selection of data that determines the growth of competitiveness of business processes as a result of the intelligent use of AI connected to the Internet.



References

1. Artificial Intelligence. Merriam-Webster. <https://www.merriamwebster.com/dictionary/artificial%20intelligence>. Accessed 18 Nov (2022)
2. Artificial Intelligence: WIPO Technology Trends: World Intellectual Property Organization. Geneva (2019)
3. Artificial intelligence for industrial growth for now and for the future, IoT and AI World Summit Eurasia (2019)
4. Bughin, J., Seong, J., Manyika, J., Chui, M., Joshi, R.: Notes from the AI frontier: modeling the impact of AI on the world economy. In: MGI Discussion Paper, Vol. 1. McKinsey Global Institute, p. 56 (2018)

5. Khan, H.U., et al.: Transforming the capabilities of artificial intelligence in GCC financial sector: a systematic literature review. *Wirel. Commun. Mob. Comput.* **12**, 2–17 (2022)
6. Mendhurwar, S., Mishra, R.: Integration of social and IoT technologies: architectural framework for digital transformation and cyber security challenges. *Enterp. Inform. Syst.* **15**(4), 565–584 (2022)
7. Nosova, S., Norkina, A., Makar, S.: The collaborative nature of artificial intelligence as a new trend in economic development. *Stud. Comput. Intell.* **1032**, 367–379 (2022)
8. Nosova, S., Norkina, A., Makar, S.: Artificial intelligence technology as an economic accelerator of business process. *Stud. Comput. Intell.* **1032**, 355–366 (2022)
9. Nosova S.S., Norkina, A.N.: Artificial Intelligence and Economics. Bachelor Course. Textbook for Universities. KNORUS, Moscow, p. 400 (2023)
10. Nosova, S.S., Norkina, A.N., Morozov, N.V.: Artificial intelligence and the future of the modern economy. *Innov. Invest.* **1**, 229–234 (2023)
11. Nosova S.S., Norkina, A.N., Morozov, N.V.: Typologies of Financial Fraud Bachelor's Degree, Specialty. Textbook for Universities. M.: KNORUS, p. 476 (2021)
12. Raimundo, R., Rosário, A.: The impact of artificial intelligence on data system security: a literature review. *Sensors* **21**(21), 7029–7036 (2021)
13. Rana, N.P., Chatterjee, S., Dwivedi, Y.K., Akter, S.: Understanding dark side of artificial intelligence (A.I.) integrated business analytics: assessing firm's operational inefficiency and competitiveness. *Eur. J. Inform. Syst.* **12**, 1–24 (2021)
14. Razzaque, A.: Artificial intelligence and I.T. governance: a literature review. *Big Data Driven Dig. Econ. Artif. Comput. Intell.* **12**, 85–97 (2021)
15. Sabillon, R., Serra-Ruiz, J., Cavaller, V.: An effective cybersecurity training model to support an organizational awareness program: the cybersecurity awareness training model (RAM). A case study in Canada. In: *Research Anthology on Artificial Intelligence Applications in Security*, pp. 174–188. IGI Global (2021)
16. Samsonovich, A.V.: Science on the Verge of Creating an “Emotional” Computer, mephi.ru (2016)
17. Taeihagh, A.: Governance of artificial intelligence. *Policy Soc.* **40**(2), 137–157 (2021)
18. Wilson, H.J., Daugherty, P.R.: Collaborative intelligence: humans and AI are joining forces. *Harvard Bus. Rev.* **96**, 114–123 (2018)
19. Wisskirchen, G., von Brauchitsch, B.: Artificial intelligence and robotics and their impact on the workplace. *IBA Glob. Employ. Instit.* **11**(5), 49–67 (2017)



The Applicability of Artificial Intelligence in the Modern Global Development of Countries and Companies

Svetlana Nosova¹ , Anna Norkina¹ , Nikolay Morozov¹ ,
Olga Medvedeva² , Irina Arakelova³ , and Sergey Bondarev⁴ 

- ¹ National Research Nuclear University “MEPHI”, Kashirskoe Shosse 31, 115409 Moscow, Russian Federation
SSNosova@mephi.ru
- ² State University of Management, Ryazansky Prospekt 99, 109542 Moscow, Russian Federation
- ³ Volgograd State Medical University, 1 Pavlov Bortsov Square, 400131 Volgograd, Russian Federation
- ⁴ Plekhanov Russian University of Economics, Stremyanny Lane 36, 117997 Moscow, Russian Federation

Abstract. Artificial intelligence has a great potential of anti-turbulent conditions for the global effective activity of countries and companies in terms of maximizing benefits and satisfaction of the population with the quality of life. The purpose of the article is to analyze the conceptual foundations of the role of the use of artificial intelligence technologies in modern economic development on the basis of research on issues such as understanding the behavior of countries, companies and industry sectors in developing views on how to introduce and absorb artificial intelligence technologies, to keep a careful account of possible failures, which can be considered as economic losses and potentially hinder the growth of economic and social turbulence as a result of the use of artificial intelligence in the global activities of countries and companies. The proposed study can be considered as the impact of artificial intelligence on the development of countries and companies based on the best knowledge available at this stage of economic development in order to give a more global view of business functions, as well as the development of basic ways to overcome obstacles to the use of artificial intelligence technologies, in particular, the elimination of unproductive links between science and business, lack of personnel capable of developing artificial intelligence technologies. As a result, it is proved that the accelerated use of artificial intelligence, coupled with the robotization of business processes with all their pros and cons, will allow countries and companies to implement a strategy for the development of consensus artificial intelligence of partner countries and international associations, as well as a number of recommendations for the most effective use of artificial intelligence while preserving important human values.

Keywords: Artificial intelligence (AI) · Business functions · Turbulence · Robotics · Cybersecurity

1 Introduction

The use of artificial intelligence can boost economic activity and at the same time increase the gap between countries and companies, given that accidents caused by powerful artificial intelligence systems can be extremely destructive and, accordingly, the benefits from its use may be uneven. The size of the benefits for those companies that switch to artificial intelligence technologies early will accumulate in subsequent years at the expense of companies with limited or zero implementation [1]. In this case, the key problem is that more advanced and powerful artificial intelligence systems can be transformative with both positive and negative consequences. Artificial intelligence will require collaboration and government involvement to reduce risks and achieve global cybersecurity benefits. In this regard, it is necessary to make serious efforts, think about laying the foundations for the security of future systems, and better understand the consequences of the use of artificial intelligence, “especially for those countries and companies who want to become a world leader” ([2], 276). In order to help countries get on the positive side of the applicability of artificial intelligence, companies must invest huge investments in artificial intelligence technology. However, given that the net effect may be different, i.e. there may be winners and losers, hence the degree of applicability of artificial intelligence and its consequences is difficult to assess, because sometimes there is not enough experience to fully understand what the real benefits of using it are. The World Intellectual Property Organization notes a sharp increase in the number of scientific papers in this area and an equally sharp increase in the number of patents, which indicates the transition from theoretical research to the practical use of artificial intelligence technologies in two areas of importance and public interest: security and employment. It is necessary to delve into the near and medium-term trends and consequences of the spread of artificial intelligence in these areas. “It is important to identify the potential for significant disruptions due to the proliferation of artificial intelligence on cybersecurity, justice (criminal and civil) and labor market models” [3]. The discussion of the future of the sphere of labor and capital represents a new basis for thinking about the susceptibility of automation and ends with the definition of recommendations for increasing the productivity of production processes based on the highlighted trends in the applicability of artificial intelligence [4].

2 Materials: Applicability of Artificial Intelligence AI Productivity Growth

In the modern economic world, the applicability of artificial intelligence technologies is beginning to spread in various industries and business functions, including management, customer service, etc. The growing ability of artificial intelligence to predict for many years has been parallel with investments in artificial intelligence startups. This is acceptable. So, the leaders of the use of artificial intelligence (these are mainly leading countries) can increase their advantage over developing countries. By the way, it should be said that since March 2023, China has not been a developing country. It is important to note that leading countries can get additional benefits from the use of artificial intelligence compared to developing countries. They may have no choice but to push

the use of artificial intelligence in order to increase productivity. Productivity growth is crucial for raising wages and living standards, and also helps to increase consumers' purchasing power for goods/services ([5], 66). Moreover, wage rates are already high in these countries, which means that there are more incentives to replace labor with machines than in developing countries with low wages. Thus, the use of artificial intelligence in developed economies contributes to productivity growth due to an increase in demand for goods/services. Demand can stimulate productivity growth not only during the recovery of the country's economy and companies after the crisis, but also in terms of possible long-term structural losses. Gradually, a wide range of artificial intelligence applications acquires real value for business when three main developments are applied:

- Best algorithms,
- Mass availability of data,
- More substantial hardware.

In contrast to earlier research in this area, current research attributes to artificial intelligence a high degree of usefulness and greater autonomy in the field of "mental work" within the framework of cognitive tasks and process automation. This allowed artificial intelligence to go beyond human cognitive abilities. It is no coincidence that Davenport and Ronanki [6] conducted a study of 152 projects to assess the impact of the use of artificial intelligence on business, dividing the effect of artificial intelligence into three parts:

- Automation in processes,
- Cognitive understanding,
- Cognitive engagement, which is associated with automated communication with customers and employees.

It is no coincidence that one of the possible definitions of AI refers to cognitive processes and especially to reasoning. Before making any decision, people reason, so they naturally explore the connections between AI and decision-making. In this direction, a distinction is made between two aspects of decision-making: diagnosis and looking into the future. It is shown that, on the one hand, AI has many connections with diagnostics (expert systems, case reasoning, fuzzy set theories). On the other hand, AI does not pay enough attention to reasoning [17]. Currently, there are no uniform standards in terms of data access, data sharing, or data protection. Almost all data is proprietary by nature and is not very widely disseminated to the research community, and it is shared with the research community, and this limits innovation and systems design. AI requires data to test and improve its learning ability. Without structured and unstructured datasets, it will be nearly impossible to get the full benefits of artificial intelligence.

3 Results

3.1 The Impact of AI Application on the Country's Economy

AI has the potential to change macroeconomic activity. "AI will affect almost all areas of our economy and society, including the regions. as diverse as financial planning, policing, elections, manufacturing and transportation" [20]. From the point of view of

macroeconomics, we are talking about the economy of innovation, when there is an important difference between a narrow field of application of innovation, for example, such as robots (specially designed for narrow tasks), and with an almost limitless field of their application, such as neural networks, often called “deep learning”. Namely, deep learning opens up the prospect of changes in the very nature of economic development. Hence, developments in the field of artificial intelligence are not just examples of new technologies, but “general-purpose technologies” that can be the driving forces of long-term scientific and technological progress. From the perspective of a market economy, “artificial intelligence is a rapidly growing market. He is ready to have a transformative impact on consumer, corporate and government markets around the world.” For example, China has set a national goal to invest \$150 billion in AI and become a world leader in this field by 2030 [7]. Thus, if we talk about comparing the key technological trajectories in the framework of AI—robotics and deep learning, they play completely different roles in the future technological development of each country. Thus, deep learning is an area of research with a high degree of versatility and can change the production process itself as a result of the introduction and dissemination of general-purpose technology. Such technologies often take the form of basic inventions and have the potential to significantly improve the productivity or quality of goods or services produced. Andrew Burt argues: The key problem facing predictive analytics is transparency, which depends on how well data scientists can explain what they are doing. Thus, the problem associated with the achievements of transparency in the field of AI has a powerful impact on the implemented volume and nature of innovations. The most cost-effective application of AI was in the field with the large-scale introduction of industrial robots in production applications. These machines are precisely programmed to perform a given task in a strictly controlled environment. Innovations in robotics have had an important impact on manufacturing and automation, primarily through the introduction of more responsive robots that rely on programmed algorithms capable of responding to various stimuli ([8], 253). Continuous innovations in robotic technologies (especially in the ability of robotic devices to perceive and interact with the environment) can lead to wider application and implementation beyond industrial automation. But for now, robots are still used mainly in specialized end-use manufacturing applications. Of course, there are counterexamples to this statement: for example, robotic space probes were a very important research tool in planetary science. So, it is important to emphasize that artificial intelligence is a new general-purpose invention in the form of a “method of invention”, which helps to identify some preliminary consequences of this hypothesis for economic management. “The AI strategy should be well coordinated between AI technology and the concepts of each business function, including data selection, determining the relationship between task concepts and technology, and fine-tuning the AI system” ([9], 366). The use of AI in many industries is growing. It is used to replace people in various fields. For example, in space exploration, advanced manufacturing, transport, energy development and health-care. “Five priorities can form the basis of China’s AI strategy: creating a reliable data ecosystem, stimulating the introduction of AI in traditional industries, strengthening the portfolio of specialized AI talents, ensuring that education and training systems meet these challenges, and establishing ethical and legal consensus among Chinese citizens and in the global community” [10]. Thus, using the exceptional computing power of

computers, people with the help of artificial intelligence can complement their skills and increase the productivity of the company. In order for these achievements to become widespread, greater transparency is needed in how AI systems work. It matters how government agencies look for AI to improve the service of citizens, how policy issues are resolved, ethical conflicts are reconciled, legal realities are resolved, and how much transparency is required in AI and data analysis solutions. This helps to protect consumers and strengthen confidence in the economic system as a whole. As an example of the possibilities, the Chinese search firm Baidu has for the first time used a facial recognition application that finds missing people. In addition, cities such as Shenzhen provide up to \$1 million to support AI laboratories. This country hopes that AI will provide security, fight terrorism and improve speech recognition programs [11].

«In a 2022 IPSOS survey, 78% of Chinese respondents (the highest proportion of countries surveyed) agreed with the statement that products and services that use AI have more advantages than disadvantages. After Chinese respondents, respondents from Saudi Arabia (76%) and India (71%) are the most positive about AI products. Only 35% of Americans surveyed (among the lowest of the countries surveyed) agreed that products and services that use AI have more advantages than disadvantages» [4]. Currently, the United States, China, Russia, North Korea and other countries are investing significant resources in military AI. The US military is deploying AI “to sift through large amounts of data and video captured by surveillance, and then warn human analysts about patterns or when there is abnormal or suspicious activity” [20]. China intends to take advantage of the initiative to become the main global AI innovation center” by 2030, potentially surpassing the United States in this process, the Center for a New American Security said in a recent report [12]. For Russia, the goal of new technologies in this area is to meet the needs of our fighters from military artificial intelligence technologies and increase the speed and flexibility of their development and procurement.

3.2 The Use of AI and the Growth of Efficiency of Companies

In general, companies have formed an idea of the need to use AI technologies as one of the strategic goals of their activities in the future. The companies plan to transform the entire activity of the enterprise. Accenture’s research shows that AI has the potential to increase profitability and lead to \$14 trillion in economic growth in 16 industries in 12 countries by 2035 [13]. It is quite possible that AI technologies can lead to a productivity gap between companies that fully implement artificial intelligence tools in their enterprises within a certain period of time and companies that do not use AI technologies at all or have not fully implemented them in their enterprises, say today. On the one hand, the first companies are likely to benefit disproportionately. They can potentially increase their cash flow. And this is all because they have a strong starting base of AI, a higher propensity to invest in AI and a positive view of the business rationale for using AI. At the other end of the spectrum, companies that do not use AI technologies may experience a decrease in cash flow compared to today’s level, provided that the same cost and revenue model is used as it is today. One of the important reasons for such pressure on profits is the presence of strong competitive dynamics among companies, which can shift market share from laggards to leaders and can provoke an uneven distribution of the advantages of AI. To do this, you need to assemble a team of potential strategic

leaders with a collective task, i.e. create a fully developed solution to the problem or design a new critical potential and a way to generate it. Give them a small budget and a tentative deadline. Then carry out assessments using an in-depth analyse. Reports on the assessment of the economic impact of AI on the way of forming a mechanism for implementing the increase in the efficiency of production processes can help managers determine the timing of investment and the share of the budget for the introduction of AI. We propose to highlight several key steps that are central to the discussion of the impact of the use of AI on the growth of companies' efficiency:

- to help logistics specialists better predict the likelihood of an impact on the supply chain;
- “oblige employees to anticipate the necessary actions and more accurately predict potential problems”;
- interact more deeply with customers, better understanding what they want;
- improve global sourcing and vendor integration, accelerate and improve analysis, provide more efficient automation of recurring procurement tasks, and support more efficient return and replacement operations;
- improve the efficiency of salary and benefits management, as well as workforce planning, increase the speed and accuracy of recruitment, instantly providing a 360-degree overview of a potential candidate through social networks and other channels.

Employees will need help throughout their lives to acquire new skills and develop new job opportunities. Political reforms will be necessary to reduce polarization and restore civility so that an open and healthy debate can be held about where responsibility for economic well-being lies [19]. In order to strengthen the financial position of their organization during a period of economic turbulence, IT managers should not limit themselves to cost savings, but look for new forms of operational excellence, continuing to accelerate digital transformation. Hence the conclusion is obvious: the use of AI in companies provides a qualitatively new level of business processes and the economy as a whole, but today the role AI in national defense is relevant for the whole world and especially for Russia.

As for the growth of AI technologies in general, actions are needed both to overcome barriers to its use by large operating businesses, and to expand the introduction of digital tools by all companies. Actions that can contribute to the spread of AI include: setting an example and digitizing the public sector, using public procurement and investment in R&D, stimulating the introduction of AI technologies by small and medium-sized enterprises, investing in hard and soft digital infrastructure and clusters, commitment to training digital professionals as well as consumers, providing global connectivity and solving privacy and cybersecurity issues [18]. In addition, regulators and policy makers will need to understand the differences in the nature of digital platforms and networks from the network industries of the past and develop tools to identify uncompetitive behavior that can harm consumers. And companies need to develop a strategy for productivity growth as a result of the accelerated application of AI technologies, including robotics, which includes all their sectors and value chains. Gartner predicts that businesses will profit from “superapps” that combine the functions of an application, platform, and ecosystem [12]. So, in order to participate in global economic activities, companies must use various AI technologies. In this case, by 2030:

- average modeling shows that about “70% of companies may adopt at least one type of AI technology, but less than half will fully master all categories;
- the demand for jobs may shift from a low level of digital skills “(from about 40% to 30%), and the largest increase may be among those that require high digital skills, increasing from about “40% to more than 50% [1]. Employers in the United States are increasingly looking for workers with AI-related skills.

4 Conclusion

1. AI is changing our lives every day. Intelligent machines are gradually appearing that make life more convenient and comfortable for everyone. There is no need to be afraid of artificial intelligence. Thanks to its implementation, AI technology can be made the most useful, depending on the user’s needs. Analysis of the applicability of artificial intelligence technologies can help to correctly assess their impact on the development of the economy as a whole and companies.
2. The widespread applicability of AI (and possibly robotics) in many sectors is likely to trigger a race within each sector to create an artificial intelligence of its own advantage that uses these new approaches. The application raises questions for competition policy. In every application sector, there is a possibility that firms capable of creating an advantage at an early stage will be able to create an entry barrier based on AI that will ensure market dominance, at least in the medium term. This suggests that the rules ensuring data availability are not only a matter of increasing research productivity or aggregation, but also speaks to the potential for protection against blocking and anti-competitive behavior.
3. Currently, there are a large number of individual companies trying to take advantage of AI in a variety of fields, but this high level of activity reflects expectations about the prospects for significant market influence in the future. Ensuring that artificial intelligence does not increase monopolization and does not increase barriers to entry into various sectors will be a key topic in the future. In addition, it is extremely important for organizations to have a cybersecurity policy that sets out employee behavior standards and best practices for protecting company data. Denial of access or disruption of the cybersecurity system can lead to serious consequences for both the company and the economy as a whole.

References

1. Nosova S, Norkina A, Makar S (2022) Artificial Intelligence technology as an economic accelerator of business process. *Stud Comput Intell* 1032 SCI:355–366
2. Markoff J (2016) As Artificial Intelligence evolves, so does its criminal potential, *New York Times*, p B3
3. Notes from the AI frontier: modeling the impact of AI on the world economy. Discussion Paper. McKinsey Global Institute. <https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world> (4 Sept 2018)
4. Welcome to the AI Index Report– Artificial Intelligence Index (stanford.edu) (2023)
5. Gartner: Strategic technology trends 2023, itweek.ru. Last accessed 06 Mar 2022

6. <https://www.washingtonpost.com/business/economy/future-wars-may-depend-as-much-on-algorithms-as-on-ammunition> (2017)
7. China announced plans to become a leader in AI by 2030, hightech.fm. Last accessed 24 Mar 2022
8. Osoba O, Welser W (2017) The risks of Artificial Intelligence to security and the future of work. RAND Corp., Santa Monica, Calif, vol 7
9. Nosova, S.S., Norkina, A.N., Medvedeva, O.E., Makar, S.V.: Artificial intelligence as a driver of business process transformation. *Proc Comput Sci* **213**, 276–284 (2022)
10. Barton D, Woetzel J, Seong J, Tian Q (2019) Artificial Intelligence: implications for China, vol 7. McKinsey Global Institute, New York
11. Mehr H (2017) Artificial intelligence for citizen services and government. Harvard Ash Center Technology & Democracy Fellow. A Center for Democratic Governance and Innovation. https://ash.harvard.edu/files/ash/files/artificial_intelligence_for_citizen_services
12. Davenport TH, Ronanki R (2018) Artificial intelligence for the real world. *Harvard Business Rev* 96(1/2):108–116
13. Artificial intelligence for industrial growth for now and for the future. <https://iotsummiteur.asia.com/en/trends/artificial-intelligence-for-industrial-growth-for-now-and-for-the-future>. Last accessed 29 May 2021
14. Davenport C (2017) Future wars may depend as much on algorithms as on ammunition, report says. Washington Post
15. Gillham J, Rimmington L, Hugh D, Verweij G, Rao A, Roberts KB, Paich M (2018) Macroeconomic impact of AI. PricewaterhouseCoopers, pp 1–71



Temporal Stability of Resting State fMRI Data Analysis by Independent Components Method

V. A. Orlov¹ , S. I. Kartashov¹ , M. V. Kalmykova¹ , A. A. Poyda¹ ,
and Vadim L. Ushakov^{2,3} 

¹ National Research Center “Kurchatov Institute”, Moscow, Russia
orlov_va@rrcki.ru

² Institute for Advanced Study of the Brain, Lomonosov Moscow State University, Moscow, Russia

³ “Psychiatric Hospital No. 1 Named After N.A. Alexeev of the Department of Health of Moscow” (GBUZ “PKB No. 1 DZM”), Moscow, Russia

Abstract. This paper is devoted to the analysis of the temporal stability of the independent components obtained by analyzing data of resting state functional Magnetic Resonance Imaging (fMRI). We analyzed 25 datasets of healthy volunteers, consisting of 1000 time samples each. The fMRI data recording time was 33.3 min for each volunteer. This approach made it possible to divide the experimental session into several time ranges to assess the temporal stability of the results obtained with the independent component analysis (ICA). During the analysis, the property of additivity of independent components was discovered: the dynamics of the independent components obtained in the analysis of individual time ranges have a high level of Pearson correlation (at least 0.9) with the dynamics of the independent components obtained in the analysis of the full experimental session, i.e., the result of ICA is robust to the choice of window size when analyzing a representative data sample. It was also shown that the time series of independent components, which topology corresponds to resting state networks, have a correlation with the global signal at the level of 0.4–0.5.

Keywords: Independent component analyses · ICA · fMRI · Resting state · Rs-fMRI

1 Introduction

Independent component analysis (ICA) is a computational method for separating a multivariate signal into additive subcomponents. It is widely applied to various signal processing problems. The method dates back to the 1980-th [1] and was refined and widely known by Pierre Common papers in 1990-th [2, 3].

This method allows to decompose the signal into independent non-Gaussian signals. In mathematical point of view ICA transforms the observed random vector $X = (x_1, \dots, x_m)^T$ into a vector $C = (c_1, \dots, c_n)^T$, which is made of maximally independent components. The ICA extracts the sources by exploring the independence underlying the measured data. Thus, it involves higher order statistics to recover statistically independent

signals from the observations of an unknown linear mixture [4]. ICA extracts independent components by maximizing their statistical independence. There are a number of ways to define independence. Among them there are three generally used: minimization of mutual information, maximization of non-Gaussianity and maximum likelihood estimation method [5, 6]. As a measure of independence the first group uses Kullback-Leibler Divergence or maximum entropy, the second group uses kurtosis and negentropy and the third group is based on Pearson distribution or on extended generalized lambda types of it. ICA has some drawbacks, for example, the supposition that the fundamental sources are non-Gaussian. Additionally, ICA can be considered computationally expensive and can have some convergence issues if input data are not accurately pre-processed [7]. Even considering such limitations, ICA is still a powerful and broadly used method in machine learning and signal processing.

ICA is used in a variety of applications, such as signal processing, image analysis, biomedical signal processing and data compression. Our research is based on functional Magnetic Resonance Imaging (fMRI) analysis. It is a non-invasive tool to study neural representation of different cognitive processes and to investigate brain function and organization. ICA can extract independent spatial maps and their corresponding time courses from fMRI data without a priori specification of time courses. The basic concept of ICA can be expressed as $X = M \cdot C$, where X is the observed data (i.e., the data matrix of fMRI signal), C is a component map (a matrix of voxel values), and M is a mixing matrix determining the time-varying contribution of each component map to the observed fMRI data. To estimate M and C simultaneously, ICA determines the unmixing matrix W (a permuted version of the inverse of the mixing matrix M) through iterative calculation. The component maps and corresponding time courses can be acquired using the following equation: $C = W \cdot X$ [8].

Resting-state functional magnetic resonance imaging (rs-fMRI) has been gradually applied to pre-surgical functional mapping. It provides essential information for intra-operative localization of brain regions. ICA-based mapping has shown advantage, as no a priori information is needed [9]. Even though ICA has been widely used for the analysis of rs-fMRI data, there are still several issues that need to be addressed. Among them: (1) the lack of gold standard or commonly accepted resting-state functional connectivity network (RFN) template, (2) the impact of the input number of independent components or the ICA results' dimensionality, (3) the elimination of components with artifacts. Classification of RFNs from the ICA results is still a tedious but important step when obtaining ICA of resting-state fMRI data [10].

ICA is used in the analysis of fMRI data, in particular, to identify functionally homogeneous regions and study patterns in their temporal behavior by analyzing functional connectivity between them [11]. Changes in functional connectivity are often associated with structural changes in the brain and clinical symptoms, for example, when analyzing patients with Alzheimer's disease [12]. When analyzing functional connectivity, data sets are generally used with a very small number of time samples: of the order of 250–300 [13–15]. In our opinion, an additional analysis of the dependence of the obtained results on the size of the analysis interval is required, while its value should be set at a level, sufficient to obtain stable results.

2 Materials and Methods

The experimental data were obtained at the National Research Center “Kurchatov Institute” on a Magnetom Verio tomograph with a magnetic field strength of 3 T. fMRI data were acquired with the following scan parameters: 42 slices, repetition time (TR) 2000 ms, echo time (TE) 20 ms, field of view (FOV) $192 \times 192 \text{ mm}^2$, voxel size $3 \times 3 \times 3 \text{ mm}^3$. As part of the study, 1000 time samples for functional data were scanned, with a total duration of about 33.5 min. The total study time was 40 min. The study involved 25 healthy volunteers aged from 18 to 31, average age 24. The research was approved by the local ethical committee of the National Research Center “Kurchatov Institute”.

The anatomical and functional data center was moved to the anterior commissure (AC). Magnetic inhomogeneity artifacts were removed from the functional data using the magnetic inhomogeneity protocol recorded during the scan (`gre_field_mapping`). A slice-by-slice correction for the phase shift caused by technical features of scanning (slices of a three-dimensional image were obtained not simultaneously, but sequentially over time TR) was made. With the help of BROCOLLI terminal scripts, the subject’s head movement artifacts were calculated and corrected. After that, structural and functional data were normalized into the Montreal Neurological Institute (MNI) atlas space. Anatomical data were segmented into 3 possible types of brain tissue (grey and white matter and cerebrospinal fluid). After normalization, the blurring of functional data was carried out with a Gaussian filter with a core of $6 \times 6 \times 6 \text{ mm}^3$.

To test the stability of independent components of the human brain regions identified by the method, the entire data array was divided into 3 time intervals:

1. Complete record of the experiment—1000 time samples;
2. The first half of the experiment—500 time samples;
3. The second half of the experiment—500 time samples.

For each selected interval, brain regions were identified using the method of independent components. The number of components varied from 11 to 17. Different number of components is explained by the strict condition to the percentage of the total time series variability described by the components: we chose the smallest number of components describing at least 80% of them. After selection of independent components, they were classified according to their association with resting state networks [16] based on an expert assessment on the analysis of topology and time–frequency characteristics of each selected component. As a result, main sensory networks of a resting state—visual, motor, auditory—were distinguished in each data set, as well as the default mode network (DMN).

3 Results

Figure 1 shows examples of topological maps of the DMN network obtained over three studied time ranges for a randomly selected subject.

For all independent components selected on the basis of expert analysis, we carried out analysis of topological intersection (the ratio of the number of common voxels to the total number of voxels from a smaller component). Thus, it was shown that intersection in all three analyzed ranges is at least 73%, and in some cases exceeds 95%.

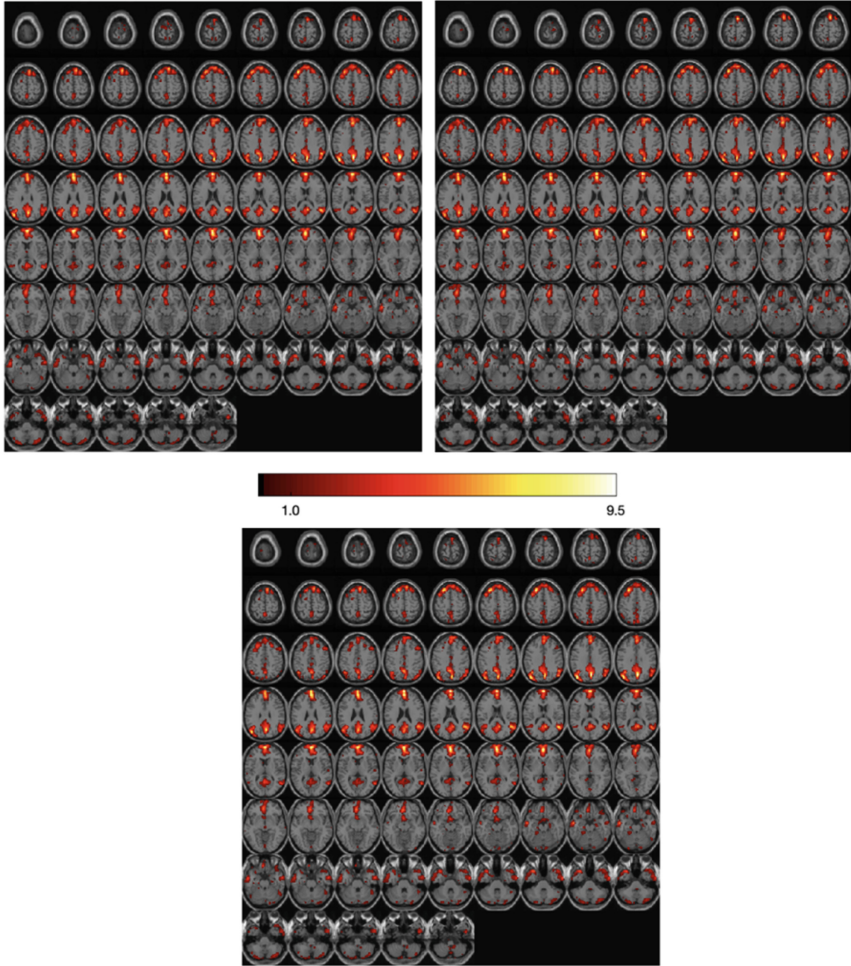


Fig. 1. An example of a topological map of a DMN network for a randomly selected subject. Top left—obtained over the entire time range, top right—obtained over the first half of the time range, bottom—obtained over the second half of the time range

Supplementing time dynamics from the first half of the experiment by corresponding time dynamics from the second half of the experiment and subsequent correlation analysis showed convincingly (with Pearson correlation coefficient $r > 0.98$) the additive properties of the method of independent components. It should be taken into account, however, that this result was obtained when applied to data containing 500 time samples.

An additional study was made of the relationship between the global signal and the dynamics of components obtained by the ICA method. The global signal reflects the overall signal from the entire brain and is calculated as the average of all brain signals. There is an assumption that the process, which is distributed throughout the brain, cannot be associated with the neural activity of individual areas. The global signal is associated

with physiological fluctuations, such as respiration, movements during scanning, and scanning features [17]. Since, on these grounds, global signal is considered as a noise contribution to the useful signal, a correlation analysis of temporal dynamics of the independent components with the global signal was carried out. Figure 2 shows a typical example of a correlation analysis obtained over a random time range for a randomly selected subject.

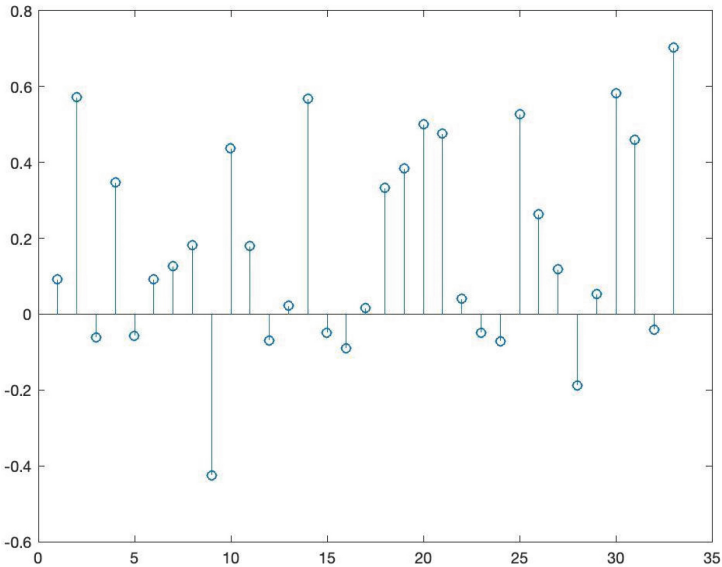


Fig. 2. An example of the results of correlation analysis of the ICA dynamics and the global signal obtained over the entire random time range for a randomly selected subject. DMN network—number 10, visual area—number 21, auditory area—numbers 25 and 33, motor artifacts—number 9. The remaining highly correlated components are either motor or vascular artifacts on the cortical surface

An analysis of the results obtained on all the data sets under study showed that average correlation between the global signal and the temporal dynamics of the components corresponding to DMN reaches 0.4, and between the global signal and other sensory networks of a resting state (in particular, visual and auditory)—0.5.

4 Discussion

As a result of the analysis of experimental data, the property of ICA additivity was established—with a sufficient number of time points their further increase does not lead to significant changes in spatiotemporal characteristics of the components obtained, and the components obtained by analyzing individual ranges have corresponding representation in the full data set. Thus, we can conclude that the size of an experimental rs-fMRI session of 500 time samples can be considered sufficient for the application of ICA.

Relatively high (of the order of 0.4) correlations between the components corresponding to the DMN and the global signal can lead to erroneous estimates of functional connectivity during the resting state. The influence of the global signal on other sensory networks of the resting state (in particular, auditory, visual, etc.) turned out to be even greater than on the DMN (about 0.5). This may be due to motion artifacts and the relative proximity of these networks to the boundary of the experimental scanning area. A significant correlation (about 0.4) was also found between the global signal and the displacement vector calculated by correcting for the subject's movements during the study.

Thus, we can conclude that, in addition to the necessity to choose correctly the sample size and the size of the analyzed time range, it is very important to analyze connectivity of the time series for selected neuronal networks of the brain with a noisy global signal.

Acknowledgments. The study was supported by a government task in the National Research Centre “Kurchatov Institute” and carried out using computing resources of the federal center for collective use “complex of modelling and data processing of mega-class research facilities NRC Kurchatov institute”.




References

1. Héroult, J., Ans, B.: Réseau de neurones à synapses modifiables: décodage de messages sensoriels composites par apprentissage non supervisé et permanent [Neuronal network with modifiable synapses: decoding of composite sensory messages under unsupervised and permanent learning]. *C. R. Acad. Sci. III* **299**(13), 525–528. French. PMID: 6437617 (1984)
2. Comon, P.: Independent Component Analysis. J-L.Lacoume. Higher-Order Statistics, Elsevier, pp. 29–38 (hal-00346684) (1992)
3. Comon, P.: Independent component analysis, a new concept? *Signal Process.* **36**(3), 287–314 (1994). ISSN 0165-1684. [https://doi.org/10.1016/0165-1684\(94\)90029-9](https://doi.org/10.1016/0165-1684(94)90029-9)
4. Calabrese, B.: Data Reduction. *Encyclopedia of Bioinformatics and Computational Biology*. Academic Press, pp. 480–485 (2019). ISBN 9780128114322
5. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. *Neural Netw.* **13**(4–5), 411–430 (2000). ISSN 0893-6080. [https://doi.org/10.1016/S0893-6080\(00\)00026-5](https://doi.org/10.1016/S0893-6080(00)00026-5)
6. Karvanen, J., Eriksson, J., Koivunen, V.: Maximum likelihood estimation of ICA model for wide class of source distributions. In: *Neural Networks for Signal Processing X. Proceedings of the 2000 IEEE Signal Processing Society Workshop (Cat. No.00TH8501)*, Sydney, NSW, Australia, vol. 1, pp. 445–454 (2000). <https://doi.org/10.1109/NNSP.2000.889437>
7. Le, Q., Karpenko, A., Ngiam, J., Ng, A.: ICA with reconstruction cost for efficient overcomplete feature learning. *Adv. Neural Inf. Proc. Sys.* **24** (2015)
8. Wei, P., Bao, R., Fan, Y.: Comparing the reliability of different ICA algorithms for fMRI analysis. *PLoS ONE* **17**(6), e0270556 (2022). <https://doi.org/10.1371/journal.pone.0270556>
9. Lu, J., Zhang, H., Hameed, et al.: An automated method for identifying an independent component analysis-based language-related resting-state network in brain tumor subjects for surgical planning. *Sci. Rep.* **7**, 13769 (2017). <https://doi.org/10.1038/s41598-017-14248-5>
10. Wang, Y., Li, T.-Q.: Dimensionality of ICA in resting-state fMRI investigated by feature optimized classification of independent components with SVM. *Front. Human Neurosci.* **9** (2015). <https://doi.org/10.3389/fnhum.2015.00259>

11. Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., Raichle, M.E.: The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 9673–9678 (2005)
12. Greicius, M.D., Srivastava, G., Reiss, A.L., Menon, V.: Default-mode network activity distinguishes Alzheimer’s disease from healthy aging: evidence from functional MRI. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 4637–4642 (2004)
13. Greicius, M.D., et al.: Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biol. Psychiatry* **62**, 429–437 (2007)
14. Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V.: Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 253–258 (2003)
15. Harrison, B., Pujol, J., Ortiz, H., Fornito, A., Pantelis, C., Yücel, M.: Modulation of brain resting-state networks by sad mood induction. *PLoS ONE* **3**, e1794 (2008). <https://doi.org/10.1371/journal.pone.0001794>
16. Biswal, B.B.: Resting state fMRI: a personal history. *Neuroimage* **62**(2), 938–944 (2012). <https://doi.org/10.1016/j.neuroimage.2012.01.090>
17. Aguirre, G.K., Zarahn, E., D’Esposito, M.: The inferential impact of global signal covariates in functional neuroimaging analyses. *Neuroimage* **8**(3), 302–306 (1998). <https://doi.org/10.1006/nimg.1998.0367>. PMID: 9758743



Analysis of Resting-State fMRI Data by CAPA Method

Vyacheslav A. Orlov¹ (✉) , Sergey I. Kartashov¹ , Alexey A. Poyda¹ ,
and Vadim L. Ushakov^{2,3} 

¹ National Research Center “Kurchatov Institute”, Moscow, Russia
orlov_va@nrcki.ru

² Institute for Advanced Study of the Brain, Lomonosov Moscow State University, Moscow,
Russia

³ “Psychiatric Hospital No. 1 Named After N.A. Alexeev of the Department of Health of
Moscow” (GBUZ “PKB No. 1 DZM”), Moscow, Russia

Abstract. The paper describes application of the co-activation patterns analysis (CAPA) method for analyzing resting-state fMRI data obtained in order to detect stable substates. The research involved 25 healthy volunteers. The analysis revealed that inside a resting-state we could distinguish 8 alternating stable substates. Their average duration was estimated at about 20–25 s.

Keywords: fMRI · Co-activation patterns · Functional organization of the brain · Resting state

1 Introduction

The research aimed at extraction and quantitative assessment of time-changing information contained in resting-state fMRI data provided a new subject for analysis—dynamic functional connectivity (DFC) of a resting-state—as well as corresponding technique. Among known set of DFC techniques, there is an approach radically different from traditional ones, because it analyses separate spatial fMRI-volumes at every time-instance, rather than analysing time series. Such approach concentrates on finding repetitive co-activation patterns (CAPs—i.e. coinciding changes of fMRI signal in a certain spatially separated group of voxels) in brain as well as on their changes in time. Most of the resting-state fMRI connectivity researches employ sliding window analysis and time-frequency connectivity [1–3]. These and related approaches evaluate functional connectivity between selected areas of interest within time windows (usually about 1–2 min), which are significantly shorter than typical scanning period, in order to find out transient interactions. However, most of the approaches to such “dynamic” connectivity analyses are still based on the calculation of pairwise links between the studied time series. Although these methods have limitations for quantitative assessment of fMRI data, which typically have by orders of magnitude more voxels (N) than time points (T). If pairwise correlations were calculated for all possible voxel pairs, then the resulting cross-correlation matrix would have by orders of magnitude more elements

($N*(N - 1)/2$ given its symmetry) than the actual $N * T$ measurements, with a maximum rank of $T - 1$, which is much smaller than its size N . Such a matrix would be a very redundant quantification of the covariance of the data. The existence of covariance between large sets of voxels justifies the practice of using much larger brain regions from atlases (or networks identified by spatially independent component analysis (ICA) rather than voxels for the analysis of functional connectivity [4, 5]. This approach, however, degrades (to some extent) the spatial resolution of fMRI due to the transition to larger areas. In addition to dimensionality reduction, it is necessary to use methods that reveal joint deviations of activity in more than two areas of the brain. A few years ago, a unique algorithm was developed that allows to evaluate instantaneous patterns of coactivation of neural networks in the brain—CAPA (co-activation patterns analysis). The CAPA method, unlike most of the DFS methods, allows to analyse data within 1 time frame, which makes it possible to exclude the dependence of the results on the choice of certain parameters, such as the window size, percentage of intersection, etc. In order to define unique brain states in experimental fMRI—the CAPA method was applied to the data [6].

2 Materials and Methods

The experimental data were obtained at the National Research Center “Kurchatov Institute” on a Magnetom Verio tomograph with a magnetic field strength of 3 T. fMRI data were acquired with scan parameters: 42 slices, repetition time (TR) 2000 ms, echo time (TE) 20 ms, field of view (FOV) $192 \times 192 \text{ mm}^2$, voxel size $3 \times 3 \times 3 \text{ mm}^3$. As part of the study, 1000 time samples for functional data were scanned, with a total duration of about 33.5 min. The total study time was 40 min. The study involved 25 healthy volunteers aged from 18 to 31, average age 24. The research was approved by the local ethical committee of the National Research Center “Kurchatov Institute”.

The anatomical and functional data center was moved to the anterior commissure (AC). Magnetic inhomogeneity artifacts were removed from the functional data using the magnetic inhomogeneity protocol recorded during the scan (`gre_field_mapping`). A slice-by-slice correction for the phase shift caused by the technical features of scanning (slices of a three-dimensional image were obtained not simultaneously, but sequentially over time TR) was made. With the help of BROCOLLI terminal scripts, the subject’s head movement artifacts were calculated and corrected. After that, structural and functional data were normalized into the Montreal Neurological Institute (MNI) atlas space. Anatomical data were segmented into 3 possible types of brain tissue (grey and white matter and cerebrospinal fluid). After normalization, the blurring of functional data was carried out with a Gaussian filter with a core of $6 \times 6 \times 6 \text{ mm}^3$.

For reorientation, normalization and spatial filtering of images, as well as correction of the time shift of signals caused by the specifics of data acquisition, a software package for the MATLAB—SPM12 environment was used in the work. The BROCCOLI library for the Mac OS bash environment was used to calculate and correct motion artifacts. Correction of magnetic field inhomogeneity artifacts, frequency filtering of the signal, decomposition into independent components and removal of “noise” components using a regression filter were performed in the FSL program.

3 Results

Due to physiological characteristics of the signal, a smooth change in connectivity between brain regions is assumed. Therefore, the states assessed using the CAPA method do not change quickly and cannot last 1 TR period. Based on these statements, an estimate of the number of “outliers” was introduced when classifying data into states—the number of transition points or single points in which the state lasts one TR period. This parameter made it possible to calculate the optimal number of states into which the entire fMRI time series can be divided. In the case of an excessive number of states (an example in Fig. 1 is 20 states), the smooth structure of transitions is broken.

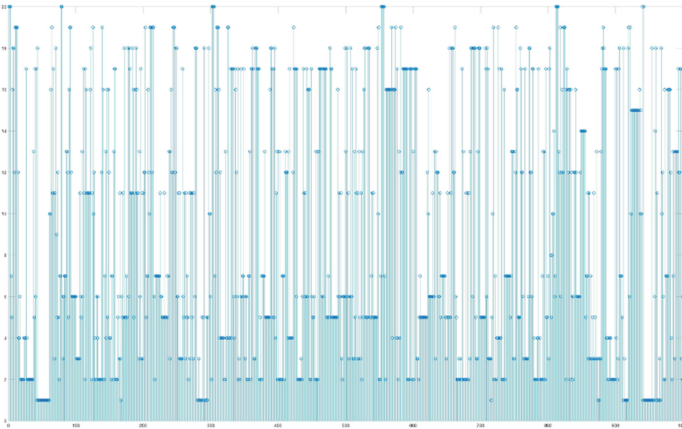


Fig. 1. Temporal scheme of the human brain states changes for an excessive number of states.

The number of outliers in the above example was more than 16% of the total number of points (163 outliers). An excessive number of examined states leads to the fact that some states (obtained with a smaller number of CAP's) are divided into 2 or more states. Low value of CAP leads to the mixing of all states into one. At this time, the remaining states exhibit abrupt or high-amplitude changes of the global signal. On Fig. 2 there is an example of data decomposition into 3 states (top) and the temporal dynamics of the global signal, which is the arithmetic average of the dynamics of the analyzed voxels (bottom).

Thus, we can conclude that the optimal number of states under study can be determined empirically, since if the number is insufficient, the specificity of the selected states disappears, while if they are redundant, the states are blurred. A total of 19 sets of results were calculated (for the number of CAPs from 2 to 20 inclusive) and in each of them the parameter (number of outliers) was estimated, as well as the degree of correlation between the occurrence of a particular state and changes in the global signal. Empirically, the optimal number of states was determined, into which it is necessary to divide the entire fMRI time series—8 (Fig. 3).

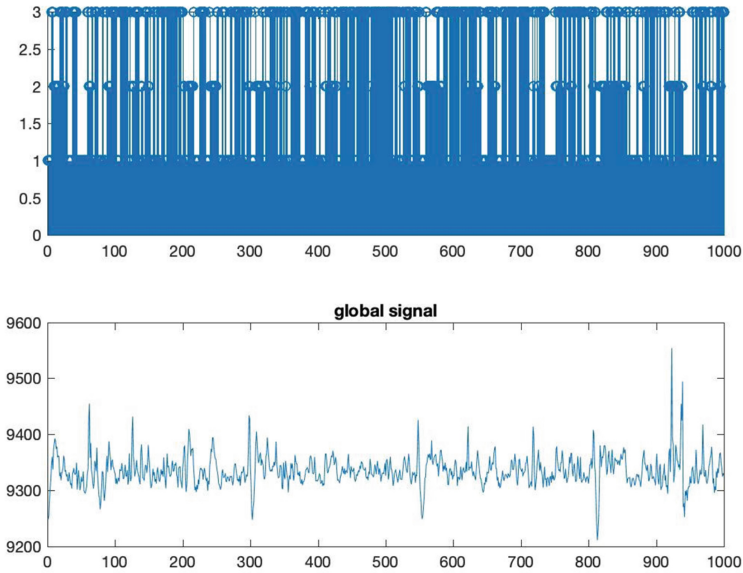


Fig. 2. An example of data decomposition into 3 states. Time diagram of changes in the states of the human brain for an insufficient number of states (top), temporal dynamics of the global signal (bottom).

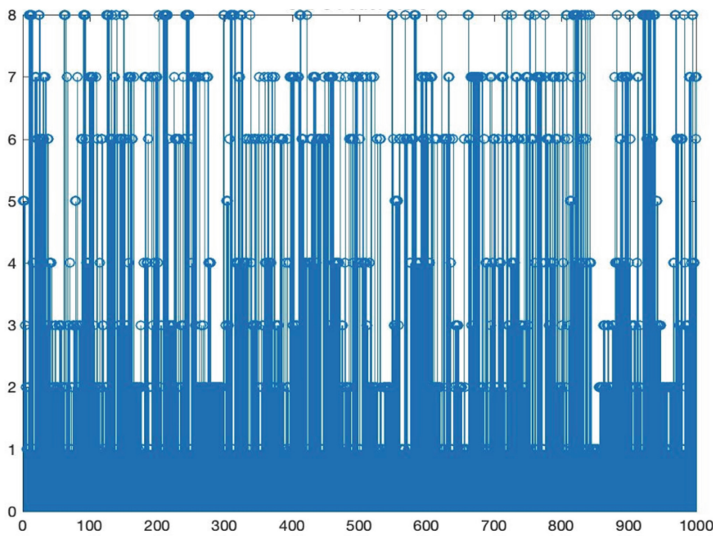


Fig. 3. The time scheme obtained by decomposing into 8 states (activation code patterns).

The average duration of a state (before its change), when decomposed into 8 patterns of coactivation, was 22 s for the entire set of volunteers. Taking into account the inertia property of the fMRI signal (the period is 15 s, while the increase is about 7 s), we can

conclude that the approximate duration of a state, without taking inertia into account, is about 15 s. Given the frequency of data recording (0.5 Hz, i.e. TR = 2 s), the duration of a state is about 7–8 time frames, which is physiological and corresponds to the time resolution of the method.

4 Conclusion

The results allow us to conclude that the detected states occur with some periodicity and last for a relatively short time—about 20 s. Empirically, an optimal (with a low number of outliers < 1%) number of CAPs for resting state analysis - 8 - was obtained. The CAP method can be applied not only to fMRI time series, but also to matrices of connectivity between regions. Such an application of the method can make it possible to obtain a temporal scheme of connectivity matrices changes. The established relationships between regions can be taken as states. As a result, a temporal scheme of changing relationships between the selected regions can be obtained.

Acknowledgments. The study was supported by a government task in the National Research Centre “Kurchatov Institute” and carried out using computing resources of the federal center for collective use “complex of modelling and data processing of mega-class research facilities NRC Kurchatov institute”.

References

1. Chang, C., Glover, G.H.: Time–frequency dynamics of resting-state brain connectivity measured with fMRI. *Neuroimage* **50**, 81–98 (2010)
2. Hutchison, R.M., Womelsdorf, T., Gati, J.S., Everling, S., Menon, R.S.: Resting-state networks show dynamic functional connectivity in awake humans and anesthetized macaques. *Hum. Brain Mapp.* **34**(9), 2154–2177 (2013). <https://doi.org/10.1002/hbm.22058>
3. Sakoglu, U., Pearlson, G.D., Kiehl, K.A., Wang, Y., Michael, A., Calhoun, V.D.: A method for evaluating dynamic functional network connectivity and task-modulation: application to schizophrenia. *MAGMA* **23**(5–6), 351–366 (2010)
4. Finn, E., Shen, X., Scheinost, D., et al.: Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* **18**, 1664–1671 (2015). <https://doi.org/10.1038/nn.4135>
5. Allen, E., Erhardt, E., Calhoun, V.D.: Data visualization in the neurosciences: overcoming the curse of dimensionality. *Neuron* **74**, 603–608 (2012)
6. Liu, X., Duyn, J.H.: Time-varying functional network information extracted from brief instances of spontaneous brain activity. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 4392–4397 (2013)



Application and Modeling of LLM in Quantitative Trading Using Deep Learning Strategies

Tiejun Pan¹ , Jinjie Yu² , Leina Zheng³ , and Yuejiao Li³ 

¹ College of Science and Technology (CST), Ningbo University, Ningbo, China

² University of Nottingham Ningbo China, Ningbo, China

2287071710@qq.com

³ Business School, Zhejiang Wanli University, Ningbo, China

Abstract. After more than 100 years of development, with the breakthrough of computer technology, deep learning and big data industry, the quantitative trading market has gradually matured, and more and more investors have begun to use quantitative trading to invest. Quantitative trading automatically executes transactions through written programs, eliminating the interference of human subjective factors on transaction execution. But the threshold for quantitative trading is high, requiring researchers to have a deep understanding of mathematics, statistics, finance, and computer technology. The newly emerged Large Language Model (LLM) can help users get started to a certain extent, by giving the general framework of the code, so that users can have a preliminary understanding of the countermeasures faster and more accurately. In terms of model training and testing, this paper adopts the CSI 300 index obtained from tushare platform to study the results of daily data, weekly data and monthly data after training. This project trained a stock price prediction model using long short-term memory (LSTM) methods. Then, the backtest model was established with the classic double moving average strategy in quantitative trading, and the backtrader platform was used to visualize the return results simulated by the backtest. Finally, we discussed the risks of using LLM codes to execute quantitative trading.

Keywords: Quantitative trading · LLM · Deep learning

1 Introduction

1.1 Research Background

The origins of quantitative trading can be traced back to the early 20th century, when researchers and traders began using statistical methods to analyze financial data and make investment decisions. By the 1990s, advances in artificial intelligence technology led to rapid growth in quantitative trading. Machine learning algorithms, which are capable of adapting and learning from data, were used to identify patterns and trends in financial data that could be used to inform trading decisions [1]. In recent years, the use of artificial intelligence based on big data analysis in quantitative trading continued to grow.

Quantitative trading refers to the use of mathematical models and algorithms instead of human subjective judgment to make trading decisions. It uses computer programs to train models on past financial data, analyze current data, predict future movements and identify trading opportunities. It automatically executes trades according to predetermined rules, which makes the trades entirely subject to objective judgment. Another advantage of quantitative trading is that it also allows traders to analyze large amounts of data accurately in a short time, and to implement complex trading strategies that might be difficult or impossible to execute manually. However, there are many problems and difficulties in quantitative trading. It requires a deep understanding of math, statistics, computer science, and finance across multiple fields, and building and maintaining the necessary infrastructure can be expensive. It is also subject to market risks and other uncertainties, and may be affected by changes in market conditions or the regulatory environment, so quantitative strategies need to be updated with the times.

In 2022, OpenAI Company took the world by storm with LLM, a heavily trained artificial intelligence that can respond to any question a user asks. It can even write codes, identify and fix errors in the codes. Its emergence has caused many occupations that could not be replaced by artificial intelligence to fall into crisis.

1.2 Research Purpose and Methods

The purpose of this research is to explore the application of LLM in quantitative trading, analyze whether quantitative analysis will be replaced by artificial intelligence, and discuss the future development direction of artificial intelligence technology in the field of quantitative trading.

Through communication and interaction with LLM, the corresponding stock price model was established according to its reply, and the CSI 300 index was imported. The feasibility of the model recommended by LLM is tested by comparing data sets with different frequencies. In the model analysis and exploration section, we will apply LSTM method. LSTM stands for Long Short-Term Memory, which is a type of recurrent neural network (RNN) architecture [2]. The double moving average strategy aims to capture trends in the price movement of a financial instrument by identifying changes in the relationship between the shorter-term and longer-term moving averages.

2 Research

2.1 Application of LLM in Stock Prediction

As we know, LLM would answer any questions you ask, and the vast majority of the responses are valuable. Some people might want to invest through LLM's answers; they believe that they could make a great profit if they follow the instruction of the answers, because the answers generated by LLM are collected from a huge scale of data. And the artificial intelligence is trained continuously; the training database is also being updated all the time, so its answer changes with the times.

We firstly asked LLM 'Please predict the top ten Chinese stocks with the best returns in the coming period'. LLM refused to give exact prediction, instead it told us

ten well-known Chinese stocks, and suggested us to conduct a research before investing in stocks. It also gave us a warning about the volatility of the stock market.

Does it mean LLM is useless for stock forecasting? Looking up to its answer, at the beginning LLM referred that it could provide information and analysis based on historical data. As a researcher major in mathematics, I want to build a model based on the data and do analysis. So the second question is ‘What method can be used to train a model based on historical data and trends?’ In the answer, LLM referred about six machining learning and statistical techniques which are Regression Analysis, Time Series Analysis, Neural Networks, Decision Trees and Random Forests, Support Vector Machines and Bayesian. From the history, we know that deep learning is an important method in quantitative trading [3], so we will focus on the third method LLM recommended. LLM told us ‘Recurrent neural networks (RNN) and long short-term memory (LSTM) networks are often used for sequential data, including time series analyses. In the next section, we choose to have a deep insight of LSTM.

2.2 Model Training Using LSTM, Comparing Different Data Sets

As we know, LLM is able to write codes. Before trying to write codes by hands, let’s firstly ask LLM to write some codes using LSTM strategy ‘please show me an example of using LSTM to train a model for stock prediction, obtaining stock data from tushare platform and visualize the training result’. Tushare provides access to a wide range of financial and economic data from various Chinese markets, it also offers powerful tools to manipulate and process financial data efficiently.

LLM provided instructions step by step, and wrote the codes for the general framework. There are some comments with the codes which could help users understand the meaning of codes in each step. It clearly shown how to use the functions from library such as pandas, tensorflow, scikit-learn. If the codes were copied directly from the answer and ran, error messages would display. The codes wrote by LLM were the simplest version of using LSTM strategy, hence the trained model using these codes would not be satisfactory. Therefore, there is a need to correct those errors and improve the exact learning strategy by hand. For example, codes from LLM only divide the data set into train set and test set. In practice, we also want to see how accurate is it going to be if we use the trained model to predict the stock. Hence, some changes were made to the codes. The dataset was divided into three parts, the train set which contained the first 80% of the data, the prediction set which contained the last 10% of the data and the test set contained rest of the data. Other variables like learning rate, recurrent training times, and length of the sequences are added to improve the degree of model fitting.

The project uses the data of the CSI 300, we focus on 399300.SZ, which is obtained from tushare platform. The obtained values include time, opening price, highest price, lowest price, closing price and trading volume, as shown in Table 1.

The time range of this project is between 2010/01/01 and 2023/06/10. The model was trained using daily, weekly and monthly data respectively using LSTM strategy. The target of this test is to compare the results and find the best dataset for model training. The model adopts mean square error (MSE) as the model evaluation index, and the formula

Table 1. CSI 300 data

Trading date	Open	High	Close	Low	Volume
20230609	3822.4088	3836.7026	3836.7026	3811.2189	134319863
20230608	3791.0826	3834.3888	3820.1867	3777.7752	118700934
20230607	3815.4823	3823.1124	3789.3418	3780.1021	98862432

is as follows,

$$MSE = \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2 \tag{1}$$

where, m is the number of samples, y_i is the stock price, and \hat{y}_i is the model forecast stock price. After training, MSE for each dataset are shown below. We also plot the predict value of stock price and the true value. By comparing those plots (see Fig. 1), the accuracy of the models can be found directly by eyes.

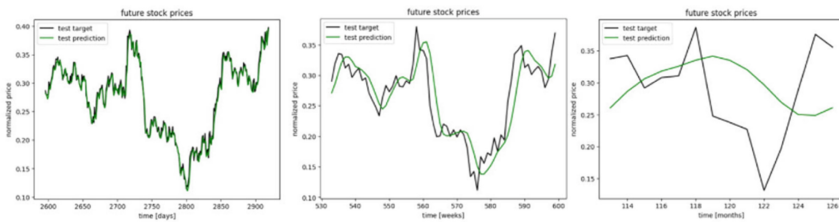


Fig. 1. Training results of the stock price prediction models recommended by LLM using LSTM method. From left to right are the results from daily, weekly and monthly data

- I. Daily data: $MSE = 0.000193$ which is very small. The prediction value of the test is quite fitted to test target using daily data. However, it takes about 2 min (long time) to finish training the stock prediction model.
- II. Weekly data: $MSE = 0.001167$ which is about 6 times larger than the MSE of daily data. It is easy to find that the whole trend is similar to test target. However, the predicted curve is smoother than the true value curve. For the highest and the lowest value, there is a huge difference. Besides, the green line (prediction line) looks like the black line (true value line) shifted to the right, which means the prediction has a little time delay. But for training time, it takes about 30 s, so the cost of training is cut down a lot.
- III. Monthly data: $MSE = 0.011480$ which is about 10 times larger than the MSE of weekly data. Since the same time period is used to train models, for monthly data, the amount is about 1/30 of daily data. Hence the accuracy of the prediction is low. Green line in the figure above is quite smooth, so the model trained by monthly data performs terrible when a shape change occurs. The value of this model is low though it only takes about 10 s to finish training.

In conclusion, for LSTM strategy, the best training data for stock prices prediction model depends on the need. If users want to see the trend as accurate as possible, they may need to take costs for training models with daily data. If users want to see the general trend in short time or with a little training cost, weekly data could be better. In this example, LLM helps to construct the basic structure of codes, though it cannot do all works for training a prediction model, it is a beginner-friendly tool in quantitative trading. The only thing the user has to do is adjust the codes to fit their training demands.

2.3 Backtest with Double Moving Average Strategy

We will discuss the application of LLM in quantitative trading strategies in this section by building a backtest model with a double moving average strategy as an example, and again compare the results of backtesting using daily, weekly and monthly data.

The double average strategy is a simple trend-following strategy commonly used in financial markets. It involves using two moving averages of different time periods to generate buy and sell signals. The first step is to calculate two moving averages of the asset's price, typically the closing price. One moving average is calculated over a shorter time period (for example 20 days), referred to as the 'fast' moving average, while the other moving average is calculated over a longer time period (for example 50 days), referred to as the 'slow' moving average. The second step is generating signals. Generate buy and sell signals based on the relationship between the fast and slow moving averages. The common approach is to consider a 'golden cross' and a 'death cross' as the signals. 'Golden cross' appears when the fast moving average crosses above the slow moving average, it generates a buy signal, indicating a potential uptrend or bullish momentum. In contrast, 'death cross' appears when the fast moving average crosses below the slow moving average, it generates a sell signal, indicating a potential downtrend or bearish momentum. The last step is trade execution. When a buy signal is generated, a long position is initiated, and when a sell signal is generated, the long position is closed. This step is the unique feature and main advantage of quantitative trading, it will strictly execute the transaction according to the signals generated by the strategy, and effectively prevent the transaction from being interfered by human subjective factors.

Backtrader is an open-source Python framework for building and testing quantitative trading strategies. It provides a flexible and comprehensive platform for developing, backtesting, and deploying trading strategies using historical market data. There are many common indicators built in, such as simple moving average(SMA), index average (EMA), smooth moving average (SMMA), etc. users can also define their own indicators. Backtrader offers built-in plotting and visualization capabilities using Matplotlib. Traders can generate interactive charts and visualizations to analyze strategy performance, visualize trade executions, and assess risk and position management. Backtrader is very easy to install, only need the pip command to download the package, without any other configuration, which is very friendly for non-computer professional investment researchers. Researchers with a certain understanding of python codes can easily get started.

Similar to the previous example, we start our project by asking LLM 'Could you please show me an example of using double average strategy to build a backtest model on backtrader platform and visualize the results, using the CSI 300 data obtained from

tushare?’ LLM provides the simplest version of double average strategy, but the model is not matching to real world trading case. When a trade occurs, the trader needs to pay a transaction fee of 0.01% to 0.30%. In this model, we set the proportion of the transaction fee to 0.05%. We also need to determine the position when the ‘gold cross’ or ‘dead cross’ appears.

In this project, we use the data of the CSI 300 stock dataset again and focus on 399300.SZ, which is obtained from tushare platform. Total initial funding is set at 100,000. After performing the backtest using daily data, weekly data and monthly data, we will compare the profits and get a conclusion (Fig. 2).



Fig. 2. Backtest results of the quantitative trading models recommended by LLM using double moving average strategy. From left to right are the results from daily, weekly and monthly data.

For model using daily data, the final total funding is 148480.57, which means we gain about 48.5% profits. Many golden crosses and dead crosses are shown in the figure. The model executed many trades automatically. However, for models using weekly data and monthly data, the number of golden cross and dead cross are small. Since the whole number of data are less than daily data. It is difficult for the model to identify signals of trading opportunities, thus many trading opportunities would be ignored. For all three cases, the running time is short. This means that using more frequent data doesn’t consume a lot of computing power. Instead, it helps us avoid missing trading opportunities and generate higher returns.

In this experiment, LLM helped to build a backtest model using double average strategy. We added some parameters to make the model more realistic. Using the model, we performed backtests with daily data, weekly data and monthly data. Finally, we found that the amount of revenue increases with the frequency of data over the same time period.

3 Random Cases of Stock Market

There are many factors affecting the stock market, such as changes in political situation, sudden natural disasters and artificial manipulation, which can cause stock prices not to change according to the expected trend. Therefore, there are certain risks in trading with the model recommended by the LLM.

As we all know, many large enterprises will take the initiative to assume corresponding social responsibilities and make contributions to social development. On the one hand, this will lead to an increase in expenses, which will lead to a decline in stock

prices. On the other hand, it can also indicate the reputation of the company, which will lead to a rise in stock prices. Therefore, it is difficult to draw a direct conclusion about the impact of a company's social responsibility on its stock price. We will study the relationship between the Social Responsibility Index (SRI) and the annual net profit of three famous Chinese Internet enterprises in the past eight years.

The China Corporate Social Responsibility Development Index is compiled by the Corporate Social Responsibility Research Center of the Faculty of Economics, Chinese Academy of Social Sciences, aiming to assess the development status and level of corporate social responsibility practices in China and provide reference for the government, enterprises and stakeholders. The index divides corporate social responsibility practices into four aspects: responsible management, market responsibility, social responsibility and environmental responsibility, and subdivides them into four sub-indexes. We collected data of Tencent, Baidu and Alibaba from 2015 to 2022, Table 2 is an example.

Table 2. SRI and annual net profit data of Tencent

Year	2015	2016	2017	2018	2019	2020	2021	2022
SRI	53.8	13.4	46.5	49.8	56.6	76.7	76.2	78.3
Net profit	291.08	414.47	724.71	787.19	933.10	1598.47	2248.22	1882.43

Overall, annual net profit increased year by year except last year. SRI of Tencent oscillated with a general upping trend. In 2016, SRI reached the least in the past 8 years, however, its annual net profit increased naturally. In 2020, the SRI increased about 20 due to its contribution in COVID-19, and annual profit growth rate was high in that year. It looks like high SRI could insure high profit growth rate. But in 2022, SRI reached the highest in the past 8 years, net profit falls for the first time. In this example, we couldn't get a conclusion of how the SRI affects annual net profit.

Similarly, for Baidu and Alibaba, there isn't clear relationship between SRI and net profit. From this example, we realize that the influencing factors of stock price are very complicated, and it is very risky to conduct quantitative trading only by modifying the code recommended by LLM. We need to consider many other influencing factors and establish different models to adapt to different scenarios.

4 Conclusion

Quantitative trading began to develop in foreign countries more than 100 years ago, but it is still in its infancy in China, and there is a lot of room for future development. Quantitative trading helps investors to have an objective cognition of the future trend by analyzing historical data and trends, and executes transactions according to procedures so that the transactions are not interfered with by subjective factors.

LLM, as a newly emerging chat artificial intelligence, is of great help to beginners to understand the basic knowledge related to quantitative trading. It can also help researchers majoring in mathematics, statistics and finance to quickly get started with the

writing of relevant programs and have a quick understanding of the code framework of strategy execution. However, its ability to write specific strategies is limited, for example, it can provide the basic framework code using long short-term memory modeling, but the improvement and optimization of specific details still need researchers' own understanding of quantitative transactions to achieve.


Acknowledgement. This paper is supported by Zhejiang Province's 14th Five Year Plan Teaching Reform Project (jg20220738), Ningbo Science and technology innovation Fund (2023Z228, 2023Z213), Zhejiang College Student Innovation and Entrepreneurship Training Program (S202310876001), Zhejiang Provincial Basic Public Welfare Fund Research Project (LGF20G020002), Zhejiang Provincial Philosophy and Social Science Planning Project (22NDJC127YB), Ningbo Municipal Basic Public Welfare Fund Research Project (2021S070).

References

1. Marti, G.: From Data to Trade: A Machine Learning Approach To Quantitative Trading. <https://ssrn.com/abstract=4315362> (2022)
2. Moghar, A., Hamiche, M.: Stock Market Prediction Using LSTM Recurrent Neural Network (2020)
3. Yan, Y., Yang, D.: A Stock Trend Forecast Algorithm Based on Deep Neural Networks



A Study of Conversational Intentionalities Expressed in Natural Language Using ChatGPT

Ivan A. Pavlenko, Arthur D. Zakirov, Andrei N. Yakovlev,
and Alexei V. Samsonovich^(✉) 

National Research Nuclear University MEPhI, Kashirskoe Shosse 31, Moscow 115409, Russian Federation

avsamsonovich@mephi.ru

Abstract. The goal of this study is two-fold: (1) to evaluate the usefulness and reliability of ChatGPT as a tool for detecting and generating nontrivial semantic categories of text, characterized by various conversational intentionalities, and (2) to build a semantic map of intentionalities and characterize its topological and geometric properties. ChatGPT 3.5 was used in this work. Results demonstrate reproducibility and reasonable accuracy. Furthermore, it was found that most intentionalities are highly correlated with each other and therefore can be expected to belong to a low-dimensional subspace on the semantic map.

Keywords: LLM · Semantic Mapping · Intentionalities · Affective Computing

1 Introduction

1.1 A Subsection Sample

When integrating artificial intelligence and language models into our daily lives, one of the key aspects of the interaction between Artificial Intelligence (AI) systems and humans is the system's ability to understand users' emotions and respond appropriately using available means of emotional communication. According to the present consensus in the field [1–6], there are six basic emotions that can also coexist and intertwine with each other. An example is given by the following sentence: “After a lot of hard work and difficulties overcoming, I finally achieved my goal, but found that the goal no longer matters to me”.

Human speech communication is not limited to the six basic emotions and frequently is characterized by a rich spectrum of intentionalities. Examples are complex social emotions, such as shame, empathy, love, jealousy, etc., and other nontrivial semantic categories, such as expression of moral support or an invitation to speak informally.

In order to enable the creation of social agents capable of meaningful usage of these subtle communicational intentionalities in a dialogue, it is necessary that such agents be able to recognize and express at least a certain predefined set of them.

The main goal of this study is to evaluate the ability of a large language model (LLM) to detect the presence of intentionalities from a given list in a given text. Specifically, the following questions are addressed by the present study:

1. Can LLM be used as a tool to determine the likelihood of an emotion in a given sentence?
2. With what accuracy an LLM can recognize emotions characterizing a sentence?
3. Can LLM recognize multiple emotions at the same time?

Results of the presented study give definite answers to these questions.

2 Materials and Methods

As the LLM to test, ChatGPT 3.5 was chosen in this study. It was easy to use as a tool for analyzing emotions, since ChatGPT technology provides convenient capabilities for processing text data. Because this LLM has already been trained on an extensive corpus of texts, it is of interest to study its abilities in solving problems of this nature.

The labeled database Hugging Face [7] was used in this study. To elicit ChatGPT responses, specially engineered prompts were used, the most successful of which were selected. Prompts were written in Russian (example translations given in Tables 1 and 2).

Table 1. Example prompt for ChatGPT for basic emotion analysis.

Field	Content
Model	gpt-3.5-turbo
Role: system	“You are a mechanism for determining human emotions.\nYour main task is to determine the probability of each emotion from the spoken sentence from 0 to 1.\nYou have only 6 emotions at your disposal to guess - joy, surprise, fear, disgust, anger, sadness.\nUse emotions only from the specified list!\nTRY NOT TO REPEAT THE SAME REMARKS TWO OR MORE TIMES IN A ROW\nDo not write anything in response to the sent proposal, except for listing the probabilities of emotions\n”

Through analysis of responses generated by ChatGPT to basic emotions we found that ChatGPT does a good job in identifying intentionalities and providing appropriate responses to them. This implies that ChatGPT is likely to be able to successfully recognize complex intentionalities, provided the appropriate prompt is used.

When developing the prompt, we used a learning method based on a small number of examples, known as “few-shot learning”. This method allows one to achieve more accurate generation of ChatGPT responses. To do this, along with the system message sent to ChatGPT, a short dialogue in the question-answer format between the user and ChatGPT was conducted, preceding the actual test interaction. This dialogue was necessary to demonstrate to ChatGPT the desired format of the response, which greatly simplified post-processing of responses generated by ChatGPT. Logical model of the experiment is shown in Fig. 1.

In the main part of our study, the *ru-go-emotions* database was used [7]. This database has been carefully adapted to meet the requirements for ChatGPT prompts. With the help

Table 2. Example prompt for ChatGPT for complex emotion analysis.

Field	Content
Model	gpt-3.5-turbo
Role: system	“You are a mechanism for detecting human emotions.\nYour main task is to determine the probability of each emotion from a spoken sentence from 0 to 1.\nYou have only 26 emotions at your disposal to guess - admiration, fun, anger, irritation, approval, care, misunderstanding, curiosity, desire, disappointment, disapproval, disgust, embarrassment, fear, gratitude, grief, joy, love, nervousness, optimism, pride, awareness, relief, remorse, sadness, surprise.\nUse only the emotions from the specified list!\nTRY NOT TO REPEAT THE SAME REMARKS TWO OR MORE TIMES IN A ROW\nDo not write anything in response to a sent proposal other than listing the probabilities of emotions\n”

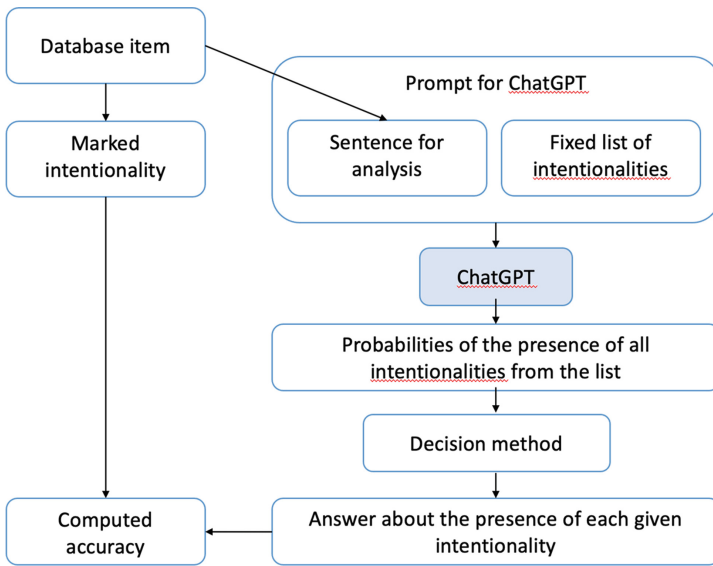


Fig. 1. Logical model of the experiment.

of participating experts, sentences containing clearly labeled emotions were selected from the database. Based on the selected data, two separate databases were created. The first database included 6 basic emotions (hereinafter referred to as basic [3–6, 12]), while the second database contained 20 compound intentionalities (hereinafter referred to as complex).

The actual ChatGPT response consists of a listing of each intentionality and the corresponding probability of the presence of this intentionality in the text, expressed as a number ranging from 0 to 1. Thus, the ChatGPT response contains results of analysis for the presence of each specified intentionality in the given text.

We tested several methods of interpretation of ChatGPT responses used evaluate the accuracy of these responses.

The first method is to select the intentionality with the highest probability as the final answer. However, this approach has the following disadvantages.

- ChatGPT does not always accurately generate a probability value for each intentionality, and the probability may vary from response to response.
- Sentences may have different interpretations, including several intentionalities, which is not always reflected in the labeled data. This circumstance makes the answer based on the maximal probability not always correct.
- Intentionalities can be similar in meaning or include each other, making it difficult to select one correct answer.

The second method is to select the first N most probable intentionalities. However, this method also suffers from drawbacks. While the sentence may not contain the given intentionality, with a low probability it can still be wrongly detected.

Here we used the following approach. A ChatGPT response was considered correct if the probability of the specified intentionality exceeds a certain threshold, the value of which was optimized.

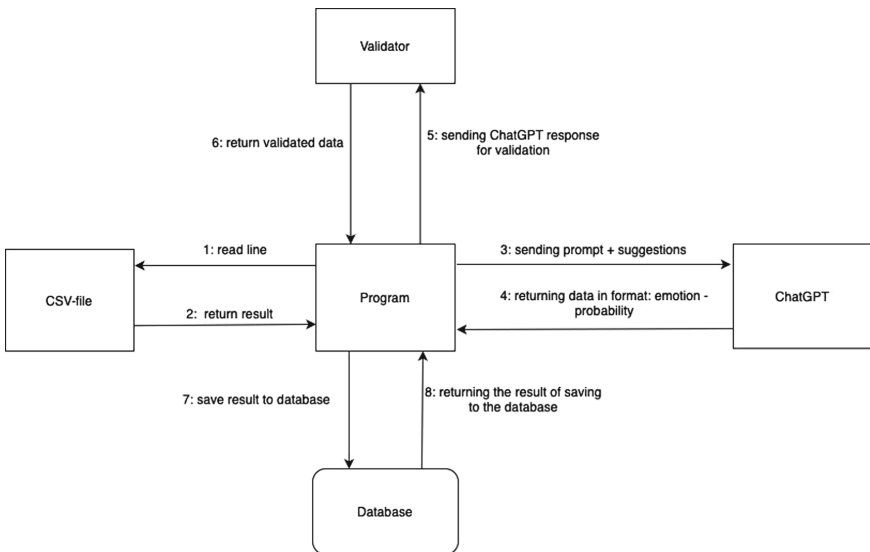


Fig. 2. The diagram explaining the operation of one processing cycle including 8 steps.

To interact with ChatGPT, a script was developed, which is structured as follows. The program reads a pre-generated and annotated CSV file line by line, where each line is a sentence characterized by an emotion, followed by the annotation with the name of this emotion. After that, the script generates the final prompt for the chat and sends data in string format to determine the emotion to ChatGPT. After issuing a response, it is read and validated in *json* format, after which the response data is validated for compliance

with the format, analyzed for correctness and sent to the database for further study. The program works cyclically as shown in Fig. 2.

- 1 line is read from the csv file in the format: “sentence - emotion”;
- a line without the true meaning of the emotion is sent to ChatGPT using an additional sentence for correct interaction with ChatGPT;
- the data is checked for the presence of additional characters that may have appeared when working with ChatGPT;
- data is recorded in the structure in the format: “emotion: probability”;
- data is written to the database.

The program works in such a way that if there is an error at one of the stages, the data is no longer considered, and the program moves on to the next cycle. This is necessary to prevent poorly processed data from being stored and further analyzed. That is, if the system makes an error in steps 1–8 (Fig. 1), the system moves on to the next step of the cycle.

During the final tests, no gaps were identified in the data, which indicates that it was possible to achieve the correct output of ChatGPT messages.

The current database layout does not allow obtaining all types of errors to collect statistics on errors of the first and second types. To solve this problem, the following method was used: since in the selected intentionalities there are intentionalities that are opposite in shades, such as sadness and joy, we can make the assumption that a sentence marked as “anger” can not be designated as “joy”, and vice versa. This assumption allows us to obtain the necessary data for statistical analysis, including not only false positives, but also false negatives. The following antonym pairs of emotions were selected.

- Admiration—disgust
- Fun—sadness
- Joy—anger
- Irritation—relief
- Approval—disapproval
- Curiosity—Fear

3 Results and Analysis

To analyze the correctness of ChatGPT responses, statistics of false positives and false negatives were computed at different threshold values. As a result of optimization, the final fixed threshold was set to the value of 0.5. At this level, computed overall test results are the following: Accuracy = 91.76%, Recall = 94.11%, Precision = 89.88%. Basic emotions were recognized in 93% of all cases, and complex intentionalities were recognized in 86% of all cases containing complex intentionalities.

To check that ChatGPT does not overinflate emotions for sentences, statistics were computed on the average and median probability values for all emotions for each sentence, after which histograms were built using these metrics (Fig. 3). The histograms show that the majority of values are lower than 0.1, meaning that ChatGPT does not overinflate emotions for sentences.

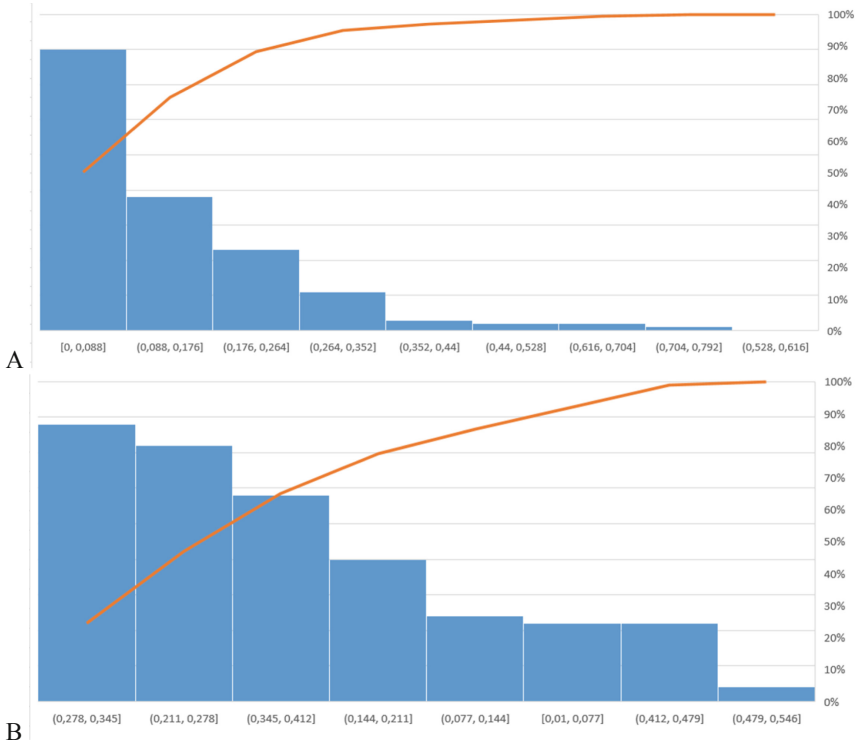


Fig. 3. Histograms of probability values for intentionalities. a: medians, b: means. The red line shows the cumulative sum.

In our analysis, data were obtained for the probability of each emotion in each sentence sent to ChatGPT. This was done separately for basic and for basic plus complex emotions. Computed correlation matrices are shown in Fig. 4.

In Fig. 4a, a significant negative correlation of the following emotions is visible:

- Anger and joy
- Anger and disgust
- Fear and Joy
- Disgust and joy
- Surprise and sadness
- Surprise and anger

This result shows that emotions with positive and negative connotations are opposed to each other.

In Fig. 4b, the following significant correlations can be found. Emotions of disapproval and anger, anger and irritation, as well as joy, admiration, approval and optimism are correlated. Anti-correlated are the emotions of anger and nervousness with approval, pride and approval with anger, grief and sadness with approval.

From this observation we conclude that significantly correlated emotions can be considered as shades of basic emotions, as in the case of the correlation of disapproval

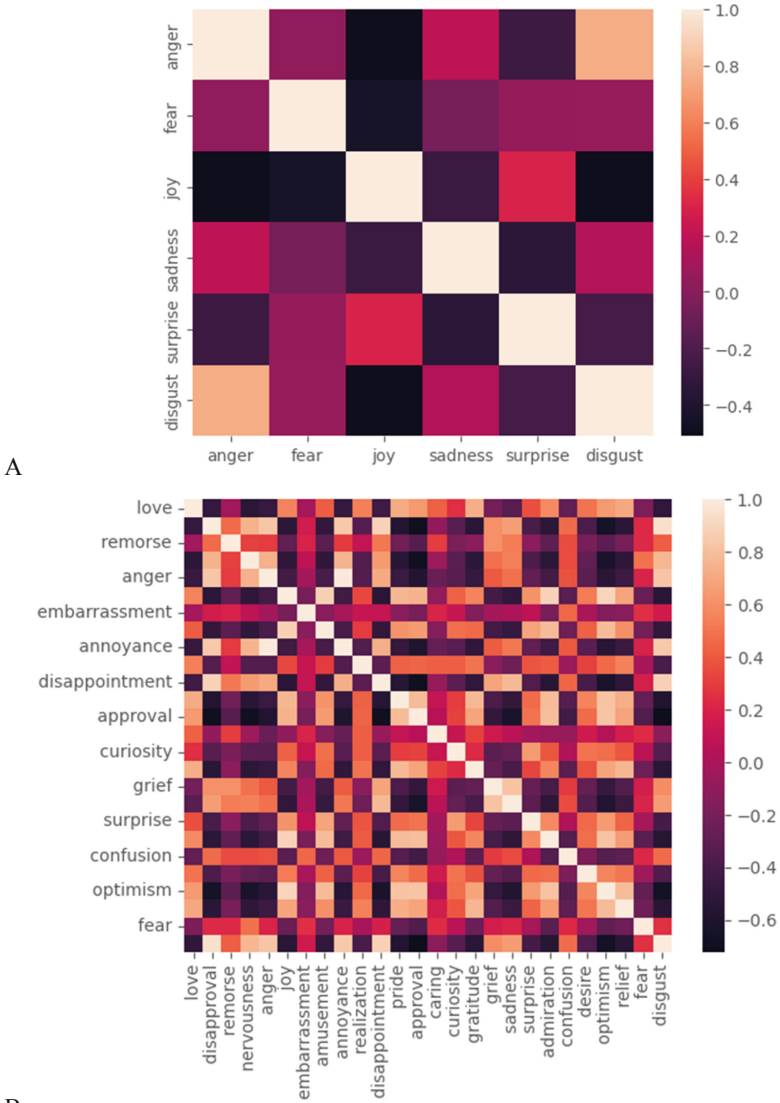


Fig. 4. Correlation matrices for emotion probabilities found in sentences by ChatGPT. a: correlation matrix for basic emotions; b: correlation matrix for basic plus complex emotions.

and irritation with anger. Since anger is a basic emotion, we can, based on this analysis, say that disapproval and irritation are shades of anger. The situation is similar with the emotions of admiration, approval and optimism, where they are shades of joy.

From all the above, one can also notice that anti-correlation, as in the first case, is most clearly expressed between emotions in which the tonality can be clearly identified, be it positive or negative.

4 Concluding Remarks

In general, the topic of emotional text analysis acquires high importance today [8–10]. Therefore, our preliminary results should be of interest to the research community.

The goal of our study was two-fold: (1) to evaluate the usefulness and reliability of ChatGPT as a tool for detecting and generating nontrivial semantic categories of text, characterized by various conversational intentionalities, and (2) to build a semantic map of intentionalities and characterize its topological and geometric properties. The first objective was fully achieved. Results demonstrate reproducibility and reasonable accuracy of emotion identification with ChatGPT. The validation of the method for basic emotions allows us to assume its usefulness and reliability in the case of complex social emotions and even more subtle nontrivial intentionalities.

Furthermore, it was found in this study that most intentionalities are highly correlated with each other. Therefore, we may expect that they belong to a low-dimensional subspace on the semantic map (cf. [11]). This hypothesis should be tested in our future studies.

The developed here approach based on intentionalities will find practical applications in many important domains, including tutoring systems controlled by cognitive architectures [13].

Acknowledgments. This work was supported by the Russian Science Foundation Grant #22-11-00213, <https://rscf.ru/en/project/22-11-00213/>.



References

1. Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)
2. Ekman, P.: An argument for basic emotions. *Cogn. Emot.* **6**(3–4), 169–200 (1992)
3. Russell, J., Mehrabian, A.: Evidence for a three-factor theory of emotions. *J. Res. Pers.* **11**, 273–294 (1977)
4. Plutchik, R.: A psychoevolutionary theory of emotions. *Soc. Sci. Inf.* **21**, 529–553 (1982)
5. Lövheim, H.: A new three-dimensional model for emotions and monoamine neurotransmitters. *Med. Hypotheses* **78**(2), 341–348 (2012)
6. Russell, J.: Core affect and the psychological construction of emotion. *Psychol. Rev.* **110**(1), 145–172 (2003)
7. Hugging Face Database. https://huggingface.co/datasets/seara/ru_go_emotions
8. Rzepka, R.: Emotional information retrieval for a dialogue agent. *Informatica* **27**, 205–211 (2003)
9. Moors, A., Ellsworth, P.C., Scherer, K.R., Frijda, N.H.: Appraisal theories of emotion: state of the art and future development. *Emot. Rev.* **5**(2), 119–124 (2013)
10. Lieto, A., Pozzato, G.L., Striani, M., Zoia, S., Damiano, R.: DEGARI 2.0: a diversity-seeking, explainable, and affective art recommender for social inclusion. *Cogn. Syst. Res.* **77**, 1–17 (2023). <https://doi.org/10.1016/j.cogsys.2022.10.001>
11. Gadzhiev, I.M., Knyshenko, M.P., Dolenko, S.A., Samsonovich, A.V.: Inherent dimension of the affective space: analysis using electromyography and machine learning. *Cogn. Syst. Res.* **78**, 96–105 (2023)

12. Samsonovich, A.V., Ascoli, G.A.: Cognitive map dimensions of the human value system extracted from natural language. *Front. Artif. Intell. Appl.* **157**, 111–124 (2007). ISSN: 09226389
13. Samsonovich, A.V., De Jong, K.A., Kitsantas, A., Peters, E.E., Dabbagh, N., Kalbfleisch, M.L.: Cognitive constructor: an intelligent tutoring system based on a biologically inspired cognitive architecture (BICA). *Front. Artif. Intell. Appl.* **171**(1), 311–325 (2008). ISSN: 09226389



Symbiotic Artificial and Human Cognitive Architectures Managing Human Attention

Thomas Pederson¹(✉)  and Amit Kumar Mishra^{1,2} 

¹ University West, Trollhättan, Sweden

thomas.pederson@hv.se, akmishra@ieee.org

² University of Cape Town, Cape Town, South Africa

Abstract. Wearable digital technologies such as Augmented Reality glasses offer a unique platform not only for monitoring proxies of individual human behaviour data (e.g. eye and body limb movements, posture, location, skin conductivity) but also for affecting behaviour, as instances of persuasive technologies often used to achieve personal human goals, e.g. for integrating physical exercise into everyday life. For artificial computational systems to gracefully affect in-situ human behavior is however associated with several challenges. It requires carefully interfacing digital processes running on the wearable device(s) with biological processes taking place inside the human body (e.g. cognitive, perceptual, motoric). It also needs to work within the constraints of engineering viability, user experience, and ethical constraints. We present our initial attempts to synchronize human biological visual attention processes with eye tracking-based visual stimuli generation in two application domains: for counteracting racial discrimination in the assessment of job applications, and for facilitating assembly tasks. Characteristic for both approaches is that the persuasion is aimed at being completely unnoticeable, at least in the long-term. We then move on to initial ideas for a more general model for integrating perceptual and cognitive functions across the biological-digital border to optimize the system as a whole. More specifically, today's AI architectures have a hard time achieving human-like high level cognition and perception which, we would argue, could potentially be addressed by a carefully designed symbiotic information exchange between existing human biological symbolic processes run inside the human brain with digital Machine Learning ones being tasked with the simpler sub-symbolic processing. Ethical concerns are, of course, also discussed including the potential reduction of “free will” and the consequences of system failure.

Keywords: Brain-inspired computing · Augmented reality · Human-computer interaction · Artificial intelligence · Symbolic-AI · Sub-symbolic-AI

1 Introduction

As digital sensors and actuators become increasingly embedded into stationary, mobile, and wearable devices and become linked to digital processes of agency (e.g. systems to control the light in a building; systems that guide novices into successful task completion, systems that reminds us to do things) human perception, cognition, and action that once was governed by biology alone (our body and minds) are now in fact often digitally augmented or at least influenced. Both sporadic and deliberately designed combinations of digital and biological (human) perception/cognition are likely to become both more frequent and more intricate in the future, given more powerful biometric sensors and Machine Learning (ML) approaches combined with better understanding of the human brain. In this paper we present our initial ideas for creating stronger connections between Biologically Inspired Cognitive Architectures (BICA) with the human brain itself, focusing on how the nature of the two different kinds of architectures could complement each other. It can be mentioned here that the authors are acutely aware of the high-risks attached to biometric sensors in the European AI Regulation [1, 13].

2 Symbiotic Management of Human Visual Attention

The human brain processes an immense amount of information originating from phenomena external to the body, that is perceived by our senses (vision, auditory, olfactory, etc.) as well as from higher-level cognitive processes (our wishes, emotions, intentions) each second of our life. It has been estimated that around 95 percent of this brain processing is completely unconscious [7, 14]. Furthermore, it is widely accepted that there is a regulatory mechanism, an “attention” mechanism, which selects among these many unconscious processes and lifts one or a few of them to our conscious awareness at a given point in time. Using the model given by Daniel Kahneman [4], when a process gets higher attention it is then handled by the slow-brain. The exact nature of this mechanism, and how it combines higher-level intentions with lower-level external stimuli to direct our conscious focused attention is still debated among neuropsychologists. It is clear however, that the human attention mechanism is affected by, and affects, the visual perception system. Human eye movements in everyday life form an important source of information to the attention mechanism for deciding where to place our visual attention (the bottom-up information flow in the perception vertical in the cortex [5]). Human eye movements also act as indicators of what our current intentions are in a more general sense (the top-down information flow in the action vertical in the cortex).

2.1 Motivation, Applications, and Challenges

By tracking eye fixations in the environment that surrounds a given human agent, a digital system could, in theory, given the presence of an adequate semantic

model of the world, get an indication of what entities in that environment that matters to that human agent, in that moment. We tend to look at things relevant to what we intend to do. However, the opposite is also often true: What we have in front of us influences what we actually decide to do [10]. The latter means that if a digital system could influence what we visually attend to, that system could to some degree, for better or worse, in principle, alter what we choose to do in that given situation or soon thereafter.

We are currently working on a Head-Mounted Display (HMD) Augmented Reality (AR) system intended to draw the user’s attention to certain objects in the surrounding physical environment and away from others. An important application area is the training of novice personnel in for instance industry where beginners sometimes get distracted by things that experts completely ignore. Results from our ongoing “subtle AR guidance” experiments in the lab is expected to find their way into, and complement, our more classic AR operator guidance approaches already tested in industrial assembly [12] and inspection [11] tasks.

We are also investigating the use of this attention manipulation approach for promoting equality where the idea is to contribute to a more fair and unbiased ranking of job applicants in multi-ethnic contexts (discrimination in such situations is a well documented global problem), by having a system make human evaluators attend to what actually matters in the applicant CVs and ignore aspects that are associated with racism and/or irrelevant. Our experiments with subtle guidance in both mentioned cases are ongoing and results not yet ready for dissemination.

We are facing two major challenges. (1) The engineering challenge of making the artificial visual attention guidance work gracefully together with its biological counterpart residing in the brain, e.g. so it doesn’t disturb or cause fatigue, (2) the ethical implications in case we would actually succeed: Under what conditions is it ethically viable to deploy and use systems that unnoticeably manipulate people’s perceptions of, and actions in, the world? While the second challenge might be the most important one in the long term, we focus on the first challenge (in a general sense) in the remaining parts of this paper, given this conference community’s focus on cognitive architectures.

3 A New Symbiotic Unconscious Human-Computer Interaction Loop

Inspired by early Subtle Gaze Direction (SGD) work by Bailey et al. [2] on desktop computer platforms, we aim at addressing the first challenge mentioned earlier by letting our AR-based system generate attention-drawing stimuli at “the right” level of visual intensity: strong enough to be perceived by a given individual’s visual sensory system, weak enough to not cause conscious reflection. As such, our design joins an increasingly populated class of interactive systems targeting the periphery of attention [3] sometimes also referred to as implicit Human-Computer Interaction (HCI) [8]. There are two aspects of this

type of interactive systems that are fundamentally new and of relevance for this workshop, when compared to the system we have designed for the past 50+ years in the HCI community: 1) They aim at inducing behaviour change in users without giving rise to conscious cognitive processing, 2) the time window in which interaction cycles between system and human agent take place is probably best measured in milliseconds rather than seconds. To actually work, both of these system characteristics demand a very tight integration, and a full two-way symbiotic relationship, between a select set of biological brain functions (e.g. visual perception at different levels of abstraction, attention filtering, short term memory) and its complementing digital Artificial Intelligence (AI) component(s) driving stimuli generation. We use the term “symbiotic” because we foresee a mutual need for collaboration (for the lack of a better term) between biological and digital cognitive processes in order to arrive to good micro-level decisions such as for instance an unconscious decision to move a specific object one centimeter to the left to facilitate an upcoming manual operation. Biological processes will react to the digital and the digital need to then within milliseconds adapt its behaviour accordingly to achieve the desired effect. Figure 1 illustrates this high-speed unconscious interaction cycle for users carrying wearable context-aware subtle guidance systems, highlighting the role of the biological attention filter in the brain for determining whether conscious reflection will be asked for or not, partially depending on the nature and intensity of incoming stimuli.

4 Perception-Centric AI-in-the-loop to Achieve a Symbiotic Perception and Action Cycle

Crucial to our approach in guiding human attention is the inclusion of AI in the perception-cognition-action loop, primarily for monitoring the state of the user, the surrounding context, and for generating guidance stimuli. For a smooth operation of the system, the processing of the digital blocks (like sensor data processing, decision regarding changes to be made to the sensors etc.) need to happen in synchronization with relevant operations in the brain that influence user behaviour. Hence, we need a unified model that represent both the processes of the biological brain as well as of the digital AI modules together. In tackling this challenge, we are developing an architecture inspired by Fuster’s model of prefrontal cortex [5].

The layered processing of information in the mammalian brain is a well-accepted model. Figure 2 shows the model popularized by Fuster [5]. It is an elegant model which is both simple to understand and is very close to the way the human brain works.

In our solution we are working on embedding the digital-AI part in our perception-action cycle using such a layered architecture. A layered architecture has the advantage of being able to enable the development of a unified architecture of a system with multiple different and disjoint sub-systems. In spite of being an elegant model, this model has multiple challenges when it comes

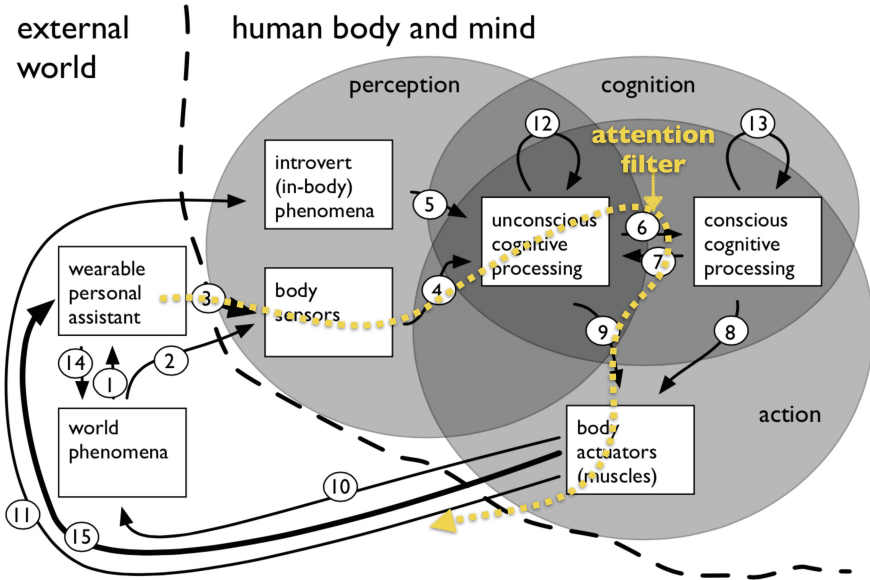


Fig. 1. How wearable subtle attention guidance systems (denoted wearable personal assistant in the figure) are imagined to produce stimuli which cause behaviour change (action) despite not triggering conscious cognitive processing, due to a carefully designed symbiotic relationship with the attention filter in the brain of the user (adapted from [6])

to implementing it to create artificial cognition in artificial systems. The infamous challenge of linking symbolic with non-symbolic layers [9] is one of the foremost ones. Similarly, the feedback paths from the perception vertical to the action vertical and vice versa are equally challenging to implement in a realistic engineering system. It can be noted that these feedback paths are crucial in the emergence of cognition as they enable the creation of internal information (independent of the information extraction from the sensory inputs).

However, the non or sub-symbolic algorithms have become really powerful because of the accelerating advances in the field of deep learning. Figure 3 delineates the architecture into two parts, viz.

- the sub-symbolic parts which are relatively easy to implement using deep learning networks, and
- the symbolic and cognitive layers which are extremely challenging to implement as a module.

We plan to use this architecture to model the complete system presented in Fig. 1. We call this digital-AI-in-the-loop based symbiotic attention modulation or DiL-SymAtMo.

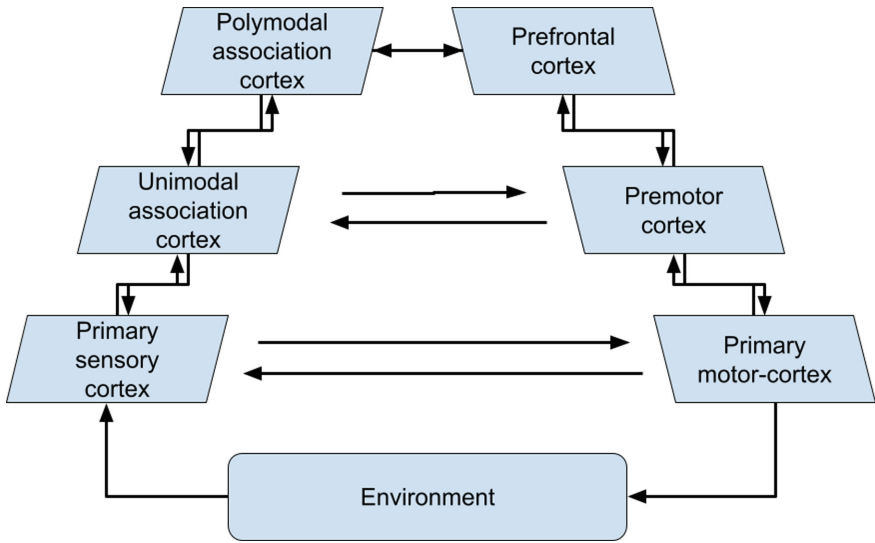


Fig. 2. Schematic model of cortical perception-action layers as presented by Fuster [5]. The inter layer feedback paths are specifically noteworthy. These feedback paths are difficult to incorporate using computers. These might be one of the origins of the emergence of cognition and consciousness.

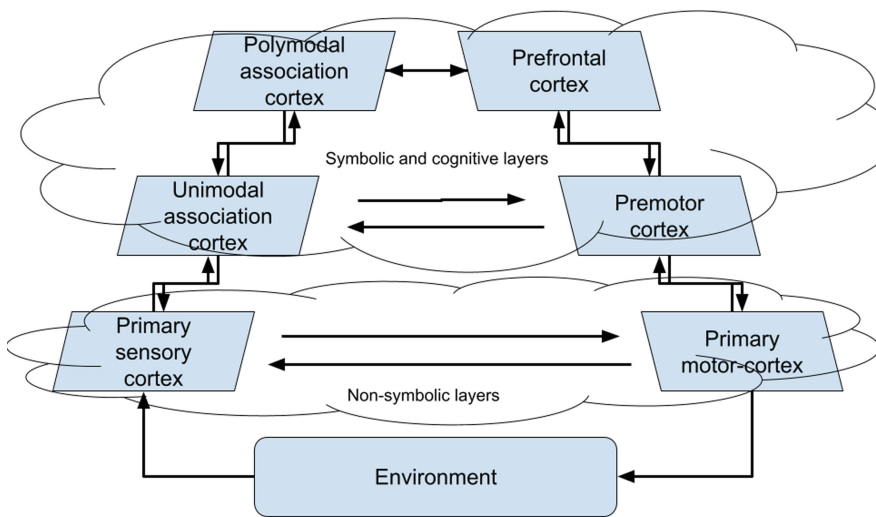


Fig. 3. Schematic model of cortical perception-action layers as presented by Fuster [5]. In this the non-symbolic layers have been marked. These are the blocks that are implementable using existing machine learning and deep learning architectures. The symbolic and cognitive layers are more difficult to implement. Even when they are implemented combining these two parts has been an existing challenge for AI researchers.

5 Conclusions

We have provided our initial thoughts and described our initial attempts in designing a combined and mutually dependent (symbiotic) computational and biological attention management architecture for the purpose of simplifying both everyday and more specific human activities, based on emerging wearable Augmented Reality systems.



Acknowledgments. This work has been partially funded through the ReStart project, KK foundation (kks.se), grant no: 20210093.

References

1. The proposed AI act from the EU: an engineer's expounding. www.linkedin.com/pulse/proposed-ai-act-from-eu-engineers-expounding-amit-kumar-mishra-phd/. Accessed: 03 Mar 2023
2. Bailey, R., McNamara, A., Sudarsanam, N., Grimm, C.: Subtle gaze direction. *ACM Trans. Graph.* **28**(4) (2009). <https://doi.org/10.1145/1559755.1559757>
3. Bakker, S.: Design for Peripheral Interaction. Ph.D. Thesis (2013)
4. Daniel, K.: Thinking, fast and slow (2017)
5. Fuster, J.M.: Upper processing stages of the perception-action cycle. *Trends Cogn. Sci.* **8**(4), 143–145 (2004)
6. Jalaliniya, S., Pederson, T.: Designing wearable personal assistants for surgeons: an egocentric approach. *IEEE Pervasive Comput.* **14**(3), 22–31 (2015). <https://doi.org/10.1109/MPRV.2015.61>
7. Lakoff, G., Johnson, M.: Philosophy in the flesh: the embodied mind and its challenge to western thought. Collection of Jamie and Michael Kassler. Basic Books (1999). https://books.google.nl/books?id=KbqxnX3_uc0C
8. Schmidt, A.: Implicit human computer interaction through context. *Pers. Technol.* **4**(2–3), 191–199 (2000). <https://doi.org/10.1007/bf01324126>
9. Son, J., Mishra, A.K.: A survey of brain inspired technologies for engineering. In: 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), pp. 1–6. IEEE (2016)
10. Suchman, L.A.: Plans and situated actions—the problem of human-machine communication. In: *Learning in Doing: social, Cognitive, and Computational Perspectives* (1987)
11. Tobisková, N., Gull, E.S., Janardhanan, S., Pederson, T., Malmköld, L.: Augmented reality for AI-driven inspection?—A comparative usability study. *Procedia CIRP* **119**, 734–739 (2023). <https://doi.org/10.1016/j.procir.2023.03.122>
12. Tobisková, N., Malmköld, L., Pederson, T.: Head-mounted augmented reality support for assemblers of wooden trusses. *Procedia CIRP* **119**, 134–139 (2023). <https://doi.org/10.1016/j.procir.2023.02.130>
13. Veale, M., Zuiderveen Borgesius, F.: Demystifying the draft EU artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Rev. Int.* **22**(4), 97–112 (2021)
14. Young, E.: Lifting the lid on the unconscious. www.newscientist.com/article/mg23931880-400-lifting-the-lid-on-the-unconscious/ (2018). Accessed 27 June 2023



The Impact of Internet Media on the Cognitive Attitudes of Individuals on the Example of RT and BBC

Alexandr Y. Petukhov^{1,2} , Sofia A. Polevaya² , Dmitry I. Kaminchenko²,
and Evgeniy A. Gorbov² 

¹ Lomonosov Moscow State University, Leninskie Gory, 1, Moscow 119991, Russia

² Nizhniy Novgorod Lobachevski State University, Gagarin Ave. 23, Nizhny
Novgorod 603950, Russia
Lectorr@yandex.ru

Abstract. Objective: Revealing of the presence or absence of changes in the cognitive attitudes of the individual under the conditions of external information impact on the example of the political-value content of the individual and groups consciousness. Background. Against the background of the growing digitalization of modern society and the networkization of the political space, the informational influence on individual and collective consciousness is increasing. It is extremely important to identify the features and effects of the influence of various information impulses on the cognitive attitudes of the individual and groups. Study design. The authors studied the impact of external information influences from watching videos of popular media («RT» and «BBC») on the cognitive attitudes of an individual using a special experiment. The authors used analysis of variance (ANOVA) and the Wilcoxon T-test to verify the presence of a real information impact. Experiment participants. Experimental sample: 21 people (86% women, 14% men). Among the participants in the experiment there were 11 people with higher education (52%) and 10 students receiving higher education (48%). Eight participants in the experiment were receiving or have already received the first speciality “psychologist” (38%) and 13 people who chose another speciality or speciality “psychologist” as a second higher education (62%). Measurements. The authors have monitored the heart rate variability of experiment participants using event-related heart rate telemetry technology. The participants have been tested to assess the level of emotional maladjustment (EED), as well as a questionnaire to identify the level of inclination towards conservatism and liberalism. Results. Watching videos from popular media has influenced the change in the level of conservatism, as well as the level of emotional maladjustment of the participants in the experiment. Higher education contributes to greater stability of cognitive attitudes, since after the informational impact, the inversion of the cognitive attitude occurred only in 16% of cases, and most of those who changed attitudes are students (67%). Conclusions. External informational influence from popular media has an impact on the change in cognitive attitudes and political-value content of the individual and group consciousness.

Keywords: Informational influence · Cognitive attitudes · Political values · Liberalism · Conservatism · Mass media

1 Introduction

The fast development of modern information communications together with the significant use of electronic devices in everyday life make studies in the field of Human-Technology interaction as relevant as ever. The same information obtained from different sources may be perceived differently by the same individual. This phenomenon is often used in so-called information wars.

This, in turn, increases the relevance of studies of the psychophysiological registration of the ways the cognitive attitudes of an individual are deformed by the external informational influence. Such studies can allow identifying specific mechanisms, algorithms, and patterns of such processes, which will, in turn, not only help with their correct explanation and definition but also can be used for forecasting in certain particular cases [1–6].

The study is aimed at solving a fundamental scientific problem, which consists in identifying the characteristic patterns of change in the psycho-physiological parameters of an individual during the deformation of their cognitive attitudes through external informational influence [4].

The information influence in the modern globalizing world is a serious challenge to the security of any state. Improving methods and the development of communication networks makes this problem one of the most pressing issues of today's world.

Among most impactful institutions which broadcast certain views to society one can name the Internet mass media and social media. Considering the intensive convergence of traditional and modern media, they have a significant impact on the consciousness of a modern people, their political values and attitudes. Therefore, studies are needed to assess the degree of influence of certain media on the cognitive attitudes of individuals and groups. For this study, we chose the two largest international channels, and namely RT and BBC, which are believed to represent polar political views.

2 Method

2.1 Procedure

The study was conducted at the Department of Psychophysiology of the Nizhny Novgorod State University. N.I. Lobachevsky during 2020. Each participant was provided with general unbiased information about a well-known event (problems with the admission of Russian athletes to the 2018 Winter Olympics), which was covered by the world media. After that, the participant put on a sensor for HRV monitoring. The participant completed a specially designed survey of his liberalism—conservatism cognitive attitudes, and then completed EED test to determine the level of emotional maladjustment. Then the participant was subjected to information influence in the form of specially selected news videos from BBC (RT), and then repeated the survey and the EED test. His HRV indices were recorded in the whole process of the study. After a certain amount of time, the experiment was repeated, using modified information stimulus, namely, a news story dedicated to the same problem, but from the second media representative RT (BBC) (Fig. 1).

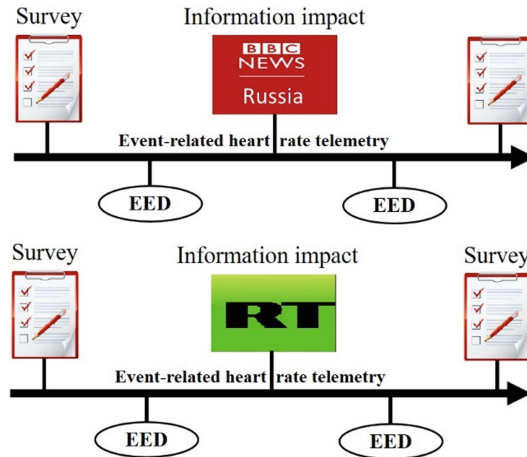


Fig. 1. Research design.

A total of 21 people from 19 to 36 years old volunteered for the experiment, with an average age of 24. In view of the fact that the largest audience of modern communication networks is young people, they were the desired sample group for this study. Among the participants there were 18 women (86%) 3 men (14%); 11 subjects were alumni (52%), 10 were students (48%); 8 people were psychologists (38%), and another 13 people of diverse professional background pursuing psychology degree as their second education (62%); 10 participants unemployed (48%), 11 employed (52%).

For the study, a survey was conducted to identify the level of inclination towards conservatism and liberalism in participants. Each pair of survey questionnaires (RT and BBC) included 11 statements. The respondents were to grade these statements on a five-point scale:

- strongly agree;
- agree;
- neither agree nor disagree;
- disagree;
- strongly disagree.

Each scale was graded in points. (−2; −1; 0; 1; 2).

The statements were of two types, the first type representing evident conservative views, the second liberal views. The sum of the positive and the negative points was calculated separately, then divided by 22 or -22, respectively, and multiplied by 100. There were 4 questionnaires where Questionnaire 1 reflected the questions of Questionnaire 2, and Questionnaire 3 reflected the questions of Questionnaire 4, thus making it possible to trace how the information impact affects the level of perception of the conservative and liberal views. 22 points arguably show that the individual has 100% inclination to conservatism at the time of the experiment; −22 points show inclination towards liberalism.

Comparing the results of Questionnaire 1 with the results of Questionnaire 2, as well as Questionnaire 3 with Questionnaire 4, one can trace how the level of inclination towards conservatism and liberalism changes after information impact.

2.2 Event-Related Heart Rate Telemetry

The event-related heart rate telemetry technology was used for monitoring and analysis of the HRV. The sequence of time intervals between R-R peaks (rhythmogram) is recorded using chest plastic electrodes. Primary signal processing and data transmission to a smartphone is carried out by the Zephyr Smart Heart Rate Monitor (HxM, Zephyr Technology) sensor platform via Bluetooth; special application Stress Monitor for Android OS (version 4.4 or higher) facilitates real-time monitoring and data transfer to a cloud server; visualization of rhythmograms, spectral analysis, and detection of stress episodes are implemented on the specialized Internet platform cogninn.ru [5].

A personalized analysis of the dynamics of autonomic regulation was carried out on the basis of HRV spectral indicators. Using the method of dynamic Fourier analysis with a window of 100s and a step of 10 s, the following indicators were calculated: the total power of the HRV spectrum—TP (ms²), characterizing the adaptive potential; spectrum power in the frequency range from 0.04 to 0.15 Hz—LF (ms²), characterizing the activity of the sympathetic nervous system in terms of heart rate modulation; spectrum power in the frequency range from 0.15 to 0.4 Hz—HF (ms²), characterizing the activity of the parasympathetic nervous system; the ratio of LF to HF is an index of autonomic balance that characterizes the tension of regulatory systems [6].

To evaluate the EED, the participant was asked to point the current estimation on a circular space. The boundaries are sets of synonymous adjectives that describe emotions in accordance with the modality (positive/negative) and the level of activity (tension/relaxation) in relation to four basic personal needs: a) security; b) independence; c) sense of achievement; d) unity (proximity). Depending on the position of the specified zone, the number of points scored by a person for each need is determined.

The average score indicates the degree of emotional maladjustment as follows: 0 points—no emotional maladjustment (physiological relaxation); 1 point—mild emotional maladjustment (physiological stress); 2—moderate emotional maladjustment (pathological stress); 3—strong emotional maladjustment (pathological relaxation) [7].

3 Results and Discussion

3.1 Assessing the Information Impact on Cognitive Attitudes

In the course of the study, an assessment of the levels of conservatism and liberalism was made before and after information exposure to RT and BBC news videos. After watching RT videos, the level of conservatism is decreasing; After watching BBC videos, the level of conservatism increases ($p < 0.05$) (Fig. 2).

We single out a number of factors that could affect the results of the experiment. In particular, the targeted age group may be characterized by certain distrust and skepticism towards the information offered [8]. RT, according to most experts [9, 10], expresses

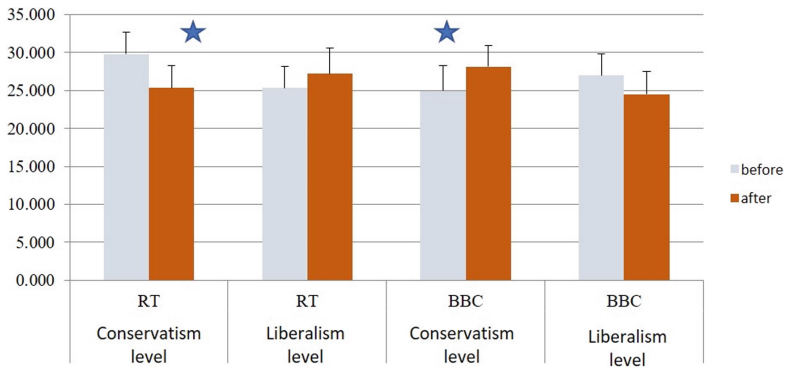


Fig. 2. Comparing conservatism and liberalism levels before and after information impact for RT and BBC; Significant differences indicated by asterisk (Wilcoxon signed ranks test) ($p < 0.05$).

a relatively conservative, pro-state views, which are normally critically assessed by the population. At the same time, BBC is a relatively liberal media, which is often criticizing conservative values in Russia [11–13]. A direct expression of such polar views, conservative or liberal, can lead to their rejection and thus result in shift in initial in the level of conservatism and liberalism in participants.

Such ambiguous results associated with changes in the level of conservatism and liberalism before and after the informational impact of various media, in our opinion, may be associated with the manifestation of the so-called network identity in the minds of young people [14]. The content-semantic content of the network identity is characterized by ideological-content multidimensionality and situationality. The diversity of the ideological and content level of network identity is manifested in the heterarchy of its value content, the absence of an unchanging value core and the lack of readiness to adhere to one general value in continuously reproducible practices of social and political behavior. Therefore, the commitment to state-centrism and conservative values prevails, while an individual with a network type of identity develops a counter intention to decentralize the value-normative content of the main meaning of the news story and actualize not conservative, but liberal values, and vice versa.

Readiness for such a levelling reaction of the individual's consciousness to information impact, in which some value is explicitly or latently expressed, generalizing the entire content-semantic component of the impact, opens up opportunities for reaching a limited consensus in society on a certain number of social and political issues. In this regard, one of the possible directions for research could be not just testing the hypothesis of a decentralizing (leveling) effect of networkization on identity by series of experiments assessing the level of conservatism—liberalism under the information impact, but also analysis of the psycho-physiological reaction of individuals to information messages expressing the idea of achieving a consensus in society on resolving certain socially significant issues.

The results of the experiment give a new insight into the understanding the young people political values in the context of networkization and the predominance of the so-called clip consciousness. The expression of a political value within the network

identity of an individual or a group is largely mediated by the specifics of the informational message, acquiring a situational character, which prompts us to question the role of political values for young people in general. Mimicking the nature of the Internet as a phenomenon, the political value picture in an individual with a network identity acquires an interactive, changeable, and unstable character, where the manifestation of the content of this personal picture is often aimed at overcoming the dominance of a certain value content within the information message sent to the individual. However, these conclusions are to be tested by a series of future experiments to allow a deeper understanding of the nature of the political views in a modern individual.

3.2 Assessing the Information Impact on the Level of Emotional Maladjustment

An assessment of emotional maladjustment level was made before and after information exposure to RT and BBC news videos. After watching both RT and BBC videos the emotional maladjustment level increases ($p < 0.05$) (Fig. 3).

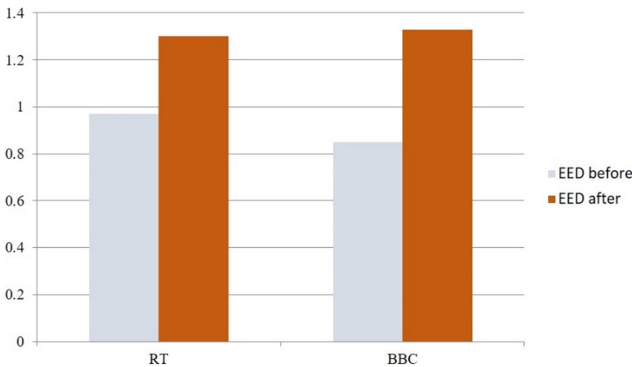


Fig. 3. Comparing emotional maladjustment level (EED) before and after information impact for RT and BBC.

An increase in the level of emotional maladjustment may be associated with the active and purposeful information impact of RT and BBC channels. In addition, despite the neutral nature of the news stories shown to the participants in the experiment, their psychophysiological reaction could be influenced by the name, logo and other data characterizing the mass media considered in the work. The opinion about a particular media that formed before the experiment could initially affect the final reaction of the subjects, however, in our opinion, the neutrally colored nature of the news story should neutralize the resulting effect of such an impact.

It is important to note that in the conditions of information overload, which is caused by the active use of modern Internet technologies, the potential of information to change the level of emotional maladjustment is questionable. At the same time, the experiment demonstrated the presence of the information impact of news videos on the individual. As noted previously by Evstifeeva and Tsurkan, the loss of reflexivity through psychological immersion in virtual cyberspace, in non-reflective virtual neuroreality, in the

ritual process of meaninglessness of meanings, in the “network path” necessarily entails the effect of dependence, the narrowing of individual consciousness [15]. The level of emotional maladjustment indicates certain experiences which, in turn, prove a certain response of the individual’s consciousness to the news given as stimuli. Such reaction may be explained by the nature of the topic of the news, where Russian athletes face obstacles before participating in the Olympics. This can arouse the interest of the individual at the cognitive-mental level, indicating the manifestation in his reaction of signs of civic identity, his association with the Russian state and the athletes representing it. The use of Russian symbols in the videos most likely became a factor that influenced the manifestation of signs of civic identity. And the symbols themselves can be considered as important markers that reveal the presence, or absence, of civic identity, as well as the degree of manifestation of signs of civic identity in the minds of individuals and groups.

3.3 Changes in Psychophysiological Characteristics as a Result of Information Impact, Taking into Account the Factor of the Presence/Absence of Higher Education

According to the individual data collected in the survey, we divided the participants in two groups. The first group included students without a degree (students), and the second group included the graduates (alumni). The HRV indices, the level of conservatism and liberalism, and the level of emotional maladjustment were compared for each group before, during and after the information influence by a BBC video and the RT video. The HRV indices only within normal physiological values were taken into account.

For the BBC video, the distribution is as follows: 10 participants (53%) are students, and 9 people (47%) alumni (Fig. 4).

For the RT video, 9 subjects (47%) are students, 10 (53%) alumni (Fig. 4).

The significant effects are associated with the factor of education (student, alumni) (ANOVA: $F = 5.9$; $p = 0.04$), context factor (BBC, RT) (ANOVA: $F = 4.09$; $p = 0.02$) and the stage of the experiment procedure (before, stimulus, after) (ANOVA: $F = 3.05$; $p = 0.02$). Any combination of these factors does not produce significant effects.

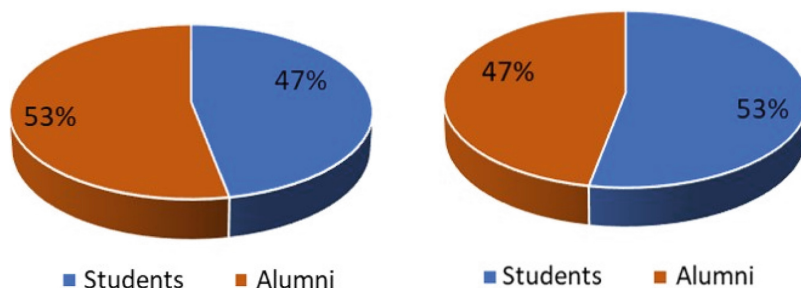


Fig. 4. Participant distribution by education level when watching BBC and RT news

TP in alumni is higher than in students (Fig. 5).

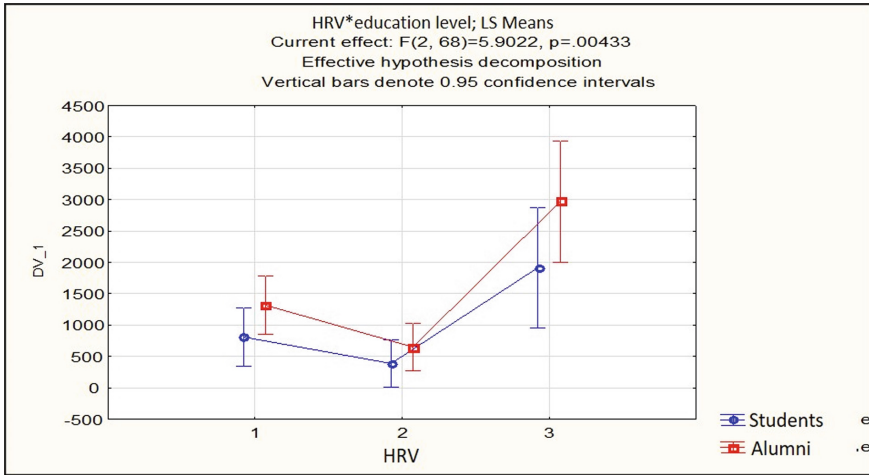


Fig. 5. Comparing the HRV measures in students and alumni before, during, and after information impact

The autonomic regulation indices associated with the stage of the experiment procedure were revealed only in alumni after watching RT video: LF and HR decrease (Figs. 6 and 7).

Thus, the decreasing activity of the sympathetic nervous system and, accordingly, the decrease in HR indicates that the RT had a calming effect on alumni. This nature of the impact can be explained by the fact that in the RT video, most likely, the symbols of the Russian state are more actively used, which are, in parallel, the determinants of civil, macropolitical and national identity. In the context of the use of the state-centric approach by the specified media, the use of such markers should not cause a serious perturbing effect on the activity of the sympathetic nervous system of the individual, i.e., such videos cause a positive, soothing effect on individuals with manifestations of signs of civic identity. Given that, according to public opinion polls, the older the respondent, the more likely he adheres to etatist-powerful values [16], and alumni are generally older than students. Therefore, it is not surprising that the RT video caused a calming effect on the activity of the sympathetic nervous system and a decrease in HR in alumni.

Before watching the videos, thanks to the results of the primary survey among the subjects, it is possible to distinguish people with a greater inclination to conservatism—liberalism. After the information impact, the inversion occurred only in 16% of participants. Among of those who changed attitudes were mostly students (67%), and it can be assumed that higher education contributes to greater stability of cognitive attitudes (Fig. 8).

Alumni are older than students. In this regard, it is also appropriate to assume that the age of the participants played a certain role in changing the cognitive attitudes. The older the individual, the more likely he is to have a more established set of political values and behavioral patterns, and vice versa. The degree of (un)stability of the value-normative set in the mind of an individual is also influenced by the presence/absence of

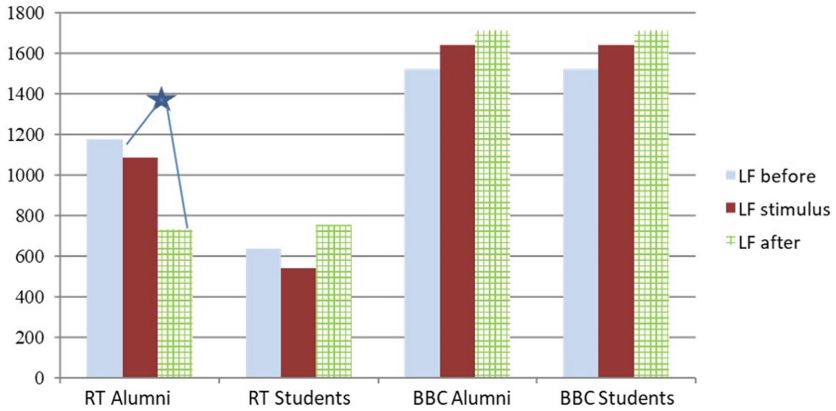


Fig. 6. Comparing the LF measures before, during, and after BBC and RT news in students and alumni; Significant differences indicated by asterisk (Wilcoxon signed ranks test) ($p < 0.05$)

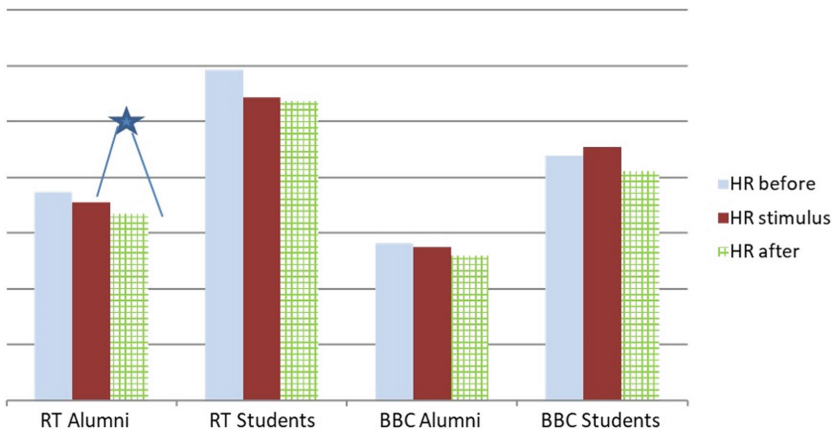


Fig. 7. Comparing the HR measures before, during, and after BBC and RT news in students and alumni; Significant differences indicated by asterisk (Wilcoxon signed ranks test) ($p < 0.05$)

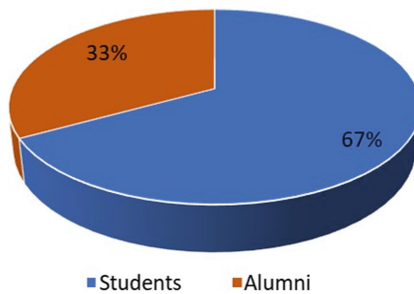


Fig. 8. The inversion of cognitive attitudes in participants by education level (students, alumni)

higher education. The presence of higher education often makes the value-normative set of the individual more stable. Moreover, younger people tend to use modern information and communication technologies more actively and, as a result, more intensively immerse themselves in the information field of network interactions. This, in turn, contributes, in our opinion, to a greater manifestation of the so-called network identity. As previously noted, the content-semantic content of network identity is characterized by ideological-content multidimensionality and is situational, which results in its ability to fluctuate under informational influence, manifesting in the change in cognitive attitudes [17–19]. However, this conclusion requires further using other media and different storylines. Consistent verification of assumptions made will allow a better insight into the manifestation of network identity among young people.

4 Conclusion

We assessed the influence of modern media on the cognitive attitudes of an individual using the news from RT and BBC, the two leading TV channels, as stimuli. The research design has been developed for measuring psychophysiological impact of shift in cognitive attitudes under external information influence, as well as the survey scale for determining the level of conservatism and liberalism attitudes.

Based on a group analysis, we conclude the following:

RT stimuli make the level of conservatism decrease; after watching BBC news, there is an increase in the level of conservatism. It is important to note that a number of factors affect the results of this experiment. In particular, most of the participants who took part in the study were students and recent graduates under 30 y. o., and may be characterized by distrust and skepticism towards the information offered. RT, according to most experts, expresses a relatively conservative, pro-government views, which is often critically assessed by the young. While BBC is a relatively liberal media, that often criticizes conservative values in Russia. A direct expression of a certain view, conservative or liberal, may lead to its repudiation, which may be reflected by the results in question.

We compared the level of emotional maladjustment and the degree of change in cognitive attitude. Both RT and BBC increase the level of emotional maladjustment in participants. Such dynamics may be associated with an active and straightforward information impact.




Comparing the students with alumni, judging by results of our experimentation, we see that for the alumni, the cognitive attitudes prove to be more stable. After the information impact, the inversion of the cognitive attitudes was observed in 16% participants only, 67% of which were students. RT videos inverted the attitudes towards liberalism. The inversion for BBC cannot be characterized by a specific tendency on a liberalism-conservatism scale. Only in alumni, RT news decrease LF and HR indices. Thus, the deactivation of the sympathetic nervous system after RT news shows its calming effect on alumni.

References

1. Kooi, B.W.: Modelling the dynamics of traits involved in fighting-predators-prey system. *J. Math. Biol.* **71**(6–7), 1575–1605 (2015). <https://doi.org/10.1007/s00285-015-0869-0>
2. Faugeras, O., Inglis, J.: Stochastic neural field equations: a rigorous footing. *J. Math. Biol.* **71**(2), 259–300 (2015)
3. Petukhov, A.Y., Polevaya, S.A.: Modeling of communicative individual interactions through the theory of information images. *Curr. Psychol.* **36**(3), 428–433 (2017). <https://doi.org/10.1007/s12144-016-9431-5>
4. Sebastian, G., Gaskell, G.M., Zwitserlood, P.: Stroop effects from newly learned color words: effects of memory consolidation and episodic context. *Front. Psychol.* **6**(278), 14 (2015)
5. Eremin, E.V., Kozhevnikov, V.V., Polevaya, S.A., Bakhchina, A.V.: Web service for visualization and storage of heart rate measurement results. Russian patent. Certificate of state registration of the database №2014621202 from 08.26.2014
6. McCraty, R., Shaffer, F.: Heart rate variability: new perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk. *Glob. Adv. Health Med.* **4**(1), 46–61 (2015)
7. Grigorieva, V.N., Tkhostov, A.Sh.: Method of assessment of emotional state of an individual. RF Patent RU 2291720 C1. Published 20.01.2007 in Patent Database no. 2
8. Kassikhina, V.E.: On legal education of younger generations. *State Law XXI* **2**, 23–28 (2016)
9. Solomatin, A.N.: Communicative strategies of RT (Russia Today). *Bull. Electronic Printed Media* **2**, 60–76 (2014)
10. Babayeva, S.: Free from morality, or what Russia believes in today. *Russ. Glob. Aff.* **5**(3), 34–45 (2007)
11. Hosseini, F.: BBC versus Euro news: discourse and ideology in news translation. *Russ. Linguistic Bull.* **3**(7), 128–132 (2016)
12. Subbotkin, V.D.: Global media and their influence on international policy (based on CNN and BBC). In: Vilshinskaya-Butenko, M.E. (ed.) *Collection of Works: Articles of the Institute of Business Communications*. Scientific Publication, Saint-Petersburg, pp. 49–54 (2017)
13. Prokofieva, V., Kostromina, S., Polevaia, S., Fenouillet, F.: Understanding emotion-related processes in classroom activities through functional measurements. *Front. Psychol.* **10**, 2263 (2019). <https://doi.org/10.3389/fpsyg.2019.02263>
14. Kaminchenko, D.I., Seliverstova, Yu.A.: Reflection of network identity of youth in the communicative space of social media. *PolitBook.* **3**, 6–27 (2020)
15. Evstifeeva, E.A., Tsurkan, D.A.: Philosophical-anthropological analysis of the clipped consciousness of youth. In: *Proceedings of Voronezh State University. Series: Philosophy*, vol. 3, issue 37, pp. 129–133 (2020)
16. Gorshkov, M.K.: Russian society and time challenges. Institut sotsiologii RAN. M.: Izdatel'stvo «Ves' Mir», 336 p (2015)
17. Petukhov, A.Y., Polevaya, S.A.: Modeling of cognitive brain activity through the information images theory in terms of the bilingual Stroop test. *Int. J. Biomath.* **10**(6), 1750092 (2017). <https://doi.org/10.1142/S1793524517500929>
18. Petukhov, A.Y., Polevaya, S.A., Gorbov, E.A.: Modelling the influence of RT and BBC on cognitive attitudes and psychophysiological indicators of individuals. *Opera Medica et Physiologica* **8**(3), 34–41 (2021). <https://doi.org/10.24412/2500-2295-2021-3-34-41>
19. Petukhov, A.Y., Polevaya, S.A., Yakhno, V.G.: The theory of information images: modeling based on diffusion equations. *Int. J. Biomath.* **09**(06), 1650087 (2016). <https://doi.org/10.1142/S179352451650087X>



Simulation Model of the Neurocognitive System Controlling an Intellectual Agent Displaying Exploratory Behavior in the Real World

Inna Pshenokova^(✉) , Kantemir Bzhikhatlov , Sultan Kankulov , Artur Apshev,
and Boris Atalikov

The Federal State Institution of Science Federal Scientific Center Kabardino-Balkarian Scientific
Center of Russian Academy of Sciences, I. Armand Street, 37-a, 360000 Nalchik, Russia
iipru@rambler.ru

Abstract. The paper presents the neurophysiological mechanisms of human exploratory behavior. The possibility of modeling such behavior in an intelligent agent based on a multi-agent neurocognitive architecture is shown. An autonomous software agent has been developed, in the control multi-agent neurocognitive architecture of which a system of intrinsic motivation is provided for the implementation of exploratory behavior. This behavior is aimed at forming the completeness of the functional representation of the fragments of the “intelligent agent - environment” system observed with the help of the agent’s sensors. Intrinsic motivation is an algorithm for stimulating an intelligent agent to perform certain behavioral programs by forming an internal stimulus (reward) to perform this program in a multi-agent neurocognitive architecture. It is shown that in the process of exploratory behavior, an intelligent agent immersed in a communicative environment forms a speech and information model of the world due to the dynamic formation of functional systems based on the cooperation of neural agents.

Keywords: Artificial General Intelligence · Intelligent agent · Multi-agent neurocognitive architectures

1 Introduction

Exploratory behavior is one of the fundamental forms of interaction between living beings and the real world, aimed at its study and cognition. In human activity, exploratory behavior displays indispensable functions in the development of cognitive processes at all levels. The concept of exploratory behavior is on a par with such fundamental concepts as learning, intelligence, creativity, forming an inextricable link with them [1].

An integral part of personality is the need for new information and knowledge. At birth, a child has certain instincts and tendencies that support the ability to survive. In the process of growing up, the child begins to explore the world around him, and his brain is stimulated by feelings as he learns and gets acquainted with life. Through exploratory behavior and curiosity, it develops and grows [2].

Although research is an elementary and fundamental form of understanding the structure of the world [3], little is known about the mechanisms and processes involved in it. Research does not directly reduce any physical need. This process is triggered by curiosity and unknown or obscure reward mechanisms [4]. There is a causal relationship between curiosity and exploratory behavior. Curiosity is a thought that motivates a person to engage in exploratory behavior leading to some outcome. This outcome may be negative or positive and may result in punishment and/or reward. Motivation refers to a goal, idea, or situation that induces action in response to reward or recognition. Rewards can be internal, such as gaining knowledge, which is considered a reward in itself, or external, which are more tangible and visible, such as winning a game or achieving a desired social status. Curiosity may be caused by innate, instinctive biological stimulation to survive, or it may have a secondary source that is cognitive and more to do with filling gaps in the person's knowledge.

Children in preschool exploring their environment are an example of the relationship between exploratory behavior and curiosity. A new environment filled with toys and other young children stimulates their curiosity. They begin to look around the environment, becoming familiar with who is in the environment, what the boundaries of the environment are, and what things they can safely interact with. They may ask many questions of other children and adults as they seek to understand the world around them, the differences between men and women, and the social norms they must learn.

While playing with toys and children, the child may exhibit affective exploration, which is determined by an emotional state or attitude. If the child becomes bored, he moves to diversified exploratory behavior, that is, to new activities that stimulate his senses and intelligence.

In this paper, we consider the neurophysiological mechanisms of human exploratory behavior and the possibility of modeling such behavior in an intelligent agent based on a multi-agent neurocognitive architecture.

The object of research is the neurophysiology of exploratory behavior.

The purpose of the work is to develop a simulation model of a neurocognitive control system for an autonomous robot that performs exploratory behavior in a real environment.

The task of the study is to develop an autonomous software agent, in the control multi-agent neurocognitive architecture of which a system of internal motivation to perform exploratory behavior is provided.

2 Material and Methods

Consider a simulation model of an autonomous software agent whose control multi-agent neurocognitive architecture provides for a system of intrinsic motivation to perform exploratory behavior. In works [5, 6], a multi-agent neurocognitive architecture is defined as a recursive cognitive architecture that allows agents and functional systems to be nested in each other. In [7], the control neurocognitive architecture of an intelligent agent was presented, which is an invariant of the organizational and functional structure of the intelligent decision-making process. An invariant based on a multi-agent neurocognitive architecture consists of software agents-neurons (agneurons) of varying

degrees of complexity that perform a sequence of mandatory operations: recognition of input patterns, emotional evaluation, goal setting, synthesis of an action plan, proactive modeling, plan execution management (see Fig. 1. at [7]). Each of these operations is performed on the basis of a multi-agent algorithm based on the exchange of messages between agneurons of various types. Since the system is recursive, each agneuron, in turn, consists of software agents-actors that also perform a sequence of mandatory operations. Communication between agents-actors is carried out in the same way as between agneurons, based on a multi-agent algorithm. The target function of agneurons is to find a path from the initial vertex of the graph of the problem situation, which describes the current state of the “intelligent agent-environment” system, to the final vertex, which describes some state of this system in the future, characterized by a higher value of the complex objective function of the intelligent agent. As a measure of activity and motivation, an abstract quantity – “vital energy” - was chosen as a measure of the viability of an intelligent agent, since we believe that its “life” continues only as long as a sufficient amount of energy is stored in it [8]. An agent can replenish his life energy only if he interacts with other agents to buy or sell the information he has. And such interaction is possible if, when searching for the optimal path in the decision tree, the agent exhibits exploratory behavior. An intelligent agent is immersed in a real environment with the help of sensors and effectors that provide an interface with users (social environment) who “talk” to him in natural language.

The system of intrinsic motivation means an algorithm for stimulating an intelligent agent to perform certain behavioral programs by generating an internal stimulus (reward) for the implementation of this program in the form of obtaining additional energy in a multi-agent neurocognitive architecture.

Consider the neurophysiology of exploratory behavior in the brain.

Like the brain of any highly developed animal, the human brain is curious, as the pursuit of new information is as important as food, sleep, or safety. Correct behavior is based on the orientation of nerve cells and neural networks of the brain, first of all, to the selection and analysis of new events and signals in the outside world. If something changes in the world, you need to detect it and take it into account in your reactions. During the evolution of the brain, the centers of curiosity formed a fairly complex and characteristic hierarchy. They occur in the midbrain, then in the diencephalon and in the cerebral cortex [9].

The first level of curiosity is to look at a new object and gather information. In the midbrain there is a zone—the quadrigemina, associated with the processing of new information and with the recognition of new events. The quadrigemina is an important integrative center that analyzes visual, auditory, and vestibular signals. In the anterior colliculi there are visual neurons that detect new events: movements in the field of view, the appearance of new objects. In the posterior colliculus of the quadrigemina there are auditory neurons that respond to the appearance of a new sound, the movement of the sound source, and a change in tone. After processing visual and auditory information, automatic responses to sensory signals are triggered, such as eye movements and head rotation in the direction of visual and auditory stimuli.

The nonspecific thalamus, midbrain reticular formation, and hippocampus contain neurons called novelty detectors [10]. These neurons compare sensory signals entering the brain with signals that were a few tenths of a second before.

After detecting a new signal, the quadrigemina, at the level of the midbrain, triggers an orienting reflex [11]. The orienting reflex is aimed at ensuring the optimal entry of new information into the brain and manifests itself in the form of a reaction of turning the eyes, head and whole body towards the new signal.

The next level of exploratory behavior is associated with active movement in space in order to make contact with a new object. Here the key role is played by the subthalamus, located on the border of the diencephalon and midbrain. The subthalamus is known as the motor region of the diencephalon. This zone contains the nuclei of the extrapyramidal motor system, which directs involuntary motor functions such as reflexes, locomotion, postural control, etc.

Signals from various centers of needs converge to the subthalamus. When some new events are detected, signals are sent to the subthalamus that start the research process. Gathering new information is very important. As much information as possible is collected about a new place or object. The question of which of them are useful and how to use them is decided later. The information is collected ahead of time. In the book [12], such work is called self-development programs. The cumulative approach to information is very significant because it allows you to adapt your behavior to a complex environment.

In an intelligent agent based on a multi-agent neurocognitive cognitive architecture, the function of the quadrigemina is performed by the cognitive node of recognition. The composition of agneurons of this cognitive block and their work are described in [14]. The main result of the recognition process is an emotionally colored event (represented as an agneuron-event and an agneuron of emotional evaluation) of the detection of some object, which is matched with the agneuron-object available in the multi-agent architecture of an intelligent agent (Fig. 1). The figure shows a neurocognitive reflection of the context and content of the dialogue, in which the user explains to the intelligent agent that the object observed by both of them in the external environment is a ball. To this end, this user responding to a question from an intelligent agent says: "This is a ball".

The dark stripes of Fig. 1 represent agneurons, while the light stripes show messages from them. The direction of "movement" of messages—from the agneuron-sender to the agneuron-receiver is indicated by a dotted arrow. The order of actions for sending messages is indicated by a number in a circle. Various agneurons located, respectively, in different functional layers, are depicted in the figure with pictograms. For example, triangles are sensors and effectors, circles are object agneurons, hexagons are action agneurons, shamrocks are event agneurons, pointed flags are goal-setting agneurons, and wide slanted arrows are action control agneurons.

An analogue of the work of neurons-detectors of novelty is the process of establishing cause-and-effect relationships between event agneurons. When an intelligent agent receives some information about its current state at the input, conceptual agneurons (4, 6, 8, 10th layer of Fig. 1) are formed in the architecture, which state the occurrence of this event. This information enters the input to event-type agneurons (layer 13) and

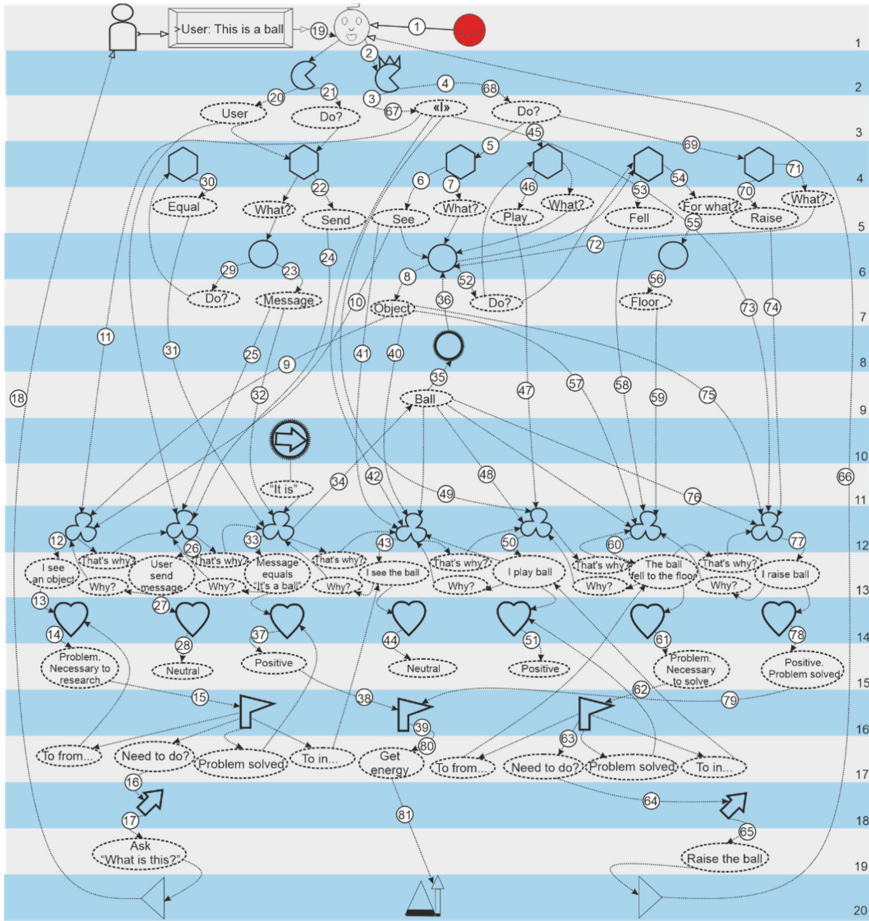


Fig. 1. Scheme of a multi-agent algorithm for motivating exploratory behavior.

is presented as a multi-agent fact. To establish its cause or effect, event agneuron by mass mailing messages like “Why?”, “Therefore?” requests from other agneurons of this type information about the previous and target events (formed using the goal-setting functional unit). Agneurons of an event that fired earlier than the current event responds to the “Why?” request. And the agneurons of the event, which are motivated to receive energy and are in search of their cause, respond to the request “Therefore?”. If there is no consequence event in the system, it is formed by polling emotional (layer 14) and event agneurons about the desired assessment and target event. After such a distribution between event agneurons, a set of multi-agent connections (contracts) is formed and information about the environment is collected. The question of which ones are useful and how to use them is decided by the value of the correlation coefficient. The correlation coefficient is calculated as the ratio of the number of positive triggers of counterparty events to the total number of events. The correlation coefficient is recalculated according to the following algorithm. If some input event corresponds to several consequence

events, the agneurons responsible for these consequences signal to receive a reward. When the agneuron responsible for the cause event announces a reward, it finds its counterparties, then modeling agneurons are formed containing the cause event and the corresponding effect event. As a result of experiments testing these solutions, the agent is trained and the degree of correlation of the acquired knowledge is assessed. With an increase or decrease in the correlation coefficient, there is a dynamic strengthening or breaking of links between agents. Disconnection means that the current event is not a consequence or cause of the event with which the multi-agent contract was concluded.

If the event-cause has received an emotional assessment in the direction of increasing its own energy, the goal-setting cognitive node (layer 16) is activated, in which the goal-setting agneuron is formed. This agneuron contains knowledge about the current event and its assessment, the desired assessment and the corresponding end event, as well as the planning horizon.

In this case, the situation of the context of the interactive interaction of an intelligent agent with a user is reflected in several events at once, recorded by a multi-agent neurocognitive architecture:

I see the object [Therefore]
User send message [Therefore]
Message is "This is a ball" [Therefore]
I see the ball...

Figure 1 shows that between the agneurons that perform the functional representation of events recorded by the multi-agent neurocognitive architecture, there is an increase in connections aimed at reflecting the cause-and-effect relationships between events that describe the context of the situation, the subject of the statement, and the statement itself. Thanks to these links, the agent understands that the object in his field of vision is a ball.

The third level of exploratory behavior is already associated with the manipulation of certain objects, when we take an object and not just look at it, but try to break it, unwind it, see what is inside. Manipulation with objects is a very important component of our mental (not just motor) activity. A small child masters these manipulations, including as a source for collecting new information. This activity is carried out by the frontal cortex, in which the premotor and motor zones are distinguished. The first formulates the movement program as a whole, the second is responsible for the contraction of specific muscles [15].

The motor and premotor areas in the neurocognitive architecture are formed by the functional nodes of the synthesis of the action plan, proactive modeling, and control over the implementation of the plan. On these layers of the architecture, action agneurons are formed (the 18th layer of Fig. 1), which transmit information about the actions that must be performed to achieve the target state, determined in the process of multi-agent interaction of event, emotional and target agneurons (Fig. 1). The algorithm of actions is passed to the input of the control agneurons, which in turn pass this information to the effectors of the intelligent agent for execution.

All levels of exploratory behavior are supported by centers of positive emotions. At the neurochemical level, dopamine is responsible for this. Every time a person learns

something new, he experiences a dopamine surge, depending on the degree of novelty and the significance of the information. That is, novelty is an important source of positive emotions. Due to positive emotions, the brain pushes the recognition of the world, to the collection of new information, to the formation of adaptive behavior. Energy serves as a system of intrinsic motivation for performing exploratory behavior for an intelligent agent. At all levels of the neurocognitive architecture, in exchange for the information they have, agents receive energy from counterparties, thus “prolonging” their lives. In turn, this coordinated behavior of agneurons allows the intelligent agent to reach the target state, colored by a positive intellectual assessment, which leads to the acquisition of new knowledge and additional energy from the user or other intelligent agent.

The highest level of exploratory behavior is the formation of a speech or information model of the environment. As we get to know the world, we learn about the existence of words that denote certain objects. It occurs in the association parietal cortex. Thus, the frontal cortex is responsible for movement, and the associative parietal cortex is responsible for speech, at the level of remembering words, their meanings and thinking. We accumulate words that denote objects, actions, signs, but they are not accumulated one by one, but arise in the form of a connected information network. Each word is associated with many other words and associations. In the process of growing up and accumulating knowledge, a speech model of the external world is formed, with the help of which a person thinks. It is a separate source of novelty and positive emotions, since with the help of this model it is possible to form new knowledge, concepts and associations.

According to Fig. 1, in the process of exploratory behavior, an intelligent agent forms a speech model of the world in the form of conceptual agneurons of various types, which are combined into multi-agent facts in the form of event agneurons, which in turn form a connected information network. This is possible due to the fact that the exchange of messages between agneurons, intelligent agents and communication with the user takes place in natural language. In the process of such an exchange, the ontology of the “intelligent agent-environment” system is formed. The works [16, 17] describe the capabilities of a multi-agent neurocognitive architecture for the synthesis of statements and speech recognition.

3 Theory and Calculation

To conduct experiments on training control neurocognitive architectures in exploratory behavior, a prototype of an autonomous agent control system was developed, which has two information processing channels: visual (video camera) and verbal (keyboard). The autonomous agent works in interaction with the user to identify the object that they observe in the external environment.

Figures 2 and 3 show a three-dimensional image of some parts of this architecture. The architecture is built in the graphic visualization subsystem of the developed software package.

Figure 2 shows conceptual agneurons objects and actions.

Figure 3 shows agneurons-events and their corresponding agneurons of emotional evaluation.

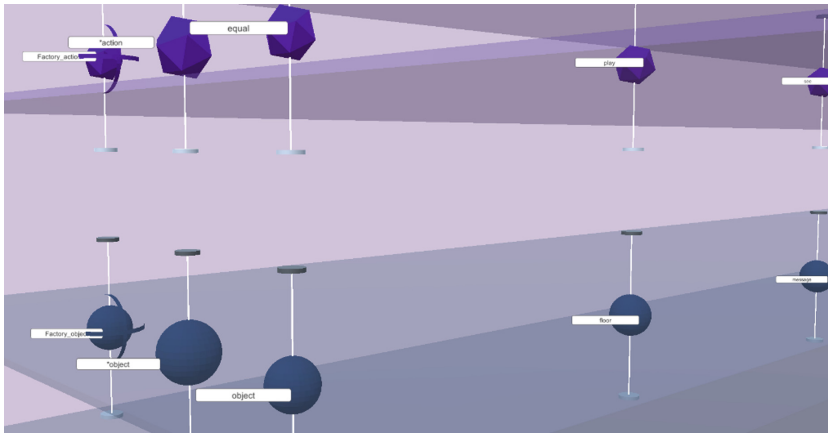


Fig. 2. Multi-agent neurocognitive architecture of an intelligent agent (layers 4, 6).

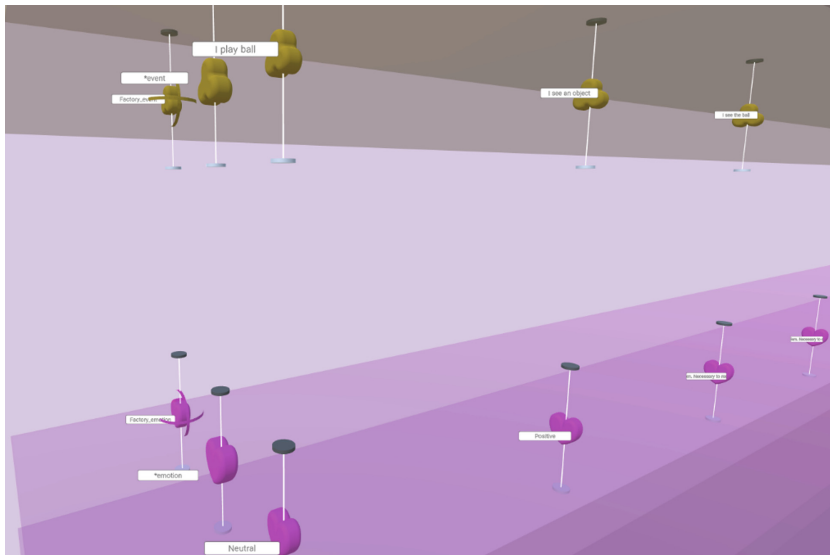


Fig. 3. Multi-agent neurocognitive architecture of an intelligent agent (12, 14 layers).

In the presented architecture, a functional representation of communication subjects is performed, as well as messages received by an intelligent agent.

As a result of the experiment, a system of functional representations of the object by the “Ball” agneuron, events by agneurons “I see an object”, “User send a message”, “Message is “This is a ball”, “I see a ball”, “I play a ball”, “The ball fall to the floor”, “I pick up the ball”, actions “See”, “Send”, “Play”, “Fall”, “Raise” and cause-and-effect relationships between them. These connections reflect the experience of interaction of an intelligent agent with the environment in which he is immersed. In particular, the

belief of an intelligent agent that in the system “intelligent agent - environment” the events: “*I see an object*”, “*The user sent a message*”, “*Message is “This is a ball”*” are a consequence of the event “*I see a ball*” means that the object he observes in the outside world is a ball. And the event “*The ball fell to the floor*”, with a negative emotional assessment, is the cause of the action “*Raise the ball*”, which, in turn, is the cause of the event “*I play ball*”, with a positive emotional assessment. Therefore, in order to receive positive emotions, an intelligent agent needs to perform the “*Raise the ball*” action, which is fed to the effectors for execution.

Based on the results of the experiment, it was concluded that a multi-agent algorithm for motivating exploratory behavior based on the growth and development of neurocognitive architectures can be used to model the neurophysiological mechanisms of human exploratory behavior in intelligent systems.

4 Conclusions

The main result of the work is the conclusion that the approach to the design of intelligent systems based on multi-agent neurocognitive architecture is able to provide simulation of the neurophysiological mechanisms of human exploratory behavior.

An autonomous software agent has been developed, in the control multi-agent neurocognitive architecture of which a system of internal motivation is provided for displaying exploratory behavior aimed at forming the completeness of the functional representation of fragments of the agent-environment system observed with the help of the agent’s sensors. One of the essential features of a holistic functional representation is a natural language description of the current situation, thanks to which the agent forms an interconnected information network “intelligent agent-environment”. The agent extracts this information from dialogues with other agents (users, software agents and robots) in the communicative environment.

Acknowledgements. This work was supported by the Russian Science Foundation grant no. 22-19-00787.

References

1. Poddyakov, A.N.: Methodological foundations for the study and development of research activities. *Res. Activities Students Modern Educ. Space* **3**, 51–58 (2006)
2. Eyestad, G.: *Self-esteem in Children and Adolescents: A Book for Parents*. Alpina Publisher, Moscow (2014)
3. Birke, L.I., Archer, J.: Some issues and problems in the study of animal exploration. In: *Exploration in Animals and Humans*, pp. 1–21 (1983)
4. Becker, J.B., Meisel, R.L.: Neurochemistry and molecular neurobiology of reward. In: *Handbook of Neurochemistry and Molecular Neurobiology: Behavioral Neurochemistry, Neuroendocrinology and Molecular Neurobiology*, pp. 739–774. Springer, US (2007)
5. Nagoev, Z.V.: *Intelligence, or Thinking in Living and Artificial Systems*. KBNTs RAS Publishing House, Nalchik (2013)

6. Nagoev, Z.V.: Multiagent recursive cognitive architecture. In: *Biologically Inspired Cognitive Architectures 2012: Proceedings of the Third Annual Meeting of the BICA Society, AISC*, vol. 196, pp. 247–248. Springer, Heidelberg (2013)
7. Nagoev, Z., Pshenokova, I., Nagoeva, O., Sundukov, Z.: Learning algorithm for an intelligent decision-making system based on multi-agent neurocognitive architectures. *Cogn. Syst. Res.* **66**, 82–88 (2021)
8. Nagoev, Z., Nagoeva, O., Anchokov, M., Bzhikhatlov, K., Kankulov, S., Enes, A.: The symbol grounding problem in the system of general artificial intelligence based on multi-agent neurocognitive architecture. *Cogn. Syst. Res.* **79**, 71–84 (2023)
9. Dubynin, V.: *The Brain and Its Needs. From Nutrition to Recognition*. Alpina Non-fiction, Moscow (2020)
10. Vinogradova, O.S.: *Hippocampus and Memory*. Nauka, Moscow (1975)
11. Sokolov, E.N.: Nervous model of stimulus and orienting reflex. *Questions Psychol.* **1**, 61–73 (1960)
12. Simonov, P.V.: *Lectures on the Work of the Brain. Need-information Theory of Higher Nervous Activity*. Institute of Psychology RAS, Moscow (1998)
13. Aleksandrova, Yu.I.: *Psychophysiology*, 4th edn. Peter, St. Petersburg (2014)
14. Nagoev, Z., Pshenokova, I., Bzhikhatlov, K., Kankulov, S., Atalikov, B.: Multi-agent neurocognitive architecture of an intelligent agent pattern recognition system. In: Corchado, F.F.R., Samsonovich, A.V. (eds.) *Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: The 13th Annual Meeting of the BICA Society 2022*, *Procedia Computer Science*, vol. 213, pp. 504–509. Elsevier, Amsterdam (2022)
15. Ramachandran, V.S.: *Encyclopedia of Human Behavior*. Academic Press, Cambridge (2012)
16. Nagoev, Z.V., Nagoeva, O.V.: Justification of Symbols and Multi-agent Neurocognitive Models of Natural Language Semantics. KBNTs RAS Publishing House, Nalchik (2022)
17. Makoeva, D., Nagoeva, O., Gurtueva, I.: Formal representation of natural language elements in multi-agent system based of self-organization of distributed neurocognitive architectures. In: Corchado, F.F.R., Samsonovich, A.V. (eds.) *Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: The 13th Annual Meeting of the BICA Society 2022*, *Procedia Computer Science*, vol. 213, pp. 631–635. Elsevier, Amsterdam (2022)



The Embodied Intelligent Elephant in the Room

Saty Raghavachary^(✉)

University of Southern California, Los Angeles, CA 90089, USA
saty@usc.edu

Abstract. The central point made in this paper is this: human-level grounded meaning in an agent can only result from directly experiencing the world, which in turn can only be possible via embodiment (coupled with ‘embrainment’ - a suitable brain architecture). Via embodiment, we humans are able to internalize our direct interactions the world, in addition to being able to associate symbols with them—this allows us to communicate via symbols, thereby externalizing our individual representations for mutual, collective benefit. Lacking embodiment, in contrast, today’s AI agents are able to only operate at a derivative, symbolic level, without being able to experientially understand, or relate to, the meaning behind the symbols they process. The only way to enable artificial agents to acquire ‘first-hand’ meaning for symbols, would be to provide suitable embodiment for them. Also, a case is made for machines capable of exhibiting non-symbolic intelligence via analog embodiment, to complement symbol-originated intelligence.

Keywords: AGI · Artificial General Intelligence · Artificial Intelligence · analog intelligence · physical intelligence · non-symbolic intelligence · embodiment

1 Introduction

‘BICA is a transdisciplinary study that aims to design, characterise and implement human-level cognitive architectures’, as per Wikipedia [1]. The BICA Society’s ‘challenge’ to researchers is to come up with cognitive architectures that are far more general and capable compared to much of the ‘narrow’ AI today, by basing such architectures on biology—after all, biological brains have been honed by evolution to enable their hosts to survive and reproduce. A variety of BICA architectures have been proposed over the years, eg. as documented in [2].

In this brief paper, it is argued that physical embodiment is essential for realizing a crucial aspect of robust cognition: grounded meaning. We start with some background material related to natural phenomena, and the notion of a symbol. After that, we compare intelligence capabilities of non-human animals, humans, and disembodied AI systems. Finally, we consider the prospect of embodied AI, relating it to biological embodiment—what advantages it could provide, and how it could co-exist with its disembodied counterpart; we also consider entirely analog embodiment, and how it could lead to agents that display non-symbolic intelligence.

2 Phenomena

Phenomena are ubiquitous in nature—behavior arising from the interplay of matter, energy and information. From pulsating quasars to subatomic particles, matter at every spatial scale undergoes phenomena; likewise in the time dimension, from terahertz vibrational modes of molecules to evolution of continents and even galaxies, phenomena span the range. In between are phenomena more common to us - vibrations of a guitar string, water boiling in a pot, light refracting through a prism, heat conduction, fracture, capillary action and a whole lot more.

‘Engineering’ could be defined as the practice of exploiting phenomena for useful purposes. Even the simplest hourglass that ‘measures’ time via gravity-driven fall of tiny sand particles, or the windup timer that ‘counts down’ time by releasing energy stored in its spring, are examples of useful devices that display basic intelligence, directly on account of their structures and phenomena.

Living organisms can be said to be comprised of structures and associated phenomena that help keep the organism alive (survival) and multiply (reproduction). Biological intelligence could similarly be considered to arise from biological structures, whose phenomena help the organism display intelligent behavior.

3 Symbols

Symbols do not exist in the physical world on their own, the way matter does for example. That is because, symbols (used in language, math, musical notation, sports, games, engineering drawings, architectural blueprints and many other fields) are entirely human-originated.

Symbols have only shared meaning—whether they are squiggles of ink (language, music...), gestures/postures, smoke color or drum rhythms, they carry no inherent meaning, and often need context to be understood properly. Symbols are potent—eg. Math is best way we have to quantify and characterize natural phenomena, digital computation offers calculations that are orders of magnitude bigger in terms of size and speed, etc.

It is important to realize that all symbols are proxies—we can model nature (eg. Gravitational force, light refraction, radioactivity...) using mathematical physics formulae, but it does not mean that nature “uses math”. Our numerical simulation of metal solidifying could perfectly model reality, but it does not mean the atoms are using math to determine their cooling rate. Phenomena do not involve math/numbers explicitly! This last point is going to be useful in Sect. 5 below, when we argue for embodiment, especially the analog variety. For completeness, it is worth noting that words are proxies too—they describe and characterize things and actions (adjectives, adverbs, nouns, verbs), intangibles (eg. Emotions, death, procrastination) and more—but without shared understanding, they are inherently meaningless.

4 Embodied Animals (Including Humans), Disembodied AI

Stating the obvious, every biological brain, from the small worm’s to a mighty whale’s, is embodied—which makes sense, considering its evolutionary purpose is to protect the body it inhabits, and to help it procreate. As Rodney Brooks puts it, ‘Elephants Don’t

Play Chess' [3]—in other words, intelligence is not just about abstract manipulation of symbols that are not grounded in reality.

Animal brains most certainly vary in complexity—eg. The simple *C. Elegans* nematode (worm) has a mere 302 neurons, whose connectome has been fully mapped [4]; in contrast, the human brain is estimated to contain about 86 billion neurons and almost that number of non-neuronal cells—with almost a trillion connection between them.

Regardless of brain complexity and capacity, all brains sense their immediate environment directly and physically—eg. gravity, heat, vibration etc. are felt (sensed) by bodies, which convey appropriate signals to the brain so that it can perceive the environment, and act accordingly (eg. keep the body upright, move away (or towards) the heat source, etc.). Such direct considering and responding is a form of intelligence [5, 6] that occurs in a symbol-free manner—in lower animals there is simply no brain capacity to explicitly process symbols, and in higher animals including humans, this could be automatic/subconscious behavior, instinctual or reflexive. Additionally, in higher animals (eg. pets, humans), the higher brain capacity leads to the formation of memories of past experience, ie. memories of direct contact with the environment, including interactions with other animals. 'Experience' could be considered as the internal representation (memory), via brain structures and phenomena, of the animal's 'experiencing' the external world via structures and phenomena. Such 'physical' intelligence is hypothesized to be non-symbolic. For instance, a group of *Drosophila* (fruit-fly) neurons have been shown to perform goal-directed navigation [7] equivalent to performing vector operations (involving heading angles and traveling directions), without explicitly performing those operations. A cat estimates (often accurately) whether it can jump across a barrier several times its own height, presumably without explicitly computing the distance and comparing that with a threshold numerical value. We humans can look at two pairs of points on paper, one pair close, the other pair farther, and instantly and accurately tell which pair contains the closest point—we do not do so by imposing a Cartesian coordinate system, computing Euclidean distances and comparing them. The brain does seem wired for math [8] but that is not to say that we explicitly use calculations every time we estimate something.

We humans do have the facility to manipulate symbols in our brains—language production and processing offer the best evidence for this. We are also able to perform basic non-linguistic symbol manipulations, eg. Math operations, logical reasoning etc. We can multiply small numbers in our heads (eg. 12 times 15) but not big ones (eg. 17745 times 48005); we can compute the square root of small whole numbers (eg. 225), but not 226, although we can reason that it would be slightly higher than 15; we can mentally reverse a small list of color names, but not a list with 100 entries. Also, we seem capable of implicit symbol manipulation, ie. Fast 'system 1' thinking, and also, slower 'system 2' thinking that involves deliberate, sequential, step-by-step symbol processing [9].

Our mental prowess with symbols aside, we do use symbols to utmost advantage—we have centuries worth of collected works of literature, science, math, art, music etc., where internal individual experience has been externalized via symbols which can then be shared. Such symbol-oriented communication is precisely what has led to human civilization's astonishing advance in the world—related to food, clothing, shelter, medicine,

technology, the arts, etc. We use symbol-derived scientific instruments and devices ranging from telescopes to particle accelerators to many in between, to deepen our understanding of nature, leading to an ongoing virtuous cycle.

At the other end of the triad is the digital computer—that uses the von Neumann architecture, to be specific. This architecture, based on binary numbers, realizes symbol manipulation that is inherent in all algorithmic computation. Higher level programming languages encode such explicit computations using data structures (in contrast to natural phenomena that involve no explicit computations on account of resulting from physical structures). There is no facility to represent ‘experience’ directly (ie. Non-symbolically) in memory, the way higher animals’ brains can. This is the source of the ‘symbol-grounding’ problem that has plagued AI all along—the symbols that the machine manipulates so skillfully (quickly, accurately, tirelessly) have no meaning for the machine, since the machine lacks the ability to “internalize” it by relating it to its stored prior experience.

To summarize, we have three cases (Fig. 1 that follows):

- lower animals (simpler brains) only process direct experience
- humans process direct experience as well as symbols, and are able to fluidly ‘map’ one to the other
- disembodied digital computers only process symbols

Figure 1 shows the distinction we are making: between embodied agents that can utilize both phenomena as well as externalized symbols from other agents to build up experience, versus disembodied ones that can only process symbols input to them (and are therefore unable to acquire experience that can help ground symbols).

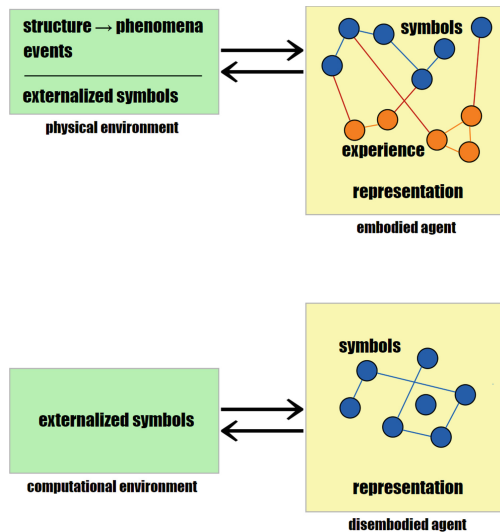


Fig. 1. Embodied vs disembodied agents

The discussion thus far leads to an interesting question—what about embodied AI? In other words, what additional benefits would occur from building an agent that has physical form?

5 Embodied AI

AI would be considered embodied, if it is expressed via an integrated body + brain architecture where the body is used for sensing and actuating, with the brain doing the intermediate processing.

Immobile robots (eg. a welding robot) would be a simple example, as would a pan-tilt-zoom (PTZ) camera mounted on a wall—the processor attached to such devices are able to perform intended sensing and actuating.

Mobile robots, including humanoid ones, can accomplish more—they have ‘agency’, ie. The ability to experience the world directly, physically, interactively and continuously, and possibly build up an internal model of their world (eg. a vacuum cleaner robot or a lawn-mowing one would build a localized map of its environment, including obstacles and perimeter). But, could this be considered ‘experience’? Not if it is based on explicit symbolic representation, as per our prior discussion!

An embodied agent in a VR world, or one in a physical world that is designed around a digital processor at the core (rather than peripherals), cannot be said to experience the world, on account of the symbol processing that occurs. This is a philosophical point, but it does help clarify, for example, why an agent that stubs its toe in VR, or a digital robot embodiment, cannot ‘really’ feel pain—there is no experience/memory, no perception—instead there is explicit calculation, which makes it a simulation of reality. VR worlds have another big limitation: they need ever-increasing storage and computation, to track the state of the environment and agents in it, eg. The damage they might sustain. Conversely, abstracting away such ‘needless’ details would result in a simplified world where the agents function as intended, but fail in unanticipated ways when transplanted into the real world (eg. this could be a problem that might plague self-driving cars (SDCs) being trained virtually).

In other words, embodied (whether mobile or immobile) robots, if based on the usual digital computational architecture using CPUs or GPUs or ML inference chips, are incapable of acquiring grounded meaning. eg. an SDC that processes LIDAR point cloud data or video, and even compares it to prior drives or other people’s drives, still cannot be said to have grounded meaning about the drive. This is not to say that such embodiment is not useful—far from it! That said, possessing grounded meaning is only bound to improve their behavior. Grounded meaning could result from analog embodiment.

By definition, analog embodiment explicitly avoids the use of symbol-processing at its core. We emphasize the ‘core’ part because, the peripherals could include digital processing. For example, we humans use myriad digital devices, digital assistive technologies, meditation apps on our smartphones—all of which involve explicit or implicit digital-to-analog conversion so that our brains process the outputs as usual (sounds, sights); neural prosthetics and seeing-eye cameras go a step further, by directly interfacing with the brain and retina respectively, in terms of analog (not digital) signals.

Analogously, our analog embodiment agent would be presumed to be able to ‘experience’ the world directly (or more accurately, negotiate its environment without explicit symbol-processing).

A human-level analog embodiment is admittedly quite far-fetched right now—we do not fully understand how brains work (eg. How memories are formed, concepts and experience are represented and accessed, exactly how emotions affect perception, and much more); we also do not have appropriate materials, structures and mechanisms that mimic how the body functions. That said, there is research in progress along these lines, including BCI (brain-computer interfacing), organoid intelligence [10], electronic skin, etc.). If and when realized, such analog embodiments might be able to explore the world directly and physically to build up individual experience, and, based on possibly complex brain architecture, feel emotions, feel pain, develop a sense of self, and consciousness [11].

Equally far-fetched would be an SDC body that builds up ‘personal experience’ related to its world: road types, potholes, weather conditions, blinding oncoming lights etc., in order to react, make decisions, and drive, more like a human driver would (but with infallible memory, attention etc. that would be beyond what a human driver could muster).

To get there from here, we could start small, and build up. By way of motivation, here are two whimsical (but useful) past examples. Braitenberg vehicles [12] are interesting thought experiments, involving the behavior of one or more simple vehicles that contain nothing but motors and sensors (specifically, there is no processor that computes what the vehicle does). It is fascinating that complex behaviors (which could be described using terms such as ‘love’, ‘fear’ ‘aggression’) result from the interactions between a vehicle and its environment, and with other vehicles. Also, Rube Goldberg [13] was a cartoonist who drew elaborate mechanisms involving plausible structures and devices (ropes, balls, buckets, candles...), which when connected together, would accomplish a rather trivial task (eg. a napkin that wipes a face). In each cartoon, the ‘mechanism is the computer’, there is no explicit symbol manipulation.

There is new work reported in the emerging field of soft robotics, where a liquid elastomer rod rolls and wiggles on a warmer-than-air surface, thereby escaping a simple maze [14]—an example of physical intelligence that does not involve a digital processor or computing. The variety of real-world phenomena should provide ample inspiration and direction, for the design and construction of analog robots, hopefully of increasing complexity over time. Many such mechanisms are found in living beings, eg. Birds’ curved beaks, maple seeds’ spinning wing, etc.—these too should be able to give us plenty of biomimetic ideas for embodied agents of the analog kind.

Also, work is underway to formalize such non-von-Neumann architectures as ‘structuring of processes’ [15].

6 Conclusions

Embodiment (a physical body, together with an appropriate brain that is integrated with it) is shown to be an essential requirement for an AGI system that needs to exhibit broad, grounded intelligence, rather than intelligence that results solely from symbol-processing. This in turn follows from the hypothesis that a body + brain is what allows


direct (non-symbolic) representation of what is sensed/experienced, along with subsequent symbolic equivalent representations for them, and allows for interchange between the two: such interchange is what provides grounded meaning for symbols, where the grounding specifically comes from the experiencing. Even at a basic level, simple analog embodiment can lead to various types of non-symbolic intelligence that arise solely on account of physical structure and phenomena, in contrast with intelligence that results from digital computation that explicitly manipulates symbols.

References

1. Wikipedia Contributors: Biologically Inspired Cognitive Architectures. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Biologically_inspired_cognitive_architectures&oldid=1050259691. Accessed 8 Sept 2023
2. Samsonovich, A.V.: Toward a unified catalog of implemented cognitive architectures. *BICA* 195–244 (2010)
3. Brooks, R.A.: Elephants don't play chess. *Robot. Auton. Syst.* **6**, 3–15 (1990)
4. White, J.G., Southgate, E., Thomson, J.N., Brenner, S.: The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philos. Trans. R. Soc. B* **314**, 1–340 (1986)
5. Raghavachary, S.: Intelligence—consider this and respond! In: Samsonovich, A.V., Gudwin, R.R., Simões, A.S. (eds.) *BICA 2020*. AISC, vol. 1310, pp. 400–409. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-65596-9_48
6. Raghavachary, S.: A physical structural perspective of intelligence. In: Klimov, V.V., Kelley, D.J. (eds.) *BICA 2021*, SCI 1032, pp. 419–427 (2022). https://doi.org/10.1007/978-3-030-96993-6_46
7. Lyu, C., Abbott, L.F., Maimon, G.: Building an allocentric travelling direction signal via vector computation. *Nature* **601**(7891), 92–97 (2021). <https://doi.org/10.1038/s41586-021-04067-0>
8. Grace, R.: Arithmetic has a Biological Origin—It's an Expression in Symbols of the 'Deep Structure' of Our Perception. <https://theconversation.com/arithmetic-has-a-biological-origin-its-an-expression-in-symbols-of-the-deep-structure-of-our-perception-211337>. Accessed 9/8/23
9. Kahneman, D.: *Thinking Fast and Slow*. Farrar, Straus and Giroux, New York, NY (2011)
10. Smirnova, L., et al.: Organoid intelligence (OI): the new frontier in biocomputing and intelligence-in-a-dish. *Front. Sci.* **1**, 1017235 (2023). <https://doi.org/10.3389/fsci.2023.1017235>
11. Chella, A., Cangelosi, A., Metta, G., Bringsjord, S. (eds.): *Consciousness in Humanoid Robots*. Frontiers Media, Lausanne (2019). <https://doi.org/10.3389/978-2-88945-866-0>
12. Braitenberg, V.: *Vehicles: Experiments in Synthetic Psychology*. MIT Press, Cambridge (1984)
13. Wolfe, M.F., Goldberg, R.: *Rube Goldberg: Inventions!* Simon & Schuster, New York (2000)
14. Zhao, Y., et al.: Physically intelligent autonomous soft robotic maze escaper. *Sci. Adv.* **9**(36). <https://doi.org/10.1126/sciadv.adi3254>
15. Jaeger, H., Noheda, B., van der Wiel, W.G.: Toward a formal theory for computing machines made out of whatever physics offers. *Nat. Commun.* **14**, 4911 (2023). <https://doi.org/10.1038/s41467-023-40533-1>



Approaches to Modeling Autonomous Agents with Scientific Abilities

Vladimir G. Red'ko^(✉) 

Scientific Research Institute for System Analysis of the Russian Academy of Sciences, 36/1
Nakhimovskii Prospect, Moscow 117218, Russia
vgredko@gmail.com

Abstract. An analysis of the methods of scientific knowledge that can be used in modeling autonomous agents-scientists is carried out. Different methods of cognition are considered, in which two systems analyzed by Daniel Kahneman are used: System 1 (intuitive, subconscious) and System 2 (logical, abstract). The features of cognition associated with insight are briefly characterized. The importance of searching for general principles covering a wide area of knowledge is emphasized.

Keywords: Autonomous agents · Scientific cognition · Insight

1 Introduction

Is it possible to build a model of an autonomous agent with the properties of scientific cognition? That is, to model an “agent-scientist”, to create a kind of computer artificial scientist (artificial scientist). This question arose in a number of works [1–3]. The work [1] proposes the principles of functioning of an agent-physicist, who learns the laws of nature in a similar way to well-known scientists. The paper [2] developed an approach to the construction and use of the agent’s internal control system, which consists of two neural networks: the Model and the Controller. The Model remembers useful information, effectively compresses information about previous events. The Controller is used to control the behavior of the agent. Also in this work it is noted that an agent using the Model, with the help of which the study of the external world takes place, can be considered as a precursor of a scientist who cognizes nature (artificial scientist). The paper [3] considers the ways of constructing models of agents that cognize the outside world, and also in this work a computer model was built in which smart agents with perfect cognitive abilities in the process of evolution of a population of agents can displace ordinary agents (without such abilities) from the population.

This paper considers gradually developing levels of cognitive abilities. To a certain extent, such consideration is similar to the analysis carried out in the book “The Phenomenon of Science” by Valentin Turchin [4]. This book analyzes the evolution of biological cybernetic systems, and the evolution of scientific knowledge is considered as the development of biocybernetic evolution. The analysis introduces the conceptual theory of metasystem transitions. According to this theory, the transition from the lower

levels of the system hierarchy to the upper ones occurs through metasystem transitions. Each metasystem transition can be considered as the union of a number of subsystems S_i of the lower level and the appearance of an additional control mechanism C for the combined subsystems. As a result of the metasystem transition, a system S' of a new level is formed ($S' = C + \Sigma_i S_i$), which can be included as a subsystem in the next metasystem transition.

In this paper, the levels of cognitive abilities are considered and analyzed, taking into account the greater specificity of these abilities. The role of two systems introduced by Daniel Kahneman is taken into account: System 1 (intuitive, subconscious) and System 2 (logical, abstract) [5]. The role of the processes of insight in the processes of cognition is considered. Particular attention is paid to the processes of formation of axiomatic theories. The importance of finding general principles covering a wide area of knowledge is also emphasized.

2 The Role of Guesses in Solving Mathematical Tasks

This section contains the author's personal recollection of his early experience in solving mathematical problems, so this section is written in the first person. When I was in secondary school, I was interested in solving complex mathematical tasks. I had collections of such tasks. I knew how to solve problems well. Often I solved tasks immediately after getting acquainted with its condition. But sometimes complex tasks came across and the solution was not immediately found. In this case, I memorized the condition of the task and sometimes the solution was found suddenly, after a few days. It was similar to using System 1 and System 2: the problem was memorized subconsciously in System 1, and after a sudden solution was found, the solution was checked in System 2. Although there were cases when the solution was not found for several days, then usually this task was forgotten and there was a transition to solving the next tasks. To a certain extent, these processes of solving mathematical problems are similar to solving problems by lyceum students at Plato's Academy in Ancient Greece. In the work [6], a simple computer model of the study of an agent-lyceum student at the Plato Academy was built. The main results of this model are characterized in the next section.

3 An Agent-Lyceum Student at the Plato Academy

In the computer model of the work [6], it was believed that an agent-lyceum student in the process of studying accumulates mathematical knowledge, masters methods for solving problems. With a successful solution of problems, the agent develops confidence in his mathematical abilities, while the probability of solving the following problems also increases. The dependence of the probability of solving the problem by the agent $P(t)$ on time t was determined. It was assumed that at $t = 0$ the lyceum agent had just entered the academy, so $P(0)$ was small; time t is discrete, one time step corresponds to an attempt to solve one task by an agent. An example of calculating the dependence $P(t)$ according to the model of the work [6] is shown in Fig. 1.

It can be seen that at first the agent-lyceum student studies for quite a long time and rarely solves tasks. Then the probability of solving the task grows and approaches to 1.

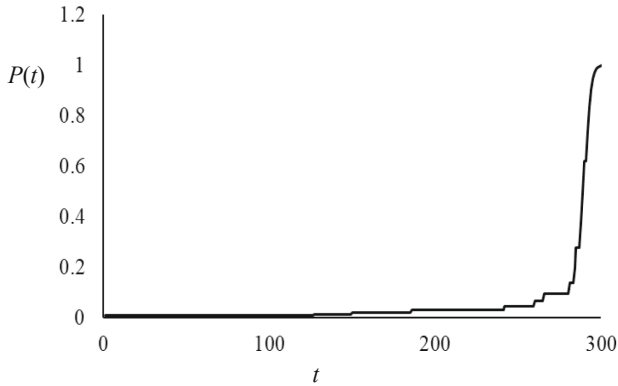


Fig. 1. Dependence of the probability of the correct solution of the problem $P(t)$ by the agentyceum student on time t .

Jumps in the growth of $P(t)$ are clearly visible at the moments of the correct solution of problems.

It is clear that the exchange of information about the methods of solving problems and about the results obtained was important for the staff of the Plato Academy. Moreover, beautiful, serious, non-trivial results were important, for example, such as the Pythagorean theorem. Therefore, reviews of these results were useful. A striking result of the development of such surveys was the “Elements” of Euclid, in which a powerful axiomatic approach was developed. The role of the axiomatic approach is briefly characterized in the next section.

4 The Role of the Axiomatic Approach, the Role of Euclid’s “Elements”

One of the brightest scientific methods is the axiomatic approach. The basis of this approach was laid in Euclid’s “Elements” (about 300 BC), see, for example, [7]. The “Elements” uses definitions, postulates, and axioms, on the basis of which numerous propositions are deductively proved.

Euclid’s “Elements” served for a long time as a model for the logical presentation of mathematical theory. For example, the structure of Isaac Newton’s Principles of Natural Philosophy [8] directly corresponded to the structure of Euclid’s Principles. Thus, Euclid’s “Elements” served as a powerful example of axiomatic theory.

Can a computerized autonomous agent “invent” an axiomatic method on its own and apply it to a particular mathematical theory? This issue will be considered in the next section.

5 Can a Computer Autonomous Agent Develop an Axiomatic Theory?

How can a computer autonomous agent create an axiomatic theory? Imagine once again the collective agent-lyceum student of the Plato Academy. There is a geometry in which this agent solves problems. The technical method of solving problems is known: use only a ruler and a compass. Mental methods are also needed: the formulation of problems and the use of evidence in solving problems. Let's assume that our agent has mastered these methods. Then it solves various problems and makes reviews of the results obtained. A particularly non-trivial process is the proof.

Let us briefly dwell on the methods of creating computer systems capable of independently producing proofs. Such systems were developed in the 1950s by Newell et al. [9].

Developing these systems, G. Gelerter and colleagues developed programs directly intended for proving theorems of Euclidean geometry [10–12]. In particular, in [11], the first theorem in Euclid's *Elements* was proved using a computer program containing 20,000 individual computer operators.

Therefore, computer programs were created that carry out the proofs of geometric theorems, including theorems from Euclid's "Elements". Although the process of proof included heuristics and the ability of the program to make independent guesses. That is, in principle, it is possible to carry out proofs of theorems by an autonomous computer agent. The work programs themselves [10–12] were quite complex, which require the ability of an autonomous agent to create effective programs.

Consider a simpler problem that an autonomous agent can solve when forming an axiomatic theory. Consider the solution of the problem of structuring the review of results. Let us analyze how the agent will be able to supplement and structure the review of the obtained results, so that in the end we get an axiomatic theory similar to Euclid's "Elements". We believe that the review of the results contains solutions to various problems, namely, problems of constructing certain figures and proving theorems. Let us now consider how the collective agent-lyceum student can transform this overview into an axiomatic theory. It is clear that now we need to add definitions to the overview, i.e. characterize the basic concepts used in solving problems. It is also necessary to add postulates and axioms to the review; for simplicity, postulates will be further called axioms. And the most non-trivial thing to do is to arrange the results of the review in such a way that the results obtained are consistently derived from the axioms and the previously stated results. At the same time, such structuring may require the introduction of new concepts, axioms, and obtaining new results so that the entire axiomatic theory is a sequentially connected chain of elements. Apparently, such structuring will occur repeatedly, since it is quite possible that a single chain will not be formed immediately, but as a result of multiple checks and necessary additions to definitions, axioms, and theorems.

At the beginning of the review, one should place definitions and axioms, striving for the completeness of the exposition of the concepts used and for a sufficiently complete system of axioms. Further, it is expedient to present the results sequentially, starting with the simplest and most general. Then it is necessary to do the structuring of the review, namely, it is necessary to check the resulting system for the fact that indeed each stated

result uses only the system of axioms and the previously stated results. This structuring is carried out as follows.

We assume that the total number of results presented in the review is n . Numbers of results $k = 1, 2, \dots, n$ are entered. Next, we consider all the results in order with the numbers $k = 1, 2, \dots$. Let us dwell on the case when a separate result with the number k^* is considered. All references to other results used in obtaining this result k^* are considered. If all references are made to results with numbers less than the number of the given result k^* , then no renumbering occurs for this result and the next result with number $k^* + 1$ is considered. If there is at least one link to a result with a number greater than k^* , then this k^* -th result is moved to the end of the list, i.e. the result is assigned the number $n + 1$. After that, all numbers of results with numbers $k^* + 1, \dots, n + 1$ are reduced by 1. Thus, the order of the results in the review is corrected. Such correction can be repeated several times to ensure the correct order of the full list of results in the review.

6 Insights, the Interaction of System 1 and System 2

As noted above, when solving problems, the result may appear suddenly. These processes have the following features. The problem is not solved immediately. The problem is remembered for some time and stored in the subconscious (in System 1). And then suddenly a solution of the problem appears. Then the solution is checked (in System 2). Such a process was very clearly characterized by Poincaré [13]. Let us describe this search process in more detail. First, Poincaré analyzed Fuchsian functions (now these functions are called “automorphic”). He did this regularly, for two weeks (every day he sat down at his desk, spent one or two hours at it), but there were no results. But one evening he drank a cup of black coffee, could not sleep, and suddenly guessed the right idea of a certain class of fuchsian functions by a sudden insight. As he writes, “ideas arose in abundance; it seemed to me that I felt how they collide with each other, until, finally, two of them, as if entangled with each other, did not form a stable connection”. The discovery came to him unexpectedly, like a sudden insight, like a reward for the long agony of searching and doubting. The next morning, he tested the idea that had arisen and formulated the results. It is important that in doing so, Poincaré had a goal: to find a solution to the problem, that is, the clashing ideas were not random, but related to this goal. Interestingly, in this process of finding a solution, the goal was modified: first, Poincaré sought to prove that there is no certain class of functions, and in the process of insight, he discovered the existence of a new class of Fuchsian functions.

Then, over a rather long period of time, Poincaré a few more times quite unexpectedly had ideas for the development of the theory of Fuchsian functions. Moreover, ideas arose suddenly, at moments not related to the main mathematical work. These ideas, which arose in the process of insights, required subsequent verification. As a result of this work, Poincaré wrote his first memoir on Fuchsian functions.

It should be noted that the processes of insight were analyzed in our work [14], in which the processes of insight were considered in biological organisms (crows, chimpanzees) and in scientific knowledge (in Archimedes, Newton, Poincaré). Also, one of the authors of our work (Alexei V. Samsonovich) built and analyzed a computer model of the process of insight, based on the consideration of the interaction of the subconscious (similar to System 1) and conscious (similar to System 2) levels. Moreover, the

model considered the chaining of schemes, similar to the chaining of ideas described by Poincaré.

7 Conclusion

Thus, in this paper, we analyzed approaches that can be used in modeling computer autonomous agents that cognize the laws of nature. The importance of using subconscious and conscious processes when creating theories of the external world is emphasized. The analysis carried out shows the complexity of the processes of cognition of the laws of nature. Nevertheless, it is clear that the considered methods can be used by a fairly efficient computer autonomous agent when creating theories of the external world.

It is important to search general principles covering a wide area of knowledge. That is the agents should have the tendency to get the clear, strong, and compact knowledge, such as Newton's laws or Euclidean axioms. Thus, we can imagine modeling autonomous agents that could come to the discovery of the laws of nature.

Acknowledgments. This work was supported by the State Program of Scientific Research Institute for System Analysis, Russian Academy of Sciences, Project No. FNEF-2022-0003.

References

1. Red'ko, V.G.: Principles of functioning of autonomous agent-physicist. In: Chella, A., Pirrone, R., Sorbello, R., Jóhannsdóttir, K. (eds.) *Biologically Inspired Cognitive Architectures 2012*, *Advance in Intelligent Systems and Computing*, vol. 196, pp. 265–266. Springer, Berlin (2012)
2. Schmidhuber, J.: On Learning to Think: Algorithmic Information Theory for Novel Combinations of Reinforcement Learning Controllers and Recurrent Neural World Models. [arXiv: 1511.09249](https://arxiv.org/abs/1511.09249). (2015) <https://doi.org/10.48550/arXiv.1511.09249>
3. Red'ko, V.G.: Towards constructing an autonomous agent-scientist. In: Velichkovsky, B.M., Balaban, P.M., Ushakov, V.L. (eds.) *Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics. Intercognsci 2020. Advances in Intelligent Systems and Computing*, vol. 1358, pp. 656–662. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-71637-0_75
4. Turchin, V.F.: *The Phenomenon of Science: A Cybernetic Approach to Human Evolution*. Columbia University Press, New York (1977)
5. Kaneman, D.: *Thinking, Fast and Slow*. Allen Lane (2011)
6. Red'ko, V.G.: How an autonomous computer agent can independently discover the laws of nature. In: *Integrated Models and Soft Computing in Artificial Intelligence. Collection of Scientific Papers of the XI International Scientific and Practical Conference*. In 2 volumes, vol. 2, pp. 108–118. RAAI, Moscow (2022) (In Russian)
7. Euclid's Elements. https://en.wikipedia.org/wiki/Euclid%27s_Elements. Last accessed 21 July 2023
8. Newton, I.: *The Principia: Mathematical Principles of Natural Philosophy*. University of California Press, Berkeley, CA (1687/1999)

9. Newell, A., Shaw, J.C., Simon, H.A.: Report on a general problem-solving program. In: Proceedings of the International Conference on Information Processing. UNESCO, Paris, 15–20 June 1959. Published in 1960 by UNESCO (Paris), R. Oldenbourg (München) and Butterworths (London), pp. 256–264. See also: http://bitsavers.informatik.uni-stuttgart.de/pdf/rand/ipl/P-1584_Report_On_A_General_Problem-Solving_Program_Feb59.pdf. Last accessed 21 July 2023
10. Gelernter, H., Hansen, J., Loveland, D.: Empirical explorations of the geometry theorem proving machine. In: Proceedings of the Western Joint Computer Conference, vol. 17, pp. 143–147 (1960). See also: <https://dl.acm.org/doi/pdf/10.1145/1460361.1460381>. Last accessed 21 July 2023
11. Gelernter, H.: Realization of a geometry theorem proving machine. In: Proceedings of the International Conference Information Processing, Paris, 15–20 June, pp. 273–282 (1959)
12. Gelernter, H.L., Rochester, N.: Intelligent behavior in problem-solving machines. *IBM J. Res. Dev.* **2**(4), 336–345 (1958)
13. Poincaré, H.: *The foundations of Science: Science and Hypothesis, The Value of Science, Science and Method* (translated by G.B. Halsted). Benediction Classics: Oxford, UK. (1913/2012)
14. Red’ko, V.G., Samsonovich, A.V., Klimov, V.V.: Computational modeling of insight processes and artificial cognitive ontogeny. *Cogn. Syst. Res.* **78**, 71–86 (2023). <https://doi.org/10.1016/j.cogsys.2022.12.004>. Last accessed 21 July 2023



Are Associations All You Need to Solve the Dimension Change Card Sort and N -bit Parity Task

Damien Rolon-Mérette^(✉) , Thaddé Rolon-Mérette , and Sylvain Chartier

University of Ottawa, 75 Laurier Ave E, Ottawa, ON K1N 6N5, Canada
drolo083@uottawa.ca

Abstract. When problem-solving, humans can cycle between learned rules to solve tasks. Yet, in artificial neural networks, this cognitive strategy is replaced by learning the entire solution space, making it far less effective. This work aimed to emulate the basis of this human strategy by using a recurrent neural associative memory model. To achieve this, two networks interacted; one served as a task Identifier and the other as a memory Extractor, giving the desired behavior influenced by the Identifier. Each network was trained on sets of interacting associations to represent behavior, such as recognizing shape, color, parity, and when a task started and ended. Once learned, the proposed model was subject to the dimension change card sort (DCCS) and the N -bit parity task. Results showed that the model could switch between behaviors to solve both tasks in linear time once the associations were learned. Moreover, this was possible with 93.3% fewer inputs and no retraining.

Keywords: Set-shifting · Associative memory · ANN · BAM · DCCS · N -bit

1 Introduction

Human cognition is very intricate to study. Different avenues have been used to comprehend how mental processes can emanate from networks of neurons [1, 2]. Nowadays, a popular approach has been to use cognitive modeling to emulate behaviors found in cognition [3–7]. In particular, formal models, known as artificial neural networks (ANNs), have seen a boom in their use due to their great success in learning complex tasks [8]. However, a significant question relates to how they accomplish these feats, especially from a biological and cognitive perspective.

For instance, ANNs usually require information prior hand for learning to be successful. Error signal at the output layer (i.e., global information) containing the solution typically needs to be backpropagated to direct the model’s learning. From a biological perspective, evidence points to local changes, not global ones, steering this process [9, 10]. Furthermore, ANNs usually accomplish significant feats, like multi-task or transfer learning, by combining various types of sophisticated mechanisms [8, 11]. Hence, forming a heterogeneous model. Although it is undebatable that the brain has diverse

functioning, findings suggest that it comes from recursive structures (i.e., homogenous networks) [12]. Additionally, ANNs need training data to span most of the solution space [13]. From a cognitive perspective, this can be akin to humans learning everything by heart. Although this is common, most tasks don't require individuals to absorb and retain an astronomical amount of data. Instead, signs point to humans using strategies they can build upon, forming heuristics rather than learning everything by heart [14]. This can be seen in individuals facing the Wisconsin Card Sorting Test (WCST) [15]. Success in this task usually depends on working memory, attention, and set-shifting, to name a few [16]. The latter is significant, as it suggests that different abilities are learned and reused as "rules". For instance, the ability to recognize shapes and colors and how to count is necessary for solving the WCST. Rather than learning all possible combinations by heart (rote association), individuals apply these "rules" (e.g., shape, color, count, etc.) in different permutations based on the appropriate context to solve tasks. Integrating this strategy to ANNs could be fruitful, as it would lower the number of training data and aid in its multi-task and transfer learning capabilities.

Interestingly, this "rule" learning strategy can be represented by associative mechanisms. In fact, the associative memory (AM) framework states that these mechanisms enable various cognitive abilities and would function by interacting with each other [11, 17–21]. The relevance of this information comes from the existence of an interesting subgroup of ANNs that are concerned with cognitive plausibility and that try to model AM. These are the recurrent neural associative memories (RNAMs) [22]. Like ANNs, they are composed of artificial nodes and weighted connections. However, these networks typically create attractors due to their internal dynamics, which are often said to correspond to those found in the brain's neuronal state space. Consequently, RNAMs find themselves in various cognitive theories [23]. An example of a popular RNAM that can learn auto- and hetero-association is the bidirectional associative memory (BAM). BAMs are nothing new and date back to the 80s [24]. Yet, they remain one of the most widely used types of RNAM to study AM. Over the years, various versions have been proposed to overcome its original limitations [25–30]. Of interest, the use of the Multiple Feature extracting bidirectional associative memory (MF) prior to a BAM has been proposed as a cognitively inspired model with interesting properties [31]. This homogenous multi-network can learn any association while remaining plausible [32]. It combines unsupervised and supervised learning and is based on simple local Hebbian and anti-Hebbian learning. Despite this combination, the MF-BAM alone cannot learn multiple interacting associations needed to emulate the "rule" learning strategy.

However, in a recent study, two MF-BAMs solved the N -bit parity task with 97% fewer training data and in a more general manner [33]. This was possible by teaching the joint model the interacting associations necessary to display a counting behavior. However, the proposed approach was only shown to learn "rules" for recognizing grayscale patterns, count, and parity. Suppose the goal is to eventually solve more complex tasks, like WCST, in a more cognitively plausible manner while using less training data and displaying some form of multi-task and transfer learning. In that case, models must also be able to scale up and integrate more "rules".

A proposed solution to remedy this is to use a modified version of the joint MF-BAMs model. By expanding the input and output layers to contain a distributed representation

of color, the model could integrate this novel dimension into its learning. Furthermore, by adjusting the learned associations to profit from this extra dimensionality, it is believed that more “rules” could be learned. This would differ from previous works where separate channels were required to solve these complex problems [34]. To investigate scalability from past efforts [33], the N -bit learning strategy will be reused in combination with those required to solve the Dimension Change Card Sort [35]. Both tasks are interesting as they can be seen as precursors to solving more complex ones like the WCST. Thus, this body of work aimed to teach the model rules related to recognizing shape, color, parity, counting, and when a task started and finished, to enable it to cycle through them to solve the N -bit parity and DCCS tasks in different succession.

To better illustrate this, the remainder of the paper divides itself as follows: 2. Model, 3. Methodology, 4. Learning the DCCS and N -bit parity, and 5. Discussion.

2 Model

Learning to solve the N -bit parity and DCCS requires interaction between two modules (MF-BAMs). The first module, the Identifier, functions by generating the appropriate instruction set (contextual information) from the incoming stimuli (input patterns). The second module, called the Extractor, retrieves the corresponding stored behaviors from the concatenation of its own internal state and the contexts. Figure 1a) shows the overall implementation of the proposed model. In short, an input pattern, \mathbf{p} , is sent to the Identifier, which will recall the according context, \mathbf{c} (instruction). This pattern will then be concatenated with a portion of the outputted behavior from the Extractor from the previous step. This concatenated pattern $\mathbf{o} = (\mathbf{s} \circ \mathbf{c})$ is sent to the second module to generate the desired behavior \mathbf{b} . The Extractor’s output comprises two concatenate parts $\mathbf{b} = (\mathbf{s} \circ \mathbf{z})$. The output \mathbf{s} keeps track of which state the model is currently in and will be sent back to be concatenated to form \mathbf{o} of the next time step, while \mathbf{z} gives the final behavior.

Each MF-BAM has two components, the MF, which includes several unsupervised network layers (ULs), and the BAM, which contains a single supervised network layer (SL). These are all based on the unsupervised and supervised version of a modified BAM [36, 37]. Figure 1b) shows the architecture of the UL (white) and SL (grey). The MF-BAM uses the same transmission function and learning rules for all its layers. The model’s behavior changes only by modifying the MF’s number of y -units in each hidden layer (denoted by u) and the number of hidden UL (represented by l) used. For all units, the same cubic transmission function, and for each connection, Hebbian/anti-Hebbian learning rules were used. For details, see [31].

3 Methodology

Instead of learning all input-target pairs required in rote associations, the model will first learn to perform five different “rules” selected to be sufficient to solve the N -bit and the DCCS. These are: To recognize the shape of 1 and 0, to recognize red and blue, to count from 0 to 9, to know parity, and to know when a task starts and ends. Multiple associations will represent each “rule”. All the simulations were performed using Python installed from Anaconda [38, 39] and were run on an RTX3090 graphics card.

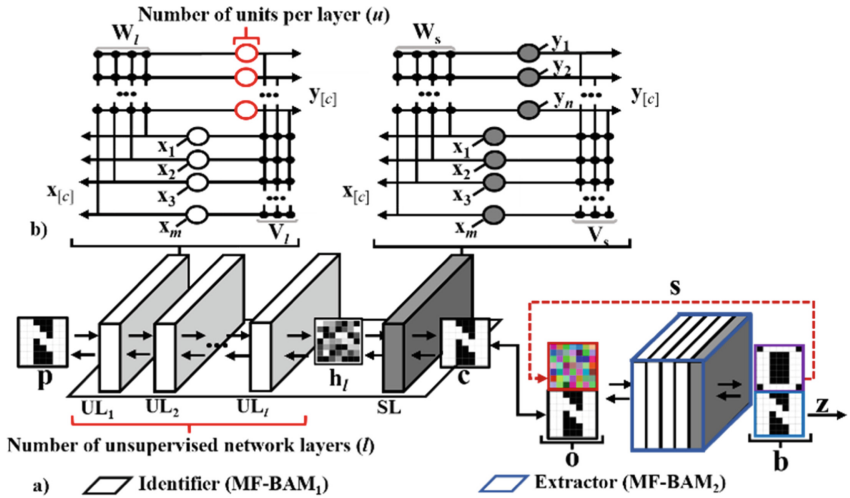


Fig. 1. a) The two MF-BAM composed of Unsupervised network Layers (UL, white) and a single Supervised network Layer (SL, grey), where l represents the number of ULs and u the number of hidden units. b) The detailed underlying network that composes each layer.

For each pattern, the 3rd dimension represented an RGB format, where values of 1 exclusively found in the first dimension produced a color of red, in the second, green, and in the third, blue. If 1 was present in all 3 dimensions, white was obtained, whereas if -1 populated this space, the output was black. Figure 2 shows an example of this RGB format compared to a simple grayscale one.

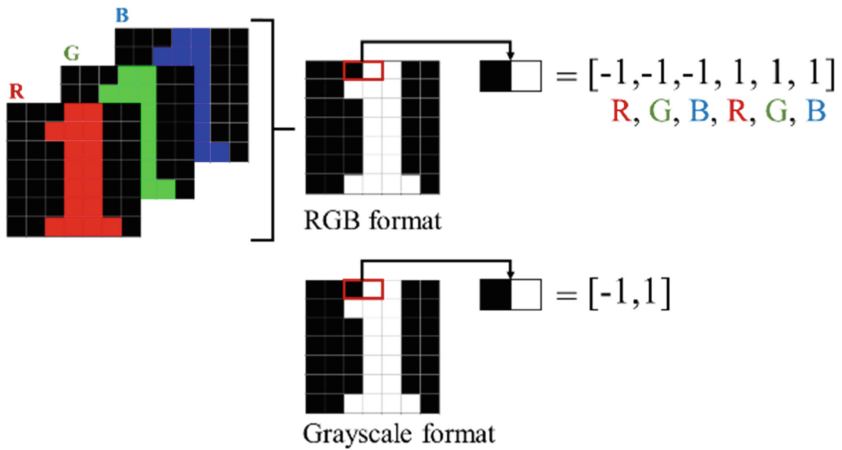


Fig. 2. Example of RGB format compared to grayscale.

Different input-target pairs were created to represent each of the five “rules” (to recognize shape, color, parity, count, and when a task starts and finishes). Thus, a total

of 9 pairs for the Identifier and 60 for the Extractor were created thought to the modules. Figure 3 shows an example of these associations (see [33] for more details).

Once learning is accomplished, the model can be used to solve the N -bit parity task and DCCS in different orders. Each task was represented by different time series. For the DCCS, a similar protocol to [35] was used. However, an input sequence started with the pattern D, followed by 14 patterns depicting either a blue 1 or red 0 and a corresponding target sequence depicting either the class blue 0 or red 1. For the first half, the task required classifying by color, and for the remainder, a pattern S was introduced to signal the condition changed (classifying by shape). For the N -bit parity task, a given bit size was portrayed as a time series starting with the pattern N, followed by the desired numbers of blue 1s and red 0s, and ending with the pattern P [33]. This recall process functioned by having the Identifier receive the input stream and determine the instructions that should be sent to the Extractor to determine the appropriate response. As the inputs are processed, the network should be able to output the correct behavior. Figure 4a shows this for the 2-bit and 4b for a short version of the DCCS.

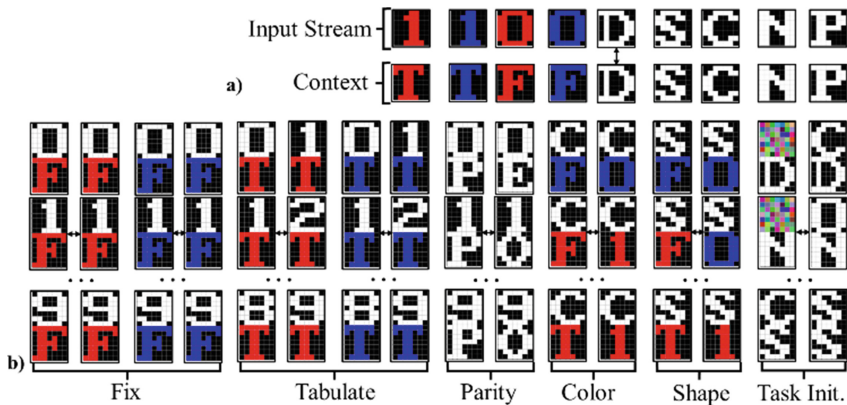


Fig. 3. Example of interacting association “rules” for a) Identifier and b) extractor.

4 Learning the DCCS and N -bit

The model was tested in two paradigms. First, it had to learn only the interacting associations required for a specific task. As such, it was trained with a specific set for the DCCS (color and shape) and a different set for the N -bit (Fix, Tabulate, Parity). This was performed to assess if the model could solve both tasks independently and establish a type of baseline. Perfect performance for the DCCS was only possible if the model matched the 14 input patterns to their corresponding targets. For the N -bit, the entirety of the 2- to 9-bit ($1020 = \sum_{N=2}^9 2^N$) time series were presented during recall, and success was only possible if the correct parity was given for each.

Furthermore, the model was compared to an out-of-the-box multilayer perceptron (MLP) from scikit-learn. Only the batch size was specified to ensure each pattern was

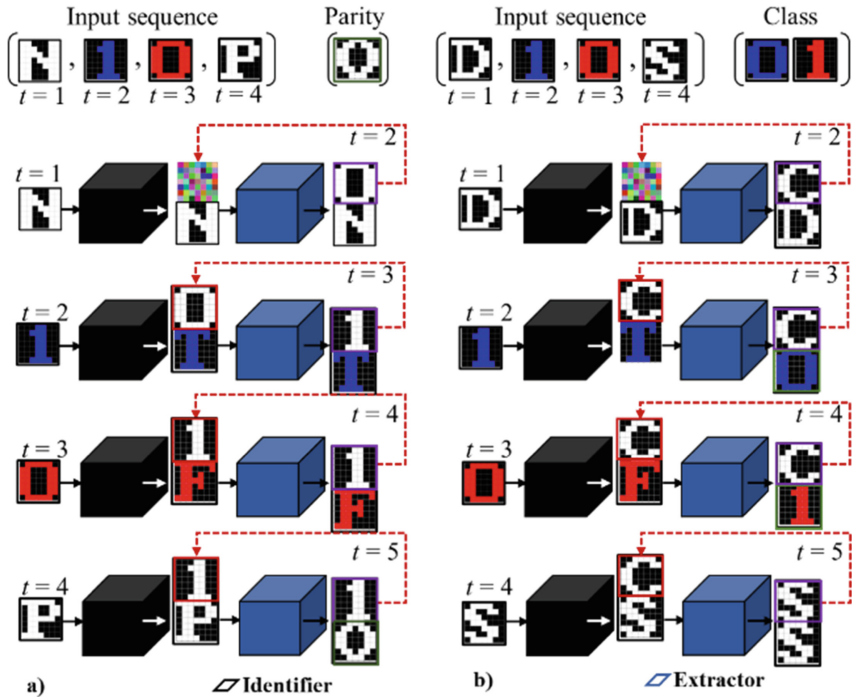


Fig. 4. Example of the model on a) the 2-bit and b) a short version of the DCCS.

presented once during an epoch. For the MLP, the rote association learning strategy was used on both the N -bit parity and DCCS task. Second, it had to learn all interacting associations and perform both tasks in random order. This was sought to investigate whether scalability had any drawbacks during learning and whether the model could perform both tasks in harmony. After each training period, performances were measured by either presenting the DCCS followed by a random sequence of the 9-bit or vice versa and calculating their respective score. For each run, this was calculated in the same fashion for each task, where the generated behavior was compared to the original target on a pixel-per-pixel basis. If the generated and target patterns were identical, it was recorded as a success. As such, average mean squared error (MSE), performances, and 95% confidence intervals were reported. For all simulations, each MF-BAM’s learning was conducted in the same fashion as in [31].

The transmission function parameter (δ) was set to 0.2, weights were randomly initialized $\sim U(-0.1, 0.1)$, and the number of units (u) per layer was set to 100, which is about the size of the input patterns of the Extractor. To assess the complexity of the associations, the number of unsupervised network layers (l) ranged from 0 to 5, where 0 signified removing the MF component of the Identifier and Extractor (i.e., SL/BAM only for each). The model was trained for each l condition and set of interacting associations 50 times and for 3000 epochs.

4.1 Results

The model was tested on two paradigms: learning a subset of interacting associations for a specific task and learning all of them. Figure 5a shows the SL layer's MSE for both the Identifier and Extractor when learning the associations related to the DCCS task only. Typically, very low MSE suggests successful learning. For the Identifier, all l conditions indicated this at the end of the 3000 epochs. However, for the extractor, $l = 0$ was unable to arrive at low values and was stuck at ≈ 0.105 . On average, for $l > 0$, $\text{MSE} < 0.001$ after 436 epochs (std = 22.193). Figure 5b shows the same measurements but for learning all interacting associations. The same previous general trend can be observed. However, for the Extractor, the $l = 0$ condition was stuck at a higher MSE value (≈ 0.254), and to achieve $\text{MSE} < 0.001$ for $l > 0$, it took, on average, 1313 epochs (std = 43.046). Figure 5c shows the average performance and 95% confidence interval for the 2- to 9-bit parity task when learning only the corresponding associations. In general, as l increased, the performance also increased consistently for the model. For $l = 0$, the model was unable to solve any bit size. Interestingly, as the number of bits increased, more l s were needed to achieve consistent perfect performances.

For the MLP, as N increased, performance plummeted. Figure 5d shows the average performances for solving the DCCS (Purple line) and 9-bit (black line) when learning all interacting associations. The results from learning with only the specific interacting association for the DCCS (Purple dash) and N -bit parity (black dash) are also shown for comparison purposes. Results indicated that as l increased, performance increased as well. Furthermore, it showed that performances did not deteriorate when more interacting associations were learned for conditions where $l \geq 3$.

5 Discussion

In this work, it was shown that a RNAM composed of two modules (MF-BAM) could solve the DCCS and the 2- to 9-bit parity task in different order solely by associative mechanism. This was achievable by the model's capability to learn multiple interacting associations resembling "rules" related to recognizing shape, color, parity, count, and when to start and finish a task. Furthermore, this was done by learning these "rules" outside the solution space of each task. Consequently, 93.3% fewer training data were needed to solve both tasks (69 inputs rather than $1028 = 8 + \sum_{N=2}^9 2^N$).

Contrary to the MLP and the rote association strategies that require constant adjustments to the architecture and retraining, the proposed model solved these tasks without these drawbacks and in a more general manner. Once the model learned the 69 patterns, both the 2- to 9-bit and DCCS could be solved regardless of the order of presentation. Of particular interest, it was shown in Fig. 5d that increasing the number of interacting associations had no significant drawbacks in terms of performance once $l \geq 3$. This is significant, as it indicates that the model can indeed scale up towards learning more "rules" and that these don't interfere with one another, a trait needed for lifelong learning [40]. However, results shown in Fig. 5a, b indicated that longer learning times should be expected if more pairs of patterns are to be learned. Furthermore, results in Fig. 5c, d showed that the model had to possess a certain number of ULs for consistent perfect performances. Interestingly, for both tasks to be learned in random order, the model

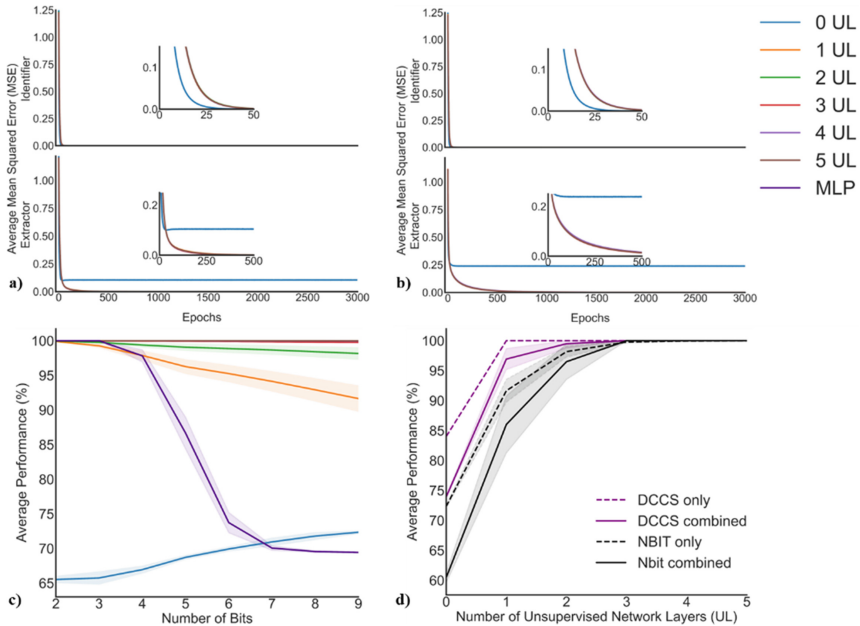


Fig. 5. Average MSE for the Identifier and Extractor during learning of the interacting associations of the **a)** DCCS only, **b)** DCCS and *N*-bit. **c)** Average performance and 95% confidence interval on the 2- to 9-bit parity task for learning the interacting associations for the *N*-bit only. **d)** Average performance and 95% confidence interval after learning the interacting associations of the DCCS (purple dash) and *N*-bit (black dash) only, and when combined for the DCCS (purple line) and *N*-bit (black line).

required $l \geq 3$. This can be explained by the ability of deeper MFs to consistently create representations of the original pairs (input-targets) that are linearly separable [30]. If this process is poorly executed, different patterns eventually fall in the same attractor, producing a linear solution when a non-linear one is needed. It would be interesting to evaluate if other similar cognitive tasks require the model to have a minimum shape. This could help explain why younger children struggle in tasks such as the DCCS and WCST [34, 41, 42]. Results seem to hint towards requiring structural maturation to perform the appropriate computations to deal with nonlinearity. The performance of Fig. 5c also showed that despite integrating color into the *N*-bit parity task, the model could also solve it similarly to previous findings [33]. This is important, as it gives substance to believe that the proposed approach can scale up and that RGBA and other types of color representation could be used. Humans can represent objects as multidimensional [43]. Thus, a model must also be able to accommodate this.

It is essential to highlight some future directions that need to be addressed. First, all “rules” had to be presented at once. Although this method did show how transfer learning could be represented by the interacting associations, it did not address the problem of catastrophic forgetting. Individuals can learn new “rules” without eliminating those

already in memory. A possible avenue could be having multiple MF-BAMs functioning as short- and long-term modules. Interacting associations could dictate important behaviors, such as when to extract novel ‘rules’ from the short-term module and when to perform replays to avoid catastrophic forgetting in the long-term component [44]. Second, using a bipolar representation was problematic, as the absence of signal was represented by -1 s. This meant an increase in correlation due to the higher dimensional representations. High correlation typically deteriorates performances. An interesting solution would be to modify the model to function only in a binary format, as the properties of 0 would help eliminate this correlation. Lastly, reinforcement behaviors were omitted. Tasks like the WCST depend on individuals to receive feedback and determine if a rule selection change should occur. It would be interesting to see if interacting associations and the attractor properties of the model could be used as a reinforcement mechanism to dictate this process of exploitation.

In summary, by using associative mechanisms, it was possible to have different modules interact, and by teaching the appropriate associations, the desired behaviors could be obtained. Although the method was geared towards the DCCS and N -bit, results showed that its learned “rules” could be used harmoniously and potentially towards other similar tasks. This opens the door to solving complex tasks using a distributed AM model more efficiently than traditional rote learning. Learning rules or sets of actions that can be reused are interesting avenues, especially to diminish the amount of data required to train models. As such, by following approaches that try to respect biological and cognitive plausibility, significant improvements can be made to ANNs, while also taking new steps toward better understanding human cognition.

References

1. Arsalidou, M., Pascual-Leone, J.: Constructivist developmental theory is needed in developmental neuroscience. *NPJ Sci. Learn.* **1**(1), 1–9 (2016)
2. Goldwag, J., Wang, G.: DishBrain plays Pong and promises more. *Nat. Mach. Intell.* 1–2 (2023)
3. Anderson, J.R., Matessa, M., Lebiere, C.: ACT-R: a theory of higher level cognition and its relation to visual attention. *Human-Comput. Interaction* **12**(4), 439–462 (1997)
4. Galotti, K.M.: *Cognitive Psychology In and Out of the Laboratory*. Sage Publications (2017)
5. Lamb, R.L., Vallett, D.B., Akmal, T., Baldwin, K.: A computational modeling of student cognitive processes in science education. *Comput. Educ.* **79**, 116–125 (2014)
6. Stella, M., Kenett, Y.N.: Knowledge modelling and learning through cognitive networks. *Big Data Cogn. Comput.* **6**(2), 53 (2022)
7. Sun, R.: The CLARION cognitive architecture: extending cognitive modeling to social simulation. In: *Cognition and Multi-agent Interaction*, pp. 79–99 (2006)
8. Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Silver, D., et al.: Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **575**(7782), 350–354 (2019)
9. Manneschi, L., Vasilaki, E.: An alternative to backpropagation through time. *Nat. Mach. Intell.* **2**(3), 155–156 (2020)
10. Hebb, D.O.: *The Organization of Behavior: A Neuropsychological Theory*. Psychology Press (2005)
11. Zhang, Y., Yang, Q.: An overview of multi-task learning. *Natl. Sci. Rev.* **5**(1), 30–43 (2018)

12. Buckner, R.L., DiNicola, L.M.: The brain's default network: updated anatomy, physiology and evolving insights. *Nat. Rev. Neurosci.* **20**(10), 593–608 (2019)
13. Marcus, G.: Deep Learning: A Critical Appraisal. arXiv preprint [arXiv:1801.00631](https://arxiv.org/abs/1801.00631) (2018)
14. Dale, S.: Heuristics and biases: the science of decision-making. *Bus. Inf. Rev.* **32**(2), 93–99 (2015)
15. Nyhus, E., Barceló, F.: The Wisconsin card sorting test and the cognitive assessment of prefrontal executive functions: a critical update. *Brain Cogn.* **71**(3), 437–451 (2009)
16. Coulacoglou, C., Saklofske, D.H.: Executive function, theory of mind, and adaptive behavior. *Psychometrics Psychol. Assessment* 91–130 (2017)
17. Anderson, J.R., Bower, G.H.: *Human Associative Memory*. Psychology Press (2014)
18. Darcey, M.: Rethinking associations in psychology. *Synthese* **193**(12), 3763–3786 (2016)
19. Dacey, M.: Associationism in the Philosophy of Mind. *The Internet Encyclopedia of Philosophy*. <https://iep.utm.edu/associat/#H7> (2020)
20. Lansner, A.: Associative memory models: from the cell-assembly theory to biophysically detailed cortex simulations. *Trends Neurosci.* **32**(3), 178–186 (2009)
21. Mandelbaum, E.: Associationist theories of thought. In: Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy* (2015)
22. Anderson, J.A.: Cognitive and psychological computation with neural models. *IEEE Trans. Syst. Man Cybern.* **5**, 799–815 (1983)
23. Knoblauch, A.: Neural associative memory for brain modelling and information retrieval. *Inf. Process. Lett.* **95**(6), 537–544 (2005)
24. Kosko, B.: Bidirectional associative memories. *IEEE Trans. Syst. Man Cybern.* **18**(1), 49–60 (1988)
25. Acevedo-Mosqueda, M.E., Yanez-Marquez, C., Acevedo-Mosqueda, M.A.: Bidirectional associative memories: different approaches. *ACM Comput. Surv. (CSUR)* **45**(2), 1–30 (2013)
26. Rolon-Merette, T., Rolon-Merette, D., Chartier, S.: Generating cognitive context with feature-extracting bidirectional associative memory. *Proc. Comput. Sci.* **145**, 428–436 (2018)
27. Rolon-Mérette, D., Rolon-Mérette, T., Chartier, S.: Distinguishing highly correlated patterns using a context based approach in bidirectional associative memory. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2018)
28. Rolon-Mérette, T., Rolon-Mérette, D., Calderini, M., Chartier, S.: Different brain, same prototype? Cognitive variability within a recurrent associative memory. In *Proceedings of the 17th International Conference on Cognitive Modeling*, pp. 192–197 (2019)
29. Rolon-Mérette, D., Rolon-Mérette, T., Chartier, S.: Learning and recalling arbitrary lists of overlapping exemplars in a recurrent artificial neural network. In: *Proceedings of the International Conference on Cognitive Modelling*, pp. 186–191 (2019)
30. Rolon-Merette, T., Rolon-Merette, D., Chartier, S.: Towards binary encoding in bidirectional associative memories. In: *The International FLAIRS Conference Proceedings*, vol. 36 (2023)
31. Rolon-Mérette, D., Rolon-Mérette, T., Chartier, S.: A multilayered bidirectional associative memory model for learning nonlinear tasks. In: *Neural Networks* (2023)
32. O'Reilly, R.C.: Six principles for biologically based computational models of cortical cognition. *Trends Cogn. Sci.* **2**(11), 455–462 (1998)
33. Rolon-Mérette, D., Rolon-Merette, T., Chartier, S.: Using bidirectional associative memory neural networks to solve the N-bit task. In: *The International FLAIRS Conference Proceedings*, vol. 36 (2023)
34. Kaplan, G.B., Şengör, N.S., Gürvit, H., Genç, I., Güzeliş, C.: A composite neural network model for perseveration and distractibility in the Wisconsin card sorting test. *Neural Netw.* **19**(4), 375–387 (2006)
35. Zelazo, P.D.: The dimensional change card sort (DCCS): a method of assessing executive function in children. *Nat. Protoc.* **1**(1), 297–301 (2006)

36. Chartier, S., Boukadoum, M.: A bidirectional heteroassociative memory for binary and grey-level patterns. *IEEE Trans. Neural Netw.* **17**(2), 385–396 (2006)
37. Chartier, S., Giguère, G., Renaud, P., Lina, J.M., Proulx, R.: FEBAM: a feature-extracting bidirectional associative memory. In: 2007 International Joint Conference on Neural Networks, pp. 1679–1684. IEEE (2007)
38. Anaconda Software Distribution: Anaconda Documentation. Anaconda Inc. Retrieved from <https://docs.anaconda.com/> (2020)
39. Rolon-Mérette, D., Ross, M., Rolon-Mérette, T., Church, K.: Introduction to Anaconda and Python: installation and setup. *Quant. Methods Psychol.* **16**(5), S3–S11 (2020)
40. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: a review. *Neural Netw.* **113**, 54–71 (2019)
41. Moriguchi, Y., Hiraki, K.: Longitudinal development of prefrontal function during early childhood. *Dev. Cogn. Neurosci.* **1**(2), 153–162 (2011)
42. Rosselli, M., Ardila, A.: Developmental norms for the Wisconsin card sorting test in 5- to 12-year-old children. *Clin. Neuropsychologist* **7**(2), 145–154 (1993)
43. Kloo, D., Perner, J., Aichhorn, M., Schmidhuber, N.: Perspective taking and cognitive flexibility in the dimensional change card sorting (DCCS) task. *Cogn. Dev.* **25**(3), 208–217 (2010)
44. Hayes, T.L., Kafle, K., Shrestha, R., Acharya, M., Kanan, C.: Remind your neural network to prevent catastrophic forgetting. In: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 Aug 2020, Proceedings, Part VIII* 16. Springer International Publishing, pp. 466–483 (2020)



Image and Audio Data Classification Using Bagging Ensembles of Spiking Neural Networks with Memristive Plasticity

Roman Rybka^{1,2,3(✉)}, Yury Davydov^{1,4}, Alexander Sboev^{1,2,3}, Danila Vlasov¹, and Alexey Serenko¹

¹ National research Centre “Kurchatov Institute”, Moscow, Russia
Rybka_RB@nrcki.ru

² National Research Nuclear University MEPhI, Moscow, Russia

³ Russian Technological University “MIREA”, Moscow, Russia

⁴ National Taipei University of Technology, Taipei, Taiwan

Abstract. Spiking neural networks (SNNs) are potentially capable of greatly reducing the energy requirements of modern intelligent systems when combined with neuromorphic computing devices based on memristors, that facilitate on-chip SNN training. Currently, the existing spiking approaches either rely on weight transfer and/or backpropagation-based training or utilize large fully-connected spiking networks, imposing high hardware requirements. In this paper, we study the application of the bagging ensembling technique coupled with SNN-based models to the audio and image classification problems. In our experiments, we use a three-layer spiking neural network with Logistic Regression decoding and consider three local plasticity rules—spike time-dependent plasticity and its nanocomposite and poly-p-xylylene memristor counterparts. Using the Digits and FSDD datasets for training and evaluation, we show that bagging yields a performance increase of up to 20% in terms of the F1-score metric, while substantially reducing the total number of connections in the network.

Keywords: Spiking neural networks · Spike-timing-dependent plasticity · Memristors · Audio classification · FSDD · Ensembles of SNN

1 Introduction

The development of machine learning methods, especially neural networks, leads to substantial energy costs associated with the creation and validation of models. In this regard, one of the topical areas of research is to reduce energy consumption when performing calculations both at the training and inference stages. A promising technology for this purpose is neuromorphic computing devices,

where the exchange of information is implemented through the exchange of short impulses—spikes—analogue to the brain of living organisms. Modern neuroprocessors, such as TrueNorth [1], Loihi [2], and Altai¹, can perform various tasks related to image and audio data analysis and agent management with power consumption in the milliwatt range [3].

There are many examples of implementations of neural networks on neuromorphic chips trained using error backpropagation [4], or reward-modulated plasticity where the weight change is modulated by some reward signal [5], or by controlled plasticity models, where plasticity changes from Hebbian to anti-Hebbian and vice versa depending on the class of the input [6–8].

However, memristor-based SNNs are difficult to train with error backpropagation [9], where information about the state of the entire system is required to compute each weight update. In contrast, learning algorithms based on local synaptic plasticity only need information from pre- and postsynaptic neurons [10, 11], allowing each synapse to learn in a self-organizing manner. Potentially, this simplifies the deployment of memristor-based equipment not only for the inference stage of the trained network but also for its training [12–14]. Additionally, since memristive devices are currently harder to produce than conventional chips, there is a need to develop SNN-based approaches that do not require large fully-connected spiking architectures to provide acceptable accuracy.

In this work, we study the possibility of creating an ensemble of spiking neural network models of small size, which can reduce connectivity and increase accuracy. Ensemble building is done through the bagging technique, where neural network models are trained on a portion of the available training data. We use the spiking neural network based on a three-layer architecture (see Sect. 2): input, excitatory and inhibitory layers. Connections between the input and the excitatory layers are trainable and use local plasticity rules, several models for which are compared, including the classical additive STDP, as well phenomenological models of conductance change of real nanocomposite [15] and poly-pylylene [16] memristors. The latter two showed comparable scores when solving classification problems [17–20].

Verification of the capabilities of the created variants of ensembles of spiking neural network models is performed based on two classification datasets described in Sect. 3: the images of handwritten digits and the audio recordings of the pronunciation of digits. Thus, the main contributions of this work are:

- (1) We investigate the bagging ensembling technique of spiking neural network models with plastic memristive connections.
- (2) We demonstrate the possibility of reducing the overall connectivity of the ensembled models when solving problems of classifying audio and image data.

¹ <https://motivnt.ru/neurochip-altai>.

2 Materials and Methods

2.1 Memristor Plasticity Models

Along with the additive Spike Timing-Dependent plasticity (STDP) [21], we use two phenomenological models of memristive plasticity described below.

Nanocomposite memristor The dependence of synaptic conductance change Δw on the value of the conductance w and on the time difference Δt between presynaptic and postsynaptic spikes, as observed in nanocomposite memristors [15], is defined in Eq. 1:

$$\Delta w(\Delta t) = \begin{cases} A^+ \cdot w \cdot \left[1 + \tanh \left(-\frac{\Delta t - \mu^+}{\tau^+} \right) \right] & \text{if } \Delta t > 0; \\ A^- \cdot w \cdot \left[1 + \tanh \left(\frac{\Delta t - \mu^-}{\tau^-} \right) \right] & \text{if } \Delta t < 0; \end{cases} \quad (1)$$

Here $A^+ = 0.074$, $A^- = -0.047$, $\mu^+ = 26.7$ ms, $\mu^- = -22.3$ ms, $\tau^+ = 9.3$ ms, $\tau^- = 10.8$ ms.

Poly-p-xylylene memristor In PPX memristors [16], the timing and initial weight dependence curves are significantly different from those for NC memristors: the dependence on the synaptic weight w is nonlinear and asymmetric. The plasticity of PPX memristors is modelled using Eq. 2:

$$\Delta w(\Delta t) = \begin{cases} \frac{|\Delta t|}{\tau} \alpha^+ e^{-\beta^+ \left(\frac{w_{\max} - w}{w_{\max} - w_{\min}} \right)} e^{-\gamma^+ \left(\frac{\Delta t}{\tau} \right)^2} & \text{if } \Delta t > 0; \\ \frac{|\Delta t|}{\tau} \alpha^- e^{-\beta^- \left(\frac{w - w_{\min}}{w_{\max} - w_{\min}} \right)} e^{-\gamma^- \left(\frac{\Delta t}{\tau} \right)^2} & \text{if } \Delta t < 0. \end{cases} \quad (2)$$

Here $\tau = 10$ ms, $\alpha^+ = 0.316$, $\alpha^- = 0.011$, $\beta^+ = 2.213$, $\beta^- = -5.969$, $\gamma^+ = 0.032$, $\gamma^- = 0.146$, $w_{\max} = 1$, $w_{\min} = 0$.

2.2 Network Model

An SNN consisting of three layers [22] was chosen for the study (see Fig. 1).

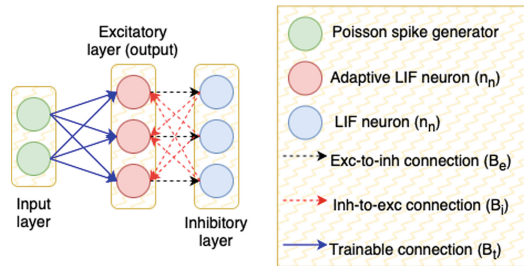


Fig. 1. SNN architecture

The input layer consists of Poisson spike generators that interpret incoming samples of size N_{inp} as frequency vectors, generate Poisson spike trains, and pass them to the excitatory layer.

The excitatory layer consists of leaky-integrate-and-fire (LIF) neurons with adaptive thresholds, whose average spiking frequencies are used to determine the class of the sample. The connections from the input to the excitatory neurons are trainable and use one of the three synaptic plasticity types discussed above. The input and the excitatory layers are connected in an each-to-each manner (see the blue arrows in Fig. 1).

The dimension of the layer of excitatory neurons N_e is a hyperparameter; typically, a larger number of neurons corresponds to better accuracy. The total number of connections between the input and excitatory layer is, therefore, equal to $B_t = N_{inp}N_e$.

The layer of inhibitory LIF neurons has the same number of neurons as the excitatory layer ($N_i = N_e = n_n$). Thus, each neuron of the excitatory layer is associated with a partner neuron in the inhibitory layer. The connections from excitatory to inhibitory neurons are formed in a one-to-one manner and have fixed weights $w_{ei} > 0$ (black arrows in the Fig. 1), $B_e = (N_e)^2$ connections overall. The connections from each inhibitory neuron run to every excitatory neuron except for its partner, ensuring competition among the neurons. These connections are also static and have weights $w_{ie} < 0$ (in the Fig. 1 are indicated by the red arrows), the total number of connections being $B_i = N_i(N_e - 1)$.

The network was implemented in the NEST [23] simulator. For the handwritten digits dataset, we used $\tau_{ref,e} = 4$ ms, $\tau_{ref,i} = 9$ ms, $\tau_{m,e} = 60$ ms, $\tau_{m,i} = 50$ ms, $w_{ei} = 20$, and $w_{ie} = -15$. For the spoken digits dataset, the hyperparameters were: $\tau_{ref,e} = 4$ ms, $\tau_{ref,i} = 3$ ms, $\tau_{m,e} = 130$ ms, $\tau_{m,i} = 30$ ms, $w_{ei} = 13$, and $w_{ie} = -12$ (here τ_{ref} denotes the refractory period of the neuron, τ_m – the membrane time constant, e and i indices denote excitatory and inhibitory neurons).

We used the LogisticRegression classifier imported from the Scikit-Learn [24] library to map the output frequencies of the network to the class labels. This model was chosen for its simplicity, due to which the accuracy of the classification is mainly provided by the spiking neural network. The model was used with the C parameter value of 1000, all other hyperparameters were set to default values.

3 Datasets

Audio classification. We use the Free Spoken Digits Dataset (FSDD) [25] consists of 3000 audio recordings of the spoken numbers from 0 to 9 in English. The digit recordings were made by 6 people; each digit was repeated by each speaker for 50 times with different speed and intonations. The FSDD dataset contains 10 classes of 300 WAV-format audio records, up to 1 s long. This dataset has a recommended method of splitting into the training and testing sets: the first 5 out of 50 audio recordings by each person and for each class are assigned to the test set, while the rest (90% of the data) goes to the training set.

Image classification. Our experiments are based on the Digits dataset that is considered to be a baseline for the image classification problems. It contains 1797 8×8 images of handwritten digits (near 180 samples per class, 10 classes in total). This dataset was imported from the scikit-learn [24] library.

4 Experiments

The input data was preprocessed as follows:

- Digits: normalization with a reduction to zero means and one standard deviation (Standard Scaler) and processing with Gaussian receptive fields (7 fields); the final dimension of the input vector is 448,
- FSDD: extracting useful features from raw audio using MFCC (30 frame-averaged components), Standard Scaler, processing with Gaussian receptive fields (7 fields); the final dimension of the input vector is 210.

The resulting vectors were converted into spike sequences using a Poisson spike generator with frequencies of 500 and 550 Hz for the Digits and FSDD datasets, respectively.

We chose bagging as an ensemble technique, in which several identical classifiers are trained on subsets of the input data, after which their predictions are aggregated by voting. This method has a number of advantages over using a single larger network, in particular, the ensemble has lower connectivity and is therefore less computationally expensive. In addition, it helps to break unwanted correlations in the training data and thus reduces overfitting. We considered the following configurations: (1) 550 excitatory neurons in one network, 1 network (no bagging); (2) 25 neurons, 41 networks in the ensemble; (3) 50 neurons, 11 networks; (4) 50 neurons, 21 networks; (5) 100 neurons, 5 networks; (6) 100 neurons, 11 networks.

In all experiments, the number of inhibitory neurons was equal to the number of excitatory neurons. The size of the training subsample for each network in the ensemble was fixed and equal to 70% of the total training set volume (with a random selection of examples, without returning to the general sample after extracting the example, i.e. without bootstrapping; was not applied to the experiments with a single network). The ensemble was implemented using the BaggingClassifier method of the scikit-learn library.

5 Results and Discussion

We obtained the results using the hold-out cross-validation technique according to the micro F1-score (see Table 1). The Table 1 shows the dependency of the classification F1-score on the total number of inhibitory connections in the ensemble.

It can be seen from Table 1 that bagging improves the classification accuracy almost for all considered cases: for the nanocomposite memristive plasticity— from 90 to 92% (Digits), for the PPX plasticity – from 78 to 94% (FSDD),

Table 1. Classification accuracy of the bagging ensembles of the SNN models on the Digits and FSDD datasets. n_e denotes the number of estimators in the ensemble, n_n —the number of excitatory neurons in a single network, N_n —the total number of LIF neurons in the ensemble, B_i —the total number of inhibitory-to-excitatory connections, B_e —the total number of excitatory-to-inhibitory connections, B_t —the total number of trainable connections (input-to-excitatory), and F1-score is multiplied by 100

Plasticity	n_e	n_n	N_n	B_i	B_e	Digits: B_t	Digits: F1	FSDD: B_t	FSDD: F1
NC	1	550	1100	301,950	550	246,400	90	115,500	93
NC	5	100	1000	49,500	500	224,000	91	105,000	90
NC	11	50	1100	26,950	550	246,400	91	115,500	90
NC	11	100	2200	108,900	1100	492,800	91	231,000	93
NC	21	50	2100	51,450	1050	470,400	92	220,500	92
NC	41	25	2050	24,600	1025	459,200	89	215,250	89
PPX	1	550	1100	301,950	550	246,400	92	115,500	78
PPX	5	100	1000	49,500	500	224,000	89	105,000	84
PPX	11	50	1100	26,950	550	246,400	86	115,500	92
PPX	11	100	2200	108,900	1100	492,800	91	231,000	94
PPX	21	50	2100	51,450	1050	470,400	90	220,500	91
PPX	41	25	2050	24,600	1025	459,200	90	215,250	90
STDP	1	550	1100	301,950	550	246,400	83	115,500	95
STDP	5	100	1000	49,500	500	224,000	82	105,000	93
STDP	11	50	1100	26,950	550	246,400	85	115,500	92
STDP	11	100	2200	108,900	1100	492,800	87	231,000	96
STDP	21	50	2100	51,450	1050	470,400	88	220,500	94
STDP	41	25	2050	24,600	1025	459,200	86	215,250	90

for STDP—from 83 to 88% (Digits) and from 95 to 96% (FSDD). In all other cases, the accuracy either stays the same or slightly drops. Notably, although the accuracy gains are often within 2–4% (most likely due to the relatively small ensemble sizes), in certain situations bagging can significantly improve the model’s performance, like in the case of the PPX-trained network on the FSDD dataset. This result is consistent with the effect of bagging on formal classifiers, i.e., in the Random Forest algorithm. Additionally, better accuracies are achieved using the networks with less weight parameters: for example, 21 estimators with 50 neurons each (2100 total neurons, 51,450 inhibitory-to-excitatory connections) is smaller than the equivalent fully-connected network with 2100 neurons, which would have 1,101,450 connections. The accuracy does not scale with the number of connections in the network and has a clear maximum in the neighborhood of 50–100 thousand connections.

6 Conclusion

In this paper, we demonstrate that applying the bagging ensembling technique to the SNN-based classifiers with memristive plasticity improves the classification accuracy while reducing the number of synaptic connections by a factor of 3–10 and facilitating stronger parallelization schemes. On the Digits dataset, the proposed approach yields the F1-score up to 92% (NC and PPX plasticity), and up to 96% on the FSDD dataset (STDP plasticity). The observed performance gains from bagging are mostly within 1–3%, however, for the PPX plasticity on FSDD they reach 20% (78% F1-score without bagging and 94% with 11 estimators with 100 excitatory neurons).

Acknowledgments. The study has been supported by the Russian Science Foundation grant No. 21-11-00328 <https://rscf.ru/project/21-11-00328/> and has been carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC “Kurchatov Institute”, <http://ckp.nrcki.ru/>.

References

1. Merolla, P.A., Arthur, J.V., Alvarez-Icaza, R., Cassidy, A.S., Sawada, J., Akopyan, F., Jackson, B.L., Imam, N., Guo, C., Nakamura, Y., et al.: A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**(6197), 668–673 (2014)
2. Davies, M., Srinivasa, N., Lin, T.H., Chinya, G., Cao, Y., Choday, S.H., Dimou, G., Joshi, P., Imam, N., Jain, S., et al.: Loihi: a neuromorphic manycore processor with on-chip learning. *IEEE Micro* **38**(1), 82–99 (2018). <https://doi.org/10.1109/MM.2018.112130359>
3. Rajendran, B., Sebastian, A., Schmuker, M., Srinivasa, N., Eleftheriou, E.: Low-power neuromorphic hardware for signal processing applications: a review of architectural and system-level design approaches. *IEEE Signal Process. Mag.* **36**(6), 97–110 (2019). <https://doi.org/10.1109/MSP.2019.2933719>
4. Lee, J.H., Delbruck, T., Pfeiffer, M.: Training deep spiking neural networks using backpropagation. *Front. Neurosci.* **10** (2016)
5. Mozafari, M., Ganjtabesh, M., Nowzari-Dalini, A., Thorpe, S.J., Masquelier, T.: Combining STDP and reward-modulated STDP in deep convolutional spiking neural networks for digit recognition. *arXiv preprint arXiv:1804.00227* (2018)
6. Güttig, R., Sompolinsky, H.: The tempotron: a neuron that learns spike timing-based decisions. *Nat. Neurosci.* **9**(3), 420–428 (2006)
7. Yu, Q., Tang, H., Tan, K.C., Yu, H.: A brain-inspired spiking neural network model with temporal encoding and learning. *Neurocomputing* **138**, 3–13 (2014). <https://doi.org/10.1016/j.neucom.2013.06.052>
8. Wang, X., Hou, Z.G., Lv, F., Tan, M., Wang, Y.: Mobile robots’ modular navigation controller using spiking neural networks. *Neurocomputing* **134**, 230–238 (2014)
9. Li, C., Wang, Z., Rao, M., Belkin, D., Song, W., Jiang, H., Yan, P., Li, Y., Lin, P., Hu, M., et al.: Long short-term memory networks in memristor crossbar arrays. *Nat. Mach. Intell.* **1**(1), 49–57 (2019). <https://doi.org/10.1038/s42256-018-0001-4>

10. Saïghi, S., Mayr, C.G., Serrano-Gotarredona, T., Schmidt, H., Lecerf, G., Tomas, J., Grollier, J., Boyn, S., Vincent, A.F., Querlioz, D., La Barbera, S., Alibart, F., Vuillaume, D., Bichler, O., Gamrat, C., Linares-Barranco, B.: Plasticity in memristive devices for spiking neural networks. *Front. Neurosci.* **9**, 51 (2015). <https://doi.org/10.3389/fnins.2015.00051>
11. Serrano-Gotarredona, T., Masquelier, T., Prodromakis, T., Indiveri, G., Linares-Barranco, B.: STDP and STDP variations with memristors for spiking neuromorphic learning systems. *Front. Neurosci.* **7**, 2 (2013)
12. Zhang, Y., Wang, X., Friedman, E.G.: Memristor-based circuit design for multi-layer neural networks. *IEEE Trans. Circuits Syst. I Regul. Pap.* **65**, 677–686 (2018). <https://doi.org/10.1109/TCSI.2017.2729787>
13. Tao, T., Ma, H., Li, D., Li, Y., Tan, S., Liu, E.X., Schutt-Aine, J., Li, E.P.: Modeling and analysis of spike signal sequence for memristor crossbar array in neuromorphic chips. *IEEE Trans. Circuits Syst. I Regul. Pap.* **70**(6), 2271–2282 (2023). <https://doi.org/10.1109/TCSI.2023.3250699>
14. Bordanov, I., Antonov, A., Korolev, L.: Simulation of calculation errors in memristive crossbars for artificial neural networks. In: 2023 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), pp. 1008–1012 (2023). <https://doi.org/10.1109/ICIEAM57311.2023.10139308>
15. Demin, V., Nekhaev, D., Surazhevsky, I., Nikiruy, K., Emelyanov, A., Nikolaev, S., Rylkov, V., Kovalchuk, M.: Necessary conditions for STDP-based pattern recognition learning in a memristive spiking neural network. *Neural Netw.* **134**, 64–75 (2021)
16. Minnekhanov, A.A., Shvetsov, B.S., Martyshev, M.M., Nikiruy, K.E., Kukueva, E.V., Presnyakov, M.Y., Forsh, P.A., Rylkov, V.V., Erokhin, V.V., Demin, V.A., et al.: On the resistive switching mechanism of parylene-based memristive devices. *Org. Electron.* **74**, 89–95 (2019). <https://doi.org/10.1016/j.orgel.2019.06.052>
17. Sboev, A., Vlasov, D., Rybka, R., Davydov, Y., Serenko, A., Demin, V.: Modeling the dynamics of spiking networks with memristor-based STDP to solve classification tasks. *Mathematics* **9**(24), 3237 (2021)
18. Sboev, A., Davydov, Y., Rybka, R., Vlasov, D., Serenko, A.: A comparison of two variants of memristive plasticity for solving the classification problem of handwritten digits recognition. In: Klimov, V.V., Kelley, D.J. (eds.) *Biologically Inspired Cognitive Architectures 2021*, pp. 438–446. Springer, Cham (2022)
19. Vlasov, D., Davydov, Y., Serenko, A., Rybka, R., Sboev, A.: Spoken digits classification based on spiking neural networks with memristor-based stdp. In: 2022 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 330–336. IEEE (2022)
20. Matsukatova, A.N., Iliasov, A.I., Nikiruy, K.E., Kukueva, E.V., Vasiliev, A.L., Goncharov, B.V., Sitnikov, A.V., Zhanaveskin, M.L., Bugaev, A.S., Demin, V.A., et al.: Convolutional neural network based on crossbar arrays of (co-fe-b) x (linbo3) 100-x nanocomposite memristors. *Nanomaterials* **12**(19), 3455 (2022)
21. Song, S., Miller, K.D., Abbott, L.F.: Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nat. Neurosci.* **3**(9), 919–926 (2000)
22. Diehl, P.U., Pedroni, B.U., Cassidy, A., Merolla, P., Neftci, E., Zarella, G.: True-happiness: Neuromorphic emotion recognition on truenorth (2016). <https://doi.org/10.1109/IJCNN.2016.7727758>
23. Gewaltig, M.O., Diesmann, M.: NEST (Neural Simulation Tool). *Scholarpedia* **2**(4), 1430 (2007)

24. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
25. Jackson, Z., Souza, C., Flaks, J., Pan, Y., Nicolas, H., Thite, A.: Jakobovski/free-spoken-digit-dataset: v1.0.8 (2018). <https://doi.org/10.5281/zenodo.1342401>



An Episode Tracker for Cognitive Architectures

Eduardo Yuji Sakabe¹, Anderson Anjos da Silva¹, Luiz Fernando Coletta¹,
Alexandre da Silva Simões², Esther Luna Colombini¹,
Paula Dornhofer Paro Costa¹, and Ricardo Ribeiro Gudwin¹(✉)

¹ University of Campinas (UNICAMP), Campinas, SP, Brazil
gudwin@unicamp.br

² University of the State of São Paulo (UNESP), Sorocaba, SP, Brazil

Abstract. This paper introduces the Episode Tracker Module, an encoding mechanism that tracks sensory information through space and time, building up high-level semantic representations called episodes. This module is aimed to extend the Cognitive Systems Toolkit (CST) as a reusable framework for building different cognitive models for episode detection. We created two instances of the episode tracker with two different mechanisms for identifying property categories (geographical regions). Each mechanism correctly induced a different episode detection dynamic. Overall, the Episode Tracker architecture provides a robust and flexible framework for episode detection.

Keywords: Episodic memory · Cognitive architectures · CST

1 Introduction

Episodic memory is an essential component of the human cognitive system. While there is evidence of episodic-like memory in other animals [1], human episodic memory is a cornerstone in the evolution of the species [16].

For example, when searching for our car in a vast parking lot, we use episodic memory to recall where we parked the vehicle and find its current location. We also use it to remember previously inspected places to avoid searching redundant areas. We cannot recollect past experiences or understand our current context without episodic memory. A notable example is the case of Kent Cochrane, widely known as patient K.C., who suffered a brain injury that did not interfere with his understanding of the facts and concepts of the world (semantic memory) but severely damaged his episodic memory, turning impossible for him, for example, to recall his visits to the lab where he was being attended [16]. Therefore, episodic memory is essential to model human intelligence as we know it.

Nuxoll and Laird also highlight several cognitive capabilities that an agent would benefit from episodic memory, such as action modeling, decision-making, retroactive learning, and virtual sensing [13].

One key issue in designing artificial episodic memory relies on how the episodes should be encoded (i.e., the mechanism which assembles the episodes) and, consequently, their representation. Although a few attempts exist to model and implement episodic memory in artificial agents and cognitive architectures, the usual encoding process suffers from a limited scope for semantic interpretation. Typically, an episode representation is just a sequence of snapshots of other memory components, resulting in inappropriate interoperability between episodes and high-level cognitive functions. Our motivation for building yet another artificial episodic memory system is surpassing these limitations and enhancing those capabilities.

In this work, we present a first approach to the problem of modeling episodic memory in cognitive architectures, addressing this concern for semantic interpretability. In particular, we focus on the problem of detecting relevant events and building episodes from perceptual data across time. We employ Cognitive Systems Toolkit (CST) [14] to build an Episode Tracker Module, the most fundamental building block towards a full-featured episodic memory computational model.

The Episode Tracker Module is an encoding mechanism, translating sensory data through space and time into high-level semantic representations called scene-based episodes. Also, as a framework within a cognitive architecture, our implementation has a flexible design that enables the adoption of different mechanisms to identify relevant events that compose an episode.

The paper is organized as follows. We first draw the reader's attention to the difference between a simpler encoding strategy, considering what we call state-based episodes, and a more elaborated (and suitable for semantic interpretation) representation of episodes, which we call scene-based. Then, we describe the main details regarding our Episode Tracker Module, followed by a description of our experiments and their results, before a final conclusion.

2 State-Based and Scene-Based Episodes

There are two ways of representing an episode: state-based or scene-based [5]. A state-based approach represents an episode as a sequence of time-stamped internal states. State-based episodes are easier to encode since they are simply copies of information from other components' states.

This approach has two main downsides. First, it consumes more memory since the episodes are simple copies of the full agent's experiences. Second, these experiences are not interpreted, creating difficulties while interoperating with high-level cognitive functions, as there is no interpretation of what happened during this state sequence.

For instance, suppose a decision-making mechanism relying on previous episodes lived by the agent to choose between actions in a certain context. If these episodes are merely sequences of states, they will not carry the semantic knowledge of what of importance occurred during these states, regarding the agent's actions. This knowledge would require an additional processing cost. This

overhead will be necessary whenever a high-level cognitive mechanism is connected with a state-based episodic memory. Scene-based representations bypass this hurdle by directly providing high-level semantic representations.

Unlike state-based episodes, scene-based episodes encode a spatiotemporal segment into a more elaborated representation carrying semantic content, which we term a *scene*. A scene comprises high-level elements, such as objects, their properties, and performed actions. Thus, scene-based episodes require complex perception mechanisms that interpret multi-dimensional sensory data to encode conceptual information (e.g., properties, objects, and actions). In this way, scene-based episodes are analogous to high-level interpretations of state-based episodes. For this reason, developing such a mechanism with a general (task-independent) approach is challenging. As a trade-off, scene-based episodes have the advantage of better interfacing with high-level cognitive functions since they have compatible high-level representations. Nevertheless, state-based episodes are the most frequent approach in cognitive architectures.

For example, in the episodic memory systems developed by Nuxoll and Laird [13], Kuppaswamy et al. [11], and Dodd and Gutierrez [6], all of them present prototypical state-based episodes, where an episode is a sequence of snapshots from other memory components, e.g., working memory.

Also, in the case of Brom et al. [4], the agent’s base memory is a tree-like structure with all possible tasks the agent can perform. The episode representation is a path traversing previously executed tasks. This path is analogous to a sequence of the agent’s states. Therefore, this can be considered, also, a state-based representation.

The case that most resembles what we call here a scene-based episodic representation is given in the work of Martin et al. [12], which introduces a bio-inspired model of episodic memory. The proposed model interprets the agent’s sensory states into a high-level visuo-spatial representation termed a *frame*¹ formed by objects, their spatial configuration, and assigned categories from a determined instant. Two subsequent frames from the agents’ experience form frame associations. In this model, an episode is a chain of frame associations. Although the episodes contain high-level visuospatial elements within an instant, there is no temporal interpretation between instants, e.g., change in object’s properties. Therefore, we still consider that the episodes have a state-based representation, where each state is a *frame*.

3 The Episode Tracker Module

In this work, we describe the Episode Tracker Module, a sub-system of a cognitive architecture, which is inspired by the cognitive models of Baddeley [2], Tulving [16] and Gärdenfors [9], from Cognitive Psychology. The main task of

¹ Martin et al. [12] originally used the term *scene* to what we are calling here a *frame*. We are using *frame* here to avoid ambiguity with using the term *scene* in scene-based episodes.

the Episode Tracker Module is to encode perceptual data over time into scene-based episodes, which are made available for further processing in the Working Memory and eventually stored in a future Episodic Memory Module, which is still being developed.

From the neuropsychological point of view, the Episode Tracker Module is analogous to a structure called *Episodic Buffer*, within Baddeley's multi-component Working Memory model, as they both integrate sensory information across time and space into high-level representations [2,3].

Episodes generated by the Episode Tracker Module might be directly stored in the Episodic Memory Module. The Episodic Memory Module is functionally equivalent to Tulving's concept of Episodic Memory [15,16]. Our model presumes that Tulving's long-term Episodic Memory receives the encoded episodes from the Episodic Buffer. We consider two pieces of evidence for proposing this approach. First, the Episodic Buffer presumably interfaces information with long-term memory. Second, the episode representation is similar in both systems. Baddeley [2] also mentions this resemblance. Thus, the Episode Tracker Module is the Episodic Memory Module's encoding mechanism for perceptual data. Our scene-based representation of episodes, using objects and properties, is deeply inspired in Gärdenfors [9] conceptual spaces. Properties are bundles of quality dimensions, and objects are composed of multiple properties. However, we have a different description of events. Our definition of events is similar to actions in conceptual spaces, while events in conceptual spaces are closer to our definition of episodes. We define an event in the Episode Tracker Module as a relevant modification in a single object's property instance between two timesteps. Previous experiences and attentional mechanisms under the effect of conscious awareness select which events are relevant to track. Finally, we define an episode as an interpreted version of multiple events containing context between events from different objects.

3.1 The Episode Tracker Module's Architecture

The Episode Tracker was constructed using the Cognitive Systems Toolkit (CST), and after completion, will be available as a cognitive module in the toolkit. CST is a toolkit implemented in Java for building cognitive architectures under development by our research group [14]². A cognitive architecture built under CST is basically constructed relying on two fundamental components: codelets³ and memory objects⁴.

² CST's source is available in <https://github.com/CST-Group/cst>.

³ Codelets, first introduced in [10] and later enhanced in [7], are small segments of non-blocking code executed in a loop. Codelets run in parallel and are responsible for all data processing within the architecture.

⁴ Memory objects in CST hold any type of data structure to store information. Memory Objects are the canonical storage for data in the cognitive architecture. Different knowledge representation schemes might be used in each Memory Object.

The Episode Tracker Module’s architecture can be seen in Fig. 1, which describes how information is processed by a sequence of tasks, from receiving perceptual data to the final delivery of scene-based episodes. Rounded rectangles are codelets, yellow circles are memory objects, red circles are memory objects containing long-term memory categories, and the arrows represent the information flow.

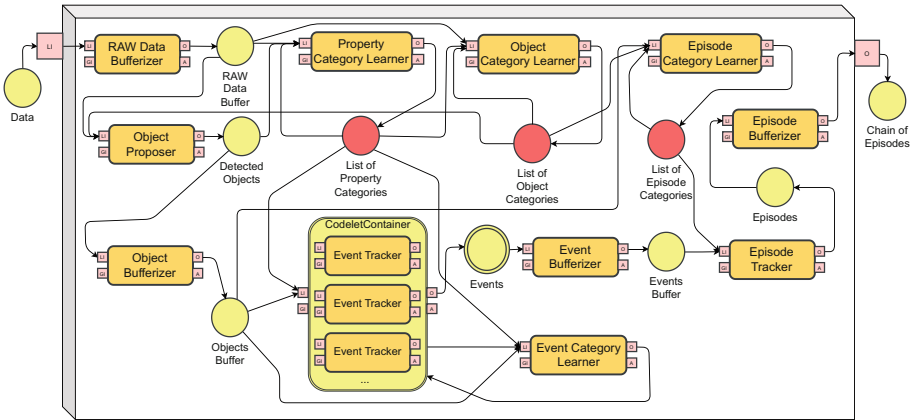


Fig. 1. Episode Tracker Module diagram.

First, the Raw Data Bufferizer receives the sensory data and creates a buffer containing the last n data states; all Bufferizers produce this same behavior. Then, the Object Proposer reads the buffer’s information detecting objects and their respective property values. Furthermore, the Object Proposer can assign object and property categories using the information provided by the List of Object Categories and the List of Property Categories, respectively.

The codelet container receives the object’s information along time, as a buffer of objects’ states. The codelet container is a CST class that allows the operation of a dynamic number of codelets to be included, operating on the same inputs and outputs. All its codelets receive the same input information and manage access to the output. Each codelet container’s Event Tracker detects a specific change in property values within one object. It can also track a shift in an object’s property categories using the List of Property Categories’ information. We consider an event an atomic episode since our episode representation is a chain of multiple events.

Finally, the Episode Tracker codelet interprets sequences of events into episodes. Episodes can group sequences of events and capture the cause-and-effect relations from multiple objects’ events. This codelet can also detect pre-existent categorized patterns of episodes through the List of Episode Categories’ information.

The resultant episode representation is scene-based since it encodes high-level visuospatial and temporal information. Visuospatial interpretation shapes objects and their properties at each instant of time. Temporal interpretation takes the form of changes in objects' properties through time and their relations (i.e., events and episodes).

4 Experiments

Our experiments used GPS data from mobile devices as sensory input to the Episode Tracker Module. Subjects from our research group commuted in the neighborhood around the UNICAMP campus, collecting GPS data (latitude, longitude, timestamp) from their smartphones using a rate of 1 Hz frequency. The Episode Tracker Module then uses this sensory input to represent the entity *user* and detect how the user changes its property *location* over time, creating relevant episodes. The challenge is inferring and categorizing relevant regions and then evaluating the movement from one relevant region to the next, assembling meaningful movement episodes. In the end, the Episode Tracker Module outputs sequences of episodes describing meaningful episodes describing both staying in a relevant position and moving from one relevant position to the next along the day.

We introduced two methods for recognizing property categories, objects, and episodes. The first method is the pheromone algorithm that learns the relevant regions online using the subject's permanence in a determined position as criteria. The second solution uses an offline clustering solution for generating the relevant regions and then inputs the regions as property categories in the Episode Tracker.

4.1 Pheromone Algorithm for Event Detection

We designed an algorithm, termed the pheromone algorithm, for detecting relevant regions. We employ the algorithm in the Property Category Learner codelet. The codelet outputs the relevant regions as property categories stored in the Property Categories memory object.

The main idea behind the algorithm is that it deems a region relevant if the subject has spent substantial time parked in that region. The representation of each region is a list of circles with a relevance value. The region is relevant if its relevance value exceeds a pre-setted relevance threshold value. The algorithm receives the subject's current location and updates the detected relevant regions. For each received location, the algorithm creates a circular area centered around the location's coordinates. If there is an overlap between any existing region and the new circle, the overlapping region's relevance increases, and the new circle is appended to the region. If there is no overlap, it creates a new region containing solely the new circle. After this evaluation, it removes regions below a pre-setted minimum value, and regions below the relevance threshold value have their relevance value multiplied by a decay rate parameter ranging from 0 to 1.

4.2 Location Clusterization Algorithm for Event Detection

Clustering algorithms are data grouping processes. They are an unsupervised learning method where the goal is to divide a dataset into clusters such that the elements within a cluster are similar to each other and different from the elements in other clusters. There are several clustering algorithms, and in this experiment, we used the MeanShift algorithm [8]. This density-based clustering algorithm works by identifying dense regions of points in the data and assigning each point to the cluster whose density is closest without defining the number of clusters beforehand. The clustering result represents relevant regions through GPS sensor data, represented by latitude, longitude, and timestamp, and processed offline to determine the regions. Latitude, longitude, timestamp, distance, and speed data were defined as input features in the cluster. Based on these characteristics, the MeanShift algorithm calculates relevant regions. After grouping, outliers were removed to keep only the concentrated locations, resulting in the relevant regions, which are the relevant property categories, being stored in the List of Property Categories memory object.

5 Results

The results of the experiments demonstrate the capability of the Episode Tracker Module to generate property categories and utilize them for event detection. The architecture shows its flexibility by employing both online and offline identification of relevant regions through the pheromone and clustering algorithms in Fig. 2a and b, respectively. This allows for tracking events over time, enabling the creation of a trajectory history. The Episode Tracker Module successfully marked the relevant regions for users using both the Pheromone Algorithm and the Clustering Algorithm. Additionally, it effectively detected events such as entering, leaving, and staying within these regions.

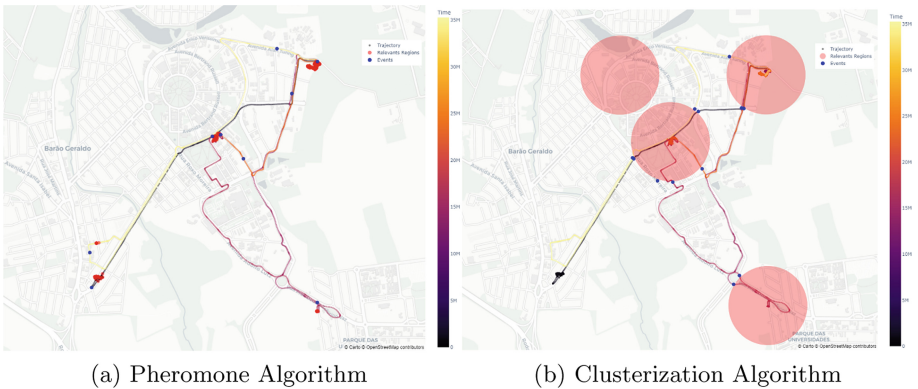


Fig. 2. Event detection algorithms: pheromone and location clusterization.

The offline identification of relevant regions can be seen as a mechanism that queries information in long-term memory. This process is similar to focusing attention on salient features or events that are considered important as a reference. By storing these relevant regions offline, the Episode Tracker Module uses memory information for event detection and tracking. Furthermore, the online identification of relevant regions aligns with the concept of selective attention, where the system dynamically focuses on specific regions of interest in real time. This mechanism enables the detection and immediate response to events, replicating the attentional processes observed in cognitive models. By incorporating both offline and online identification of relevant regions, the Episode Tracker Module combines the advantages of long-term memory and selective attention. This association with attentional models enhances the system's ability to adapt and respond to changing environments, making it a powerful tool for event detection and tracking tasks.

6 Conclusion

Our work introduces the Episode Tracker Module, a cognitive module within the Cognitive Systems Toolkit that interprets sensory information through time and space into scene-based episodes. Its modular design allows the association of different algorithms or learning models with the same cognitive module, which can be effective in identifying events from sensor data on mobile devices. Two algorithms were implemented and tested: a pheromone algorithm that identifies relevant regions online and a clustering algorithm that identifies relevant regions offline. The results of the experiments showed that the architecture can deal with different ways of storing category properties and that both algorithms were able to identify events accurately. The modular cognitive architecture allows different approaches to be experimented with and evaluated, enabling the identification of better solutions for the problem in question.

Acknowledgements. This project was supported by the Ministry of Science, Technology, and Innovation of Brazil, PPI-Softex grant # [01245.013778/2020-21]. The authors also thank CEPID/BRAINN (Proc. FAPESP 2013/07559-3).

References

1. Allen, T.A., Fortin, N.J.: The evolution of episodic memory. *Proc. Natl. Acad. Sci.* **110**(supplement_2), 10379–10386 (2013)
2. Baddeley, A.: The episodic buffer: a new component of working memory? *Trends Cogn. Sci.* **4**(11), 417–423 (2000). ISSN: 1364-6613
3. Baddeley, A.D., Allen, R.J., Hitch, G.J.: Binding in visual working memory: the role of the episodic buffer. *Neuropsychologia* **49**(6), 1393–1400 (2011). ISSN: 0028-3932

4. Brom, C., Lukavský, J., Kadlec, R.: Episodic memory for human-like agents and human-like agents for episodic memory. *Int. J. Mach. Conscious.* **02**(02), 227–244 (2010). ISSN: 1793-8430. Publisher: World Scientific Publishing Co
5. Castro, E.C., Gudwin, R.R.: A scene-based episodic memory system for a simulated autonomous creature. *Int. J. Synth. Emot. (IJSE)* **4**(1), 32–64 (2013). ISSN: 1947-9093. Publisher: IGI Global
6. Dodd, W., Gutierrez, R.: The role of episodic memory and emotion in a cognitive robot. In: *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005*, pp. 692–697 (2005). ISSN: 1944-9437
7. Franklin, S., Kelemen, A., McCauley, L.: IDA: a cognitive agent architecture. In: *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218)*, vol. 3, pp. 2646–2651 (1998). ISSN: 1062-922X
8. Fukunaga, K., Hostetler, L.: The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inf. Theory* **21**(1), 32–40 (1975)
9. Gärdenfors, P.: *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press (2014). ISBN: 978-0-262-02678-9
10. Hofstadter, D.R., Mitchell, M.: The copycat project: a model of mental fluidity and analogy-making. In: *Analogical Connections. Advances in Connectionist and Neural Computation Theory*, vol. 2, pp. 31–112. Ablex Publishing, Westport, CT (1994). ISBN: 978-1-56750-039-4
11. Kuppaswamy, N.S., Cho, S.H., Kim, J.H.: A cognitive control architecture for an artificial creature using episodic memory. In: *2006 SICE-ICASE International Joint Conference*, pp. 3104–3110 (2006)
12. Martin, L., Jaime, K., Ramos, F., Robles, F.: Bio-inspired cognitive architecture of episodic memory. *Cogn. Syst. Res.* **76**, 26–45 (2022). ISSN: 1389-0417
13. Nuxoll, A.M., Laird, J.E.: Enhancing intelligent agents with episodic memory. *Cogn. Syst. Res.* **17–18**, 34–48 (2012). ISSN: 1389-0417
14. Paraense, A.L.O., Raizer, K., de Paula, S.M., Rohmer, E., Gudwin, R.R.: The cognitive systems toolkit and the CST reference cognitive architecture. *Biol. Inspired Cogn. Archit.* **17**, 32–48 (2016). ISSN: 2212-683X
15. Tulving, E.: Episodic and semantic memory. In: *Organization of Memory*, pp. xiii, 423. Academic Press, Oxford, England (1972)
16. Tulving, E.: Episodic memory: from mind to brain. *Annu. Rev. Psychol.* 1–25 (2002)



Application of Machine Learning to Construct Solitons of Generalized Nonlinear Schrödinger Equation

A. G. Sboev^{1,2}(✉), N. A. Kudryashov¹, I. A. Moloshnikov^{1,2}, D. R. Nifontov¹,
S. V. Zavertyaev², and R. B. Rybka^{1,2}

¹ National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, Russian Federation
sag111@mail.ru

² National Research Centre “Kurchatov Institute”, Moscow, Russia
zavertyaev.sv@phystech.edu

Abstract. This paper presents an application of the PINN method to solving the non-integrable soliton generalized nonlinear Schrödinger equation, with evaluations on the solution accuracy of the influence of introducing in the PINN algorithm the conservation laws. An effect of choice of different activation functions is also analyzed. Despite the fact that obtained results demonstrate the best accuracy for the model with the SIN activation function without a use of conservation laws, experimental results show that they may be useful in cases of non-optimal choices of activation functions or collocation points.

Keywords: Nonlinear differential equation · Exact solution · Arbitrary refractive index · Optical soliton

1 Introduction

Currently, a development of fiber optic communication lines widely used to transmit information among technological devices is a highly relevant and practically meaningful task. One of the most popular equations used in the description of the propagation of impulses in nonlinear media - the nonlinear Schrödinger equation, looking like

$$i q_t + a q_{xx} + \alpha |q|^2 q = 0, \quad (1)$$

where $q(x, t)$ is a complex function, a and α are parameters.

Well known, Eq. (1) belongs to the class of integrable equations, and the Cauchy problem for it is solved by the method of the inverse scattering problem. This is the main attractiveness of the nonlinear Schrödinger equation with the Kerr nonlinearity taking into account of dependence of the refractive index in the optical medium. It has soliton solutions for the wave packet envelope and has numerous applications for analysis and interpretation of numerous experimental data emerging during the propagation of pulses in optical media. However, in a

number of practically important cases, the refractive index in a medium depends on the radiation power in a more complex form. In this paper, we study the processes described by the generalized nonlinear Schrödinger equation having the form

$$i q_t + a q_{xx} + \alpha |q|^{2n} q - \beta |q|^{4n} q = 0, \tag{2}$$

where $q(x, t)$ is a complex function, n is real number, x and t are independent variables, α , β and a are parameters of mathematical model. In contrast to the nonlinear equation Schrödinger, Eq. (2) does not belong to the class of completely integrable partial differential equation. Currently, physics-informed neural networks (PINN) models become a commonly used tool to solve a wide circle of physical domain problems [1]. Despite the confirmed applicability of PINN for solving many physical equations, a set of cases are mentioned in the literature the [2], when the PINN approach fails. In this paper, we present the application of the method to solving the non-integrable soliton equation and investigate an influence of choice of different activation functions and the use of conservation laws on the solution accuracy. Approaches of solving nonlinear Schrödinger equations such as [3–8] by PINN methods has been broadly considered in recent papers. However the effect of applying external physics laws (such as Power, Moment and Energy conservation) and the limitations of classic PINNs to such problems is still to be discovered. Applying additional losses of a conserved quantity to other problems such as Cahn-Hilliard equation is stated to improve accuracy and convergence [9]. Another method considered is BC-PINN [10] that is stated to reach convergence in problems, where classic PINN failed.

2 Optical Soliton of Eq. (2)

Let us look for the solution of Eq. (2) in the form

$$q(x, t) = y(z) e^{i(kx - \omega t + \theta_0)}, \quad z = x - C_0 t, \tag{3}$$

where k and C_0 are parameters of solution, θ_0 is an arbitrary constant.

Substituting (3) into Eq. (2) and equating an imaginary and a real parts to zero, we and solving for the system of resulting equation we get a solution of Eq. (2) as follows:

$$q(x, t) = \left[\frac{4\mu e^{(x-2ak t-z_0)\sqrt{\mu}}}{1 + 2\lambda e^{(x-2ak t-z_0)\sqrt{\mu}} + (\lambda^2 - 4\mu\nu) e^{2(x-2ak t-z_0)\sqrt{\mu}}} \right]^{\frac{1}{2n}} e^{i(kx - \omega t + \theta_0)}, \tag{4}$$

$$\lambda = \frac{4\alpha n^2}{a(1+n)}, \quad \nu = \frac{4\beta n^2}{a(1+2n)}, \quad \mu = \frac{4n^2(\omega - ak^2)}{a}. \tag{5}$$

3 Solving a Differential Equation Using Neural Networks

The original Eq. (2) is solved in the range:

Constant value: $n = 1, a = 1, \alpha = 1, \beta = 1, k = 1, z_0 = 0, \theta_0 = 0, \omega = 9/8$. For the initial condition t_0 , the points are obtained using the exact solution (4) ($q_{exact}(x, t_0)$ from (4)). Boundary and initial condition:

$$q(x, -5) = q_{exact}(x, -5), \quad x \in [-45, 45], \quad t \in [-5, 5] \tag{6}$$

$$q(45, t) = q(-45, t) = q_x(45, t) = q_x(-45, t) = 0 \tag{7}$$

In neural network we use decomposed form of equation with $q = u + iv$.

4 Neural Network

A neural network (MLP) with two inputs x and t and two outputs for u and v is used. In all experiments, except for the experiments with small model, the network consists of 4 hidden fully connected layers of 100 neurons each with a selected activation function and the output layer of 2 neurons with a linear activation function to predict u and v . For experiments with small model, a model with 2 layers of 32 neurons was used. We tried different activation functions: sin, cos, sigmoid, tanh, the comparison results are shown in Table 2 (sin - sine, cos - cosine, tanh - hyperbolic tangent, sigmoid - sigmoid activation function, K - function proposed by Kudryashov [11], that when transferred to neural networks can be reduced to sech or $K(x) = \frac{1}{e^x + e^{-x}}$). To form collocation points, we tried several options: **random** - generating once for the entire training 20k points uniformly randomly from the entire domain of definition, **reinit rand** - generating 20k collocation points uniformly randomly every 1k Adam training steps, **high loss** - generating 20k collocation points every 1k Adam steps, of which 10k points are uniformly random, 10k with the maximum error on the differential equation, **like bc** - a simplified version of bc-pinn [10], gradually increase the domain of definition, while always choosing 20k collocation points. To do this, we limit the domain of definition by t for the first 1k training iterations to 1/50 of the entire measurement, and every 1k steps we increase the domain of definition by 1/50 and so on 50 times, **mesh** - generating once 25k points on a uniform grid in the domain of definition, 50 slices of t each slice of 500 points of x . Adam and L-BFGS were used as an optimizer. Further “use lbfgs = True” - means that after Adam the L-BFGS optimizer was used, “use lbfgs = False” - only Adam was used. L-BFGS hyperparameters: max_iterations = 50000, tolerance = float32 lower limit, max_line_search_iterations = 50, num_correction_

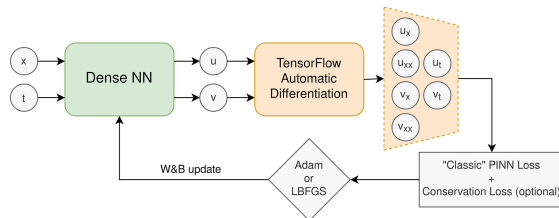


Fig. 1. Schematic diagram of the proposed neural network

pairs = 50. The L-BFGS implementation from TensorFlow Probability¹ was used. For the Adam optimizer, a decaying learning rate (lr) was used in all runs, every 100 training iterations the learning rate decreased by a factor of decay rate ($lr = lr * decay_rate$). Further, under the learning rate, we present the initial value of lr.

5 Conserved Quantities

Equation (2) can be rewritten in so-called conservation form in three different ways:

$$\frac{\partial T_j}{\partial t} + \frac{\partial X_j}{\partial x} = 0, \quad j \in 1, 2, 3 \tag{8}$$

where T_j called density and X_j called flow. T_j and X_j are:

$$T_1 = i |q|^2, \quad X_1 = a (q^* q_x - q q_x^*), \quad T_2 = \frac{i a}{2} (q_x^* q - q_x q^*), \tag{9}$$

$$X_2 = \frac{i}{2} (q^* q_t - q_t^* q) + a |q_x|^2 + \frac{\alpha}{n + 1} |q|^{2n+2} - \frac{\beta}{2n + 1} |q|^{4n+2}, \tag{10}$$

$$T_3 = \frac{i \alpha |q|^{2n+2}}{2n + 2} - \frac{i \beta |q|^{4n+2}}{4n + 2} - i a |q_x|^2, \quad X_3 = i (q_x^* q_t + q_x q_t^*). \tag{11}$$

From equations in form (8) three conserved quantities C_j can be derived:

$$P = \int_{-\infty}^{\infty} \text{Im} (T_1) dx = C_1, \quad M = \int_{-\infty}^{\infty} \text{Re} (T_2) dx = C_2, \quad H = \int_{-\infty}^{\infty} \text{Im} (T_3) dx = C_3 \tag{12}$$

As shown in the section, there are three conserved quantities for this equation, after substituting $q = u + iv$ into Eq. (12) we get:

$$C1 = \int_{-\infty}^{\infty} (u^2 + v^2) dx, \quad C2 = \int_{-\infty}^{\infty} (uv_x - u_x v) dx \tag{13}$$

$$C3 = \int_{-\infty}^{\infty} a(u_x^2 + v_x^2) - \frac{\alpha(u^2 + v^2)^{n+1}}{2n + 2} + \frac{\beta * (u^2 + v^2)^{2n+1}}{2n + 4} dx \tag{14}$$

Let us denote the resulting integrands as functions F_{c1}, F_{c2}, F_{c3} :

$$F_{c1} = u^2 + v^2, \quad F_{c2} = uv_x - u_x v \tag{15}$$

$$F_{c3} = a(u_x^2 + v_x^2) - \frac{\alpha(u^2 + v^2)^{n+1}}{2n + 2} + \frac{\beta * (u^2 + v^2)^{2n+1}}{2n + 4} \tag{16}$$

Then Eq. (14) can be rewritten as:

$$C1 = \int_{-\infty}^{\infty} F_{c1}(u, v) dx, \quad C2 = \int_{-\infty}^{\infty} F_{c2}(u, v, u_x, v_x) dx, \\ C3 = \int_{-\infty}^{\infty} F_{c3}(u, v, u_x, v_x) dx. \tag{17}$$

Provided that the points are generated using a grid with a dx step along the x axis, one can numerically calculate the predicted constants for each slice in t by replacing

¹ <https://www.tensorflow.org/probability?hl=en>.

the integral with the corresponding sum and substituting the corresponding outputs of the neural network and their derivatives into Eq. (17): $I\hat{1}_j = \sum_{i=1}^{Nx} \hat{F}_{c1}(x_i, t_j) \cdot dx$, $I\hat{2}_j = \sum_{i=1}^{Nx} \hat{F}_{c2}(x_i, t_j) \cdot dx$, $I\hat{3}_j = \sum_{i=1}^{Nx} \hat{F}_{c3}(x_i, t_j) \cdot dx$, where $x_i = x_{min} + i \cdot dx$, t_j - time slice ($t_j = t_{min} + j \cdot dt$), dx, dt - coordinate and time grid steps respectively, i, j - step number by coordinate and time respectively, $\hat{F}_{c1}, \hat{F}_{c2}$ and \hat{F}_{c3} - function values F from formula (16) predicted by the neural network. Having calculated the predicted $I\hat{1}_j, I\hat{2}_j, I\hat{3}_j$, we can implement a loss function to minimize the difference between these values and $C1, C2, C3$ calculated from the analytical solution at the initial time. This idea underlies the term of the loss function $loss_{cl}$ (Eq. 23).

6 Loss Function for Neural Network with Conserved Quantities

Loss function for the case when conservation laws are not used ($nlaws = 0$)

$$MSE_{t_0} = \frac{1}{N_{t_0}} \sum_{i=1}^{N_{t_0}} (q_{exact}(x_i, t_0) - pred_val_i(x_i, t_0))^2 \tag{18}$$

$$MSE_{bc} = \frac{1}{N_{bc}} \sum_{j=1}^{N_{bc}} (0 - pred_val_i(x_{bc}, t_j))^2 \tag{19}$$

$$MSE_c = \frac{1}{N_c} \sum_{k=1}^{N_c} (re_{pred}(x_k, t_k))^2 + \frac{1}{N_c} \sum_{k=1}^{N_c} (im_{pred}(x_k, t_k))^2 \tag{20}$$

$$Loss = MSE_{t_0} + MSE_{bc} + MSE_c \tag{21}$$

where t_0 - initial time, bc - boundary points, N_{bc} - number of boundary points, N_c - number of collocation points, N_{t_0} - number of initial points. re_{pred} and im_{pred} are calculated by substituting the predicted values and derivatives taken using tensorflow automatic differentiation into decomposed equations.

Loss function according to conservation laws:

$$lwloss_n = \frac{1}{Nt} \sum_{j=1}^{Nt} (I_n - \sum_{i=1}^{Nx} \hat{F}_{cn}(x_i, t_j) \cdot dx)^2, n \in \{1, 2, 3\} \tag{22}$$

where Nt is the number of slices by t in the method of generating collocation points by the grid (mesh), Nx is the number of points by x in one slice. $|q_pred(x_i, t_j)|^2$ - value calculated based on neural network outputs. $I1, I2, I3$ - numerical integral over the initial slice taken using the quad function from the scipy² library, $I \approx C$ from Eq. (17).

$$loss_{cl} = th \left(\frac{\sum_{i=1}^{nlaws} lwloss_i}{\sum_{i=1}^{nlaws} I_i^2} \right), nlaws \in \{1, 2, 3\} \tag{23}$$

where th is the hyperbolic tangent function, $nlaws \in \{1, 2, 3\}$ is the number of laws used, I_i is the numerically calculated value of the constants C_i from Eq. (12).

² <https://scipy.org/>.

7 Experiments

We test several hypotheses how different methods affect for accuracy: 1 - how activation functions and methods of generating collocation points affect, 2 - which gives the use of the L-BFGS optimizer, 3 - which gives the use of conserved quantities. To evaluate models, we use the Mean Squared Error (**MSE** $|h|$) metric of the module of the function. All estimates are made on 5 time slices t (slices for t values from the list: $-5, -2.5, 0, 2.5, 5$) with a fixed number of points (1k points uniformly from -45 to 45) in x .

The results of comparing different **activation functions** and **methods of generating collocation points** are given in Table 1. The model with the activation function K showed the best result. To check what **L-BFGS** gives, we compared runs with 50 thousand Adam iterations and then L-BFGS with a maximum number of 50 thousand iterations and a variant with only 100 thousand Adam iterations. The L-BFGS method showed better results compared to Adam. For all activation functions, the L-BFGS showed an increase of exactly an order of magnitude. The best run with using L-BFGS and activation function sin showed $\text{MSE } |h| = 1.312\text{e}-09$. The results of comparison of different activation functions and **influence of conserved quantities** on accuracy are given in Table 2, where laws - how many conservation laws were used. The difference between the results when using **conserved quantities** is clearly visible when using **small models** in terms of the number of neurons and network layers. Such models do not provide high accuracy, but they give an idea of how the conservation laws work when learning neural network. Figure 2 clearly show the difference in predictions. The model that was trained without laws (Fig. 2) shows a gradually fading peak, but falling within the boundaries of what should be. And the model that was trained with conservation laws (Fig. 2) predicts a peak of almost the desired shape, but with a slight shift. By metric “MSE $|h|$ ” these two launches are the same ($\text{MSE } |h| = 6\text{E}-03$), but the values of $loss_{cl}$ differ greatly, for the variant without laws $loss_{cl} = 4$, and for the variant with laws $loss_{cl} = 9\text{E}-05$.

Table 1. Comparison of various activation functions and methods for choosing collocation points. c gen - collocation point generation method

Activ./c gen	reinit rand	random	high loss	like bc	Avg dur
K	4.92e-07	4.82e-07	2.56e-06	2.73e-06	2.9 h
cos	4.50e-05	2.30e-05	4.45e-04	2.37e-04	1.4 h
sigmoid	1.92e-05	9.15e-05	4.05e-04	5.34e-05	1.4 h
sin	1.30e-06	8.07e-06	1.66e-06	7.34e-06	1.4 h
tanh	5.23e-05	1.72e-05	7.42e-05	2.71e-05	1.3 h

The calculations were carried out on machines with an Nvidia K80 GPU. On average, one experiment took about 2 h. In the tables, the duration is indicated in hours. Hyperparameters for examination of the effect of conservation laws in loss function, Table 2: collocation point generation method - mesh with 50 slices in t , 500 points in x , 50k iterations Adam, $lr = 0.001$, lr decay rate = 0.99, 1k points for t_0 initial conditions, 500 points for boundary conditions (for each condition), use `lbfgs = True`. Params for examination of influence of the activation function and methods of generating collocation points (Table 1) are similar to runs for examination of the effect of conservation

Table 2. Comparison of the impact on the accuracy of conservation laws, detailing for various activation functions. Collocation points are generated by **mesh**.

Activ./nlaws	0	1	2	3
cos	1.51E-07	1.31E-07	6.60E-08	1.02E-07
sigmoid	1.64E-03	1.86E-03	1.91E-04	1.25E-05
sin	7.87E-08	1.85E-07	1.50E-07	3.10E-06
tanh	1.40E-06	3.02E-06	8.32E-06	6.02E-06

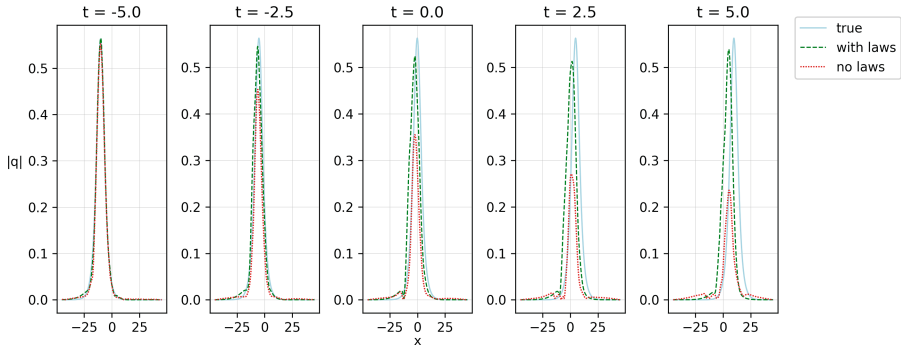


Fig. 2. Small models performance trained with and without laws.

laws, except: lr = 0.005, use lbfgs = False, lr decay rate = 0.997, 50 points for boundary conditions (for each condition), 50 points for initial conditions t_0 . For examination of L-BFGS effect params are similar to runs for examination of the influence of the activation function and methods of generating collocation points. Except use_lbfgs and the number of Adam iterations changed. For a small neural network params are similar to ones for examination of the influence of the activation function and methods for generating collocation points, except the number of layers in the network is 2×32 neurons.

8 Conclusion

Presented results of realized PINN method with the use of conservation laws for generalized nonlinear Schrodinger’s equation show its applicability to solve this class of equations and a possibility to use conservation laws, as added mean, to control the solution.

Acknowledgements. The study has been supported by the Russian Science Foundation grant No. 23-41-00070 <https://rscf.ru/project/23-41-00070/> and has been carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC “Kurchatov Institute”, <http://ckp.nrcki.ru/>.

References

1. Markidis, S.: The old and the new: can physics-informed deep-learning replace traditional linear solvers? *Front. Big Data* 92 (2021)
2. Wang, S., Yu, X., Perdikaris, P.: When and why PINNs fail to train: a neural tangent kernel perspective. *J. Comput. Phys.* **449**, 110768 (2022)
3. Wu, G.Z., Dai, C.Q., Wang, Y.Y., Chen, Y.X.: Propagation and interaction between special fractional soliton and soliton molecules in the inhomogeneous fiber. *J. Adv. Res.* **36**, 63–71 (2022). <https://doi.org/10.1016/j.jare.2021.05.004>. <https://www.sciencedirect.com/science/article/pii/S2090123221000904>
4. Dai, C.Q., Wu, G., Li, H.J., Wang, Y.Y.: Wick-type stochastic fractional solitons supported by quadratic-cubic nonlinearity. *Fractals* **29**(07), 2150192 (2021). <https://doi.org/10.1142/S0218348X21501929>
5. Kivshar, Y.S., Agrawal, G.P.: *Optical Solitons: From Fibers to Photonic Crystals*. Academic Press (2003)
6. Kudryashov, N.A.: Solitary wave solutions of hierarchy with non-local nonlinearity. *Appl. Math. Lett.* **103**, 106155 (2020)
7. Kudryashov, N.A.: Method for finding highly dispersive optical solitons of nonlinear differential equations. *Optik* **206**, 163550 (2020)
8. Kudryashov, N.A.: Mathematical model of propagation pulse in optical fiber with power nonlinearities. *Optik* **212**, 164750 (2020)
9. Wang, H., Qian, X., Sun, Y., Song, S.: A modified physics informed neural networks for solving the partial differential equation with conservation laws. Available at SSRN 4274376 (2022)
10. Matthey, R., Ghosh, S.: A novel sequential method to train physics informed neural networks for Allen Cahn and Cahn Hilliard equations. *Comput. Methods Appl. Mech. Eng.* **390**, 114474 (2022)
11. Kudryashov, N.A.: Solitary waves of the generalized Sasa–Satsuma equation with arbitrary refractive index. *Optik* **232**, 166540 (2021)



Spoken Digits Classification Using a Spiking Neural Network with Fixed Synaptic Weights

Alexander Sboev^{1,2,3(✉)}, Maksim Balykov^{1,4}, Dmitry Kunitsyn^{1,2},
and Alexey Serenko¹

¹ National Research Centre “Kurchatov Institute”, Moscow, Russia
Sboev_AG@nrcki.ru

² National Research Nuclear University MEPhI, Moscow, Russia

³ Russian Technological University “MIREA”, Moscow, Russia

⁴ Moscow Institute of Physics and Technology, Moscow, Russia

Abstract. The paper evaluates the applicability of an approach based on the usage of a spiking neural network with synaptic weights fixed from a uniform random distribution to solving audio data classification problems. On the example of the Free Spoken Digits Dataset pronounceable digit classification problem using a linear classifier trained on the output frequencies of spiking neurons as a decoder, an average accuracy of 94% was obtained. This shows that the proposed spiking neural network performs such a transformation of the audio data that makes it linearly separable. Numerical experiments demonstrated the stability of the algorithm to the parameters of the spike layer, and it was shown that the constants of the threshold potential and the membrane leakage time can be both equal and different for different neurons.

Keywords: Spiking neural networks · Reservoir computing · Audio classification

1 Introduction

The opportunities of implementation of spiking neural networks (SNNs) on neuromorphic computing hardware [1, 2] with ultra-low power consumption [3] make SNNs relevant for application to machine learning problems that are to be solved on autonomous devices. The advantage of neuromorphic hardware in terms of energy efficiency is most pronounced in streaming data analysis problems [4] such as sound classification.

Solving a classification problem using a spike neural network requires the following:

- encoding the input data into sequences of spikes to present to the input synapses of the network;
- obtaining the synaptic weights of the network;
- decoding output spike sequences into class labels.

There are several approaches to encoding input data [5]: time encoding is often used, where information is contained in the relative arrival times of input spikes, and rate encoding, when spikes arrive at the network at random times, and information is contained in the number of spikes.

Synaptic weights of an SNN can be obtained by transforming a pre-trained conventional neural network [6,7], or by training the SNN directly with error backpropagation [8,9]. Such approaches allow building multilayer CNNs that solve classification tasks with accuracy on par with artificial neural networks (ANNs). However, developing a method for training such networks that could be implemented on a neurochip still remains a problem.

Training spiking neural networks deployed on neuromorphic computing devices [10] is possible using local synaptic plasticity rules, in particular, Spike-Timing-Dependant Plasticity (STDP) [11]. The possibility of solving classification tasks by STDP-based learning is shown with rate [12,13], temporal [14] and correlation [15] encoding of input data. However, these CNNs do not have a multilayer topology and therefore require computationally expensive data pre-processing to solve nonlinear classification tasks.

One of the approaches to overcome this problem is to use a layer of spiking neurons with fixed weights as a data preprocessing layer. An example of this approach are reservoir networks, or Liquid State Machines (LSM). The reservoir, a layer of a large number of neurons with randomly given recurrent synaptic connections, performs some random non-linear transformation of the input data into a higher-dimensional space for further recognition by the trainable readout layer. Of the multitude of the reservoir neurons, such a subset of neurons is supposed to exist that classes can be distinguished by a linear combination of its outputs (spike rates or spike times). However, the hardware implementation of the LSM is hampered by the large number of its neurons and connections.

However, instead of a reservoir with recurrent connections, a fully-connected layer can be used with weights fixed based on logistic functions or from a uniform random distribution. This has been shown for conventional neural networks [16] and then for spiking networks [17].

The purpose of this work is to evaluate the applicability of a fully connected reservoir layer with random weights as a preprocessor for the classification of sound recordings using the Free Spoken Digits classification problem as an example. For this, the following has been done:

- optimal parameters of the spiking neuron layer are found;
- the robustness of solving the classification problem to the neuron parameters is studied by comparing the parameters that are fixed the same for all neurons to those distributed uniformly within a given range.

The search for optimal parameters and the study of their influence is carried out on two data sets described in Sect. 2: the initial search is carried out on Fisher’s Iris due to its small size, then FSDD is considered. The reservoir layer scheme is described in Sect. 3.1; it consists of Leaky Integrate-and-Fire spike neurons, described in Sect. 3.3. The efficiency of the reservoir layer is evaluated by the accuracy of solving the classification problem by the decoders described in Sect. 3.2.

2 Input Data

2.1 Fisher’s Iris Dataset

Fisher’s Iris dataset is used for initial parameter search due to its small size. The dataset consists of 150 records of iris flowers of 3 classes: Iris setosa, Iris versicolor and Iris virginica. A record is a vector of 4 features: sepal length, sepal width, petal length and petal width.

2.2 Free Spoken Digits Dataset

As a benchmark classification task of audio processing we consider the Free Spoken Digits Dataset (FSDD). The dataset contains 3000 recordings of pronunciations of digits from 0 to 9, 300 records of each digit. The pronunciations are split between 6 different people, and the task is to classify the recording by the person who pronounces it, thus 6 classes. For the sake of comparison to existing literature, we use a pre-determined splitting of the dataset into training and testing sets from University of Toronto School of Continuing Studies¹.

3 Methods

3.1 Reservoir Layer Structure

The proposed reservoir layer consists of N Leaky Integrate-and-Fire (LIF) neurons, to which P input Poisson spike generators are connected in the all-to-all fashion. Here N is an adjustable parameter, and P is the number of features of the input dataset. The connection weights w are randomly distributed in the range $[0, 5]$ and are constant throughout the whole classification process. The output of the reservoir layer are its output spiking rates: a vector of the number of spikes emitted by each neuron during the processing of the input vector.

3.2 Classification Algorithm

The efficiency of the proposed reservoir layer is assessed by evaluating the accuracy of solving the classification task by the following algorithm, depicted in Fig. 1.

¹ <http://github.com/ravasconcelos/spoken-digits-recognition>.

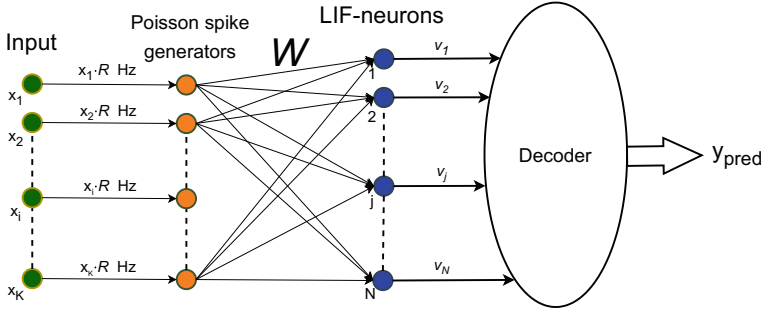


Fig. 1. The scheme of processing an input vector x into the class label y_{pred} using a layer of Leaky Integrate-and-Fire neurons and a linear decoder

Input data vectors are presented to the spiking network with rate encoding: each Poisson generator emits spikes with the mean rate $x \cdot r$ proportional to its corresponding component x of the input vector, where $r = 129 \text{ Hz}$ is kept as in earlier studies [17].

Input vectors of the dataset thus encoded are presented to the spiking network, and spike counts of all neurons in response to every vector are recorded. Then, solving the classification tasks is performed by a decoder, which is trained on the output spike rates in response to training set vectors, and then classifies the output spike rates in response to testing set vectors.

Two classifiers are used for decoding: Gradient Boosting (with 6000 estimators, the learning rate of 0.002, and maximum depth of 250), in order to assess how much information about the classes can be extracted from the spiking layer output, and Logistic Regression (with default parameters from the scikit-learn library), so as to assess whether the spiking layer output is linearly separable.

3.3 Neuron Model

The dynamic state of the LIF neurons is described by the following equation:

$$\frac{dV}{dt} = \frac{V_{rest} - V}{\tau_m} + \frac{I_{syn}(t)}{C_m}, \tag{1}$$

where $C_m = 406.62 \text{ pF}$ is the membrane capacity, $V_{rest} = -70 \text{ mV}$ is the resting potential, τ_m is the membrane leakage time constant adjusted separately for each of the two classification tasks, I_{syn} is the input synaptic current. I_{syn} is described as follows:

$$I_{syn} = \sum_i \sum_{t_{sp}^i} w_i(t_{sp}^i) \frac{q_{syn}}{\tau_{syn}} e^{\frac{t_{sp}^i - t}{\tau_{syn}}} \theta(t - t_{sp}^i),$$

where $q_{syn} = 5 \text{ fC}$, $\tau_{syn} = 5 \text{ ms}$, and the sum is performed over all input synapses i neurons and input spike timings t_{sp}^i .

As soon as $V(t) > V_{\text{th}}$, the neuron emits a spike. The neuron threshold V_{th} and membrane leakage time constant τ_m are adjusted separately for each of the two classification tasks. In order to assess the robustness of classification to these parameters, they are either the same for all neurons (the values found are in columns “Identical” in Table 1) or distributed uniformly within an adjustable range (the ranges found are in columns “Varying” in Table 1).

Table 1. Optimal spiking network parameter values found for each dataset, and optimal ranges found for parameters varying in different neurons

	Iris Fisher dataset		FSDD	
	Identical	Varying	Identical	Varying
V_{th} , mV	- 69.1	[- 69.864, - 68.36]	- 47.1	[- 48.4, - 35.8]
τ_m , ms	73.48	[1.74, 147.84]	76.52	[2.26, 150.79]
N	22		50	

4 Experiments

In order to test the feasibility of the approach under consideration, we first apply it to Fisher’s Iris classification. Classification performance is measured by the F1-macro score and obtained using 5-fold cross-validation: the Fisher’s Iris dataset is split into 5 equal parts of 30 vectors, 10 of each class, and one part is used for testing while the other ones are used for training the decoder. Random synaptic weights of the spiking network are drawn anew for each fold.

Adjustable parameters listed in Sect. 3 have been found using the Hyperopt library, using the training set F1-macro on Fisher’s Iris as the optimization objective. In order to decrease the computation cost and assess the robustness of the classification algorithm to the network parameters, for the FSDD classification task we sought to reuse as many parameters as possible from those found for Iris. The parameters that had to be adjusted separately for FSDD are presented in Table 1.

5 Results

The optimal values of the number of neurons in the reservoir layer N , as well as the ranges of the threshold V_{th} , differed substantially between the two classification tasks. On the contrary, the values of the membrane leakage time constant τ_m turned out to be similar.

F1-macro scores obtained on Fisher’s Iris and FSDD classification are presented in Tables 2 and 3 respectively. The two “Propose layer” rows correspond to the two ways of setting the neuron parameters as given in the respective columns

Table 2. F1-macro score of Fisher’s iris classification

	Gradient boosting			Logistic regression		
	Min	Mean	Max	Min	Mean	Max
Proposed layer with identical parameters of all neurons	0.91	0.94	0.97	0.94	0.96	0.98
Proposed layer with varying neuron parameters	0.92	0.94	0.97	0.93	0.96	0.98
Classifier without SNN	0.93	0.95	0.96	0.95	0.97	0.97
SNN with rate encoding [14]	0.87	0.92	0.97	–		
SNN with temporal encoding [14]	First-spike decoding					
	0.97	0.99	1.00			

Table 3. F1-macro score of Free Spoken Digits classification

	Gradient boosting			Logistic regression		
	Min	Mean	Max	Min	Mean	Max
Proposed layer with identical parameters of all neurons	0.91	0.92	0.93	0.93	0.94	0.94
Proposed layer with varying neuron parameters	0.91	0.92	0.93	0.937	0.944	0.949
Classifier without SNN	0.967	0.972	0.976	0.981	0.983	0.984

of Table 1, and scores are presented for two decoding classifiers. Mean, minimum, and maximum values are obtained over the five cross-validation folds for Fisher’s Iris; for FSDD with its fixed splitting into training and testing sets, five independent runs were performed with generating the random synaptic weights and training the decoding classifier.

For comparison, we present the accuracy of the classifiers trained on the dataset vectors directly without involving the spiking layer (denoted “Classifier without SNN”), and the accuracy of some other spiking networks from literature with different decoding.

Accuracies obtained using the proposed spiking layer are on par with those obtained by the classifiers alone, thus proving that the output rates of the spiking neurons retain all the sufficient information about the classes. Setting V_{th} and τ_m of all neurons identical or distributing them evenly along the optimal range leads to similar accuracy, suggesting that the classification performance is rather robust to these parameters within their optimal ranges.

6 Conclusion

A layer of spike neurons with synaptic weights fixed from a uniform random distribution is capable of preprocessing audio signals represented by spectral features for their subsequent classification. On the benchmark of classifying spoken digits by the speaker in the Free Spoken Digits Dataset, a classifier based on logistic regression, trained on the number of output spikes of neurons, showed accuracy of 94%. Numerical experiments show that the constants of the threshold potential and the membrane leakage time can be set either the same or different for different neurons, which proves the robustness of the classification algorithm to these parameters.

Thus, spiking neurons with non-trainable weights have been shown to be applicable within a classification algorithm for audio data processing.

Acknowledgements. The study has been supported by the Russian Science Foundation grant No. 23-11-00260 <https://rscf.ru/project/23-11-00260/> and has been carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC “Kurchatov Institute”, <http://ckp.nrcki.ru>.


References

1. Merolla, P.A., et al.: A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**(6197), 668–673 (2014)
2. Davies, M., et al.: Loihi: a neuromorphic manycore processor with on-chip learning. *IEEE Micro* **38**(1), 82–99 (2018). <https://doi.org/10.1109/MM.2018.112130359>
3. Rajendran, B., Sebastian, A., Schmuker, M., Srinivasa, N., Eleftheriou, E.: Low-power neuromorphic hardware for signal processing applications: a review of architectural and system-level design approaches. *IEEE Signal Process. Mag.* **36**(6), 97–110 (2019). <https://doi.org/10.1109/MSP.2019.293371>
4. Timcheck, J., et al.: The intel neuromorphic DNS challenge (2023)
5. Auge, D., Hille, J., Mueller, E., Knoll, A.: A survey of encoding techniques for signal processing in spiking neural networks. *Neural Process. Lett.* 1–18 (2021). <https://doi.org/10.1007/s11063-021-10562>
6. Esser, S.K., Appuswamy, R., Merolla, P., Arthur, J.V., Modha, D.S.: Backpropagation for energy-efficient neuromorphic computing. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 28, pp. 1117–1125. Curran Associates, Inc. (2015). <https://proceedings.neurips.cc/paper/2015/file/10a5ab2db37feedfdeaab192ead4ac0e-Paper.pdf>
7. Diehl, P.U., Zarella, G., Cassidy, A., Pedroni, B.U., Neftci, E.: Conversion of artificial recurrent neural networks to spiking neural networks for low-power neuromorphic hardware. In: *IEEE International Conference on Rebooting Computing (ICRC)*, pp. 1–8. IEEE (2016)
8. Bohte, S.M., Kok, J.N., La Poutre, H.: Error-backpropagation in temporally encoded networks of spiking neurons. *Neurocomputing* **48**(1), 17–37 (2002)

9. Shrestha, S.B., Orchard, G.: Slayer: spike layer error reassignment in time (2018)
10. Saïghi, S., et al.: Plasticity in memristive devices for spiking neural networks. *Front. Neurosci.* **9**, 51 (2015). <https://doi.org/10.3389/fnins.2015.00051>
11. Buonomano, D.V., Carvalho, T.P.: Spike-timing-dependent plasticity (STDP) (2009). <https://doi.org/10.1016/B978-008045046-9.00822-6>
12. Diehl, P.U., Cook, M.: Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Front. Comput. Neurosci.* (2015). <https://doi.org/10.3389/fncom.2015.00099>
13. Demin, V., et al.: Necessary conditions for STDP-based pattern recognition learning in a memristive spiking neural network. *Neural Netw.* **134**, 64–75 (2021). <https://doi.org/10.1016/j.neunet.2020.11.005>
14. Sboev, A., Serenko, A., Rybka, R., Vlasov, D.: Solving a classification task by spiking neural network with STDP based on rate and temporal input encoding. *Math. Methods Appl. Sci.* **43**(13), 7802–7814 (2020). <https://doi.org/10.1002/mma.6241>
15. Sboev, A., Serenko, A., Rybka, R.: Correlation encoding of input data for solving a classification task by a spiking neural network with spike-timing-dependent plasticity. In: *Biologically Inspired Cognitive Architectures*, No. 1032 in *Studies in Computational Intelligence*, pp. 457–462. Springer International Publishing (2022). https://doi.org/10.1007/978-3-030-96993-6_51
16. Velichko, A.: Neural network for low-memory IoT devices and MNIST image recognition using kernels based on logistic map. *Electronics* **9**(9) (2020). <https://doi.org/10.3390/electronics9091432>
17. Vlasov, D., Davydov, Y., Serenko, A., Rybka, R., Sboev, A.: Spoken digits classification based on spiking neural networks with memristor-based STDP. In: *International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 330–336. IEEE (2022)



A Brain-Inspired Cognitive Architecture (BICA) Approach to the Neurosymbolic Gap

Howard Schneider^(✉) 

Sheppard Clinic North, Vaughan, ON, Canada
hschneidermd@alum.mit.edu

Abstract. In this paper we consider a brain-inspired cognitive architecture approach to the neurosymbolic gap. The difference in the abilities of artificial neural networks (e.g., excellent perception) and symbolic systems (e.g., excellent logic) can be referred to as the neurosymbolic gap. Most attempts to combine properties of neural networks and symbolic systems are hybrid combinations of these different systems. A brain-inspired cognitive architecture (BICA), the Causal Cognitive Architecture 5 (CCA5), has both connectionist and symbolic properties. This architecture uses spatial navigation maps as the common data structure and requires spatial and temporal binding of inputs, predictive coding, innate knowledge procedures, and the ability to feed back and re-operate on intermediate results. We show how this BICA approach closes the neurosymbolic gap without the need to overtly combine separate symbolic systems and neural networks. As well, given that the BICA model presented is inspired by the mammalian and in particular the human brain, it provides insight into the mechanisms at work in cognition.

Keywords: Cognitive architecture · Artificial intelligence (AI) · Neurosymbolic

1 The Neurosymbolic Gap and Hybrid Solutions

In this paper we explore a brain-inspired cognitive architecture (BICA) approach to the neurosymbolic gap. We consider how this BICA approach closes the neurosymbolic gap without the need to overtly and often awkwardly combine a symbolic system to an artificial neural network. Given that the BICA approach presented below is inspired by the mammalian and in particular the human brain, it provides insight into the mechanisms at work in cognition and in discussions of consciousness.

In terms of perception, artificial neural networks (ANN's) perform on the level of humans [1]. However, in terms of logically making sense of a problem at hand, especially if training examples are limited, their performance barely compares to the level of a four-year old child [2]. Enhanced deep learning generative transformer models are able to write human prose or generate images in response to text. However, as Leivada, Murphy and Marcus [3] point out, such systems are not able to understand language actually logically at the level of a small child. There is an effective lack of logical abilities in neural network-based artificial intelligence (AI) systems. This difference in the abilities

of neural network AI systems and symbolic, logical AI systems is sometimes referred to as the “neurosymbolic gap” [4, 5].

Neurosymbolic AI attempts to combine properties of ANN’s and symbolic systems, i.e., neural network learning properties with the ability for symbolic reasoning. Garcez and Lamb [6] review research that attempts to integrate artificial neural networks with logical symbolic systems. Kautz [7] summarizes the field of neurosymbolic AI systems and divides it into six major designs. All of these designs are hybrid systems, i.e., part neural network and part symbolic system. As an example, one of Kautz’s groups are “Symbolic[Neuro]” systems which use a neural network recognition method within a symbolic AI system. Cingillioglu [8] classifies neurosymbolic AI from more neural to more symbolic, but again all these are hybrid systems.

Garcez and Lamb [6] note that neurosymbolic AI systems attempt to provide a bridge between localist and distributed representations. Below we describe a different type of non-hybrid neurosymbolic AI system, a brain-inspired cognitive architecture, that does not require or use any such bridge.

2 A Non-hybrid Solution to the Neurosymbolic Gap

In this paper we explore a non-hybrid (i.e., non-overtly hybrid) solution to the neurosymbolic gap that does not combine a typical artificial neural network with a typical symbolic AI system, as the previous solutions described above all use to some extent. The solution which we present is the Causal Cognitive Architecture. This architecture is a brain-inspired cognitive architecture (BICA) [9–16]. Given the existence of spatial maps in the hippocampi in mammals, as well as navigational abilities throughout the vertebrates and many invertebrates, the architecture postulates and requires the use of navigation maps not just for navigation but in the core mechanisms of the architecture. As well, the architecture requires spatial and temporal binding of sensory inputs, predictive coding (errors between what the architecture thought it would sense and actually senses are propagated forward), innate knowledge procedures concerning objects, physics, agents, numbers, and social group members, the ability to feed back and re-operate on intermediate results and optionally core analogical processing of sensory inputs and stored data.

The current version of the architecture, the Causal Cognitive Architecture 5 (CCA5), is shown in Fig. 1. The principal, recurring data element of the architecture is the “navigation map,” an example is shown in Fig. 2. The navigation map is not a typical neural network, although it is connectionist in operation. There is not one navigation map in the architecture but millions or billions of them [15].

3 Operation of the Causal Cognitive Architecture 5 (CCA5)

3.1 Cognitive Cycles

Each cognitive cycle, sensory features streaming in from different perceptual sensors are processed by the Input Sensory Vectors Shaping Modules (Fig. 1) and made compatible with the navigation map format (Fig. 2) used internally by the architecture. Each sensory system’s inputs are propagated to the Input Sensory Vectors Association Modules

(Fig. 1) where they are spatially bound, i.e., mapped onto a best-matching local (i.e., local to a particular sensory system) navigation map. Then in the Object Segmentation Gateway Module (Fig. 1), objects detected in the sensory inputs are segmented, and visual, auditory, and other sensory features of each segmented object are spatially mapped onto additional navigation maps dedicated to one sensory modality. This represents the first step in spatial object binding [13–16].

As well, a parallel sensory stream has gone through the Sequential/Error Correcting Module (Fig. 1) which converts changes with time into a vector value which is also bound along with the spatial features onto the same navigation maps, effectively representing temporal binding (not shown in Figs. 3 or 4) [14–16].

These single-sensory navigation maps are then mapped onto a best-matching multi-sensory navigation map taken from the Causal Memory Module (Figs. 1, 3 and 4). This represents the second step in spatial and temporal object binding [13–16].

The best-matching local (i.e., for each sensory system) navigation maps and the best-matching multi-sensory navigation maps effectively represent what the architecture expects to see (Fig. 3). The information mapped onto these best-matching local and multisensory navigation maps (Fig. 4) represents effectively changes or errors in what the actual sensory inputs actually were, i.e., an effective predictive coding is occurring rather than memorization of input sensory values. (If there are large differences then new navigation maps will actually be created [14, 15].)

Instinctive primitives and learned primitives are actions to perform on navigation maps, essentially acting as small rules, and also are stored within modified navigation maps. Instinctive primitives are innate knowledge procedures concerning objects, agents, numbers, and social group members. Learned primitives are procedures which are learned by the architecture. Instinctive primitives and learned primitives are selected by a process similar to Fig. 3's best-matching navigation map process matching the sensory inputs as well as results of previous intermediate results in the Navigation Module (Fig. 1). A best-matching instinctive primitive or learned primitive (termed the Working Primitive or WPR) is then applied against the best-matching multisensory navigation map (which has been updated with the changes of the sensory inputs, termed the Working Navigation Map or WNM) in the Navigation Module.

The application of WPR on WNM via the mechanisms (which can actually be quite straightforward) of the Navigation Module (Fig. 1), produces a signal to the Output Vector Association Module (Fig. 1) and then to the external world.

The instinctive primitives are based on the work of Spelke and others [17, 18] showing similar innate procedures in mammalian (mainly human) infants.

Imagine that a robot controlled by a CCA5 architecture is in a forest and has a goal (the Goal/Emotion Module in Fig. 1 will influence this [14–16]) of moving to a certain point. There is a river in front of it. The sensory inputs are processed as described above. In the Navigation Module there is a Working Navigation Map (WNM) shown in Fig. 2, representing water in front of the CCA5 controlled robot. An instinctive primitive related to water is the best-matching instinctive or learned primitive to the sensory inputs, and becomes the Working Primitive (WPR). The operation of WPR on WNM is to avoid water and as a result instead of continuing straight the robot turns left. Then a new cognitive cycle starts again—new sensory inputs are processed through the architecture,

the Navigation Module (Fig. 1) produces an output (or not, as described in the next sections), and an output action occurs, and so on.

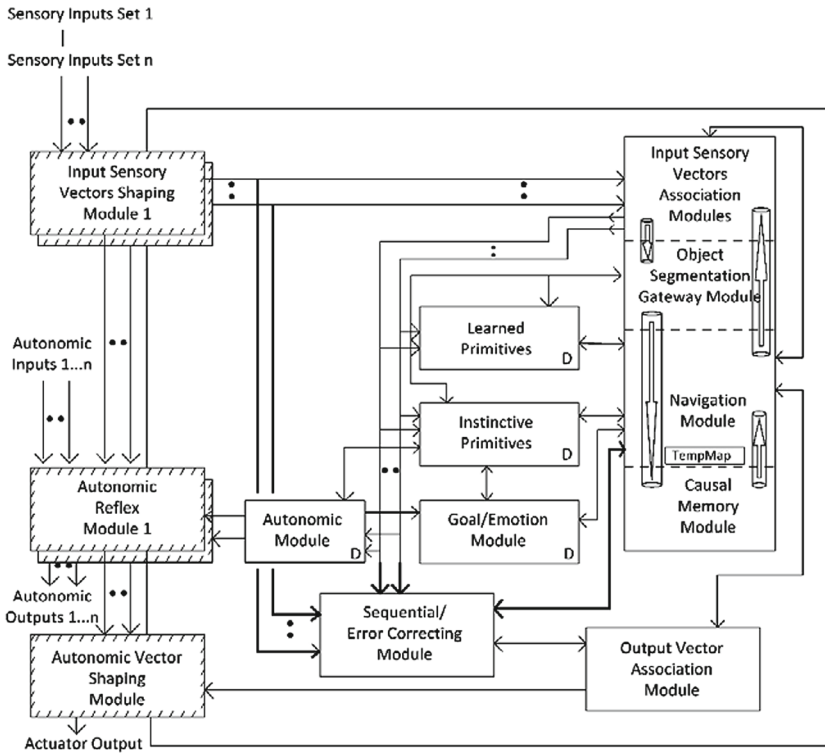


Fig. 1. Causal Cognitive Architecture 5 (CCA5). (“D” in some modules signifies that its properties change as the architecture develops, i.e., with experience and usage.)

3.2 Feedback

Feedback pathways are ubiquitous throughout the CCA5 architecture—states of a downstream module can influence the recognition and processing of more upstream sensory inputs. The differences between the expected sensory input and the actual sensory input are computed and fed forward, and influence the binding of the sensory inputs onto local sensory navigation maps in the Input Sensory Vectors Association Modules as well as the final binding of the local navigation maps onto a multisensory navigation map (Figs. 3 and 4) which becomes the working navigation map in the Navigation Module. This is described above, and is explored in more formal detail in [16].

In the CCA5 the feedback pathway between the Navigation Module and the Input Sensory Vectors Association Modules is enhanced. Normally, if the result of an operation of the Working Primitive (WPR) on the Working Navigation Map (WNM) in the Navigation Module does not produce an actionable output, then no action occurs. Perhaps

solid	solid	solid	water	solid	solid
solid	solid	solid	water	solid	solid
solid	solid	water	water, sound23	solid	solid
solid	solid	water	water	solid	solid
solid	solid	<u>water,</u> <u>link{4574}</u>	solid	solid	solid
solid, <u>iprimi-</u> <u>tive{8974}</u>	solid	water	solid	solid	solid

Fig. 2. Example of a Navigation Map—the $6 \times 6 \times 0$ spatial dimensions are shown containing sensory features (most visual but “sound23” is auditory) and some links to other navigation maps and to instinctive primitives (i.e., instructions for operations). Solid arrows represent links within the navigation maps. Dashed arrows represent links to cells in other navigation maps.

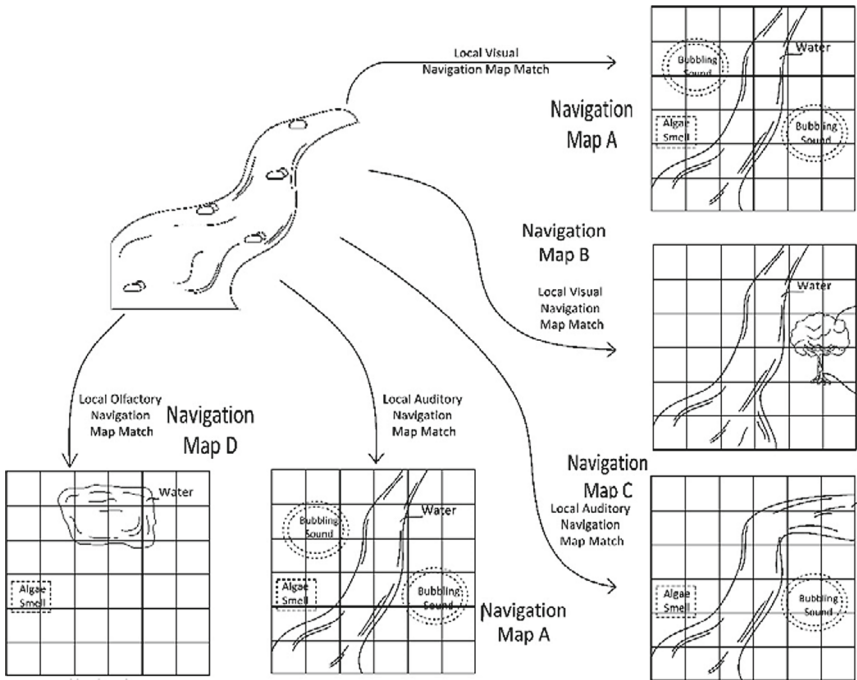


Fig. 3. Matching best-matching local input sensory navigation maps to Causal Memory Module previously stored best-matching multi-sensory maps. The best match is Navigation Map A.

in the next cognitive cycle with different sensory inputs and possibly a different instinctive primitive chosen as the best-matching instinctive primitive (WPR), the Navigation Module will produce an actionable output. However, now with the enhanced feedback pathway between the Navigation Module and the Input Sensory Vectors Association Modules, the intermediate results of the Navigation Module which did not produce any actionable output, can be fed back and stored in the Input Sensory Vectors Association

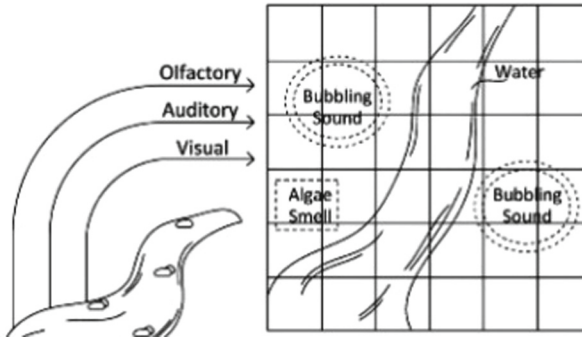


Fig. 4. Updating the Causal Memory Module best-matching multi-sensory navigation map A with features from the sensory scene different than what has been previously stored on the map. (A similar process occurs for each of best-matching local input sensory data navigation maps.)

Modules. In the next cognitive cycle these intermediate results will automatically be considered as the input sensory information and propagated to the Navigation Module and operated on again.

As shown in [14–16], by feeding back and re-operating on the intermediate results, the Causal Cognitive Architecture is able to formulate and explore possible cause and effect of actions, i.e., generate causal behavior. This is not surprising as the instinctive primitives (or the learned primitives, i.e., rules and procedures which the architecture learns) can essentially perform small logical operations and by being able to re-operate on the intermediate results the architecture can explore the effect of various actions. An example of such causal behavior is given by [14] where a robot controlled by a similar Causal Cognitive Architecture working as a patient aide has never seen this particular patient or hospital room before. The patient is using a walker to walk and asks the robot for a glass of water. As the patient's hand releases from the walker to accept the glass of water the patient starts falling down. The robot has not been preprogrammed how to stop a patient from falling but has a goal not to allow a patient to fall. The motion of the patient's arm and hand are bound onto a navigation map with what is called a motion prediction vector indicating the patient's body going towards the direction of the ground at a certain angle. The activation of a learned primitive in the Causal Cognitive Architecture robot that the patient should not fall, then triggers a physics instinctive primitive to push back against something falling or moving, in order to stop the movement. Thus, the Causal Cognitive Architecture robot pushes back against the falling patient opposite to the angle of the motion prediction vector, and stops the patient's fall [14].

More sophisticated learned primitives can allow more extensive exploration of effects. As well, a more sophisticated simulation whereby the architecture is attempting to repair a broken machine with many turning gears [14] can demonstrate cause and effect in the architecture better. However, in all cases, the same mechanism is being used—instinctive and learned primitives are being applied to a navigation map and then the intermediate results are re-operated on as necessary until a desired result is obtained. (Note that time-outs will occur after a certain number of cognitive cycles.)

This mechanism of feeding back results and then re-operating on them (instead of the new sensory inputs) may seem somewhat awkward to the computer scientist—why not instead create some temporary memory registers and a small algorithm directing results to and between these registers? The reason is that the Causal Cognitive Architecture is biologically inspired, and from an evolutionary perspective, it seems more reasonable that by enhancing feedback pathways from the Navigation Modules, intermediate results could be stored in the Input Sensory Vectors Association Modules and reprocessed in the next cognitive cycle, with few other evolutionary changes required. As well, this may be consistent with the evolutionary emergence of causal abilities and psychotic disorders [10, 11, 15].

3.3 Analogical Reasoning

Above we saw that if there is no actionable output [i.e., an output that can be propagated to the Output Vector Association Module (Fig. 1) from the Navigation Module (Fig. 1)], then we can feed back the Working Navigation Map (WNM) to the Input Sensory Vectors Association Modules. Then in the next cognitive cycle these intermediate results will automatically be considered as the input sensory information and propagated to the Navigation Module and operated on again.

Experimentation revealed that re-operating on different combinations of sensory inputs which are processed into a different Working Navigation Map (WNM) causing possibly a different selected Working Primitive (WPR), unfortunately often still does not give a causally related output or may give an output which may be actionable but not that useful. However, this experimentation [16] revealed that with a very small additional modification to the architecture [the appropriation of part of the Navigation Module to use as a temporary memory register, TempMap (Fig. 1)] and to the feedback algorithm, then more advantageous analogical processing of the intermediate results can occur.

Equations/pseudo-code (1)–(5) are adapted from [16]. If after the operation of the Working Primitive (WPR, which is a navigation map, i.e., an array) on the Working Navigation Map (WNM, also an array) there is no actionable output [i.e., no output which can be sent to the Output Vector Association Module (Fig. 1)], then as represented (i.e., *action* \neq “move*”) in (1), the results in the Navigation Module are treated as intermediate results and are fed back to be temporarily stored in the Input Vectors Association Modules, i.e., *Nav_Mod.feedback_to_assocn_mod(WNM)* (1). These intermediate results are also sent to the Causal Memory Module (Fig. 1) where they are matched to the best navigation map in the Causal Memory Module which becomes the new Working Navigation Map (WNM) (2). The most recently used link (or other scheme [16]) of this new WNM points to a navigation map which is then stored in the TempMap memory location within the Navigation Module (Fig. 1), and which we term TEMP MAP (3). Then in (4), TEMP MAP (also an array) is automatically propagated to the Navigation Module where the value of the current Working Navigation Map (WNM) is subtracted from TEMP MAP with the difference now forming the current value of the Working Navigation Map (WNM).

action \neq “move*”,

$$\Rightarrow \text{Nav_Mod.feedback_to_assocn_mod(WNM)} \quad (1)$$

$$\Rightarrow WNM = Causal_Mem_Mod.match_best_map(WNM) \tag{2}$$

$$\Rightarrow TEMPMAP = Nav_Mod.use_linkaddress1_map(WNM) \tag{3}$$

$$\Rightarrow WNM = Nav_Mod.subtract(WNM, TEMPMAP) \tag{4}$$

$action_{t-1} \neq \text{“move*”}$,

$$\Rightarrow WNM = Nav_Mod.retrieve_and_add_vector_assocn() \tag{5}$$

In the next cognitive cycle (thus, $t - 1$ represents the preceding cycle) as (5) shows, the previous intermediate results stored in the Input Vectors Association Module are propagated forward to the Navigation Module, where the pseudocode in (5) specifies they are added to the current Working Navigation Map (WNM), forming a new Working Navigation Map (WNM).

Induction by analogy effectively occurs by this process of moving, comparing, subtracting, and adding navigation maps. Consider Eqs. (6)–(9). There are two variables x and y . Variable x has properties $P_1, P_2, P_3, P_4, \dots P_n$ (6). Variable y also has properties $P_1, P_2, P_3, P_4, \dots P_n$ (7). We now see that variable y has another property N (8). Therefore in (9) we can conclude by induction by analogy that variable x also has property N .

In (1) we can refer to WNM as variable x , or perhaps as navigation map x . We want to know what this navigation map x will do next, i.e., which navigation map will it call. Consider variable y , or perhaps named as navigation map y , as referring to WNM in (2). It is the best-matching navigation map to navigation map x and thus we assume it will share many properties. We explore what navigation map y does next (i.e., what navigation map does the *linkaddress* we chose link to). We see that navigation map y links to the navigation map which we then store in TEMPMAP (3) and that the difference between navigation map y and TEMPMAP is WNM (4). We will consider this difference, i.e., current WNM in (4) to be property N (8). Since navigation map y has property N , therefore by induction by analogy, we can say that navigation map x also has property N (9). Thus, we add property N , which is actually the difference, i.e., current WNM in (4), to navigation map x , which is actually the original WNM, producing the result of navigation map x with property N as being the Working Navigation Map WNM represented in (5).

$$P_1x \text{ and } P_2x \text{ and } \dots P_nx \tag{6}$$

$$P_1y \text{ and } P_2y \text{ and } \dots P_ny \tag{7}$$

$$Ny \tag{8}$$

$$\therefore Nx \blacksquare \tag{9}$$

3.4 An Example of Analogical Reasoning in the CCA5

Induction by analogy in the CCA5, described in the previous section, proved advantageous for the utility of the architecture, not for human IQ-like tests (which the CCA5 is not developed enough to test on in any case) but rather, in the routine day to day decisions an agent operating in a relatively complex environment is required to make. Given that the CCA5 is mammalian brain inspired and that analogical reasoning emerged with relatively few changes, it is not surprising that there exists much psychological evidence that the core of human cognition relies on analogies [19].

Consider a simple example. There is a large hole in the ground filled with leaves. An embodiment controlled by a CCA5 architecture (which we will simply call “CCA5”) has the current goal of going across a field when it comes to this large hole. Normally, if there is a large, empty hole, an instinctive primitive will be triggered (via the mechanisms discussed in the preceding sections) which operates on the Working Navigation Map resulting in a decision to avoid the hole. The Navigation Module will make a decision to turn right or turn left rather than go straight. However, if there is a solid path (e.g., a bridge, although the current CCA5 does not actually know what a “bridge” is) across the hole then the CCA5 will continue along the solid path to the other side of the hole and continue going across the field.

In a new example, the CCA5 sees a large hole filled with leaves (which it knows from an internal catalog as “solid08”). The CCA5 does not know anything about leaves other than they are solids. It may not know what to do, but the feedback mechanisms will not provide much help, and eventually an actionable decision results—the CCA5 will continue in paths across solids otherwise. In such a case, the CCA5 would go forward, the leaves would not support its weight, and it falls down into the hole and becomes damaged.

Consider another example, however, where the CCA5 uses analogical reasoning. The CCA5 still knows nothing about leaves (other than being able to recognize them as “solid08” from a preprogrammed catalog), but it had the previous experience of stepping in a small hole filled with crumpled newspaper (which it recognized as “solid22”, knowing nothing about newspaper) and its foot falling into the hole. Thus, in this new case, what happens is when the CCA5 sees the hole with leaves (“solid08”) the analogy mechanism described above causes the navigation map representing stepping on “solid22” (which shares properties as being in sheets with “solid08”) to link to the navigation map of falling into the hole with newspapers, i.e., falling in. This is the result (i.e., navigation map) which is fed forward in the next cognitive cycle to the Navigation Module. This causes another instinctive primitive to be triggered about falling into a hole, which is applied against the Working Navigation Map in the Navigation Module. The result of the operation of this intrinsic primitive is not to go forward. Thus, the Navigation Module makes the decision to go left or right, and does not fall into the hole filled with leaves. Thus, even though our very simple and immature CCA5 (i.e., it has not been well developed with instinctive primitives, it has no learned primitives, and so on) had little particular knowledge ahead of time about leaves, it was able to automatically (via the core mechanism of induction by analogy) make the correct decision not to continue straight across the leaves in the hole.

4 Discussion

As noted above, the difference in the abilities of neural network AI systems and symbolic, logical AI systems is sometimes referred to as the “neurosymbolic gap” [4, 5]. As noted above, neurosymbolic AI systems which combine properties of ANN’s and symbolic systems, are almost always overtly hybrid systems [6–8]. However, in this paper we have explored the ability of a brain-inspired cognitive architecture, the Causal Cognitive Architecture 5 (CCA5), to effectively represent a more integrated solution to the neurosymbolic gap.

Above we have reviewed the operation of the CCA5, particularly with regard to the topic of the neurosymbolic gap. The reader is directed to [10–16] for a more detailed description and analysis of the architecture of the CCA5. Schneider [16] gives a detailed and more formal description of the function and data flow of the components making up the architecture. The operation of the architecture can be summarized via Fig. 1 in terms of cognitive cycles of sensory features streaming in, being formatted, propagated to the Input Sensory Vectors Association Modules and spatial binding to local (single sensory) navigation maps and then to a multi-sensory best-matching navigation map from the Causal Memory Module (Figs. 3 and 4), as well as being propagated to the Sequential/Error Correcting Module and then temporal binding onto the multi-sensory best-matching navigation map. The processed sensory inputs as well as previous results from the Navigation Module, trigger the selection of a best-matching “Working Primitive”, which is applied against the updated multi-sensory best-matching navigation map (“Working Navigation Map”), often producing an actionable output from the Navigation Module. The output from the Navigation Module is further processed at the Output Vector Association Module and sent to actuators of the embodiment. Then another cognitive cycle occurs.

If no actionable output occurs from the Navigation Module, then as discussed above, the current Working Navigation Map can be fed back to the Input Sensory Vectors Association Modules and operated on again in the next cognitive cycle. As discussed above, causal abilities as well as inductive analogical reasoning readily emerge from these feedback cycles.

The architecture can match sensory inputs in a connectionist fashion and via the predictive coding of matching against stored navigation maps and spatial and temporal binding, can perceive well even with noisy, imperfect sensory inputs. The architecture can perform varied logical reasoning through the instinctive and learned primitives, as well having causal and inductive analogical reasoning as its core mechanisms. Thus, in addition to its connectionist properties it also has symbolic ones.

It could be argued that the CCA5 non-hybrid solution, does in fact have hybrid components, hence the term “non-overtly hybrid” used above. The navigation maps containing instinctive and learned primitives indirectly represent collections of logical operations. The operations on the navigation maps to find best-matching navigation maps and other such operations, do occur in parallel, and links are strengthened and weakened between different navigation maps and different cells even in the same map. So, in this manner, there are symbolic and connectionist aspects to the architecture. However, these components are very different than what is seen in typical symbolic systems or in artificial neural networks. As well, these components cannot function

on their own—they are critical parts working together to make up the architecture. The components presented in Fig. 1 are essential to the successful workings of the CCA5. The architecture of the CCA5 is essentially that of a non-hybrid system, which as demonstrated above has neurosymbolic properties. Creating systems that effectively close the neurosymbolic gap is an active area of research [4–7]. As noted above, these are largely hybrid approaches which combine a neural network system with a symbolic AI system. There are challenges in doing so—the representations in the neural networks are very different than the representations in the symbolic AI system combined with the latter, depending on how the combination is achieved. An advantage of a non-overtly-hybrid system such as the CCA5 is that there is no such awkward combination of representations required.

Future work on the CCA5 and its succeeding architectures is to automate the acquisition of sensory inputs. Currently, simulated sensory inputs have to be manually hand-fed to the model in order for it to build up its navigation maps. Doing so will allow more experimental examples to better understand the properties of the system and possible practical uses for it. As well, the collection of instinctive primitives is very small at present. Enhancing this collection of primitives will assist with the collection of sensory inputs as well as processing of navigation maps.

As demonstrated above, one of the approaches that should also be considered with regard to better neurosymbolic integration, is a brain-inspired cognitive architecture approach, such as the example of the CCA5. As well, in discussions of cognition, artificial cognition, or consciousness [20], given that the CCA5 is inspired by the human brain and thus may reflect its mechanisms, the neurosymbolic properties of cognition should be considered.

References

1. Phillips, P.J., Yates, A.N., et al.: Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *PNAS* **115**(24), 6171–6176 (2018). <https://doi.org/10.1073/pnas.172135511>
2. Waismeyer, A., Meltzoff, A.N., Gopnik, A.: Causal learning from probabilistic events in 24-month-olds. *Dev. Sci.* **18**(1), 175–182 (2015). <https://doi.org/10.1111/desc.12208>
3. Leivada, E., Murphy, E., Marcus, G.: DALL-E 2 Fails to Reliably Capture Common Syntactic Processes. arXiv (2022). <https://doi.org/10.48550/arXiv.2210.12889>
4. Besold, T.R., d'Avila Garcez, A., Bader, S., et al.: Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. arXiv (2017). <https://doi.org/10.48550/arXiv.1711.03902>
5. Ebrahimi, M., Eberhart, A., Bianchi, F., Hitzler, P.: Towards bridging the neuro-symbolic gap. *Appl. Intell.* **51**, 6326–6348 (2021). <https://doi.org/10.1007/s10489-020-02165-6>
6. Garcez, A.A., Lamb, L.C.: Neurosymbolic AI: The 3rd Wave. arXiv (2020). <https://doi.org/10.48550/arXiv:2012.05876>
7. Kautz, H.A.: The third AI summer: AAAI Robert S. Engelmore memorial lecture. *AI Mag.* **43**(1), 93–104 (2022). <https://doi.org/10.1002/aaai.12036>
8. Cingillioglu, N., Russo, A.: DeepLogic: Towards End-to-End Differentiable Logic Reasoning. arXiv (2019). <https://doi.org/10.48550/arXiv.1805.07433v3>
9. Schneider, H.: Meaningful-based cognitive architecture. In: Samsonovich, A.V. (ed.) *Procedia Computer Science*, vol. 145, pp. 471–480 (2018). <https://doi.org/10.1016/j.procs.2018.11.109>

10. Schneider, H.: The meaningful-based cognitive architecture model of Schizophrenia. *Cogn. Syst. Res.* **59**, 73–90 (2020). <https://doi.org/10.1016/j.cogsys.2019.09.01>
11. Schneider, H.: Subsymbolic versus symbolic data flow in the meaningful-based cognitive architecture. In: Samsonovich, A. (ed.) *BICA 2019. Advances in Intelligent Systems and Computing*, vol. 948 (2020). https://doi.org/10.1007/978-3-030-25719-4_61
12. Schneider, H.: Causal cognitive architecture 1: integration of connectionist elements into a navigation-based framework. *Cogn. Syst. Res.* **66**, 67–81 (2021). <https://doi.org/10.1016/j.cogsys.2020.10.021>
13. Schneider, H.: Causal cognitive architecture 2: a solution to the binding problem. In: Klimov, V.V., Kelley, D.J. (eds.) *Studies in Computational Intelligence*, vol. 1032 (2022). https://doi.org/10.1007/978-3-030-96993-6_52
14. Schneider, H.: Causal cognitive architecture 3: a Solution to the binding problem. *Cogn. Syst. Res.* **72**, 88–115 (2022). <https://doi.org/10.1016/j.cogsys.2021.10.004>
15. Schneider, H.: Navigation map-based artificial intelligence. *AI* **3**(2), 434–464 (2022). <https://doi.org/10.3390/ai3020026>
16. Schneider, H.: An inductive analogical solution to the grounding problem. *Cogn. Syst. Res.* **77**, 74–216 (2023). <https://doi.org/10.1016/j.cogsys.2022.10.005>
17. Spelke, E.S.: Initial knowledge. *Cognition* **50**, 431–445 (1994). [https://doi.org/10.1016/0010-0277\(4\)90039-6](https://doi.org/10.1016/0010-0277(4)90039-6)
18. Kinzler, K.D., Spelke, E.S.: Core systems in human cognition. In: von Hofsten, C., Rosander, K. (eds.) *Progress in Brain Research*, vol. 164 (2007) (Chapter 14)
19. Hofstadter, D.R.: Analogy as the core of cognition. In: Gentner, D., et al. (eds.) *The Analogical Mind: Perspectives from Cognitive Science*, pp. 499–538. MIT Press (2001)
20. Bołtuć, P.: Consciousness for AGI. In: Samsonovich, A.V., Klimov, V.V. (eds.) *Procedia Computer Science*, vol. 169, pp. 365–372 (2020)



FECG: A Flexible Holter for Ambulatory Heart Rate Monitoring

Yuduo Shan¹, Tingting Liu¹(✉), and Zhen Liu²

¹ College of Science and Technology, Ningbo University, Ningbo, China
liutingting@nbu.edu.cn

² Ningbo University, Ningbo, China

Abstract. Brain-like artificial intelligence technology has been developing continuously, and its practical application has been widely popularized. Since the effect of brain-like AI technology is mainly reflected in the interaction with the user, it is necessary to collect various physiological indicators of the user during the interaction process in order to test the specific feelings of the brain-like AI technology on the user, such as emotional changes or physiological reactions. In order to realize this need, we designed and fabricated an almost completely flexible wearable ECG monitoring system using textile electrodes and Flexible Printed Circuit Board (FPCB) technology: FECG (Flex Electrocardiogram System). The FECG has been tested by user experiments and is capable of delivering reference-level physiological data (3 uV equivalent input noise, 85 dB common mode rejection and 500 Hz sampling rate) that meets medical standards while achieving long-term wearability and high comfort. The construction of this system can minimize the interference factors, while testing the user's ECG data more conveniently and accurately, providing a detailed test basis for constructing a reasonable interaction effect.

Keywords: ECG · Holter · FPC · Textile electrode · Flexible wearable devices

1 Introduction

Since the effect of brain-like AI technology is mainly reflected in the interaction with the user, it is necessary to collect various physiological indicators of the user during the interaction process in order to test the specific feelings of the brain-like AI technology on the user, such as emotional changes or physiological reactions [1]. Electrocardiogram (ECG) is just such an important physiological indicator, so much so that it can reflect changes in a person's mood, visualizing some abstracted data and displaying it as a waveform. 24H continuous recording of ECG data is currently achieved mainly using bedside ECG monitoring devices or holter. Both them require several wet electrodes and lead wires close to the body to work, thus lacking in portability and inconvenient to use. Factors that are not conducive to prolonged use are driving the development of new electrodes dedicated to bioelectricity measurements and digital holter with wireless communication capabilities and their use in practical healthcare activities, which can

allow recording of ECG data on each day and without interrupting or limiting the user's motions [2, 3], This is critical because the ECG is a very real-time data and it is difficult to predict its connection with a certain phenomenon, which must be monitored over time in order to build a database that can be used for the detection of affective levels (valence, arousal, control, familiarity, liking and basic emotions), which can in turn come in handy in the research of brain-like AI techniques [4].

1.1 Related Work

As an important development direction of Holter ECG, wearable ECG equipment has been continuously developed in recent years. As in the work of C. D. Capua, A. Meduri and R. Morello, They have improved the wearable ECG equipment from "professional" medical equipment to "home" health equipment, making it possible to wear ECG equipment on a daily basis and greatly increasing the popularity of ECG equipment [5]. Similarly, B. Liu and his team's work on Silver nanowire-composite electrodes has made technical innovations in ECG electrodes, which solves the problem of comfort for long-term wearing electrodes [6]. The innovative work of V.P. Rachim and W. Chung on the combination of Wearable noncontact armband and mobile ecg monitoring system is to optimize the data processing problems that wearable ECG devices inevitably encounter, such as unstable sampling and irregular data [7]. Obviously, in order to achieve good practical results, many scholars are working hard on the portability, accuracy, reliability, and comfort of wearable ECG devices.

Although wearable electronics have come a long way, the trade-offs between functionality, performance, and cost among the various components are still deeply debatable. Taking electrodes as an example, the traditional silver/silver chloride (Ag/AgCl) electrodes with electrolyte gel, although affordable and high-performance, are not suitable for long-term wearable use because they are not breathable, cannot be immersed in water, are uncomfortable, and are potentially skin irritating. After a period of time, normal human activities will unavoidably cause the electrodes to degrade in performance and even cause skin lesions [8, 9]. Various dry electrodes proposed in recent years, such as Textile Electrodes [10], Microtip Electrodes [11], Foam Electrodes [12] and so on, while realizing high performance, some of the shortcomings of conventional electrodes have been addressed. However, due to the use of new materials and structures, it is difficult to apply them directly to conventional ECG devices, and they need to be specifically examined and designed for their respective characteristics in order to achieve the best results.

Flexible Printed Circuit (FPC) is a technology that works well with flexible dry electrodes, as it offers the advantages of a flexible substrate and better electrical performance compared to traditional wearable ECG devices, allowing the various parts of the system to reach their full potential [13]. Since the majority of the system is composed of flexible materials, this new wearable ECG device can achieve a high degree of human fit, which in turn significantly improves wearing comfort and measurement accuracy.

In this paper, we report in detail on a body-friendly flexible holter that uses textile electrodes as measurement electrodes and integrates the rest of the electronics such as the filter chip and the processor via FPC. In addition to the description of the system principle and structure, in order to verify the feasibility and actual performance of the

system, this paper is accompanied by its experimental measurement data and some simple comparisons.

2 Research Design and Methodology

2.1 ECG Signal Analysis

Detecting the user's physiological signals and recognising the user's emotions are particularly important in the process of human-computer interaction between brain-like AI and humans, and are the cornerstones of quantitative processing and evaluation of data.

The electrocardiogram (ECG) signal basically corresponds to the electrical activity of the heart. By analysing and utilising ECG signals, we can accomplish a variety of purposes, such as measuring the heart rate, examining the rhythm of heartbeats, diagnosing heart abnormalities, emotion recognition and biometric identification [14]. A high-quality ECG signal with physiological monitoring and emotion recognition will obviously be of great help to our research in brain-like AI technology.

The origin of the electrical activity measured by ECG is in the muscle fibers of different parts of the heart. The ECG may roughly be divided into the phases of depolarization and repolarization of the muscle fibers making up the heart. The depolarization correspond to the P-wave (atrial depolarization) and QRS-wave (ventricles depolarization). The repolarization correspond to the Twave and U-wave (ventricular repolarization). The elements in the ECG-complex are shown in Fig. 1 [15]. Not just a simple rhythm, an ECG signal that meets medical standards should have a clearly visible characteristic waveform for analysis.

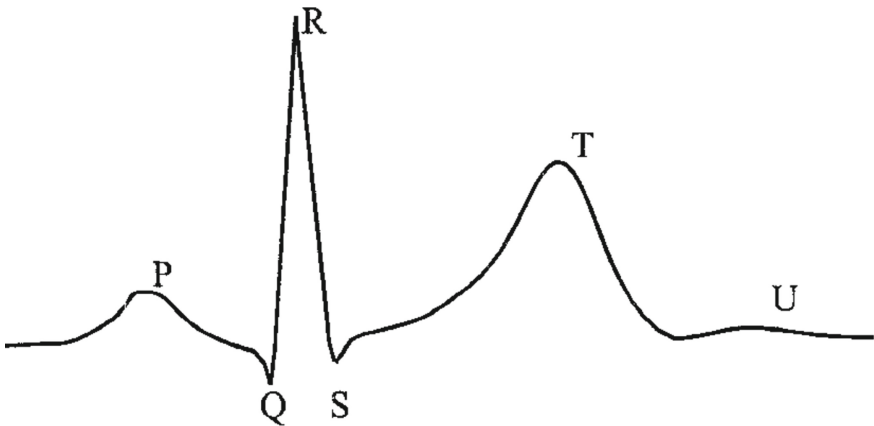


Fig. 1. A simple example, regarding the P-QRS-T waves making up the ECG signal

For ECG recordings the variation in electrical potentials in twelve different derivations out of the ten electrodes are measured. These twelve different electrical views of the activity in the heart are normally referred to as leads. The twelve leads are made up of three bipolar and nine monopolar leads. The three bipolar leads are the electrical

potentials between the right and left arm (lead I), the right arm and left foot (lead II) and between the left arm and left foot (lead III). The monopolar leads are the left arm (aVL), the right arm (aVR), the left foot (aVF) and on the six chest electrodes (V1–V6) are measured. The right foot is normally used for grounding purposes only [16]. With such a demand, we need to achieve success with at least one of the lead modalities to validate the possibility of adding more leads until full connectivity is achieved.

We chose a simplified “limb I”. This representative limb lead has very little difference from the chest lead and has a simple connection, requiring only two electrodes and a ground electrode. We scaled the measurement point of “limb I” to the front of the body from the chest to the waist (LA, RA and LL), as shown in Fig. 2, which is also a standard lead.

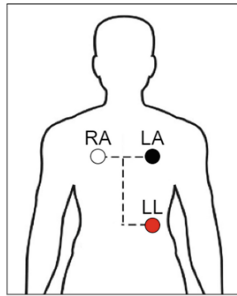


Fig. 2. I schematic, Measuring leads RA and LA, grounding lead LL

2.2 Electrode Selection

Conventional electrodes are difficult to perform the job of monitoring ECG signals stably over a long period of time in brain-like artificial intelligence systems because of their respective drawbacks: the Ag–AgCl electrodes usually need to be coated with a conductive gel to be used, which has a very short lifespan and needs to be repeated for long periods of time of wear. Metal dry electrodes need to be given some external pressure to be fixed on the body surface, and such a hard object pressed on the body for a long period of time tends to cause discomfort and may be corroded by body fluids such as sweat, which may lead to degradation of performance. Therefore, we tried to use a flexible dry electrode that is human-friendly and can be worn for a long time without causing adverse reactions in the human body: the textile electrode.

This textile electrode resembles a piece of fabric, but it both warp and weft yarns are silver-plated nylon yarns (the content of silver is 17%), whose fineness is 29.56 tex. The warp density is 360 per 10 cm, and the weft density is 260 per 10 cm. The warp density is 360 per 10 cm, and the weft density is 260 per 10 cm. The warp tightness is 72.4%, and the weft tightness is 52.3%. When a certain external pressure is applied, it will adhere to the skin, presenting its electrical conductivity and thus acting as an electrode [17]. Its actual photo is shown in Fig. 3.

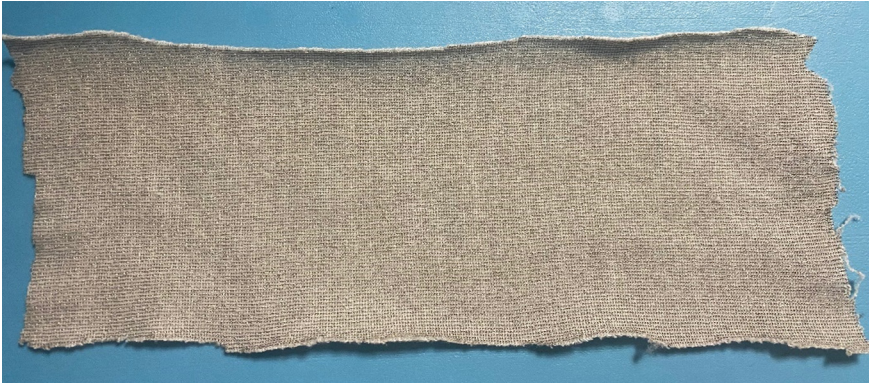


Fig. 3. The textile electrode

2.3 Circuit Design

ECG signals are extremely weak electrical signals with low amplitude (0.1–4.0 mV), narrow bandwidth, low frequency (0.5–100.0 Hz), and high internal resistance (1 M Ω –1 G Ω). When acquired, they often face a variety of interferences, most of which need to be eliminated at the circuit design level in order to ensure signal accuracy. Several typical interference noises are listed below:

Industrial Frequency. The power supply network covers most of the community and home environments, and the 50 Hz industrial frequency (IF) interference noise it introduces is one of the main interferences to ECG signals. IF interference interferes with ECG signals in the form of displacement currents through capacitive coupling in the measurement leads and in the human body itself.

Myoelectricity. When human muscles are in an active state, they generate electrical noise ranging in frequency between 20–5000 Hz, which interferes with the ECG signal.

Baseline Drift. A particularly low-frequency curve is superimposed on the original signal, giving it a slow slight tendency to go up and down. If the baseline drift is not eliminated, then the erroneous trend can be taken as the acquired raw signal, affecting the accuracy of the signal and subsequent data processing results.

Electrostatic Discharge. In human movement, the static electricity generated by friction between fabrics and fabrics, fabrics and skin is one of the main factors of the human body charged, and the human body that becomes a source of static electricity in contact with the cardiac instrument will generate a strong electromagnetic disturbance, i.e., Electrostatic Discharge (ESD). The ESD phenomenon can lead to the emergence of high-frequency disturbances as high as 1GHz, and at the same time there is a risk that the extremely high transient voltage will cause irreversible damage to cardiac instrumentation. The ESD phenomenon can lead to high frequency interference of up to 1GHz, while extremely high transient voltages carry the risk of irreversible damage to cardiac instruments.

Common Mode and Differential Mode Interference. Since the system is battery or DC powered, Common Mode (CM) interference and Differential Mode (DM) interference

are also a challenge. Since CM voltages do not provide useful information about the heart and may actually affect measurement accuracy, the system must be able to suppress CM interference while responding to the target signal, the DM ECG voltage.

Radio Frequency. Due to the application of brain-like artificial intelligence technology requires the use of a large number of high-precision electronic components and electrical equipment, the system will be working in an extremely complex Radio Frequency (RF) environment, need to ensure that the ECG signal is not affected by 5G communication signals, WiFi signals, Bluetooth signals, and the interference frequency of the electrical equipment.

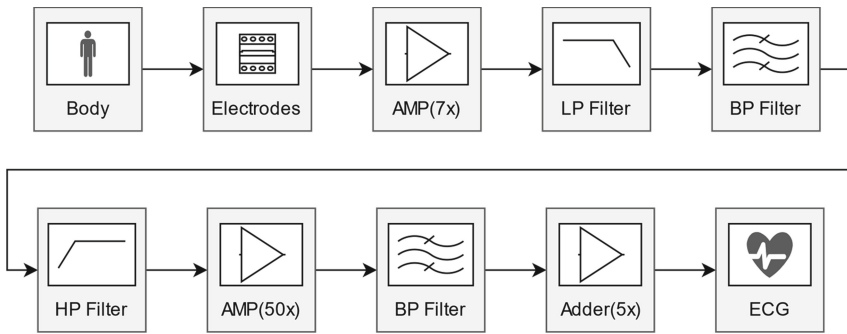


Fig. 4. Block diagram of the system

To eliminate the adverse effects of the above noise, we designed a high performance ECG sampling circuit. Its structural block diagram is shown in Fig. 4.

Front end instrumentation amplifier with right leg driver. The front-end instrumentation amplifier circuit with a gain of $7 \times$ together with the right leg driver circuit can shield most of the simple common-mode interferences at the initial stage of signal acquisition, reaching a high common-mode rejection ratio.

100 Hz second-order low-pass filter with 0.1 Hz first-order high-pass filter. Their purpose is to ensure that noise at too high and too low a frequency does not flow into the next acquisition session.

50 Hz bandstop filter amplifier. Due to the strong IF interference noise, the circuit is divided into two links to resist its effects. This is the first link, which will eliminate most of the IF interference.

Main amplifier. The ECG signal is amplified here by a factor of 50, allowing the waveform and its characteristics to be represented in the meter.

50Hz band reject filter. This is the second part of the resistance to IF interference. After the signal flows through here, its IF interference noise is basically eliminated.

Inverting adder. The ECG signal, which has been basically filtered, is amplified five times to complete the final adjustment, after which the signal is output to the MCU.

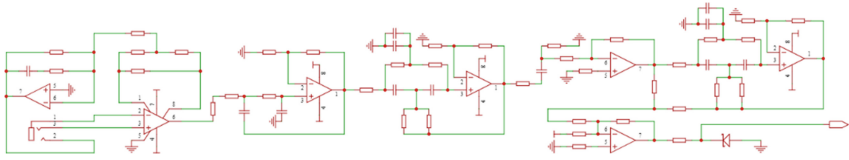


Fig. 5. Circuit schematic diagram of FECCG. It is composed of the filters mentioned above

Combining the above circuits, we have basically accomplished the noise reduction and filtering requirements presented in the previous section, and the circuit schematic is shown in Fig. 5.

2.4 Microchip Selection

The ECG signal acquisition circuit requires a stable and precise operational amplifier chip, so we chose the LM358B chip from Texas Instruments.

LM358B is an Industry-Standard Dual Operational Amplifiers, which means that it has an extremely wide range of applications for most of the scenarios that require the use of operational amplifiers, and its performance is well-tested.

The MAX44248 is an ultra-high-precision, low-noise, zero-drift dual-channel op amp that maintains ultra-low power consumption over a wide supply range. The device integrates a proprietary auto-zero circuit that eliminates drift over time and temperature and the effects of $1/f$ noise by continuously measuring and compensating for input misalignment. The device also incorporates an EMI filter to minimize high frequency demodulation on the output signal. The op amps are supplied from a single 2.7–36 V supply or ± 1.35 to ± 18 V dual supply. The devices are unity-gain stabilized with a 1 MHz gain bandwidth product and consume only 90 μA of current per op amp. The low offset voltage, low noise, and wide supply range make it ideal for use in this ECG measurement system.

The AD620 is a low-cost, high-precision instrumentation amplifier that requires only a single external resistor to set the gain and has a gain range of 1–10,000. Smaller than a discrete design and consuming less power (maximum supply current is only 1.3 mA), the AD620 offers high accuracy of 40 ppm (maximum nonlinearity), low offset voltage of up to 50 μV , and offset drift of up to 0.6 $\mu\text{V}/^\circ\text{C}$. It also features low noise, low input bias current, and low power consumption for ECG monitor applications. Performance is complemented by low noise, low input bias current, and low power consumption. The device's input stage uses SuperBeta processing to achieve low input bias currents of up to 1.0 nA. The AD620 also features low input voltage noise of 9 $\text{nV}/\sqrt{\text{Hz}}$ (1 kHz), 0.28 μV p-p (0.1–10 Hz band), and 0.1 $\text{pA}/\sqrt{\text{Hz}}$ input current noise, with a 0.01% build-up time of 15 μs . In short, it offers a high degree of accuracy and a high degree of power consumption in the ECG systems has also been quite widely used.

2.5 Wearability Research

In the intended operating scenario of this system, it is particularly important to have good usability. In the absence of specialized electrocardiographic apparel or customized

sizing in the field, the key to usability is to ensure wearability. Therefore, the system should be designed as a universal solution that can be adapted to most common clothing types.

In order to make the whole system mountable on normal clothing, we have also designed it with flexible printed circuit (FPC) technology. As shown in Fig. 8, the FPC boards we used were able to achieve excellent bending rate performance while ensuring that the circuitry functioned correctly, so the FECG accomplished the goal of all flexibility.

The flexible textile electrodes and the flexible circuit board together achieve a high degree of plasticity and comfort, allowing the system to be easily installed on everyday clothing, such as undershirts or shirts (preferably one size smaller, so as to provide the electrodes with the appropriate attachment pressure).

3 Experimental

3.1 Prototyping

In accordance with the previously mentioned research and design, a prototype of this system was built. It consists of three main parts.

Main body. The main body part includes the chips, the FPC board, the electrode interface, the test contact, and the battery on the back of the FPC. After completing the PCB design, we commissioned a manufacturing plant (Shenzhen JLC Technology Group Co., Ltd.) to carry out the production of FPC boards, including circuit etching, board shearing and other conventional PCB manufacturing processes. After the delivery of the FPC boards, we soldered the chips, capacitors and resistors, and other circuit components to the FPC boards in accordance with the design drawings.

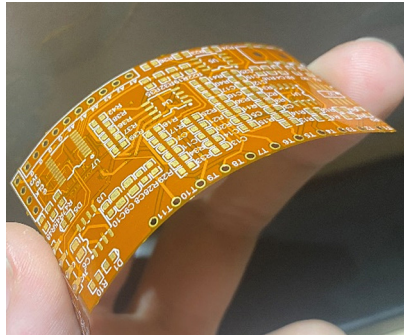


Fig. 6. Good bendability of FPC

Figure 6 shows the excellent bending rate of this FPC board.

After ensuring that the circuit components are soldered correctly, we solder the electrode interface, test contacts and battery. Its final appearance is shown in Fig. 7.

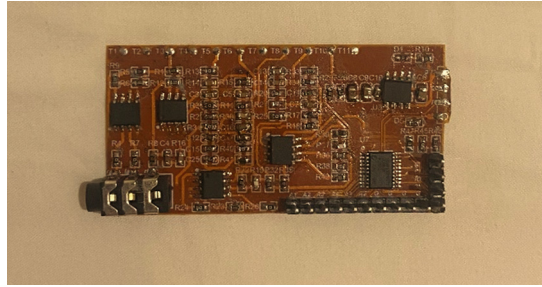


Fig. 7. Finished product

Electrodes. The electrode part includes three textile electrodes as well as a three-quarter shielded leadwire. Among the three textile electrodes, the two electrodes responsible for measuring the electrical signals are in a smaller round shape because this shape is more convenient for accurate measurement of the electrical signals at the specified points. The other grounding electrode is in a larger rectangular shape because this shape is more conducive to stabilizing the ground potential. All three textile electrodes are fitted with copper electrode clasps, through which they are connected to the main part by a three-quarter lead wire with a shield.

Carrier. The carrier part consists of an undershirt and a belt, and their photos are shown in Fig. 8.



Fig. 8. Undershirts and belt

They are ordinary undershirts and belts that are common in life, made of chemical fiber or cotton, and can feel slightly pressure on the body without additional characteristics. They are only used as a representative clothing and act as carriers in this system.

They are shown in Fig. 9 when combined as a whole. For display purposes, the undershirt is turned inside out to show the FECG.

3.2 Trial Operation

After completing the prototype, we first used it to test the ECG signal generator. The ECG signal generator was set to normal ECG with a parameter of 60 bpm. as shown in Fig. 10. The test results were accurate.



Fig. 9. FECCG showcase



Fig. 10. FECCG test on an ECG signal generator

We obtained ECG signal data on a test user (basically healthy, no cardiac-type disease). Its ECG waveform is shown in Fig. 11.



Fig. 11. The FECCG with textile electrodes

For comparison, we also used a conventional Ag–AgCl wet electrode to obtain data on the same user. Its waveform is shown in Fig. 12.

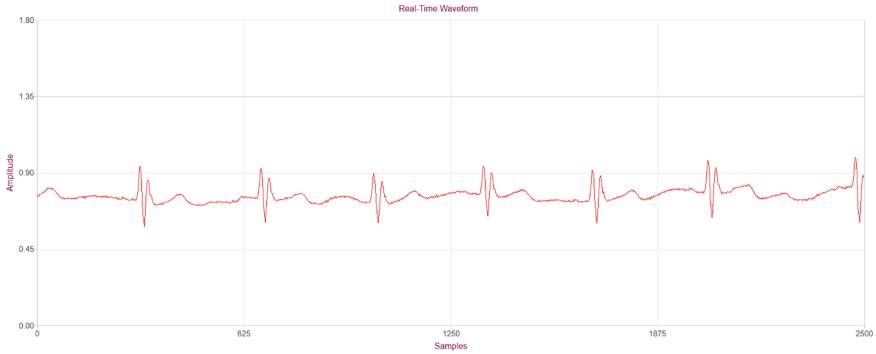


Fig. 12. The FECG with Ag–AgCl wet electrodes

The analysis shows that there is no significant difference in performance between the two, and their characteristic points are clearly visible, and even the noise of the ECG signal is lower in some zones when using the textile electrode—this is because the impedance of the textile electrode is much higher than that of the Ag–AgCl wet electrode. The appropriately high impedance can be measured without affecting the ECG signal while being insensitive to noise and interference.

3.3 Conclusion

The actual performance parameters measured for the FECG are demonstrated in the following tables, where Table 1 shows the electrode performance.

Table 1. Electrode test parameter sheet

Actual denier	Conductivity	Strength	Elongation
Approx. 140 d	< 500 Ω/m	Approx. 400 cN	Approx. 35%

Table 2 shows the system performance.

Table 2. System Test Parameter Sheet

Sample rate	Channels	EIN	CMR
500 Hz max	2 (or 6 links)	Approx. 3 uV/√Hz	Approx. 85 dB

By comparing the experimental data, we draw the following conclusions. Compared with the equipment used in the current routine medical testing process, FECG reduces the interference caused by the equipment itself to users with higher comfort and mobility under the premise of ensuring a small difference in detection accuracy. For example,

ordinary hollers often cause users to feel troubled and bored due to their complicated wires and large and rigid bodies, which greatly interferes with the effect of emotion recognition [18–20]. FECG tries its best to reduce the interference it introduces, and acts as a recorder rather than a participant in the interaction between the brain-like artificial intelligence system and humans, ensuring that the evaluation data it provides has reference value.

Research and practical tests have shown that the FECG is able to provide reference-level physiological data that meets medical standards while achieving long-term wear and high comfort. The construction of this system can minimize the interference factors while testing the user's ECG data more conveniently and accurately, and provide a detailed test basis for brain-like artificial intelligence to construct a reasonable interaction effect.

References

1. Velik, R.: AI reloaded: objectives, potentials, and challenges of the novel field of brain-like artificial intelligence. *BRAIN Broad Res. Artif. Intell. Neurosci.* **3**(3), 25–54 (2012)
2. Lin, B.S., Chou, W., Wang, H.Y., Huang, Y.J., Pan, J.S.: Development of novel non-contact electrodes for mobile electrocardiogram monitoring system. *IEEE J. Transl. Eng. Health Med.* **1**, 1–8 (2013)
3. Spanò, E., Pascoli, S.D., Iannaccone, G.: Low-power wearable ECG monitoring system for multiple-patient remote monitoring. *IEEE Sens. J.* **16**(13), 5452–5462 (2016)
4. Miranda-Correa, J.A., Abadi, M.K., Sebe, N., et al.: Amigos: a dataset for affect, personality and mood research on individuals and groups. *IEEE Trans. Affect. Comput.* **12**(2), 479–493 (2018)
5. Capua, C.D., Meduri, A., Morello, R.: A smart ECG measurement system based on web-service-oriented architecture for telemedicine applications. *IEEE Trans. Instrum. Meas.* **59**(10), 2530–2538 (2010)
6. Liu, B., Luo, Z., Zhang, W., Tu, Q., Jin, X.: Silver nanowire-composite electrodes for long-term electrocardiogram measurements. *Sens. Actuators A Phys.* **247**, 459–464 (2016)
7. Rachim, V.P., Chung, W.: Wearable noncontact armband for mobile ECG monitoring system. *IEEE Trans. Biomed. Circuits Syst.* **10**(6), 1112–1118 (2016)
8. Jung, H.C., et al.: CNT/PDMS composite flexible dry electrodes for long-term ECG monitoring. *IEEE Trans. Biomed. Eng.* **59**(5), 1472–1479 (2012)
9. Reyes, B.A., et al.: Novel electrodes for underwater ECG monitoring. *IEEE Trans. Biomed. Eng.* **61**(6), 1863–1876 (2014)
10. Paul, G., Torah, R., Beeby, S., Tudor, J.: Novel active electrodes for ECG monitoring on woven textiles fabricated by screen and stencil printing. *Sens. Actuators A Phys.* **221**, 60–66 (2015)
11. O'Mahony, C., Pini, F., Blake, A., Webster, C., O'Brien, J., McCarthy, K.G.: Microneedle based electrodes with integrated through-silicon via for biopotential recording. *Sens. Actuators A Phys.* **186**, 130–136 (2012)
12. Tseng, K.C., Lin, B., Liao, L., Wang, Y., Wang, Y.: Development of a wearable mobile electrocardiogram monitoring system by using novel dry foam electrodes. *IEEE Syst. J.* **8**(3), 900–906 (2014)
13. Gao, W., Emaminejad, S., Nyein, H., Challa, S., Chen, K., Peck, A., Fahad, H.M., Ota, H., Shiraki, H., Kiriya, D., Lien, D.H., Brooks, G.A., Davis, R.W., Javey, A.: Fully integrated wearable sensor arrays for multiplexed in situ perspiration analysis. *Nature* **529**(7587), 509–514 (2016)

14. Berkaya, K.S., Uysal, K.A., Gunal, S.E., et al.: A survey on ECG analysis. *Biomed. Sig. Process. Control* **43**, 216–235 (2018)
15. Biel, L., Pettersson, O., Philipson, L., et al.: ECG analysis: a new approach in human identification. *IEEE Trans. Instrum. Meas.* **50**(3), 808–812 (2001)
16. Webster, J.G.: *Medical Instrumentation, Application and Design*, 2nd edn., p. 814. Wiley, New York (1995)
17. Zhang, X., Zhong, Y.: A silver/silver chloride woven electrode with convex based on electrical impedance tomography. *J. Text. Inst.* **112**(7), 1067–1079 (2021). <https://doi.org/10.1080/00405000.2020.1800926>
18. DiMarco, J.P., Philbrick, J.T.: Use of ambulatory electrocardiographic (holter) monitoring. *Ann. Intern. Med.* **113**(1), 53–68 (1990)
19. Kennedy, H.L.: The history, science, and innovation of holter technology. *Ann. Noninvasive Electrocardiol.* **11**(1), 85–94 (2006)
20. Vavrinsky, E., Subjak, J., Donoval, M., et al.: Application of modern multi-sensor holter in diagnosis and treatment. *Sensors* **20**(9), 2663 (2020)



Features of Internal Pronunciation of Words by a Group of People with Rhotacism in Comparison with Normative Pronunciation

Olga Shevaldova^(✉) and Alexander Vartanov

Lomonosov Moscow State University, Moscow, Russia
shevaldovaolga@gmail.com

Abstract. Existing studies in the field of speech disorders do not provide a systematic understanding of the relationship between the bioelectrical activity of the brain and the nature of speech disorders, the characteristics of the processes of speech perception and internal pronunciation. This work is aimed at comparing the activity of the brain during the internal pronunciation of words by a group of people without speech disorders and a group of people with rhotacism. For the first time, an analysis and comparison of event-related potentials (ERP) of the brain in the process of internal pronunciation in people with and without rhotacism was carried out. The electroencephalographic (EEG) study involved 36 people, 18 of them had a speech disorder in the form of rhotacism. The subjects were presented with auditory stimuli (words) spoken by a speaker with standard sound pronunciation. The subject's task was to mentally repeat the word, maintaining the intonation and pronunciation features, as in external speech. The results obtained in this study using a new method of localization of brain activity demonstrate significant differences in ERP during the mental pronunciation of words between the studied groups of people in a number of brain structures, including cortical and subcortical formations. The group of people with rhotacism is characterized by the presence of a pronounced ERP N200 in more evolutionarily early brain structures, such as the midbrain, medulla oblongata and insular lobe on the left. The group of people without speech disorders is characterized by the presence of pronounced ERP in the following structures: caudate nuclei on the right and left, right globus pallidus, cingulate cortex, striatum, dorsomedial prefrontal cortex, anterior cingulate cortex, field 17 on the right and left, Broca's area on the right, Wernicke's area on the right, angular gyrus on the right, anterior prefrontal cortex on the right and left. All differences are obtained with an estimate of 95% confidence interval.

Keywords: EEG · ERP · Internal pronunciation · Inner speech · Rhotacism · Speech disorders · Words · Functional dyslalia

1 Introduction

Functional dyslalia - defects in sound pronunciation caused by a violation of the functioning of the cortical sections of the speech-motor, speech-auditory analyzer or incorrect speech formation. It is manifested by motor (distortion) or sensory (mixing, substitution) inaccuracy in the pronunciation of phonemes. Rhotacism is a form of articulatory-phonetic dyslalia associated with the presence of a defect in the pronunciation of the sounds *r* and *r'* [1].

The mechanism of occurrence of functional dyslalia is associated with imbalance and weakness of the dynamics of nervous processes in the brain. The cortical sections of the speech-auditory and speech-motor systems are without pathology, but the balance of excitation and inhibition in them is disturbed and not coordinated. The nature of the leading defect is determined by the localization of disturbances in cortical neurodynamics [1].

Despite ongoing research in the field of diagnosis and treatment of speech disorders in children and adults, such issues as the relationship between the features of the bioelectrical activity of the brain with the nature of the speech defect and the features of inner speech remain unresolved. The solution of these issues is important for understanding the neurophysiological mechanisms of organization of hidden speech activity, which is a necessary condition for the development of improved methods for decoding inner speech and the development of methods for correcting speech development disorders.

A study [2] has shown that the brain successfully distinguishes between internal and external speech. Simultaneously applying the EEG method and the method of functional near infrared spectroscopy (fNIRS), the authors showed that the differences between internal and external speech are due not only to specific language and motor processes, but also to inhibitory mechanisms.

Many works show great progress in deciphering motor [3] and visual neural signals [4] to help restore lost functions in patients with neurological disorders. An extension to these approaches is the development of assistive devices that restore natural communication in patients with limited verbal communication [5].

The ability to interpret unspoken or imaginary speech using electroencephalography is of therapeutic interest to people suffering from speech disorders. It is also necessary to create a brain-computer interface without reference to articulatory actions [6].

In a recent work [7], the authors showed the possibility of using EEG-based automatic speech recognition systems as a feedback tool in speech therapy for patients with aphasia. The results presented in another article show the first step towards demonstrating the feasibility of using non-invasive neural signals to develop a reliable real-time speech prosthesis for stroke survivors suffering from aphasia, apraxia and dysarthria [9].

The purpose of this study is to identify, according to EEG data, the specifics of the processes of internal pronunciation of words by a group of people with rhotacism in comparison with the normative pronunciation. The task of the work was to compare the ERP with the internal pronunciation of words containing the sound “*r*”, the external pronunciation of which is difficult for a group of people with rhotacism. At the same time, ERP are recorded based on localization according to EEG data (patent of the Russian

Federation No. 2 785 268 [10]) in the area of 41 structures, including all subcortical formations, as well as cortical areas that are somehow involved in the process of internal pronunciation. The obtained results are planned to be used for the implementation of new methods for decoding inner speech.

2 Methods

2.1 Participants and Stimuli

The study involved 36 subjects: 18 women and 18 men aged 22–38 years. The subjects were divided into two groups: 15 people (6 women and 9 men) had a peculiarity of sound pronunciation in the form of racism and formed group 1, the remaining 21 subjects (12 women and 9 men) had a standard pronunciation and were included in group 2. The subjects had higher education. All subjects did not have a history of mental illness, and also signed a voluntary informed consent to participate in the experiment, approved by the ethics committee of the Faculty of Psychology of Moscow State University named after M. V. Lomonosov No. 6, 2020.

Five words of the Russian language were used as stimuli, however, ERP was registered only for 3 of them: “kur'er” (meaning “courier”), “ograda” (meaning “fence”), “raketa” (meaning “rocket”). These stimuli were presented in audio format, recordings were used, spoken by a female voice, in which these words are presented without additional sounds, noises and the possibility of forming phrases and sentences.

2.2 Equipment

The BrainSys program was used to record and edit the EEG to exclude artifacts. Registration of the electrical activity of the brain was carried out monopolar, using a 19-channel electroencephalograph Neuro-KM (company Statokin, Russia), according to the ‘10–20%’ international system. Stimuli were presented using the Presentation program (version 18.0 from Neurobehavioral Systems, Inc.). Average ERP and corresponding 95% confidence intervals were calculated for each stimulus.

2.3 Procedure

The presentation of stimuli was carried out using a stationary computer through headphones using the Presentation program in a random order, each 50 times, the total sequence of presentations consisted of 250 stimuli, the duration of the experiment was about 25 min. The beginning of internal, soundless pronunciation was given by a special signal (short sound). During the experiment, the subject was with his eyes closed.

2.4 Data Processing

To exclude artifacts, a visual analysis of the EEG was performed using the BrainSys program.

Further analysis was carried out using the author's method for determining the localization of brain activity "virtually implanted electrode" (Vartanov, 2022; patent of the Russian Federation No. 2 785 268).

This technology makes it possible to reconstruct the electrical activity of a source with predetermined coordinates relative to scalp electrodes based on scalp EEG data. The method is based on the analysis of the dynamics and correlation of changes in the signal along the leads, but with the addition of artificially generated data calculated on the basis of the distances from the point under study to the scalp electrodes. In this case, a unipolar source model is used for low-frequency (up to 32 Hz) EEG and the law of linear decrease in electric potential with distance. Principal Component Analysis (PCA) and orthogonal rotation uniquely find one factor that is known to exist in the combined array of experimental and artificially generated data. Each potential source is located independently of the others. The obtained and denormalized factorial values for the experimental EEG can be interpreted as the electrical activity of the "local field" when the electrode is implanted in the corresponding point of the brain. The method minimizes the appearance of false activity, a special cleaning procedure is possible, which makes it possible to exclude the influence of neighboring areas and reveal the own electrical activity of this area of the brain.

As a result, the activity was studied at 41 points selected according to the MNI152 atlas in the center of the following structures: Hypothalamus, Brainstem, Mesencephalon, Medulla Oblongata, Caput n.Caudati L, Caput n.Caudati R, Globus Pallidus Medialis L, Globus Pallidus Medialis R, Putamen L, Putamen R, Thalamus L, Thalamus R, Hippocampus L, Hippocampus R, Corpus Amygdaloideum L, Corpus Amygdaloideum R, G.Cingulate Medialis, Anterior Cingulate BA32, Broca BA44 L, Insula L BA13, Insula R BA13, Ventral Striatum BA25, Parietal cortex BA7 R, Wernicke BA22 L, Dorsomedial prefrontal cortex BA9 L, Dorsomedial prefrontal cortex BA9 R, Supramarginal gyrus BA40 L, Cerebellum R, BA22 R, Angular G.BA39 L, Supramarginal gyrus BA40 R, BA44 R, Parietal cortex BA7 L, V1 BA17 L, V1 BA17 R, Cerebellum L, Angular G.BA39 R, Middle Frontal BA10 L, Middle Frontal BA10 R, Orbital Frontal BA47 L, Orbital Frontal BA47 R.

For each of these structures, the ERP was averaged over all stimuli (three words, $3 * 50 = 150$) and all subjects within the corresponding group ($21 * 150 = 3150$ and $15 * 150 = 2250$, respectively) and a 95% confidence interval was also calculated, which made it possible to assess the significance of differences between groups in the amplitude of the corresponding peaks.

3 Results

The most pronounced ERP in the group of people without speech disorders was registered in the following structures: the caudate nuclei on the right and left (Caput n.Caudati L, Caput n.Caudati R), the Globus pallidus on the right (Globus Pallidus Medialis R), the cingulate cortex (G.Cingulate Med.24), the striatum (Ventral Striatum BA25), the dorsomedial prefrontal cortex (Dorsomedial prefrontal cortex BA9), anterior cingulate cortex (Anterior Cingulate BA32), field 17 on the right and left (V1 BA17 L, V1 BA17 R), Broca's area on the right (BA44 R), Wernicke's area on the right (Wernicke BA22 R),

angular gyrus on the right (Angular G.BA39 R), anterior prefrontal cortex on the right and left (Middle Frontal BA10 L, Middle Frontal BA10 R). There is a pronounced ERP P200 in the caudate nucleus (Fig. 1 on the left), cingulate cortex, anterior cingulate gyrus, striatum, dorsomedial prefrontal cortex, anterior prefrontal cortex for the normal group. In people with rhotacism, no pronounced potentials were registered in these structures. It is worth noting the presence of ERP P150 and N200 in the area of the globus pallidus, N250 in the area of field 17 according to Brodmann, N200 in Broca's area on the right, Wernicke's area on the right (Fig. 1 on the right) and the angular gyrus also for a group of people without speech disorders.

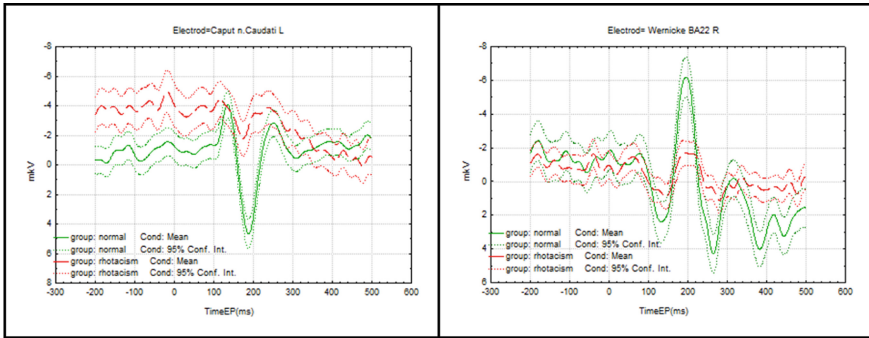


Fig. 1. ERP localized in the caudate nucleus on the left (Caput n.Caudati L) and Wernicke's area on the right (Wernicke BA22 R) for a group of people without speech disorders (group: normal) and a group of people with rhotacism (group: rhotacism). The small dotted line represents the corresponding 95% confidence intervals. On the horizontal axis, time in ms; on the vertical axis, amplitude in μ V.

In the area of the midbrain (Mesencephalon), medulla oblongata (Medulla Oblongata) and the insular lobe on the left (Insula L BA13), significant differences in ERP were recorded in groups of people with and without rhotacism. The group of people with rhotacism is characterized by the presence of pronounced ERP N200 in the areas of the midbrain, medulla oblongata and insular lobe (Fig. 2).

There were no significant differences in the following structures: brainstem (Brainstem), hippocampus (Hippocampus), insula on the right (Insula R BA13) and Wernicke's area on the left (Wernicke_BA22_L). For both groups of subjects, negative ERP was found in these structures at a latency of 200 ms (N200) (Fig. 3).

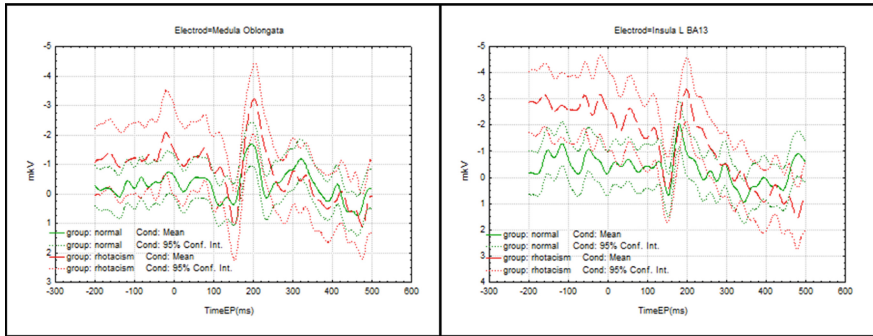


Fig. 2. ERP localized in the medulla oblongata (Medulla Oblongata) and insular lobe on the left (Insula L BA13) for a group of people without speech disorders (group: normal) and a group of people with rhotacism (group: rhotacism). The small dotted line represents the corresponding 95% confidence intervals. On the horizontal axis, time in ms; on the vertical axis, amplitude in μV .

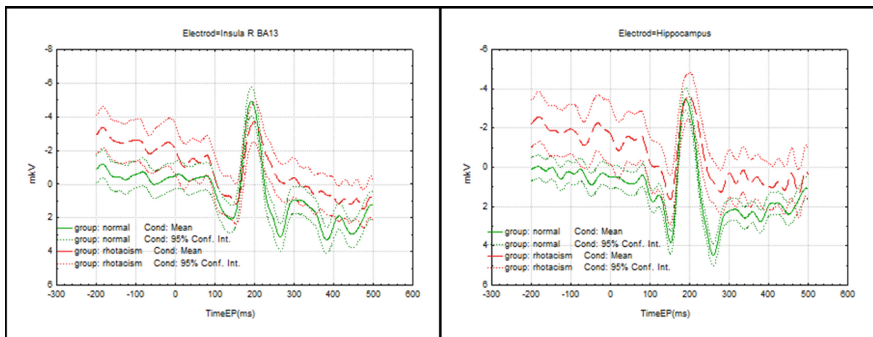


Fig. 3. ERP localized in the insula on the left (Insula L BA13) and hippocampus (Hippocampus) for a group of people without speech disorders (group: normal) and a group of people with rhotacism (group: rhotacism). The small dotted line represents the corresponding 95% confidence intervals. On the horizontal axis, time in ms; on the vertical axis, amplitude in μV .

4 Conclusion

The results obtained in the present study demonstrate significant differences in ERP in the internal pronunciation of words between groups of people with rhotacism and people without speech disorders. These differences were registered in many brain structures, including cortical and subcortical formations using the author's method for determining the localization of brain activity "virtually implanted electrode". The group of people with rhotacism is characterized by the presence of a pronounced ERP N200, compared with the group of people without speech disorders, in more evolutionarily early brain structures, such as the midbrain and medulla oblongata, which, presumably, can be explained by the fixation of the pattern of defective pronunciation of the sound 'r' in early childhood. According to the results obtained, in evolutionarily later brain structures, significant differences in ERP are either not observed (hippocampus, Wernicke's area

on the left, etc.), or a pronounced ERP is determined only in a group of people without speech disorders (anterior cingulate cortex, Wernicke's area on the right, etc.). The result obtained in relation to the insular lobe confirms some inconsistency of the participation of this structure in the process of speech production [10]. In the insula on the left, there are significant differences in ERP N200 between the group of people with rhotacism and people without speech disorders. No significant differences were found in the same structure on the right.

Acknowledgements. The research is financially supported by the Russian Science Foundation, Project II: 20-18-00067-II.

References

1. Boltakova, N.I.: Theoretical Foundations of Speech Therapy. Dyslalia: Brief Lecture Notes, p. 69. Kazan Federal University, Kazan (2013)
2. Stephan, F., Saalbach, H., Rossi, S.: The brain differentially prepares inner and overt speech production: electrophysiological and vascular evidence. *Brain Sci.* **10**(3), 148 (2020). <https://doi.org/10.3390/brainsci10030148>
3. Ajiboye, A.B., Willett, F.R., Young, D.R., Memberg, W.D., Murphy, B.A., Miller, J.P., et al.: Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration. *Lancet* **389**, 1821–1830 (2017). [https://doi.org/10.1016/S0140-6736\(17\)30601-3](https://doi.org/10.1016/S0140-6736(17)30601-3)
4. Lewis, P.M., Ackland, H.M., Lowery, A.J., Rosenfeld, J.V.: Restoration of vision in blind individuals using bionic devices: a review with a focus on cortical visual prostheses. *Brain Res.* **1595**, 51–73 (2015). <https://doi.org/10.1016/j.brainres.2014.11.020>
5. Martin, S., Iturrate, I., Millán, J.D.R., Knight, R.T., Pasley, B.N.: Decoding inner speech using electrocorticography: progress and challenges toward a speech prosthesis. *Front. Neurosci.* **12**, 422 (2018). <https://doi.org/10.3389/fnins.2018.00422>
6. Balaji, A., Haldar, A., Patil, K., et al.: EEG-based classification of bilingual unspoken speech using ANN. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1022–1025 (2017). <https://doi.org/10.1109/EMBC.2017.8037000>
7. Ballard, K.J., Etter, N.M., Shen, S., Monroe, P., Tien Tan, C.: Feasibility of automatic speech recognition for providing feedback during tablet-based treatment for apraxia of speech plus aphasia. *Am. J. Speech Lang. Pathol.* **28**(2S), 818–834 (2019). https://doi.org/10.1044/2018_AJSLP-MS18-18-0109
8. Krishna, G., Carnahan, M., Shamapant, S., Surendranath, Y., Jain, S., Ghosh, A., Tran, C., Millan, J.D.R., Tewfik, A.H.: Brain signals to rescue aphasia, apraxia and dysarthria speech recognition. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, pp. 6008–6014 (2021). <https://doi.org/10.1109/EMBC46164.2021.9629802>
9. Vartanov, A.V.: A new method of localizing brain activity using the scalp EEG data. *Proc. Comput. Sci.* **213**, 41–48 (2022). <https://doi.org/10.1016/j.procs.2022.11.036>
10. Woolnough, O., Forseth, K.J., Rollo, P.S., Tandon, N.: Uncovering the functional anatomy of the human insula during speech. *eLife* **8**, e53086 (2019). <https://doi.org/10.7554/eLife.53086>



One Robust Variant of the Principal Components Analysis

Z. M. Shibzukhov^{1,2}(✉)

¹ Moscow Pedagogical State University, Moscow, Russia

² Institute of Applied Mathematics and Automation KBSC RAS, Nalchik, Russia
intellimath@mail.ru

Abstract. A new robust variant of the formulation of the problem and the method of searching for the principal components is considered. It is based on the application of differentiable estimates of the mean value, insensitive to outliers, to construct a robust target functional. This approach makes it possible to overcome the impact of outliers in the data. The effectiveness of the proposed method is clearly demonstrated on real datasets.

Keywords: Principal components analysis · Robust estimation · Data decomposition

1 Introduction

The classical *principal component analysis* (PCA) was originally considered as the problem of the best approximation of a finite set of points by straight lines and planes [1]. PCA is one of the methods of dimensionality reduction and anomaly detection in data. Also, the presentation of training data in the basis of the *principal components* (PC) contributes to improving the reliability of the error back propagation method for finding the global minimum of the quadratic error function of a neural network of linear elements [2]. It is also used to compress the weight matrix of a neural network to simplify its structural complexity [3].

However, if part of the data is significantly distorted, then this inevitably leads to a noticeable distortion of the results of the classical PCA. For example, in Fig. 1 the angular displacement of the main components towards the outliers is clearly demonstrated.

An overview of classical and robust PCA can be found in [4]. A detailed overview of modern robust variants of PCA is contained in [5].

In this paper, we propose a new robust version of the classical formulation of the problem and an iterative reweighting method for searching for PC. It is based on the application of differentiable estimates of the average value, insensitive to outliers. In principle, this approach makes it possible to overcome the influence of on the process of searching for the PC. The classical and new robust

version of PCA is described below. The effectiveness of the proposed method is clearly demonstrated on real data. Our approach can also be considered as a generalization of methods based on the application of M-estimates to solve the PCA problem [6].

2 Classical Formulation and the Method

Before presenting a robust version of PCA, let's consider the classical formulation of the problem. The problem of finding the center vector a_0 and the orthonormal basis a_1, \dots, a_m of an m -dimensional hyperplane in \mathbb{R}^n is solved, such that the sum of the squares of the Euclidean distances to it from the points x_1, \dots, x_N minimal:

$$\sum_{k=1}^N \left(\|x_k - a_0\|^2 - \sum_{j=1}^m (x_k - a_0, a_j)^2 \right) \rightarrow \min. \tag{1}$$

The solution of this problem is reduced to a chain of problems in which the vectors a_0, a_1, \dots, a_m are sequentially located.

The vector a_0 is found as a solution to the problem:

$$a_0 = \arg \min_{a \in \mathbb{R}^n} \sum_{k=1}^N \|x_k - a\|^2. \tag{2}$$

Its solution is a sample average:

$$a_0 = \frac{1}{N} \sum_{k=1}^N x_k. \tag{3}$$

After finding a_0 , centering is performed:

$$x_k \rightarrow x_k - a_0, \quad k = 1, \dots, N. \tag{4}$$

Then the vectors a_j ($j = 1, \dots, m$) are sequentially searched for as a solution to the problem

$$a_j = \arg \min_{\|a\|=1} \frac{1}{N} \sum_{k=1}^N (\|x_k\|^2 - (a, x_k)^2). \tag{5}$$

After finding the next main component a_j , the transformation is performed:

$$x_k \rightarrow x_k - a_j(a_j, x_k). \tag{6}$$

Using the Lagrange multiplier method, we reduce the search problem a_j to the following problem:

$$a_j, \lambda_j = \arg \min_{a, \lambda} \left\{ \frac{1}{N} \sum_{k=1}^N (\|x_k\|^2 - (a, x_k)^2) + \lambda (\|a\|^2 - 1) \right\}. \tag{7}$$

Let X be a matrix composed of rows, $S = \frac{1}{N}X^\top X$ is a covariance matrix. Then

$$\frac{1}{N} \sum_{k=1}^N (a, x_k)^2 = \frac{1}{N} (Xa)^\top (Xa) = \frac{1}{N} a^\top (X^\top X)a = a^\top Sa. \tag{8}$$

In accordance with the necessary condition of the extremum, a_j and λ_j are the solution of a system of equations:

$$\begin{aligned} Sa_j &= \lambda_j a_j \\ \|a_j\| &= 1. \end{aligned} \tag{9}$$

That is, a_j and λ_j are orthonormal eigenvectors and eigenvalues of the matrix S . To search for a_j , one can apply an iterative procedure:

$$a^{t+1} = \frac{1}{\lambda^t} (Sa^t), \quad \lambda^t = \frac{(a^t)^\top Sa^t}{(a^t, a^t)}. \tag{10}$$

3 Robust Formulation and the Method

The proposed robust formulation of the problem involves replacing the minimization problem

$$Q(a) = \frac{1}{N} \sum_{k=1}^N \Phi_k(a) \tag{11}$$

with the minimization problem

$$Q_M(a) = M\{\Phi_1(a), \dots, \Phi_N(a)\}, \tag{12}$$

where $M\{z_1, \dots, z_N\}$ is a differentiable estimation of the average value, insensitive or insensitive to emissions [7, 8].

The problem of minimizing Q_M is reduced to solving a vector equation

$$\sum_{k=1}^N \frac{\partial M}{\partial z_k} \nabla \Phi_k(a) = 0. \tag{13}$$

To solve it, one can apply the iterative reweighing procedure:

$$a^{t+1} = \arg \min \sum_{k=1}^N v_k^t \Phi_k(a), \tag{14}$$

where

$$v_k^t = \frac{\partial M\{\Phi_1(a^t), \dots, \Phi_N(a^t)\}}{\partial z_k}. \tag{15}$$

This formulation and method of solving the problem, in principle, can significantly reduce the impact of outliers.

The robust version of the formulation of the problem of finding the center a_0 takes the following form:

$$a_0 = \arg \min_{a \in \mathbb{R}^n} M \{ \|x_1 - a\|^2, \dots, \|x_N - a\|^2 \}. \tag{16}$$

This problem is reduced to solving the equation

$$a = \sum_{k=1}^N \frac{\partial M \{ \|x_1 - a_0\|^2, \dots, \|x_N - a_0\|^2 \}}{\partial z_k} x_k, \tag{17}$$

that can be solved using the following iterative procedure:

$$a^{t+1} = \sum_{k=1}^N v_k^t x_k, \tag{18}$$

where

$$v_k^t = \frac{\partial M \{ \|x_1 - a^t\|^2, \dots, \|x_N - a^t\|^2 \}}{\partial z_k}. \tag{19}$$

After finding a_0 , centering is also performed:

$$x_k \rightarrow x_k - a_0, \quad k = 1, \dots, N. \tag{20}$$

The robust version of the search problem a_j takes the following form:

$$a_j = \arg \min_{\|a\|=1} M \{ \|x_1\|^2 - (a, x_1)^2, \dots, \|x_N\|^2 - (a, x_N)^2 \}. \tag{21}$$

It also boils down to solving a system of equations:

$$\begin{aligned} S_a a &= \lambda a \\ \|a\|^2 &= 1, \end{aligned} \tag{22}$$

where

$$S_a = \sum_{k=1}^N v_k(x_k)^\top x_k, \tag{23}$$

and

$$v_k^t = \frac{\partial M \{ \|x_1\|^2 - (a^t, x_1)^2, \dots, \|x_N\|^2 - (a^t, x_N)^2 \}}{\partial z_k}. \tag{24}$$

The search for the solution of (22) is based on an iterative reweighing scheme.

At the beginning, a^0 and λ^0 are selected. At step p , a^{p+1} and λ^{p+1} are searched for as a solution to a system of equations:

$$\begin{aligned} S^p a &= \lambda a \\ \|a\|^2 &= 1, \end{aligned} \tag{25}$$

where

$$S^p = \sum_{k=1}^N v_k^p(x_k)^\top x_k, \tag{26}$$

and

$$v_k^p = \frac{\partial M \{ \|x_1\|^2 - (a^p, x_1)^2, \dots, \|x_N\|^2 - (a^p, x_N)^2 \}}{\partial z_k}. \tag{27}$$

To find a solution to the system (25), the iterative procedure (10) is used.

4 Illustrative Examples

For experimental confirmation of the effectiveness of the approach proposed here, examples of the application of classical and robust PCA for several data sets are considered. In the robust version, the following function of estimating the average value is used—the censored average [7, 8]:

$$CM_\alpha\{z_1, \dots, z_N\} = \frac{1}{N} \sum_{k=1}^N \min(z_k, \bar{z}_\alpha), \tag{28}$$

where $0 < \alpha < 1$,

$$\bar{z}_\alpha = \arg \min_u \sum_{k=1}^N \rho_\alpha(z_k - u), \tag{29}$$

$$\rho_\alpha(r) = \begin{cases} (1 - \alpha)\rho(r), & \text{if } r < 0 \\ 0, & \text{if } r = 0 \\ \alpha\rho(r), & \text{if } r > 0, \end{cases} \tag{30}$$

$$\rho(r) = \sqrt{\varepsilon^2 + r^2}. \tag{31}$$

Here \bar{z}_α is the smoothed variant of α -quantile. The censored average is an arithmetic mean in which all arguments exceeding \bar{z}_α are replaced by \bar{z}_α . At the same time,

$$\frac{\partial CM_\alpha}{\partial z_k} = \begin{cases} \left(\frac{1}{M} + \frac{m}{M}\right) \frac{\partial \bar{z}_\alpha}{\partial z_k}, & \text{if } z_k < \bar{z}_\alpha \\ \frac{m}{M} \frac{\partial \bar{z}_\alpha}{\partial z_k}, & \text{if } z_k \geq \bar{z}_\alpha \end{cases} \tag{32}$$

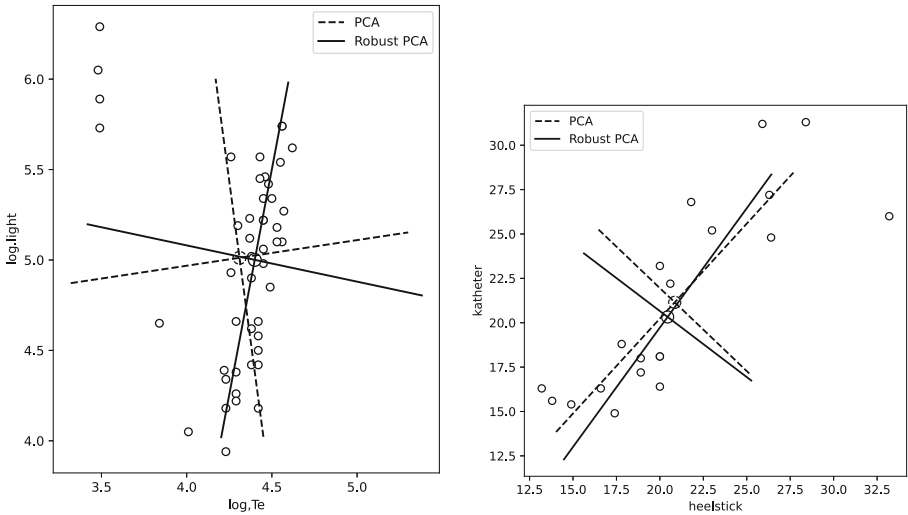


Fig. 1. Illustrative examples of robust PCA for two plain datasets. **Left:** PCs for STARS CYG dataset. **Right:** PCs for KELLY1984 dataset.

The following iterative procedure is used to find \bar{z}_α :

$$u^{t+1} = \frac{\sum_{k=1}^N \varphi(z_k - u^t) z_k}{\sum_{k=1}^N \varphi(z_k - u^t)}, \tag{33}$$

where $\varphi(r) = \rho'_\alpha(r)/r$.

Dataset starsCYG. Consider a data set for constructing a Hertzsprung-Russell diagram of the CYG OB1 star cluster [9], which describes the relationship between the logarithm of luminosity (log.light) and the logarithm of temperature (log.Te) of stars. In Fig. 1 (left) there are 7 points that can be attributed to outliers. The application of the classical component method gives the vectors of the PC rotated counterclockwise in the direction of outliers. The application of the robust PCA ($\alpha = 0.87, \varepsilon = 0.001$) makes it possible to find the unbiased position of the center and the PC that do not deviate under the influence of outliers.

Dataset Kelly1984. Consider a dataset Kelly1984 [10], which describes simultaneous pairs of measurements of serum kanamycin levels in blood samples drawn from 20 babies. In Fig. 1 (right) there are some points that can be attributed to outliers. The application of the classical component method gives the vectors of the PC rotated counterclockwise in the direction of outliers. The application of the robust PCA ($\alpha = 0.8, \varepsilon = 0.001$) makes it possible to find the unbiased position of the center and the PC that do not deviate under the influence of outliers.

Dataset HIP_stars. Consider a data set for plotting a chart for stars from a dataset [11]. Figure 2 show a projection on a pair of Vmag and B-V parameters. The classical PCA gives offset PC for a given projection. The proposed robust PCA ($\alpha = 0.95, \varepsilon = 0.001$) makes it possible to overcome the influence of outliers.

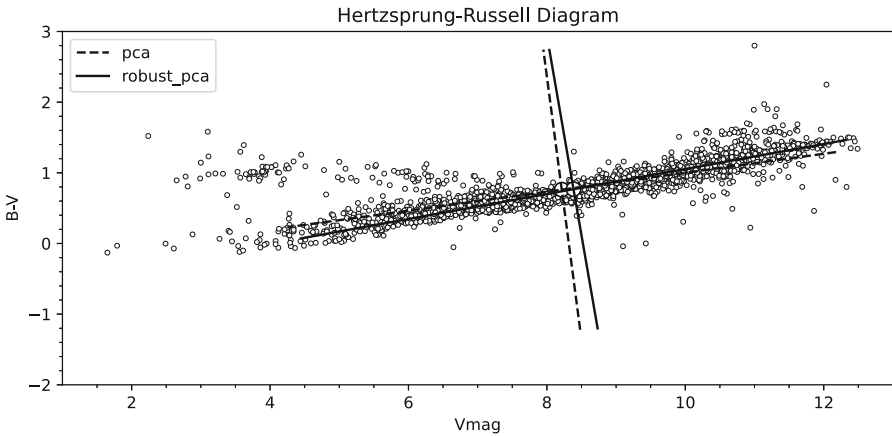


Fig. 2. The positions of the centers and PCs for HIP_STARS dataset.

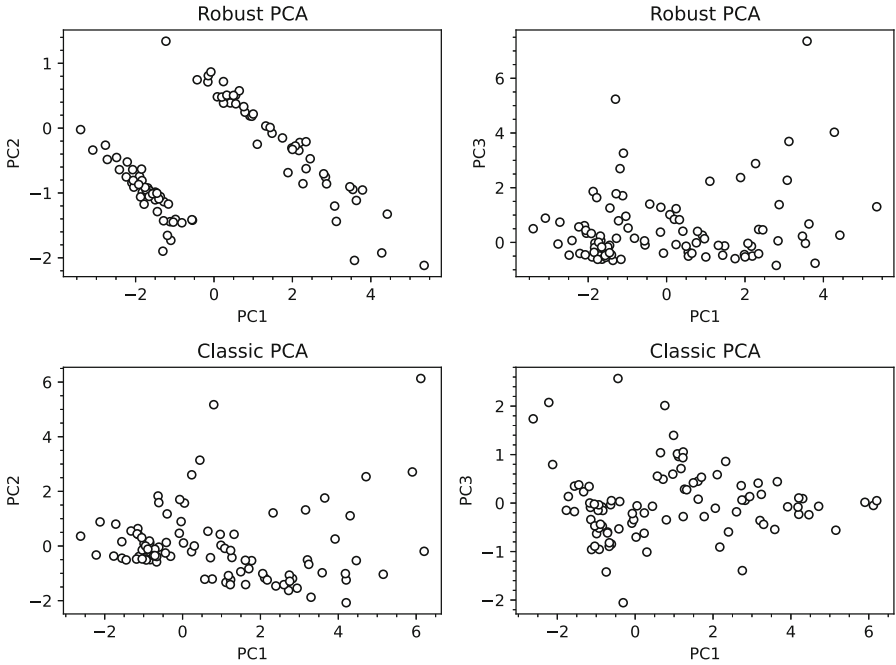


Fig. 3. Projections of the CigarettesSW dataset on PC1×PC2 and PC1×PC3 for robust and classic PCA

Dataset CigarettesSW. Consider panel data on cigarette consumption in 48 continental US states for 1985–1995 years [12]. The PCA is used for tabular data that covers 7 features (cpi, population, packs, income, tax, price, taxes; state, year are excluded). Figure 3 clearly show that the data in projections on PC1×PC2 and PC1×PC3, which are obtained on the basis of the robust PCA ($\alpha = 0.55, \varepsilon = 0.001$), have a more contrasting appearance. On the PC1×PC2 projection, the data lines up along two straight lines, on the PC1×PC3 projection, the data lines up along the PC1 axis. For the corresponding projections of data on PC1×PC2 and PC1×PC3, which are obtained on the basis of the classical PCA, the data visually have a larger spread.

5 Conclusion

A robust version of the formulation of the PCA problem, based on minimizing differentiable estimates of the mean, which can be significantly more resistant to outliers, makes it possible to find unbiased vectors of the PC. The indicator α in the smoothed variant of quantile estimate roughly corresponds to the proportion of data that are not outliers (a more accurate value is experimentally found in its neighborhood). This method also makes it possible to identify outliers by analyzing the empirical distribution of distances to straight lines passing through the center a_0 along the vectors of the PC.

References

1. Pearson, K.: On lines and planes of closest fit to systems of points in space. *Phil. Mag.* **2**, 559–572 (1901). <https://doi.org/10.1080/14786440109462720>
2. Baldi, P., Hornik, R.: Neural networks and principal component analysis: learning from examples without local minima. *Neural Netw.* **2**, 53–58 (1989). [https://doi.org/10.1016/0893-6080\(89\)90014-2](https://doi.org/10.1016/0893-6080(89)90014-2)
3. Praggastis, B., Brown, D., Marrero, C.O., Purvine, E., Shapiro, M., Wang, B.: The SVD of convolutional weights: a CNN interpretability framework. [arXiv:2208.06894](https://arxiv.org/abs/2208.06894) (2022)
4. Jolliffe, I.T.: *Principal Component Analysis*. Springer International Publishing (2002). <https://doi.org/10.1007/b98835>
5. Bouwmans, T., Sobral, A., Javed, S., Jung, S.K., Zahzah, E.-H.: Decomposition into low-rank plus additive matrices for background/foreground separation: a review for a comparative evaluation with a large-scale dataset. [arXiv:1511.01245](https://arxiv.org/abs/1511.01245) (2015)
6. Zhang, T., Lerman, G.: A novel M-estimator for robust PCA. [arXiv:1411.4863](https://arxiv.org/abs/1411.4863) (2014)
7. Shibzukhov, Z.M.: Machine learning based on the principle of minimizing robust mean estimates. In: Samsonovich, A.V., Gudwin, R.R., Simões, A.d.S. (eds.) *Brain-Inspired Cognitive Architectures for Artificial Intelligence: BICA*AI 2020. Advances in Intelligent Systems and Computing*, vol. 1310, pp. 472–477. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-65596-9_56
8. Shibzukhov, Z.M.: Minimizing robust estimates of sums of parameterized functions. *J. Math. Sci.* **260**, 249–264 (2022). <https://doi.org/10.1007/s10958-022-05689-z>
9. Rousseeuw, P.J., Leroy, A.M.: *Robust Regression and Outlier Detection*. Wiley (1987). <https://doi.org/10.1002/0471725382>
10. Kelly, B.: The influence function in the errors in variables problem. *Ann. Stat.* **12**(1), 87–100 (1984). <https://doi.org/10.1214/aos/1176346394>
11. Dataset Hipparcos. https://www.astrostatistics.psu.edu/datasets/HIP_star.html
12. Dataset CigarettesSW: Cigarette Consumption Panel Data. <https://rdrr.io/rforge/gmm4/man/CigarettesSW.html>



Using Electronic Nose in Forensic Odor Analysis

Alexander Shtanko^(✉) and Sergey Kulik^{ID}

National Research Nuclear University MEPhI, Kashirskoe Shosse 31, Moscow 115409, Russia
shtanko-mephi@yandex.ru

Abstract. Forensic odor analysis often uses biodetectors (usually dogs) to perform odor matching, such as when determining whether two odor samples correspond to each other. Biodetectors can be hard to train, expensive to maintain and time-consuming in application. A task of finding a sample corresponding to a given one out of several samples is prohibitively expensive with large enough number of samples. An alternative to biodetector is using artificial odor analyzer, so called e-nose. We propose a two-step method for finding a corresponding sample out of several samples: first picking few candidates using e-nose then applying biodetector to find the required sample out of the already chosen candidates. Provided e-nose has high enough performance characteristics, this approach can make the task solvable in practice. We also calculate theoretical performance of this method as well as generalize the method into an abstract cascade classifier similar to the one used in Viola-Jones algorithm and calculate its theoretical performance.

Keywords: Forensics · Cascade classifier · E-nose · Cognition

1 Introduction

Odor analysis can be applied in various fields including food quality control and criminal forensics. Generally, animals, so-called biodetectors, are unrivaled in odor analysis. As such, for example, criminal forensics law enforcement agencies use trained dogs and have established procedures for odor analysis and employ necessary information processing systems [1], perhaps similar to data management systems in scientific organizations [2]. Given established procedures in forensic analysis the results given by biodetectors in forensics can be considered perfect for the purposes of this article. However live animals can be expensive to train and maintain thus the scale of their application is limited. One potential way to apply odor analysis on a larger scale is using gas analyzer, also called artificial nose or e-nose.

We analyzed patents of Russian Federation and have found inventions applying e-nose for odor analysis in various fields (Table 1). Several of the patents contain designs and methods for new types of e-nose, majority of them contain specifics of the application for certain tasks such as application for food quality control or medical application.

As can be seen from the table, there are already law enforcement uses of e-nose devices, for example, patent 2761165 (row 32 from the table) describes odor recognition

Table 1. Invention patents mentioning e-nose.

No.	Patent number	Year	Description
1	2279065	2006	Processing method of signals from multisensor electronic nose
2	2327984	2008	Multichannel electronic nose on piezosensors
3	2331069	2008	Method for the quantitative assessment of meat flavor
4	2334228	2008	Detection of adulteration of milk products by artificial flavors
5	2341793	2008	Method for measuring the humidity of dry bouillons
6	2407578	2010	System for transmitting smells at a distance
7	2441233	2012	Method for the detection of amines in gas-air mixtures
8	2442158	2012	Static e-nose sensors for odors of grapes, grape material, and juice
9	2442159	2012	Static electronic nose sensors for detection of adulteration of apple juices, nectars and beverages by adding artificial flavorings
10	2452946	2012	Optimization of liquid smoke-based marinade recipes by flavor
11	2456590	2012	Detection of multicomponent gas mixtures of benzene, etc.
12	2458139	2012	Method for the medical diagnosis from cervical mucus
13	2466528	2012	Method for determining the early spoilage of seeds
14	2502997	2013	Coffee drinks recognition for different social groups and evaluation of flavor characteristics within the group
15	2510494	2014	Method for determining dysbiosis in poultry
16	2533692	2014	Multisensor acoustic array for electronic nose and tongue
17	2564877	2015	Non-invasive rapid diagnosis of respiratory organs in calves
18	2586284	2016	Method for detection of wheat disease
19	2586446	2016	Method for analyzing the composition of the gases
20	2614667	2017	Method for the rapid assessment of the quality of yeast
21	2619261	2017	Method for recognition of ethanol
22	2620343	2017	Detection of milk adulteration by diluting with piezosensors
23	2640507	2018	Method of organoleptic evaluation of the children toys
24	2649217	2018	Hybrid acoustic sensor (electronic nose and electronic tongue)
25	2659712	2018	Detection of explosive and narcotic substances in the air
26	2676860	2019	Gas multisensor for the analysis of gas mixture
27	2679409	2019	Obtaining the diagnostic information from the skin smell
28	2699366	2019	Compact electronic nose type device
29	2707621	2019	A method for analyzing samples
30	2729106	2020	A method for non-invasive monitoring of upper airways in calves

(continued)

Table 1. (continued)

No.	Patent number	Year	Description
31	2751756	2021	Security system
32	2761165	2021	Rapid instrumental recognition of human and clothing odor
33	2778205	2022	Optical absorption gas analyzer

method for humans and clothes, patent 2659712 (row 25 from the table) describes the detection method of controlled substances.

There are also academic papers regarding odor analysis devices [3], including realizations using neural networks [4]. Neural networks have been applied in various fields including our works [5], and even specifically for forensic analysis [6] and memristor-based networks [7]. Forensic analysis automation often includes considerations for operator's factors [8].

Although e-nose is not used in forensics as an established practice it is possible to suppose that at some point e-nose would be accurate enough to be used in forensics. It should be expected that these types of devices will still have lower accuracy than biodetectors. However, if the cost of their use is low enough, they can be used at least together with biodetectors.

The research in the field of artificial odor analysis will potentially aid in cognitive research, including emotion research [9], risk estimation [10], professional competence [11] and cognitive map [12].

In this article we propose two-step method for using e-nose together with biodetectors, we also generalize the approach to a cascade classifier similar to the one used in Viola-Jones algorithm [13].

2 Two-Step E-Nose Method

Let us review the task at hand. There are several examined odor samples and a given query sample. The task is to find among the examined odor samples the one (target sample) corresponding to the query sample. We suppose that among the examined samples there is one target sample.

Since during the solution of this task every examined sample is either the target or not, and every sample must be labeled as either the target or not; this task can be viewed as binary classification problem regardless of the exact procedure employed with the target sample being the positive sample and the rest of the samples being the negative samples.

We suppose that the biodetector can perform classification perfectly.

Let us examine the odor analysis from probabilistic standpoint of operation research [14] using the probability theory [15].

Thus, there are N examined odor samples and one target sample among them. There is an e-nose device, and it can perform the classification with certain accuracy. We suppose e-nose provides confidence score and could use different thresholds for positive

classification. Let us introduce the following performance characteristics of the e-nose device: p_{11} —probability of correctly labeling the positive sample as positive; p_{10} —probability of incorrectly labeling the positive sample as negative; p_{00} —probability of correctly labeling the negative sample as negative; p_{01} —probability of incorrectly labeling the negative sample as positive given certain threshold h . Naturally, $p_{11} + p_{10} = 1$ and $p_{01} + p_{00} = 1$. ROC-curve of the e-nose shows the true positive rate p_{11} and false positive rate p_{01} given varying threshold values. Thus, p_{11} can be viewed as a function of p_{01} , that is $p_{11} = ROC(p_{01})$.

The proposed method consists of the first step in which all examined samples are classified by the e-nose and the second step in which chosen (classified as positive) in the first step samples are classified by biodetectors.

Suppose there are N examined samples. The e-nose on average will positively classify the following number of samples:

$$(N - 1)p_{01} + p_{11} = Np_{01} + (p_{11} - p_{01}) \approx Np_{01}. \tag{1}$$

The probability that the target sample will be among the chosen candidates equals p_{11} .

Since the resources of biodetectors are limited, the number of chosen samples as candidates should be small. Suppose the number of candidates can not be larger than M . Then, according to the formula above $p_{01}N \leq M$ meaning the threshold of the e-nose should be chosen so $p_{01} \leq R = M/N$. As ROC-curve is monotonic, $p_{11} \leq ROC(R)$.

Thus, in the first step of the method Np_{01} samples will be chosen and the probability that the target sample will not be rejected is not greater than $ROC(R)$.

3 Cascade Classifier Performance Evaluation

The proposed method could be generalized to a cascade classified like the one used in Viola-Jones face detection algorithm.

Suppose there is a stream of objects need to be classified. These objects can be either positive (“1”) with probability p_1 or negative (“0”) with probability p_0 . Naturally, $p_1 + p_0 = 1$.

We suppose that $p_1 \ll p_0$ meaning the vast majority of objects are negative.

Let there be an initial system S consisting of a single block B to classify these objects (see Fig. 1).

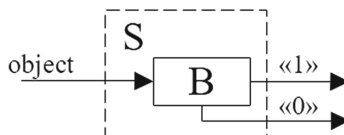


Fig. 1. Initial single-block system S .

This system, on average, spends a certain amount of resources on classification C (per object) and has associated probabilities of correct and incorrect classification

$(p_{00}, p_{01}, p_{10}, p_{11})$ as described in previous section. In practice the resource C can be, for example, time or monetary cost.

To reduce the resources spent this system can be replaced with a two-block system. Another (faster) block is appended in front of the block B to form a two-block cascade classifier (see Fig. 2).

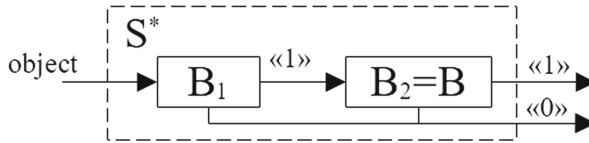


Fig. 2. Improved two-block system S^* .

The constructed two-block system S^* works the following way. An object is inputted into block B_1 where it is classified. If the object is classified as negative the system as a whole returns a negative classification result. Otherwise, the object is passed onto the block B_2 . The classification result of the block B_2 is final. Block B_1 has its own error probabilities $(p_{00}^1, p_{01}^1, p_{10}^1, p_{11}^1)$ and associated costs C^1 . And the block B_2 has the same error probabilities and costs as the initial system: $(p_{00}^2, p_{01}^2, p_{10}^2, p_{11}^2) = (p_{00}, p_{01}, p_{10}, p_{11})$, $C^2 = C$.

In practice the block needs to rarely reject positive objects (even if it hurts overall accuracy) $p_{10}^1 \ll 1$. Consequently, $p_{11}^1 \approx 1$.

The cost of the two-block system consists of the cost of the first block C^1 which is applied every time and the cost of the second block C^2 which is applied with probability p_{01}^1 in case of a negative object (appearing with probability p_0) and with probability p_{11}^1 in case of a positive object (appearing with probability p_1). Thus, the per object cost of applying system S^* equals:

$$C^* = C^1 + (p_0 p_{01}^1 + p_1 p_{11}^1) C^2 = C^1 + (p_0 p_{01}^1 + p_1 p_{11}^1) C. \quad (2)$$

Let the efficiency measure be equal to the ratio of the initial system cost to the improved system cost $E = C/C^*$ meaning how many times the improved system more cost-effective compared to the initial system. Given the expression for C^* the measure can be rewritten:

$$E = \frac{C}{C^*} = \frac{C}{C^1 + (p_0 p_{01}^1 + p_1 p_{11}^1) C^2} = \frac{1}{\left(\frac{C^1}{C^2} + p_0 p_{01}^1 + p_1 p_{11}^1\right)}. \quad (3)$$

It makes sense to use the improved system if, at least, the cost of the initial system are not less than the cost of the improved system, that is $E \geq 1$. Then:

$$\frac{1}{\left(\frac{C^1}{C^2} + p_0 p_{01}^1 + p_1 p_{11}^1\right)} \geq 1, \quad (4)$$

$$1 \geq \frac{C^1}{C^2} + p_0 p_{01}^1 + p_1 p_{11}^1, \quad (5)$$

$$\frac{C^2}{C^1} (1 - p_0 p_{01}^1 - p_1 p_{11}^1) \geq 1. \tag{6}$$

Let the savings ratio of the block B₁ compared to the block B₂:

$$\alpha = \frac{\frac{1}{C^1}}{\frac{1}{C^2}} = \frac{C^2}{C^1}. \tag{7}$$

Then the criterion of the efficient application of the improved system transforms into

$$\alpha (1 - p_0 p_{01}^1 - p_1 p_{11}^1) \geq 1. \tag{8}$$

Since it is assumed that very few objects are positive $p_0 \approx 1, p_1 \approx 0$ then the criterion can approximately be rewritten as $\alpha (1 - p_{01}^1) \geq 1$ or, consequently,

$$\alpha p_{00}^1 \geq 1. \tag{9}$$

Let us calculate the probabilities of correct and incorrect classification for the improved system $(p_{00}^*, p_{01}^*, p_{10}^*, p_{11}^*)$. It is assumed that the classification probabilities for the different blocks are independent.

Then the probability of a true positive classification equals

$$p_{11}^* = p_{11}^1 p_{11}^2. \tag{10}$$

It should be noted that $p_{11}^1 p_{11}^2 = p_{11}^1 p_{11} \leq p_{11}$ meaning this value is always no better compared to the initial system.

The probability of the correct negative classification equals $p_{00}^* = p_{00}^1 + p_{01}^1 p_{00}^2$. Since $p_{00}^1 + p_{01}^1 p_{00}^2 = p_{00}^1 + (1 - p_{00}^1) p_{00}^2 = p_{00}^2 + p_{00}^1 - p_{00}^1 p_{00}^2$ which can further be rewritten as $p_{00}^2 + p_{00}^1 (1 - p_{00}^2) = p_{00}^2 + p_{00}^1 p_{01}^2 = p_{00} + p_{00}^1 p_{01}^2 \geq p_{00}$ this value is no worse than the initial system.

The probability of a false positive classification equals $p_{01}^* = p_{01}^1 p_{01}^2$. Since $p_{01}^1 p_{01}^2 = p_{01}^1 p_{01} \leq p_{01}$ this value is always no worse than the initial system.

The probability of a false negative classification equals $p_{10}^* = p_{10}^1 + p_{11}^1 p_{10}^2$. Since $p_{10}^1 + p_{11}^1 p_{10}^2 = p_{10}^1 + (1 - p_{10}^1) p_{10}^2 = p_{10}^2 + p_{10}^1 - p_{10}^1 p_{10}^2 = p_{10}^2 + p_{10}^1 (1 - p_{10}^2)$ which can further be rewritten as $p_{10}^2 + p_{10}^1 p_{11}^2 = p_{10} + p_{10}^1 p_{11}^2 \geq p_{10}$ this value is no better than the initial system.

Generally, the number of negative answers of the system increases because if either block gives a negative classification, it is classified as negative.

The system can be further generalized to use multiple blocks (Fig. 3).

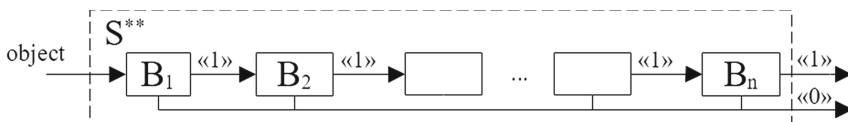


Fig. 3. Generalized *n*-block system S**.

The average cost of such a system can be calculated as:

$$C^{**} = C^1 + \sum_{i=2}^n \left(p_0 \prod_{k=1}^{i-1} p_{01}^k + p_1 \prod_{k=1}^{i-1} p_{11}^k \right) C^i. \quad (11)$$

Similarly, error probabilities can be calculated.

4 Conclusion

Thus, odor analysis (especially in forensics) is accurately performed by biodetectors (trained dogs). Artificial odor analysis devices (e-nose) exist and are applied in various fields; however, we should expect them to perform less accurately.

We proposed a two-step method for forensic odor analysis, specifically for the task of searching for a sample corresponding to a given one among a number of samples. This method makes use of both e-nose and biodetectors. True positive rate of the system is calculated to be no greater than $ROC(R)$.

This approach is further generalized to an abstract cascade classifier. Resource consumption of this method is calculated. It is shown that such an approach can be more resource efficient than using a one-step (one-block) classifier. The criterion of the efficient application is introduced. Error probabilities are calculated. Cascade classifier gives fewer positive classifications and more negative classifications.

As such artificial odor analysis devices can be applied to practical tasks even if they are not as accurate as biodetectors.

Acknowledgments. This work was supported by the MEPhI Program Priority 2030.

References

1. Rybina, T., Kulik, S., Smirnov, D.: Forensic information system for olfactory research. Proc. Comput. Sci. **213**, 175–179 (2022)
2. Artamonov, A., Ionkina, K., Tretyakov, E., Timofeev, A.: Electronic document processing operating map development for the implementation of the data management system in a scientific organization. Proc. Comput. Sci. **145**, 248–253 (2018)
3. Kuchmenko, T.A.: Electronic nose based on nanoweights, expectation and reality. Pure Appl. Chem. **89**(10), 1587–1601 (2017)
4. Adak, M.F., Yumusak, N.: Classification of E-nose aroma data of four fruit types by ABC-based neural network. Sensors **16**(3), 304 (2016)
5. Kulik, S.D., Shtanko, A.N.: Experiments with neural net object detection system YOLO on small training datasets for intelligent robotics. In: Mechanisms and Machine Science, vol. 80, pp. 157–162. Springer, Cham (2020)
6. Kulik, S.D., Shtanko, A.N.: Recognition algorithm for biological and criminalistics objects. In: Biologically Inspired Cognitive Architectures 2019. Proceedings of the Tenth Annual Meeting of the BICA Society, vol. 948, pp. 283–294. AISC (2020)
7. Danilin, S.N., Shchanikov, S.A.: Neural network control over operation accuracy of memristor-based hardware. In: Proceedings of 2015 International Conference on Mechanical Engineering, Automation and Control Systems, pp. 1–5. MEACS 2015 (2015)

8. Kulik, S., Nikonets, D.: Forensic handwriting examination and human factors: improving the practice through automation and expert training. In: *The Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC2016)*, Proceedings, pp. 221–226. Moscow, Russia, 06–08 July 2016
9. Samsonovich, A.V.: Intellectual agents based on a cognitive architecture supporting human-like social emotionality and creativity. In: *Studies in Computational Intelligence*, vol. 799, pp. 39–50. Springer, Cham, Switzerland (2019)
10. Lebedeva, A.V., Guseva, A.I.: Cognitive maps for risk estimation in software development projects. In: *Biologically Inspired Cognitive Architectures Meeting*, pp. 295–304. Springer, Cham (2019)
11. Aleshinskaya, E., Ahmad, A.: A cognitive model to enhance professional competence in computer science. *Proc. Comput. Sci.* **169**, 326–329 (2020)
12. Samsonovich, A.V., Ascoli, G.A.: Cognitive map dimensions of the human value system extracted from natural language. In: Goertzel, B., Wang, P. (eds.) *Frontiers in Artificial Intelligence and Applications*, vol. 157, pp. 111–124. IOS Press (2007)
13. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 511–518 (2001)
14. Taha, H.A.: *Operations Research: An Introduction*, 8th edn. Pearson Prentice Hall, Upper Saddle River, New Jersey (2007)
15. Feller, W.: *An Introduction to Probability Theory and its Applications*, vol. 1, 3rd edn., Wiley, New York (1968)



Hierarchical AGI from First Principles

Sergey Shumsky^(✉) 

Moscow Institute of Physics and Technology, Dolgoprudny 141701, Russian Federation

serge.shumsky@gmail.com

Abstract. The paper provides evidence based on the free energy principle in favor of the hierarchical design of AGI. A neuro-symbolic hierarchical architecture of AGI is proposed as a development of Friston's hierarchical model of the brain.

Keywords: Artificial general intelligence · Hierarchical reinforcement learning · Neuro-symbolic architecture

1 Introduction

Following Karl Friston [8] we consider intelligence as a generic property of complex adaptive systems that behave to preserve their identity, optimally and proactively responding to external threats. The key concept for such systems is the *regulator*.

According to the Good Regulator Theorem, every good regulator of a system must be a model of that system [6]. That is, the ability to adapt implies the ability to build predictive models of the world – a hallmark of intelligence.

Adaptive systems belong to a wide class of dissipative systems. Energy dissipation causes compression of the phase space and contraction of phase trajectories to low-dimensional attractors, as, for example, in self-oscillatory systems. But not all attractors are that simple. There exist *strange attractors* with a very complicated structure that can represent complex adaptive goal-directed behavior of intelligent agents. After all, such behavior is aimed at preserving one's identity, that is, at staying within the limits of the corresponding attractor. Any small perturbations return the system to the same attractor. Such systems are natural regulators, and some of their basic properties follow directly from dissipative dynamics, described by the Fokker-Planck equation:

$$\partial_t p(x, t) = -\partial_x [f(x)p(x, t) - \Gamma \partial_x p(x, t)]. \quad (1)$$

According to Boltzmann's H-theorem, the latter has a Lyapunov function called *free energy*, which does not increase with time:

$$\mathcal{F}(t) = \int dx p(x, t) \ln \frac{p(x, t)}{p(x)} \equiv D(p(x, t) || p(x)) \geq 0, \quad \dot{\mathcal{F}}(t) \leq 0, \quad (2)$$

being the Kullback-Leibler divergence between the current and asymptotic probability density distributions [13].

Because negative log probability measures information, approaching an attractor can be seen as *learning* (how to respond to all possible perturbations). It is natural to call such information, which is crucial for maintaining the integrity of the system, *knowledge* [16]. Thus, the adaptation of systems to their environment can be interpreted as the acquisition of appropriate knowledge necessary for survival. Therefore, dissipative systems with sufficiently complex attractors behave as if they are learning and acquiring new useful skills, that is, such systems behave intelligently. We will argue that the complexity of the attractor can be provided by scale-free hierarchical dynamics.

The rest of the paper is structured as follows. Section 2 uses the free energy approach to provide evidence that strong intelligence requires a hierarchical design. Section 3 goes on to propose a neuro-symbolic Deep Control architecture solving hierarchical reinforcement learning. The merits of the latter are discussed in Sect. 4, related work – in Sect. 5. The last Sect. 6 concludes the paper.

2 Hierarchical Intelligence

Following the above reasoning, the behavior of intelligent agents can be modeled using the *mental states* \mathbf{m} of the agent’s mind, as tuning parameters aimed at minimizing average expected free energy [9]:

$$\mathcal{F} = \int d\mathbf{m} dx q(x, \mathbf{m}) \ln \frac{q(x, \mathbf{m})}{p(x)} = D(q(\mathbf{m}) || e^{-\mathcal{F}(\mathbf{m})}), \quad (3)$$

$$\mathcal{F}(\mathbf{m}) \equiv \int dx q(x|\mathbf{m}) \ln \frac{q(x|\mathbf{m})}{p(x)} = D(q(x|\mathbf{m}) || p(x)). \quad (4)$$

It depends on two functions: (i) slowly varying *generative model* $p(x) = p(a, w, s)$, which describes agent’s knowledge of how his actions a change the hidden states of the world w and how the latter determine the readings of his sensors s , and (ii) *variational model* $q(x, \mathbf{m})$ of the agent’s mind with the expected next mental state \mathbf{m} , which determines his plan of actions by minimizing free energy (4).

The latter process is akin to the contraction of “thinking muscles” with each cognitive act, lasting a fraction of a second in mammals. At the same time, people are able to think on completely different time scales – hours, days and even years ($\sim 10^8$ of cognitive acts). Each of the corresponding mental states has its own time scale. That is, our thinking is (supposedly) hierarchical. We propose the same hierarchical design for AGI, since it has undeniable advantages.

Indeed, consider a hierarchical variational model $q_L(x, \mathbf{m})$ that uses L nested levels of mental states to minimize free energy:

$$\mathcal{F}_L = \int d\mathbf{m} dx q_L(x, \mathbf{m}) \ln \frac{q_L(x, \mathbf{m})}{p(x)}, \quad (5)$$

$$q_L(x, \mathbf{m}) \equiv q(x|\mathbf{m}_1)q(\mathbf{m}_1|\mathbf{m}_2) \dots q(\mathbf{m}_{L-1}|\mathbf{m}_L)q(\mathbf{m}_L).$$

One can rewrite (5) in terms of (conditional) entropies:

$$\mathcal{F}_L = \langle \mathcal{F}(\mathbf{m}_1) \rangle_{q(\mathbf{m}_1)} - \mathcal{H}[q(\mathbf{m}_1)] - \sum_{l=2}^L \langle \mathcal{H}[q(\mathbf{m}_l | \mathbf{m}_{l-1})] \rangle_{q(\mathbf{m}_{l-1})}.$$

Initial *non-informative priors* $q_0(\mathbf{m}_l)$ for all layers are independent of each other and have the maximum possible entropies. That is, the entropies of all mental states decrease in the learning process, as does the free energy:

$$\mathcal{F}_L \leq \mathcal{F}_L^0.$$

This leads to the following inequality:

$$\langle \mathcal{F}(\mathbf{m}_1) \rangle_{q(\mathbf{m}_1)} \leq D(q_0(x) || p(x)) - \sum_{l=1}^L \Delta \mathcal{H}_l \tag{6}$$

with one positive constant (initial KL-divergence) and L negative terms on the right side, as a result of learning:

$$\begin{aligned} \Delta \mathcal{H}_1 &\equiv \mathcal{H}[q_0(\mathbf{m}_1)] - \mathcal{H}[q(\mathbf{m}_1)] > 0, \\ \Delta \mathcal{H}_{l>1} &\equiv \mathcal{H}[q_0(\mathbf{m}_l)] - \langle \mathcal{H}[q(\mathbf{m}_l | \mathbf{m}_{l-1})] \rangle_{q(\mathbf{m}_{l-1})} > 0. \end{aligned}$$

Therefore, even with a small decreases in entropy on each layer, for sufficiently large L the average KL-divergence in the input space can reach its theoretical minimum: $\langle \mathcal{F}(\mathbf{m}_1) \rangle_{q(\mathbf{m}_1)} = \langle D(q(x | \mathbf{m}_1) || p(x)) \rangle_{q(\mathbf{m}_1)} \rightarrow 0$.

This is by no means a mathematical proof, as we do not consider convergence issues. Rather, it illustrates the following general principle in favor of hierarchical design: a stack of weak learners is a strong learner.

3 Deep Control Architecture

In particular, Friston et al. [11, 20] used the free energy approach with the hierarchical generative model with discrete *cognitive states* \mathbf{s}_l and a corresponding time scale $T_l = T^l$ at the l -th level. In their model, each state of a higher level generates a trajectory of states of length T one level below. This trajectory is defined by its initial state $p(\mathbf{s}_{l-1}^0 | \mathbf{s}_l^t)$ and the transition matrix for the sequence of T lower-level states, generated by a given higher-level state:

$$p(\mathbf{s}_{l-1}^{\tau+1} | \mathbf{s}_{l-1}^\tau, \mathbf{s}_l^t), \quad \tau \in [0 : T - 1]. \tag{7}$$

Although this model does a good job of explaining the basic structure of the brain [3, 19] and the nature of the psyche [10, 17, 28], it lacks the concise learning rules: how one can define the temporal depth T and transition matrices (7) from the data?

To answer these questions, we present a hierarchical neuro-symbolic architecture for reinforcement learning, represented by a stack of weak controllers that

form a strong controller with appropriate learning rules. Our approach differs in that at each level we operate not with individual states, but with their sequences $\mathbf{m}_l = \mathbf{s}_l^1 \dots \mathbf{s}_l^T$. We call them *mental states*, assuming that the agent’s brain can represent (short) plans and operate on them. Thus, expanding the set of states to the set of trajectories, we can restrict ourselves to only one step of the Markov matrix at each level: $p(\mathbf{m}'_l | \mathbf{m}_l)$.

The problem now is (i) to find from experience the set of such mental states \mathbf{m}_l and (ii) to match the subsets of this set with the corresponding cognitive states of a higher level $\{\mathbf{m}_{l-1}\} \leftrightarrow \mathbf{s}_l$.

We start with a single-level system and apply reinforcement learning, which can be derived from the free energy approach [7, 9]. Namely, the (negative) free energy of a mental state (4) is the average expected reward for that state:

$$R(\mathbf{m}) = -\mathcal{F}(\mathbf{m}).$$

The goal of an agent is to maximize the total reward for a given planning horizon T . The maximum total reward, starting from a given mental state and following the optimal strategy $\pi : \mathbf{m}^t \rightarrow \tilde{\mathbf{m}}^{t+1}$, is given by the *value* function:

$$Q(\mathbf{m}^t) = R(\mathbf{m}^t) + \max_{\pi} \sum_{\tau=1}^{T-1} \sum_{\mathbf{m}^{t+\tau}} R(\mathbf{m}^{t+\tau}) p_{\pi}(\mathbf{m}^{t+\tau} | \mathbf{m}^t). \tag{8}$$

In a Markov decision process (MDP) the optimal deterministic strategy π^* for such an agent is:

$$\tilde{\mathbf{m}}^{t+1} = \arg \max_{\mathbf{m}^{t+1}} Q(\mathbf{m}^{t+1}) p_{\pi^*}(\mathbf{m}^{t+1} | \mathbf{m}^t). \tag{9}$$

The predicted mental state $\tilde{\mathbf{m}}^{t+1}$ defines the next action \tilde{a}^{t+1} , but not the next perception \tilde{s}^{t+1} , since the agent does not fully control the environment. That is, the predicted state $\tilde{\mathbf{m}}^{t+1}$ naturally differs from the real one \mathbf{m}^{t+1} .

The agent can learn value function from experience making use of a well known Bellman equation following from (8):

$$Q(\mathbf{m}^t) = R(\mathbf{m}^t) + \gamma \sum_{\mathbf{m}^{t+1}} Q(\mathbf{m}^{t+1}) p_{\pi^*}(\mathbf{m}^{t+1} | \mathbf{m}^t), \tag{10}$$

with a discount parameter $\gamma = (T - 1)/T$. The solution can be found iteratively using *Q-learning* [30]:

$$Q(\mathbf{m}^t) \leftarrow Q(\mathbf{m}^t) + \alpha^t [R(\mathbf{m}^t) + \gamma Q(\mathbf{m}^{t+1}) - Q(\mathbf{m}^t)], \quad \mathbf{m}^{t+1} \sim p_{\pi^*}(\mathbf{m}^{t+1} | \mathbf{m}^t)$$

with (gradually reducing) learning rate $\alpha^t \ll 1$. The expected reward and transition probabilities are also updated iteratively based on real data (the number of transitions between mental states $C_{\mathbf{m}\mathbf{m}'}$ and the corresponding rewards).

Although such training is easy to implement, it is highly inefficient. The optimal Q-learning algorithm is proved to converge in $O(T^3)$ iterations [30].

That is, millions of iterations are required to learn how to reach the goal on the horizon of 100 time steps.

Therefore, we assume that the planning horizon at each level is fairly short for sufficiently fast Q-learning. After all, we only need a weak controller at each level. The resulting useful behaviors, mental states \mathbf{m} , are implicitly encoded in the transition matrix $p_{\pi^*}(\mathbf{m}'|\mathbf{m})$ learned from experience. To find them explicitly, we recursively expand the original set of cognitive states $\{\mathbf{s}\}$ along with the dimension of the transition matrix.

Namely, the transition probabilities are determined by the number of observations of successive mental states $C_{\mathbf{m}\mathbf{m}'}$:

$$p_{\pi^*}(\mathbf{m}'|\mathbf{m}) = C_{\mathbf{m}\mathbf{m}'} / \sum_{\mathbf{m}'} C_{\mathbf{m}\mathbf{m}'}. \quad (11)$$

Mental states can be defined recursively in the course of Q-learning by merging the most frequently occurring pairs of existing ones. That is, when $C_{\mathbf{m}'\mathbf{m}''}$ exceeds a certain threshold C_0 , a new mental state is defined as the concatenation of the corresponding pair:

$$\mathbf{m} \leftarrow \mathbf{m}'\mathbf{m}'', \quad C_{\mathbf{m}'\mathbf{m}''} > C_0. \quad (12)$$

These rules for the formation of mental states define a formal language – the set of all valid strings of cognitive states \mathbf{s} (*characters* of that language): $\{\mathbf{m}\} = \{\mathbf{s}^1 \dots \mathbf{s}^\tau\}$. We'll call this set *mental language*. The rules of mental language (12) allow the agent to model sequences and thereby perceive and predict temporal structures. Indeed, the optimal strategy (9) can now predict sequences of cognitive states. This allows the agent to plan its behavior several steps ahead even in the current MDP setting.

So far we have solved the first problem – learning from experience a *dictionary* of the mental language. Now we need to define a metalanguage, that is, to define higher-level symbols as subsets of lower-level mental states, considered as different ways of implementing the corresponding higher-level plan. This can be done using the semantics of the above mental language.

The semantic meaning of words is determined by the context in which they appear. In our case semantics is defined by the transition probabilities $p_{\pi^*}(\mathbf{m}'|\mathbf{m})$ and $p'_{\pi^*}(\mathbf{m}|\mathbf{m}')$, which are normalized rows and columns of the empirical matrix $C_{\mathbf{m}\mathbf{m}'}$. That is, each mental state has a corresponding *semantic vector* $\mathbf{x}_{\mathbf{m}} = p_{\pi^*}(\dots|\mathbf{m})p'_{\pi^*}(\dots|\mathbf{m})$ – concatenation of right and left context probabilities.

Behaviors with similar semantic vectors appear interchangeably in similar situations and can be thought of as different implementations of the same higher level plan step. That is, we can cluster the semantic vectors of all mental states, and map all members of the same cluster to the same symbol of the next hierarchical level $\{\mathbf{m}_l\} \rightarrow \mathbf{s}_{l+1}$. Thus, the history of mental states, encoded in the symbolic stream of the next level, forms a training set for learning the next hierarchical level according to the same algorithms described above.

Hierarchical reinforcement learning proceeds by accumulating training data at the current top layer to generate the next top layer that will control behavior

on even larger time scales. Such layered learning forms an ever-growing stack of controllers, co-creating a hierarchy of nested plans, each formulated in its own mental language.

Hierarchical control is carried out from top to bottom by building and executing nested plans. Namely, the upper layer chooses the top-level plan according to (9). This plan is decomposed using language grammar rules (12) into a sequence of top-level symbols $\tilde{\mathbf{m}}_L \rightarrow \tilde{\mathbf{s}}_L^1 \dots \tilde{\mathbf{s}}_L^T$, the steps of this plan. These symbols are then translated one-by-one to a lower level according to semantic mapping: $\tilde{\mathbf{s}}_L^t \rightarrow \{\tilde{\mathbf{m}}_{L-1}^t\}$, as a set of their possible realizations.

The lower level selects the best realization, which corresponds to the current development of events. Each step of the chosen realization, in turn, is similarly translated to the level below. Thus, a hierarchy of nested plans is maintained, which are constantly adjusted to account for the inevitable deviations from the original plan. We called this architecture Deep Control [24, 25].

4 Discussion

The proposed neuro-symbolic architecture has several advantages.

- Hierarchical planning avoids the combinatorial explosion that limits the planning horizon in current Deep RL models such as MuZero (where $T \approx 50 \div 60$) [22]. In Deep Control, the planning horizon grows exponentially as the number of layers increases.
- The complexity of learning depends linearly on the size of training dataset, in contrast to the quadratic complexity of learning in neural networks [12, 25].
- Symbolic operations require much less computation than multidimensional matrix operations in deep neural networks.
- Unlike current large language models like ChatGPT, Deep Control is based on reinforcement learning making it easy to learn conversational behavior based on $Q(\mathbf{m})$ along with the rules of the language $p(\mathbf{m}'|\mathbf{m})$.
- Last but not least, rule-based symbolic thinking and decision making is much easier to understand and debug than in the case of neural networks.

Early implementations of Deep Control proved to be computationally efficient in classical RL problems [21], learning trading strategies [15] and in natural language understanding [26]. According to preliminary experiments the new Deep Control model ADAM [23, 27] can learn a language at a rate of about 1 GB/h on a 100 GFlops CPU. Thus the 1.2 TB GPT-3 training set can be processed in approximately 50 days on a single CPU with sufficient memory. It corresponds to about $4 \cdot 10^{18}$ Flop, which is 5 orders of magnitude less than the $3 \cdot 10^{23}$ Flop needed to train GPT-3 [5]. These experiments will be presented elsewhere in a separate paper.

5 Related Work

The most promising results in the field of AGI have so far been obtained within Deep RL [22, 31]. Hierarchical RL, despite numerous attempts, has not yet achieved great success and is usually limited to two hierarchical levels [1, 2, 4, 14, 18, 29].

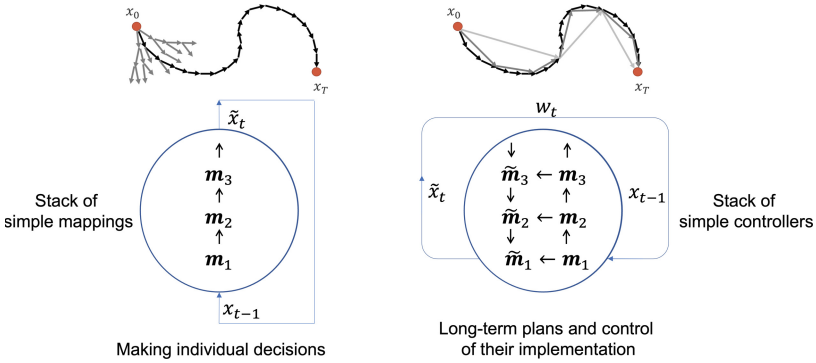


Fig. 1. Deep RL vs Deep Control. Deep RL (left) uses neural networks for individual decision making, augmented by Monte Carlo tree search. Deep Control (right) maintains a hierarchy of plans inscribed in each other.

The main problem is that the neural networks in Deep RL solve only part of the problem, approximating the value of the next action-states. Long term planning requires testing a large number of trajectories using, for example, a Monte Carlo tree search. On the contrary, Deep Control builds a hierarchy of nested plans within a single model, avoiding combinatorial explosions (see Fig. 1). In fact, deep neural networks can be effectively used as an analog I/O interface for Deep Control, leveraging the strengths of both analog and neuro-symbolic architectures, as shown in Fig. 2.

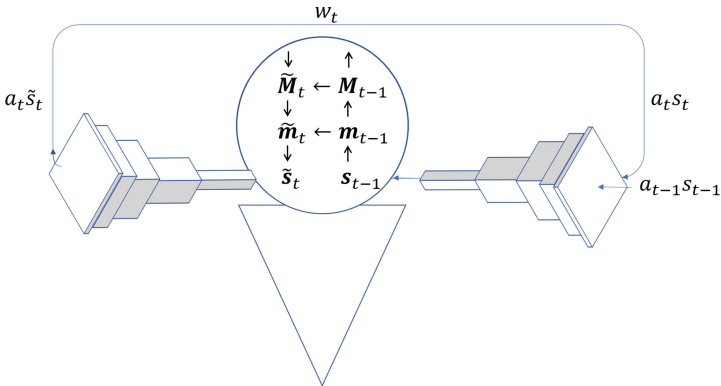


Fig. 2. Hierarchical symbolic thinking with deep neural networks as an analog interface with high-dimensional sensorimotor data.

6 Conclusion

In conclusion, we presented a neuro-symbolic cognitive architecture based on the principle of free energy, which solves the problem of hierarchical reinforcement learning. The proposed model builds a stack of controllers that maintain a hierarchy of nested plans, constantly adapting to the current situation. Each hierarchical level learns to implement the stages of plans of a higher level, using the semantics and grammar of its internal language, representing, respectively, the neuro- and symbolic part of the proposed architecture.

References

1. Bakker, B., Schmidhuber, J., et al.: Hierarchical reinforcement learning based on subgoal discovery and subpolicy specialization. In: Proceedings of the 8th Conference on Intelligent Autonomous Systems, pp. 438–445 (2004)
2. Barto, A.G., Mahadevan, S.: Recent advances in hierarchical reinforcement learning. *Discrete Event Dyn. Syst.* **13**(1–2), 41–77 (2003)
3. Bastos, A., et al.: Canonical microcircuits for predictive coding. *Neuron* **76**(4), 695–711 (2012)
4. Botvinick, M.M.: Hierarchical reinforcement learning and decision making. *Curr. Opin. Neurobiol.* **22**(6), 956–962 (2012)
5. Brown, T., et al.: Language models are few-shot learners. *Adv. Neural. Inf. Process. Syst.* **33**, 1877–1901 (2020)
6. Conant, R.C., Ashby, R.W.: Every good regulator of a system must be a model of that system. *Int. J. Syst. Sci.* **1**(2), 89–97 (1970)
7. Da Costa, L., Sajid, N., Parr, T., Friston, K., Smith, R.: The relationship between dynamic programming and active inference: the discrete, finite-horizon case. arXiv preprint [arXiv:2009.08111](https://arxiv.org/abs/2009.08111) (2020)
8. Friston, K.: The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**(2), 127–138 (2010)
9. Friston, K.: A free energy principle for a particular physics. arXiv preprint [arXiv:1906.10184](https://arxiv.org/abs/1906.10184) (2019)
10. Friston, K.J., Parr, T., Yufik, Y., Sajid, N., Price, C.J., Holmes, E.: Generative models, linguistic communication and active inference. *Neurosci. Biobehav. Rev.* **118**, 42–64 (2020)
11. Friston, K.J., Rosch, R., Parr, T., Price, C., Bowman, H.: Deep temporal models and active inference. *Neurosci. Biobehav. Rev.* **90**, 486–501 (2018)
12. Hoffmann, J., et al.: Training compute-optimal large language models. arXiv preprint [arXiv:2203.15556](https://arxiv.org/abs/2203.15556) (2022)
13. Klimontovich, Y.L.: Nonlinear Brownian motion. *Phys. Usp.* **37**(8), 737 (1994)
14. Levy, A., Platt, R., Saenko, K.: Hierarchical reinforcement learning with hindsight. arXiv preprint [arXiv:1805.08180](https://arxiv.org/abs/1805.08180) (2018)
15. Makarov, I., Fakhrutdinov, T., Kichik, M., Mamontov, K., Baskov, O., Shumsky, S.: Forecasting in financial markets using the ADAM architecture and reinforcement learning methods. In: 2021 International Conference Engineering and Telecommunication (En&T), pp. 1–7. IEEE (2021)
16. Marletto, C.: *The Science of Can and Can't: A Physicist's Journey Through the Land of Counterfactuals*. Penguin UK (2021)

17. Montague, P.R., Dolan, R.J., Friston, K.J., Dayan, P.: Computational psychiatry. *Trends Cogn. Sci.* **16**(1), 72–80 (2012)
18. Nachum, O., Gu, S.S., Lee, H., Levine, S.: Data-efficient hierarchical reinforcement learning. *Adv. Neural. Inf. Process. Syst.* **31**, 3307–3317 (2018)
19. Pezzulo, G., Parr, T., Friston, K.: The evolution of brain architectures for predictive coding and active inference. *Philos. Trans. R. Soc. B* **377**(1844), 20200531 (2022)
20. Pezzulo, G., Rigoli, F., Friston, K.J.: Hierarchical active inference: a theory of motivated control. *Trends Cogn. Sci.* **22**(4), 294–306 (2018)
21. Pivovarov, I., Shumsky, S.: Marti-4: new model of human brain, considering neo-cortex and basal ganglia – learns to play Atari game by reinforcement learning on a single CPU. In: *Artificial General Intelligence: 15th International Conference, AGI 2022, Proceedings, Seattle, WA, USA, 19–22 Aug 2022*, pp. 62–74. Springer (2023)
22. Schrittwieser, J., et al.: Mastering Atari, go, chess and shogi by planning with a learned model. *Nature* **588**(7839), 604–609 (2020)
23. Shumskii, S.: ADAM: a model of artificial psyche. *Autom. Remote. Control.* **83**(6), 847–856 (2022)
24. Shumsky, S.: Deep structure learning: new approach to reinforcement learning. In: *Lectures on Neuroinformatics. Proceedings of the XX All-Russian Scientific Conference Neuroinformatics-2018*, pp. 11–43 (2018) (in Russian)
25. Shumsky, S.: *Machine Intelligence. Essays on the Theory of Machine Learning and Artificial Intelligence.* RIOR (2019) (in Russian)
26. Shumsky, S.: Scalable natural language understanding: from scratch, on the fly. In: *2018 International Conference on Artificial Intelligence Applications and Innovations (IC-AIAI)*, pp. 73–74. IEEE (2018)
27. Shumsky, S., Baskov, O.: ADAM: a prototype of hierarchical neuro-symbolic AGI. In: *16th Annual AGI Conference AGI-23* (in press). Springer (2023)
28. Veissière, S.P., Constant, A., Ramstead, M.J., Friston, K.J., Kirmayer, L.J.: Thinking through other minds: a variational approach to cognition and culture. *Behav. Brain Sci.* **43** (2020)
29. Vezhnevets, A.S., et al.: Feudal networks for hierarchical reinforcement learning. In: *International Conference on Machine Learning*, pp. 3540–3549. PMLR (2017)
30. Wainwright, M.J.: Variance-reduced q -learning is minimax optimal. *arXiv preprint [arXiv:1906.04697](https://arxiv.org/abs/1906.04697)* (2019)
31. Wang, X., et al.: SCC: an efficient deep reinforcement learning agent mastering the game of StarCraft II. In: *International Conference on Machine Learning*, pp. 10905–10915. PMLR (2021)



The Future of International Climate Politics: An Agent-Based Approach

Anna Shuranova^{1,2} , Matvei Chistikov³ , Yuri Petrunin⁴ ,
Vadim Ushakov^{5,6,7} , and Denis Andreyuk^{5,8} 

- ¹ Laboratory for Economics of Climate Change, Faculty of World Economy and International Affairs, HSE University, Malaya Ordynka, 17, 119017 Moscow, Russia
ashuranova@hse.ru
- ² Center for Future Energy Systems “Energynet”, Marii Ulyanovoy ul., 12(114), 119331 Moscow, Russia
- ³ Centre for Comprehensive European and International Studies, Faculty of World Economy and International Affairs, HSE University, Malaya Ordynka, 17, 119017 Moscow, Russia
- ⁴ School of Public Administration, Lomonosov Moscow State University, Lomonosovskiy pr., 27(4), 119991 Moscow, Russia
- ⁵ Mental Health Clinic No. 1 named after N.A. Alexeev of Moscow Health Department, Zagorodnoe sh., 2, 117152 Moscow, Russia
- ⁶ National Research Nuclear University MEPhI, Kashirskoe sh., 31, 115409 Moscow, Russia
- ⁷ Institute for Advanced Brain Studies, Lomonosov Moscow State University, GSP-1, Leninskie Gory, 119991 Moscow, Russia
- ⁸ Faculty of Economics, Lomonosov Moscow State University, GSP-1, Leninskie Gory, 1-46, 119234 Moscow, Russia

Abstract. Climate change, being one of the most important problems that affect the modern civilization, has become a pressing issue on the world political agenda. However, the humanity cannot unite in a battle against it, with the states pursuing their own interests in how much they wish to contribute to the global mitigation efforts. Accordingly, international climate politics is undergoing increasing fragmentation, with groups of states having similar interests or characteristics unifying not only their positions in international negotiations, but also their approaches to mitigation and adaptation. We attempt to picture the future of international climate politics by using a simple model for group polarization, thus considering states as members of a large social group which are connected by bonds of different strength. The strength of these bonds increases if countries’ opinions on climate change are close and decreases otherwise. We are also considering two additional parameters: economic power and trade surplus. Economic power affects the speed with which states’ opinions change during their interactions, while trade surplus does not let the total sum of bonds’ strengths significantly increase or decrease. The results of our modelling suggest that after 1000 iterations 3 large groups are likely to be formed with high, intermediate and low levels of ambition to take action on climate change mitigation. Moreover, the arithmetic mean of countries’ opinions decreased from 1st to 1000th iterations.

Keywords: Climate change · International climate politics · Climate clubs · Agent-based modelling · Polarization

1 Introduction

Climate change is one of the most urgent global problems the world faces, encompassing all aspect of human life from economy and political ideologies to national and international security [1–4]. The anthropogenic emissions have already caused a warming of 1.1 °C compared to the preindustrial era, with the sea levels rising, extreme temperatures becoming commonplace, droughts and floodings threatening more territories and permafrost thawing, accompanied by the destruction of man-made infrastructure and additional emissions of methane [5], which has a 25 times higher global warming potential than carbon dioxide [6]. Changing ecosystem conditions, coupled with spreading natural disasters including storms, wildfires, and heatwaves, affect animal and plant species, their behaviour and habitats, in certain cases causing species loss. Increasing water stress, crop yields under threat and declining food security, as well as better environment for vector-borne diseases and deteriorating weather conditions rendering a growing number of lands uninhabitable all undermine human security [7].

With such a scope of threats and risks to the planet, climate change has become one of the most relevant issues on the world political agenda [8]. The international climate change regime established by the United Nations Framework Convention on Climate Change (UNFCCC) and shaped by Kyoto Protocol and Paris Agreement encompasses the majority of world's nations who declare it as their common goal to limit the global temperature growth “well below” 2 °C by the end of the century [9]. Meanwhile, climate change is affecting all the countries and regions differently: some, like Europe, start experiencing heatwaves [10], while for the others, such as small island developing states, the problem of flooding becomes more relevant [11], and in the Arctic region the warming rates are one of the fastest in the world – over 0.75 °C per decade [12]. Nevertheless, all these different effects provide equal incentives for the states to take active climate action. However, climate policy ambition depends on a whole number of other factors such as socioeconomic development, fossil fuel dependence, culture, political regime, etc. [13–22] Such factors define states' positions in the international climate politics and their willingness both to cooperate and to contribute to the global efforts to combat climate change.

Interstate differences determined by various factors account for the emergence of dividing lines in international climate politics, which do not always correspond with one another and with “traditional” political and economic groups and blocs on the world political arena. For one, the transition of many countries away from fossil fuels to renewable energy and hydrogen is creating a division between energy-importing and exporting nations, with the former using climate goals and investments in renewables as an additional instrument to ensure their energy security [23] and the latter widely considered as more likely to stall international climate negotiations and conduct less ambitious policies at home [13, 22, 25–28]. However, even renewable energy is not as evenly distributed on the planet as it is commonly believed [29], and wind and solar technical potential are becoming an increasingly important component in the state's economic power. Energy issues are closely related to the resource factor in terms of critical materials which are crucial for the energy transition and can slow it down in case of their scarcity [30]. Another close, but more universal division is likely to emerge between leaders in low- and zero-carbon technologies and the rest of the world: while

the race for the cheapest, most efficient and most high-end is still ongoing, it is already evident that the principal contesters are the US, China, Japan, and the European Union (EU), with the winner still unclear [31]. This can be framed in a broader context of socioeconomic development as a factor that shapes states' climate policy, with the richer having more assets available to direct at climate goals and the poorer having to cope with more urgent development problems with which the standard set of climate policy instruments is not necessarily compatible [32]. It is, in turn, an outline for another major dividing line along the various dimensions of inequality on all levels from interstate to individual [33], with primarily the former not only affecting the way states position themselves in international climate politics, but also being exacerbated by the green transition [34]. This is also connected with the problem of historical responsibility for global temperature rise, with the developed world initially causing the greatest part of anthropogenic emissions and the developing countries playing a major role in it over the past few decades [35, 36]. Adding all this to the fact that climate change is gradually acknowledged to be a security issue and can increase violent conflict potential [37, 38] as well as cause migration flows [39], we can get a broad outlook on the scope of divisions that prevent international climate governance from being effective.

The existence of such divisions is reflected in the number of groups of states in the international climate politics. Mostly these are formed as negotiation groups during the UNFCCC Conferences of the Parties (COP) [40]. They cannot be characterized by high stability, but many have endured since the first COPs and all are usually officially mentioned by the UNFCCC, with 14 groups listed in 2023. These include regional negotiation alliances such as Arab States or African Group; groupings by other criteria which are vital for their members' policies such as Small Island States or Rainforest Nations; developing countries mostly united by G77 and allied with China; Least Developed Countries; the EU; the Umbrella Group historically formed by non-European developed signatories of the Kyoto Protocol and bound by little in common; etc. [41] A newer form of interstate groupings that potentially involves deeper cooperation are climate clubs [42–45]. With their scope theoretically ranging from coordination of the members' approaches to climate policy formation to joint action with punitive measures upon non-compliance [46], in practice only one such grouping has been established to date [47]. Tendencies for climate policy alignment are also manifesting within other international organizations and cooperation formats such as ASEAN [48], BRICS [49], and EAEU [50]. Nevertheless, international climate politics largely remains an anarchic space, with few signs of strong alliances, a large number of possible lines of cooperation, and an even larger one – for division. With states primarily pursuing their national interests and accordingly choosing their paths to combat climate change, it is still unclear which of the factors that determine their climate policy will become the aligning ones, and which dividing lines will finally shape international climate political landscape.

In this study, we attempt to picture such landscape of the future by using a model for group polarization described in Chebotarev et al. [51]. It allows us to include both the states' position in international climate politics, their power and the existing bonds between them to obtain a model of how international climate politics might look while states progress on their way to carbon neutrality. Since our main purpose is to demonstrate the climate dimension of international politics, we do not apply different emissions

pathways or construct scenarios of increased global ambition; rather, we rely on current policies and extrapolate them, while taking into account economic factors that are crucial in shaping the world political and economic environment.

Due to new technological possibilities to accumulate large masses of empirical data, the last couple of decades saw significant progress in modelling group decision-making processes [52, 53]. A number of approaches has emerged that help describe how participants of a communication process arrive at consensus on certain subject of discussion and which conditions lead to divergence and polarization of opinions; some are interdisciplinary, comprising insights from social sciences, ethology, theory of evolution and even thermodynamics, the combination of which allows to perceive groups as information contours organised as neural networks [54]. In this case, the focus shifts on to bonds between group members since the bond strength is equivalent to the probability of information exchange, while the “learning” of sociomorphic neural networks progresses as these strengths are changing [55]. Here, we attempt to apply this neuroevolutionary approach to model political processes, in particular, in the area of international climate politics.

2 Methodology and Data

This study’s methodology builds upon the model of opinion polarization in social groups introduced in Chebotarev et al. [51]. We regard the international community as a social group and climate change mitigation as an issue-area which is the subject of polarization. The model is based on interstate interactions during which they exchange opinions about climate change mitigation. The rules of this interaction can be related as follows:

1. Each country is initially equipped with a set of four parameters: a) Its opinion on climate change mitigation (Op); b) The strengths of its bonds with other countries based upon their trade volumes (B); c) Its economic power represented by GDP (P); d) Its trade surplus which is equal for each state at the beginning (S).
2. Data for the states’ opinion on climate change mitigation was received from Environmental Performance Index (EPI). We intentionally used only the Climate Change variable from the EPI to reflect the states’ progress on mitigation policies. Data for the trade volumes was received from BACI database and calculated on the basis of bilateral trade flows [56]. As for the GDP, the data was obtained from the World Bank database. All data was normalized to be represented within the range from 0 to 1. It should be noted that we used data for 2021 which was the latest available data at the time of writing.
3. Interaction between states occurs in what we call iterations. During each tact each country ‘attempts’ to interact with each of the others. Whether the interaction will actually take place or not is determined probabilistically, depending on the value of parameter B - the strength of a bond between two given states.
4. During each iteration all pairs ‘exchange’ their opinions about climate change mitigation. This process leads to changes of opinion, bond strength and trade surplus of each state, while their economic power remains constant.

5. The states’ opinions tend to come closer as a result of an interaction. The exact change that will occur in each of their opinions depends on the power difference in each pair: more powerful countries change their opinion slower, while weaker countries do it faster – and the greater the inequality, the more each of the extremes will tend to zero or full opinion change, respectively. The value of new opinions (Op^*) assigned to the states after the interaction is determined by the following formula:

$$Op_1^* = \begin{cases} Op_1 - \left| \frac{1-|Op_1-Op_2|}{2} * (Op_1 - Op_2) * 0, 1 \right| - \left(1 - \frac{P_2}{P_1} \right), & \text{if } P_1 > P_2 \\ Op_1 - \left| \frac{1-|Op_1-Op_2|}{2} * (Op_1 - Op_2) * 0, 1 \right| + \left(1 - \frac{P_1}{P_2} \right), & \text{if } P_2 < P_1 \\ Op_1 - \left| \frac{1-|Op_1-Op_2|}{2} * (Op_1 - Op_2) * 0, 1 \right|, & \text{if } P_1 = P_2 \end{cases} \quad (1)$$

$$Op_2^* = \begin{cases} Op_2 + \left| \frac{1-|Op_1-Op_2|}{2} * (Op_1 - Op_2) * 0, 1 \right| + \left(1 - \frac{P_2}{P_1} \right), & \text{if } P_1 > P_2 \\ Op_2 + \left| \frac{1-|Op_1-Op_2|}{2} * (Op_1 - Op_2) * 0, 1 \right| - \left(1 - \frac{P_1}{P_2} \right), & \text{if } P_2 < P_1 \\ Op_2 + \left| \frac{1-|Op_1-Op_2|}{2} * (Op_1 - Op_2) * 0, 1 \right|, & \text{if } P_1 = P_2 \end{cases} \quad (2)$$

where Op represents the opinion on climate change mitigation and P – the economic power.

1. During an interaction the strength of bonds increases if the opinions of the countries are closer than 0.15 before it and decreases otherwise. The bond strength is also related to the trade surplus: in case it is low, the interstate bonds are changing slower. The surplus itself, in turn, grows if the bonds are decreasing and reduces if the bonds are increasing. In other words, due to this the overall sum of bonds values in the model should stay almost the same during all iterations.
2. After a specified number of steps (1000 in our case), the results are visualized in two forms: a graph representing the strengths of bonds and a histogram of opinion distribution. The graphs are visualized using Gephi software.

We have made the algorithm for the model open for access via the following link on GitHub: <https://github.com/MatthiasRex1/Future-of-climate-politics>.

3 Results and Conclusion

Figure 1 demonstrates histograms and graphs with country labels for steps 1, 500 and 1000. Table 1, in turn, presents descriptive statistics for these respective iterations.

At around 500th iteration the opinions arrive at an equilibrium and remain practically unchanged after another 500 steps. However, there are still observable changes in the bond strength. Nevertheless, the model comes to a clear equilibrium in terms of both bonds and opinions by the 1000th iteration.

The graphs with country labels for the 1000th step are presented on Figs. 2 and 3 respectively. Three clusters marked C1, C2 and C3 have been formed, with strong bonds inside each cluster and weak bonds with out-groups.

Table 2 summarises descriptive statistics for each cluster.

Before addressing the results of the model, it should be noted that authors have used the simplest variation of the agent-based models discussed in contemporary scientific

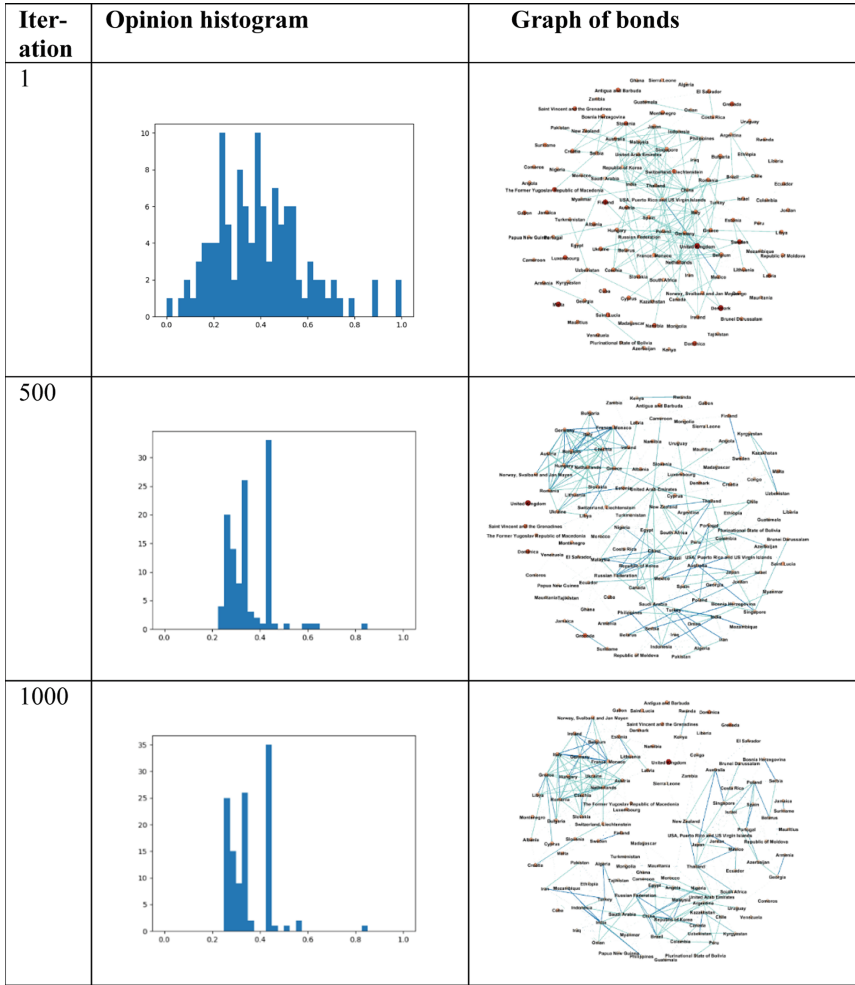


Fig. 1. Strength of bonds and opinion distribution for Iterations 1, 500 and 1000

Table 1. Descriptive statistics for Iterations 1, 500 and 1000

Iteration number	Arithmetic mean	Median
1	0.397939285	0.388440053
500	0.360047584	0.341235471
1000	0.356892944	0.333395208

literature. Therefore, the results are mostly hypothetical and serve merely as a demonstration of how agent-based modelling can be applied to studying international climate

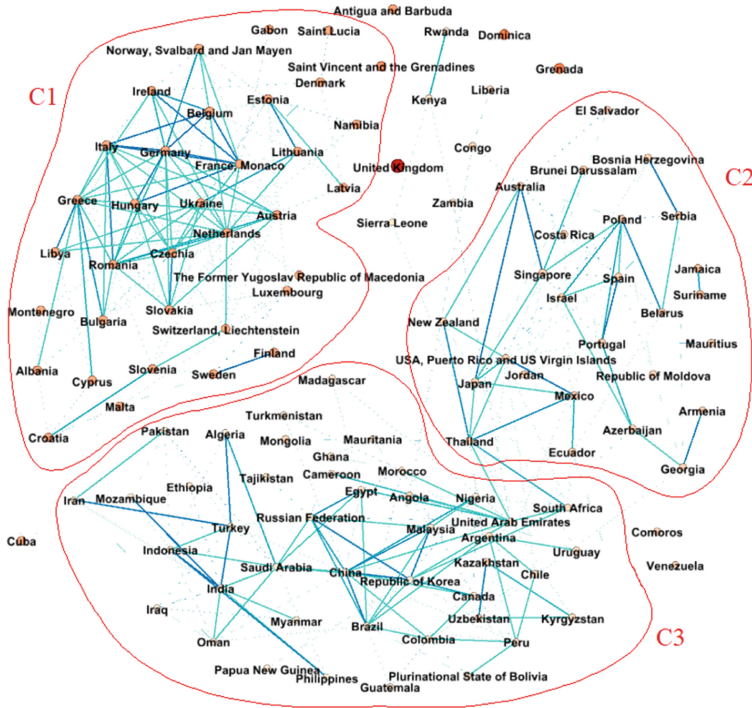


Fig. 2. Strength of bonds for Iteration 1000

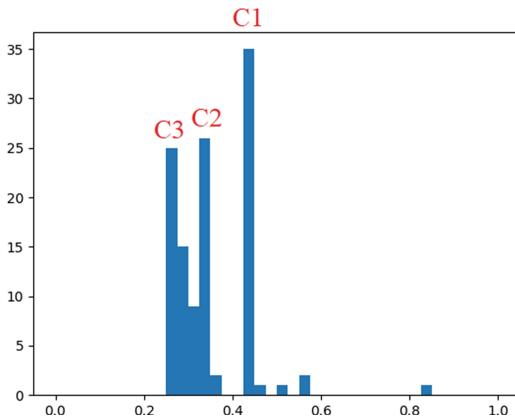


Fig. 3. Opinion distribution for Iteration 1000

politics. However, even this simplistic version of agent-based modelling allowed us to get the results which are highly correspond with reality of international climate politics.

The results of the modelling demonstrate that, if the states' opinion on climate policy, their economic bonds and power are taken into account, three clusters are likely

Table 2. Descriptive statistics for clusters formed at 1000th iteration

Cluster	Arithmetic mean	Median
C1	0.447103	0.447079
C2	0.334959	0.336737
C3	0.274351	0.280347

to emerge on the international climate arena. It is worth noting that the opinion, based on the Environmental Performance Index, is largely constructed out of emissions projections for 2050 and emission growth rates, which constitute the majority of the Climate Change part of the EPI. These, in turn, indirectly reflect the ambition of states' climate policies and their effectiveness since emissions dynamics is largely influenced by mitigation measures undertaken by states. Bonds and power, in turn, were purposefully chosen to be reflected through economic indicators since economy is inseparable from questions surrounding climate change – it is essentially human economic activity that causes anthropogenic emissions, and it is its modification that is needed to stop them. The most effective climate policy instruments are widely considered to be economic [57–59], and it is economy that lies at the core of most major dividing lines in international climate politics – between resource-abundant and resource-poor states, developed and developing countries, etc.

The clusters that are to form in international climate policy, according to the model, vary by climate policy ambition (the EPI average score on Climate Change) and in many ways coincide with both current geopolitical landscape and the emerging dividing lines outlined above.

The most ambitious cluster comprises many member states of the European Union which has one of the most far-fetching climate policies in the world and openly aims to become a global leader in combating climate change [60], as well as some states with aligned policy and comparable climate targets such as Norway. This cluster also includes several small island states whose very survival depends on the climate action they and the rest of the world take; and a few African countries whose position can be explained not by climate ambition *per se*, but by economic issues that caused emissions dynamics compatible with relatively high EPI scores.

The second cluster is perhaps the most diverse both by regional representation and economic powers of its members. It includes the “laggard” EU member states such as Poland notorious for its coal-based power sector [61]; the US which, due to its political system organization, cannot force carbon pricing and other restrictive regulatory measures upon the states and has to restrict itself to incentivizing programs [62]; as well as Australia, New Zealand and Japan which are economically linked with Western developed world but still have to catch up with it on climate. It also notably contains 5 of the 6 members of the Eastern Partnership of the EU, with the sixth (Ukraine) promoted to the first cluster: green transition is an important component of this initiative, with the EU providing funds and opportunities to enable it in exchange for political cooperation, including the promotion of democracy within the participants' political systems [63]. Some Balkan states which also fall under the EU's influence also emerged in this cluster.

Finally, many of its other members are Latin American states, which leads to a possible explanation of their position by strong economic ties with the US, which is in the same cluster.

The third cluster, although even more diverse and far more numerous, can be clearly divided into several subgroups which are all united by lesser motivation to implement ambitious climate policies than the first two clusters. The first group includes the Arab states specializing in fossil fuels export, which are joined by Indonesia, South Africa and other exporters, among which is, surprisingly, also Canada. A number of non-exporter Islamic states can also be attributed here. Russia belongs to this group too, at the same time forming its own – of the former Soviet Union republics now united by the Commonwealth of Independent Nations. It is also a part of the BRICS group which is whole in this cluster and whose presence signifies that one of the most powerful non-Western blocs has every chance to become the next “climate club” with its own agenda. The BRICS, in turn, can become the voice of the rest of the cluster – mostly developing countries where climate change mitigation cannot be applied blindly, without taking care of socio-economic and, in some cases, political challenges; within this category, a notable subgroup of Latin American states less dependent on the US can be distinguished.

Finally, some states simply do not fit into any category. Some, like the United Kingdom, appear to be standing out in their ambition, while others are more likely to be put outside of clusters due to statistical error.

It thus becomes evident that the dividing lines observable empirically in international climate policy are partly confirmed by the model. On the one hand, there emerge developed and mostly energy-poor states with high climate policy ambition and close economic and political ties which are bound to be drawn closely even further by the climate agenda as a means of advancing, among other objectives, their energy security interests. On the other hand, there are developing and energy-abundant states falling into the same cluster, despite political differences, great diversity and often – the absence of strong economic relations. This opens up the possibility to conclude that climate agenda and, in particular, promoting the position that climate goals should not hinder socio-economic development and be the reason for new barriers in the world, might serve as a unifying factor for this heterogeneous group. While disputing many other issues which, among some of them, even lead to violent conflicts, these states can align their positions in the face of what they see as a global challenge to their growth. With the climate politics further progressing and biting into international trade as does the European carbon border adjustment mechanism, it will not be too courageous to suppose that the third cluster can come closer politically – in a joint effort against both the “green menace” [64] and the Western-oriented world order it represents.

To conclude, it should be noted that the results of this study need to be interpreted carefully since international climate politics is part of a larger political landscape and can be affected by other factors than those directly related to it. An illustrative example here would be the authors’ earlier work that analysed the policies of the European Union on the Nord Stream 2 pipeline and the ideological and economic factors of decision-making, demonstrating that in most scenarios, the balance of ‘opinions’ should shift in favour of the project’ continuation and successful completion [65]; however, less than a year later a number of strong unpredictable interfering factors led to its destruction

as a result of a terrorist attack. With international climate politics, such factors are also bound to have an influence, as the 2022 energy crisis has already illustrated. Taking these limitations into account, it is nevertheless possible to conclude that the approach presented in this study can be applied to analyse long-term projections of state positions and possible alliances in international politics as well as construct scenarios regarding the most significant challenges and threats the humanity faces.

Acknowledgements. The research output by V. Ushakov was funded by the Russian Science Foundation (project No. 22-11-00213). The research of A. Shuranova and M. Chistikov is an output of a research project implemented as part of the Basic Research Program at the HSE University. Support from the Research Program of the Faculty of World Economy and International Affairs at HSE University is gratefully acknowledged.

References

1. Dzyurdzya, O.A., Burlakov, V.V., Fedotova, G.V., Skubriy, E.V., Orlova, L.N.: Climate change as a global threat to the world economy. In: Zavyalova, E.B., Popkova, E.G. (eds.) *Industry 4.0*, pp. 13–21. Palgrave Macmillan, Cham (2021). https://doi.org/10.1007/978-3-030-75405-1_2
2. Safonov, G.: Social Consequences of Climate Change: Building Climate Friendly and Resilient Communities via Transition from Planned to Market Economies. Study, Friedrich-Ebert Stiftung. <https://library.fes.de/pdf-files/id-moe/15863.pdf> (2019)
3. Chan, E.Y., Faria, A.A.: Political ideology and climate change-mitigating behaviors: insights from fixed world beliefs. *Glob. Environ. Chang.* **72**, 102440 (2022). <https://doi.org/10.1016/j.gloenvcha.2021.102440>
4. Phillis, Y.A., Chairetis, N., Grigoroudis, E., Kanellos, F.D., Kouikoglou, V.S.: Climate security assessment of countries. *Clim. Change* **148**, 25–43 (2018). <https://doi.org/10.1007/s10584-018-2196-0>
5. IPCC: Summary for policymakers. In: Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S.L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M.L., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J.B.R., Maycock, T.K., Waterfield, T., Yelekçi, O., Yu, R., Zhou, B. (eds.) *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom, New York, NY, USA, pp. 3–32 (2021). <https://doi.org/10.1017/9781009157896.001>
6. Eurostat: Glossary: Carbon Dioxide Equivalent (n.d.). https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Carbon_dioxide_equivalent
7. IPCC: Summary for policymakers In: Pörtner, H.-O., Roberts, D.C., Tignor, M., Poloczanska, E.S., Mintenbeck, K., Alegría, A., Craig, M., Langsdorf, S., Löschke, S., Möller, V., Okem, A., Rama, B. (eds.) *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom, New York, NY, USA, pp. 3–33 (2022). <https://doi.org/10.1017/9781009325844.001>
8. Keohane, R.O.: The global politics of climate change: challenge for political science. *PS Polit. Sci. Polit.* **48**(1), 19–26 (2015). <https://doi.org/10.1017/S1049096514001541>
9. United Nations: Paris Agreement. https://unfccc.int/sites/default/files/english_paris_agreement.pdf (2015)

10. Lhotka, O., Kysely, J.: The 2021 European heat wave in the context of past major heat waves. *Earth Space Sci.* **9**(11), e2022EA002567 (2022). <https://doi.org/10.1029/2022EA002567>
11. Martyr-Koller, R., Thomas, A., Schleussner, C.-F., Nauels, A., Lissner, T.: Loss and damage implications of sea-level rise on small Island developing states. *Curr. Opin. Environ. Sustain.* **50**, 245–259 (2021). <https://doi.org/10.1016/j.cosust.2021.05.001>
12. Rantanen, M., et al.: The arctic has warmed nearly four times faster than the globe since 1979. *Commun. Earth Environ.* **3**(168), 1 (2022). <https://doi.org/10.1038/s43247-022-00498-3>
13. Tørstad, V., Sælen, H., Bøyum, L.S.: The domestic politics of international climate commitments: which factors explain cross-country variation in NDC ambition? *Environ. Res. Lett.* **15**, 024021 (2020). <https://doi.org/10.1088/1748-9326/ab63e0>
14. Fredriksson, P.G., Neumayer, E.: Democracy and climate change policies: Is history important? *Ecol. Econ.* **95**, 11–19 (2013). <https://doi.org/10.1016/j.ecolecon.2013.08.002>
15. Obydenkova, A.V., Salahodjaev, R.: Climate change policies: the role of democracy and social cognitive capital. *Environ. Res.* **157**, 182–189 (2017). <https://doi.org/10.1016/j.envres.2017.05.009>
16. Buys, P., Deichmann, U., Meisner, C., That, T.T., Wheeler, D.: Country stakes in climate change negotiations: two dimensions of vulnerability. *Clim. Policy* **9**(3), 288–305 (2009). <https://doi.org/10.3763/cpol.2007.0466>
17. Piggot, G.: The influence of social movements on policies that constrain fossil fuel supply. *Clim. Policy* **18**(7), 942–954 (2017). <https://doi.org/10.1080/14693062.2017.1394255>
18. Spilker, G.: Helpful organizations: membership in inter-governmental organizations and environmental quality in developing countries. *Br. J. Polit. Sci.* **42**(2), 345–370 (2012). <https://doi.org/10.1017/S0007123411000329>
19. Tobin, P., Schmidt, N.M., Tosun, J., Burns, C.: Mapping states' paris climate pledges: analysing targets and groups at COP 21. *Glob. Environ. Chang.* **48**, 11–21 (2018). <https://doi.org/10.1016/j.gloenvcha.2017.11.002>
20. Bailer, S., Weiler, F.: A political economy of positions in climate change negotiations: economic, structural, domestic, and strategic explanations. *Rev. Int. Organ.* **10**, 43–66 (2015). <https://doi.org/10.1007/s11558-014-9198-0>
21. Lachapelle, E., Paterson, M.: Drivers of national climate policy. *Clim. Policy* **13**(5), 547–571 (2013). <https://doi.org/10.1080/14693062.2013.811333>
22. Ide, T.: Recession and fossil fuel dependence undermine climate policy commitments. *Environ. Res. Commun.* **2**(10), 101002 (2020). <https://doi.org/10.1088/2515-7620/abbb27>
23. Chu, L.K.: The role of energy security and economic complexity in renewable energy development: evidence from G7 countries. *Environ. Sci. Pollut. Res.* **30**(19), 1–21 (2023). <https://doi.org/10.1007/s11356-023-26208-w>
24. Cergibozan, R.: Renewable energy sources as a solution for energy security risk: empirical evidence from OECD countries. *Renew. Energy* **183**(15), 617–626 (2022). <https://doi.org/10.1016/j.renene.2021.11.056>
25. Johnsson, F., Kjærstad, J., Rootzén, J.: The threat to climate change mitigation posed by the abundance of fossil fuels. *Clim. Policy* **19**(2), 258–274 (2019). <https://doi.org/10.1080/14693062.2018.1483885>
26. Peszko, G., van der Mensbrugge, D., Golub, A., Ward, J., Marijs, C., Schopp, A. et al.: Diversification and Cooperation in a Decarbonizing World: Climate Strategies for Fossil Fuel-Dependent Countries. World Bank Publications (2020)
27. Policy Research Working Paper 9315. <https://openknowledge.worldbank.org/bitstream/handle/10986/34056/Diversification-and-Cooperation-Strategies-in-a-Decarbonizing-World.pdf?sequence=4&isAllowed=y>
28. Eisenack, K., Hagen, A., Mendelevitch, R., Vogt, A.: Politics, profits and climate policies: how much is at stake for fossil fuel producers? *Energy Res. Soc. Sci.* **77**(2), 102092 (2021). <https://doi.org/10.1016/j.erss.2021.102092>

29. Overland, I., Juraev, J., Vakulchuk, R.: Are renewable energy sources more evenly distributed than fossil fuels? *Renew. Energy* **200**, 379–386 (2022). <https://doi.org/10.1016/j.renene.2022.09.046>
30. Pommeret, A., Ricci, F., Schubert, K.: Critical raw materials for the energy transition. *Eur. Econ. Rev.* **141**(7645), 103991 (2022). <https://doi.org/10.1016/j.euroecorev.2021.103991>
31. Geopolitics of the Energy Transformation: The Hydrogen Factor. https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2022/Jan/IRENA_Geopolitics_Hydrogen_2022.pdf
32. Tanner, T., Horn-Phathanothai, L.: *Climate Change and Development*. Routledge, New York (2014)
33. Faus Onbargi, A.: The climate change–inequality nexus: towards environmental and socio-ecological inequalities with a focus on human capabilities. *J. Integr. Environ. Sci.* **19**(1), 163–170 (2022). <https://doi.org/10.1080/1943815X.2022.2131828>
34. Taconet, N., Méjean, A., Guivarch, C.: Influence of climate change impacts and mitigation costs on inequality between countries. *Clim. Change* **160**(11), 15–34 (2020). <https://doi.org/10.1007/s10584-019-02637-w>
35. Hickel, J.: Quantifying national responsibility for climate breakdown: an equality-based attribution approach for carbon dioxide emissions in excess of the planetary boundary. *Lancet Planet. Health* **4**(9), e399–e404 (2020). [https://doi.org/10.1016/S2542-5196\(20\)30196-0](https://doi.org/10.1016/S2542-5196(20)30196-0)
36. Total Greenhouse Gas Emissions (kt of CO₂ Equivalent). <https://data.worldbank.org/indicator/EN.ATM.GHGT.KT.CE>
37. Buhaug, H., Benjaminsen, T.A., Gilmore, E.A., Hendrix, C.S.: Climate-driven risks to peace over the 21st century. *Clim. Risk Manag.* **39**, 100471 (2023). <https://doi.org/10.1016/j.crm.2022.100471>
38. Mach, K.J., Kraan, C.M., Adger, W.N., Buhaug, H., Burke, M., et al.: Climate as a risk factor for armed conflict. *Nature* **571**(7764), 193–197 (2019). <https://doi.org/10.1038/s41586-019-1300-6>
39. Watson, T., Lenton, T., de Campos, R.S.: The climate change, conflict and migration nexus: a holistic view. *Clim. Resil. Sustain.* **2**(2), e250 (2023). <https://doi.org/10.1002/cli2.50>
40. Castro, P.: National interests and coalition positions on climate change: a text-based analysis. *Int. Polit. Sci. Rev.* **42**(1), 95–113 (2021). <https://doi.org/10.1177/0192512120953530>
41. UNFCCC Negotiating Groups Chairs and Coordinators. <https://unfccc.int/sites/default/files/resource/UNFCCC%20Group%20Chairs%20and%20Coordinators.pdf>
42. Tarr, D.G., Kuznetsov, D.E., Overland, I., Vakulchuk, R.: Why carbon border adjustment mechanisms will not save the planet but a climate club and subsidies for transformative green technologies may. *Energy Econom.* **122**(1), 106695 (2023). <https://doi.org/10.1016/j.eneco.2023.106695>
43. Tagliapietra, S., Wolff, G.B.: Conditions are ideal for a new climate club. *Energy Policy* **158**, 112527 (2021). <https://doi.org/10.1016/j.enpol.2021.112527>
44. Overland, I., Huda, M.S.: Climate clubs and carbon border adjustments: a review. *Environ. Res. Lett.* **17**(9), 093005 (2022). <https://doi.org/10.1088/1748-9326/ac8da8>
45. Nordhaus, W.: Dynamic climate clubs: on the effectiveness of incentives in global climate agreements. *Proc. Natl. Acad. Sci.* **118**(45), e2109988118 (2021). <https://doi.org/10.1073/pnas.2109988118>
46. Falkner, R., Nasiritousi, N., Reischl, G.: Climate clubs: politically feasible and desirable? *Clim. Policy* **22**(4), 480–487 (2022). <https://doi.org/10.1080/14693062.2021.1967717>
47. G7 Statement on Climate Club. <https://www.g7germany.de/resource/blob/974430/2057926/2a7cd9f10213a481924492942dd660a1/2022-06-28-g7-climate-club-data.pdf>
48. ASEAN Joint Statement on Climate Change to the 26th Session of the Conference of the Parties to the United Nations Framework Convention on Climate Change (UNFCCC COP26). <https://asean.org/wp-content/uploads/2021/10/10.-ASEAN-Joint-Statement-to-COP26.pdf>

49. Joint Statement Issued at the BRICS High-Level Meeting on Climate Change. http://brics2022.mfa.gov.cn/eng/hywj/ODMM/202205/t20220529_10694182.html
50. Climate Agenda. <https://eec.eaeunion.org/en/comission/departement/dotp/klimaticheskaya-povestka/>
51. Chebotarev, V., Andreyuk, D., Elizarova, A., Ushakov, V.: Polarization of opinions in the group: a modeling algorithm considering the dynamics of social bonds. *Proc. Comput. Sci.* **213**, 596–601 (2022). <https://doi.org/10.1016/j.procs.2022.11.108>
52. Liu, Z.: New paradigm of computational sociology. In: *The Power of Ideas: A History of Technological Thoughts on Digital Economics*, pp. 55–73. Springer Nature, Singapore (2022). https://doi.org/10.1007/978-981-19-4574-8_4
53. Macy, M.W., Willer, R.: From factors to actors: Computational sociology and agent-based modeling. *Ann. Rev. Sociol.* **28**(1), 143–166 (2002). <https://doi.org/10.1146/annurev.soc.28.110601.141117>
54. Andreyuk, D.S.: Neuroevolutionary paradigm as a model for the formation of large nanotechnological projects. *Int. J. Nanotechnol.* **16**(1–3), 12–14 (2019). <https://doi.org/10.1504/IJNT.2019.102387>
55. Andreyuk, D.S., Shuranova, A.A.: Modelling social bonds dynamics in groups: an approach to optimise interdisciplinary science projects and to analyse long-term social evolution. *Int. J. Nanotechnol.* **18**(9–10), 915–925 (2021). <https://doi.org/10.1504/IJNT.2021.118165>
56. BACI. http://www.cepii.fr/CEPII/en/bdd_modele/bdd_modele_item.asp?id=37
57. Goers, S.R., Wagner, A.F., Wegmayr, J.: New and old market-based instruments for climate change policy. *Environ. Econ. Policy Stud.* **12**(1), 1–30 (2010). <https://doi.org/10.1007/s10018-010-0161-x>
58. Bailey, I.: Market environmentalism, new environmental policy instruments, and climate policy in the United Kingdom and Germany. *Ann. Assoc. Am. Geogr.* **97**(3), 530–550 (2007). <https://doi.org/10.1111/j.1467-8306.2007.00562.x>
59. Coria, J., Köhlin, G., Xu, J.: On the use of market-based instruments to reduce air pollution in Asia. *Sustainability* **11**(18), 4895 (2019). <https://doi.org/10.3390/su11184895>
60. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions ‘Fit for 55’: Delivering the EU’s 2030 Climate Target on the Way to Climate Neutrality. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021DC0550>
61. Poland 2022. Energy Policy Review. <https://www.iea.org/reports/poland-2022>
62. U.S. Climate Change Policy. <https://crsreports.congress.gov/product/pdf/R/R46947>
63. Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the regions. https://www.eeas.europa.eu/sites/default/files/1_en_act_part1_v6.pdf
64. Smeets, N.: The Green Menace: unraveling Russia’s elite discourse on enabling and constraining factors of renewable energy policies. *Energy Res. Soc. Sci.* **40**, 244–256 (2018). <https://doi.org/10.1016/j.erss.2018.02.016>
65. Andreyuk, D.S., Yu, P.Y., Shuranova, A.A.: New approach to analyse ideological and economic factors in the politics of the European Union. *Riv. Stud. Polit. Int.* **88**(3), 415–430 (2021)



Memory Based Reinforcement Learning with Transformers for Long Horizon Timescales

Shweta Singh¹, Sudaman Katti², and Vedant Ghatnekar³(✉)

¹ IIIT, Hyderabad, India

² VIT, Pune, India

³ MIT WPU, Pune, India

1032190997@mitwpu.edu.in

Abstract. The most well-known sequence models make use of complex recurrent neural networks in an encoder-decoder configuration. The model used in this research makes use of a transformer, which is based purely on self-attention mechanism, without relying on recurrence at all. More specifically, encoders and decoders which make use self-attention and operate based on a memory are used. In this research work, results for various 3D visual and non-visual reinforcement learning tasks designed in Unity software were obtained. Convolutional neural networks, more specifically, nature CNN architecture is used for input processing in visual tasks and comparison with standard long short-term memory (LSTM) architecture is performed for both visual tasks based on CNNs and non-visual tasks based on coordinate inputs. This research work combines the transformer architecture with the proximal policy optimization technique used popularly in reinforcement learning for stability and better policy updates while training, especially for continuous action spaces, which are used in this research work. Certain tasks in this paper are long horizon tasks which carry on for a longer duration and require extensive use of memory-based functionalities like storage of experiences and choosing of appropriate actions based on recall. The transformer, which makes use of memory and self-attention mechanism in an encoder-decoder configuration proved to have better performance when compared to LSTM in terms of exploration and rewards achieved. Such memory-based architectures can be used extensively in the field of cognitive robotics and reinforcement learning.

Keywords: Transformers · Models of learning and memory · Reinforcement learning

1 Introduction

In various sequence-to-sequence problems such as the neural machine translation, the popular approaches were based on the use of RNNs in an encoder-decoder fashion. However, these architectures have a great limitation when working with long sequences often seen in tasks such as robotic navigation, their ability to retain information from the first elements gets lost when new elements are incorporated into the sequence. In the encoder, the hidden state in every step is associated with a certain element in the input

sequence, usually based on how recent it is. Therefore, if the decoder only accesses the last hidden state of the decoder, it will lose the important information about the first elements of the sequence. Thus, to deal with this problem, a novel concept was introduced: the attention mechanism. Instead of paying attention to the last state of the encoder as in the case of RNNs, in each step of the decoder, all the states of the encoder which are able to access information about all the elements of the input sequence are considered. This is the main working principle of the attention mechanism. This mechanism allows the decoder to assign greater weight or importance to a specific element of the input for each element of the output.

2 Methodology

This research work makes use of the ML agents toolkit in unity for implementation of reinforcement learning algorithms through python. The environments for all the tasks were created in Unity. A model of a simple Biped robot which can transverse across planes was used as the agent for all the tasks. Continuous action spaces were used. Nature CNN architecture was used for input preprocessing in case of visual tasks.

2.1 Proximal Policy Optimization

Proximal Policy Optimization (PPO) is a fairly recent advancement in the field of Reinforcement Learning, which provides an upgrade on Trust Region Policy Optimization (TRPO). PPO aims to strike a balance between key factors like ease of implementation and tuning, sample complexity, sample efficiency and trying to compute updates at each step that minimizes the cost function while making sure that the deviations from the previous policy are relatively small. This paper makes use of a custom policy which combines PPO with transformer and memory-based attention. This policy is implemented through the stable baselines python package. PPO was chosen over other methods like DQN due to its efficacy in case of the continuous action spaces used here for various experiments.

2.2 Episodic Memory Transformer

The episodic memory buffer consists of all past observations per time step in an embedded form encoded by the attention-based encoder. The decoder of the attention-based policy network makes use of the episodic memory and current observation to compute a distribution over actions using PPO. The memory grows linearly with the episode length but as each observation is embedded into a low dimensional vector, hundreds of time steps can easily be stored in the hardware memory. While RNNs are restricted to a fixed-size state vector, which usually can only capture short-term dependencies.

2.3 Attention-Based Policy Network

The policy network $\pi(a|o, EM)$ makes use of the embedded current observation and the encoded memory to compute a distribution over the action space. The Transformer

architecture consists of an encoder and decoder which use the attention block (Fig. 1).

$$\text{Att}(U, K, V) = \text{softmax}(UK^T)V \quad (1)$$

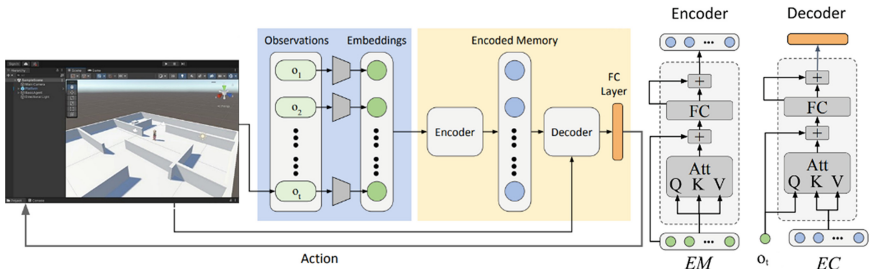


Fig. 1. Transformer architecture [8]

Encoder. The Transformer model makes use of self-attention to encode the embedded memory EM which is used to calculate the Query (Q), Key (K) and Value (K) in the Att Block. This transforms each embedded observation by using its relations to other past observations. Self-attention has a potential to capture the spatio-temporal dependencies in the environment.

Decoder. The decoder produces actions based on the current observation, given the context EC which is the encoded memory. It applies similar machinery as the encoder, with the notable difference that the Query (Q) in the attention layer is the embedding of the current observation o_t and the Key (K) and Value (V) are calculated from EC.

3 Results and Discussions

3.1 Maze Task

Maze Task Environment is shown in Fig. 2.

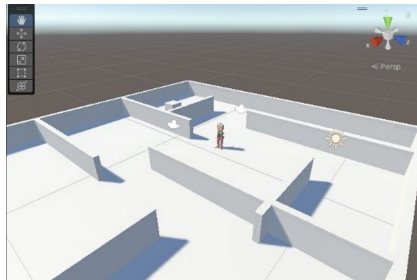


Fig. 2. Maze task environment

The goal of the agent in this maze navigation environment is to navigate to the target through the maze. The position of the agent is passed as the observation which helps the agent learn the limits of its movement and navigate the maze. A small reward is given when the agent touches the wall adjacent to the target while a final reward of 2000 is given when the agent reaches the target.

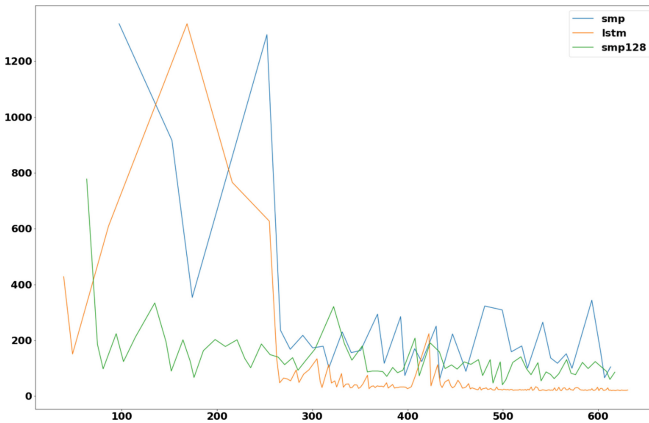


Fig. 3. Comparison of episode lengths during training

The above graph (Fig. 3) represents the episode durations during training for the Transformer and LSTM where the X-axis is the Episode number and the Y-axis is the Episode length during training. The blue line corresponds to Transformer with memory size of 1 whereas the green line corresponds to Transformer with memory size of 128. The orange line corresponds to LSTM. Convergence to a fixed path was observed in case of LSTM. The transformer tried multiple variations of paths and tried to touch the adjacent wall multiple times as compared to LSTM in order to maximize the reward. As a result, convergence to a fixed path was not observed but the transformer model achieved higher rewards.

3.2 Exploration Task

The main aim of this task is to move through the arena and explore as much of it as possible. Black tiles totaling to 242 are present throughout the arena and provide a reward of 1 when the agent touches them. This task tests the agent’s ability to explore. Visual inputs of size 300 * 300 with 3 color channels (RGB) were used here and nature CNN architecture was used for input preprocessing. Both the transformer and LSTM models, were trained for 50,000 timesteps and the number of tiles covered were plotted. Exploration task environment is shown in Fig. 4. Summary of results is shown in Table 1.

In this coverage task, the transformer agent performed higher exploration as compared to LSTM. LSTM episodes were shorter and the number of variations between paths followed and actions taken was comparatively less, as a result exploration was lower.

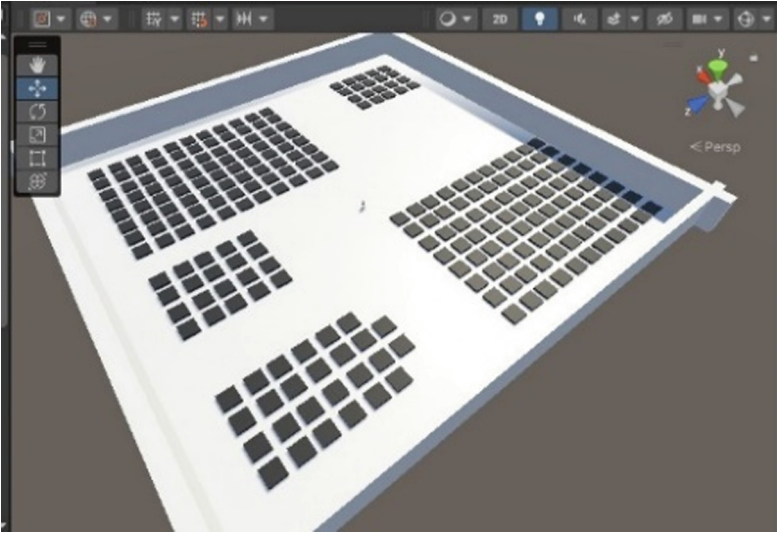


Fig. 4. Exploration task environment

Table 1. Percentage of coverage

Algorithm	Percentage of covered tiles
Transformer based memory	90.4% (219)
LSTM	86.3% (209)

3.3 The Long Horizon Search Task

In this task, the main goal of the agent is to discover new objects. For this purpose, the agent makes use of 300 by 300 visual observations with 3 color channels (RGB). 4 categories of objects with different colors are used—cylinder, capsule, sphere and cube. The arena consists of 4 rooms and no room consists of all 4 categories. The agent receives a reward when it discovers a new category. The model was trained for 1 million steps. Since it was observed in the first two tasks that the Transformer-based model performed significantly better than LSTM, we only tested the transformer model and its generalization ability in the further tasks.

Results for the Long Horizon Search Task (Fig. 5) with Transformer based memory are represented in Table 2.

Test 1. The first test involved performing the search task in the same environment with all the objects being at the same place as in training. Testing was run for 5 episodes to validate that the model was trained.

Test 2. In this scenario, the arrangement was the same as during training with the exception that the position of two objects was changed slightly. Since the visual observations,

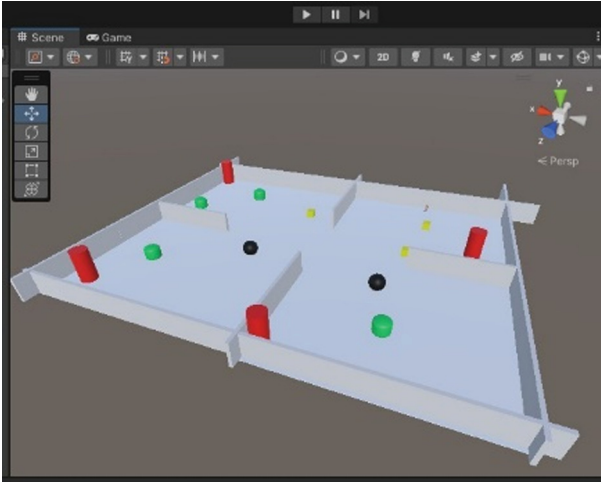


Fig. 5. Long horizon search task environment

Table 2. Results of long horizon search task with transformer based memory

Test	All 4 classes discovered	3 classes discovered
Test 1	5 (100%)	0 (0%)
Test 2	9 (60%)	6 (40%)
Test 3	26 (87%)	4 (13%)
Test 4	10 (67%)	5 (33%)

especially from a distance, were not that different from training in this case, the agent decided to stick to its training trajectory. The testing was run for 15 episodes.

Test 3. In this scenario, the positioning and arrangement of objects for all 4 rooms was changed internally. Testing was done for 30 episodes. In this case, since the visual observations for all the rooms were considerably distinct, the agent decided to not go with its default training trajectory. The movements were not swift and a lot of time was spent in observation and alignment with respect to targets.

Test 4. In this scenario, the spawn location of the agent was changed, the arrangement of the objects in each room was also changed drastically. The episode duration was doubled in this case. Swapping of objects between rooms was also done. Testing was done for 15 episodes.

3.4 The Long Horizon Multistage Task

The multistage task (Fig. 6) consists of 3 phases out of which, the middle one is a distractor phase. In the first phase the agent is supposed to go to any of the 4 capsules in

front of it. After this, the agent is transported to the distractor phase wherein it has to do an exploration task. After it has explored a certain number of tiles, it is transported to the third phase where it has to recall the color of the capsule it touched in the first stage and navigate to that capsule. The distractor phase is used to make the agent forget its main goal by making the agent complete a different task and acts as a test of its memory.

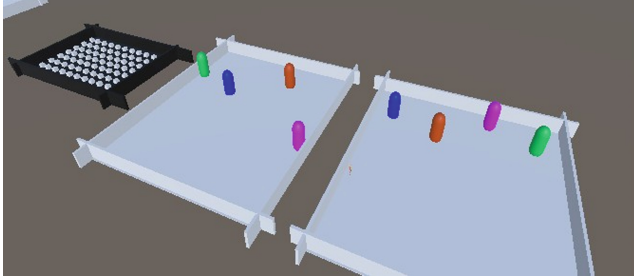


Fig. 6. Multistage task environment

Results. The model, which works based on visual inputs with 3 color channels and nature CNN preprocessing was trained for 1 million steps with a continuous action space. The trained model was tested for 30 episodes. The agent correctly identified the object in phase 3 for 11 out of 12 episodes (Fig. 7).

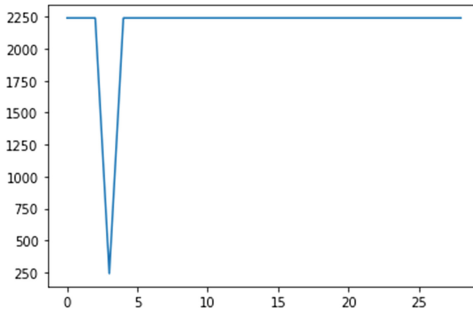


Fig. 7. Multistage task test rewards

Similar to search task, testing with different conditions was done for the multistage task as well. The location of objects in the third phase was shuffled and testing was done for 30 episodes.

The agent correctly identified the object in phase 3 for 28 out of the 30 episodes (Fig. 8). Even with changing of scenarios and the forgetting effect of the distractor phase, the agent was consistent in its behavior.

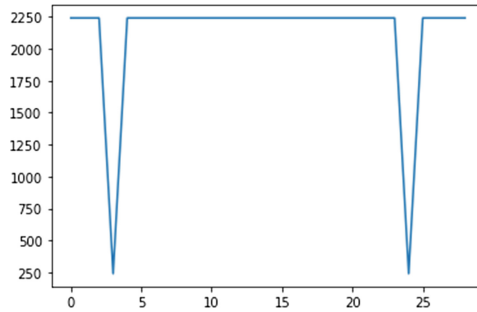


Fig. 8. Multistage task test rewards after shuffling

4 Conclusion

The transformer-based memory displayed higher performance than LSTM in the exploration and long horizon tasks due to its ability to store more detailed memory embeddings of the past. Even when the scenarios were changed during testing, the transformer showed great performance proving that transformer-based memory in combination with proximal policy optimization is generalizable. It also showed the ability to remember previous goals even when other learning is performed intermediately during distractor phases. These abilities of the transformer-based memory model prove it to be a better choice for robotic locomotion and navigation tasks over conventional RNNs and LSTMs.



Further improvements can be made to the memory architecture by introducing forgetting mechanisms and choosing only relevant memories when encoding instead of the entire memory buffer. We are working on implementing these improvements in our future work to better the performance of robotic agents and achieve human-like cognition.

References

1. Mirowski, P., et al.: Learning to navigate in cities without a map. In: *NeurIPS* (2018)
2. Xia, F., et al.: Gibson env: real-world perception for embodied agents. In: *CVPR* (2018)
3. Savinov, N., et al.: Semi-parametric topological memory for navigation. In: *ICLR* (2018)
4. Mirowski, P., et al.: Search on the replay buffer: Bridging planning and reinforcement learning. In: *NeurIPS* (2019)
5. Pritzel, et al.: Neural episodic control. In: *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 2827–2836. *JMLR.org*, Sydney (2017)
6. Racanière, S., Weber, T., et al.: Imagination-augmented agents for deep reinforcement learning. In: *Advances in Neural Information Processing Systems*. *CoRR* (2017)
7. Miyake, A., Shah, P.: *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. Cambridge University Press (1999)
8. Image source: Liu Junjia. https://zhuanlan.zhihu.com/p/157316200?utm_id=0



Investor Bot for Business Process

Sergey Kulik  and Ivan Sofronov 

National Research Nuclear University MEPhI, Moscow, Russia
ivanse99@yandex.ru

Abstract. Many people are thinking about accumulating and increasing their capital. The most obvious and, at first glance, the easiest way is to invest in securities. However, competent investment implies deep knowledge in the field of economics and the work of stock markets. The purpose of this work is to create an intelligent system that provides automated interaction between investors and private professional traders. In the future, it is planned to use this system when opening an investment hedge fund. Such a system will allow individuals to not worry about complex mathematical formulas and charts for making money, but simply entrust their capital to professionals who will take a certain percentage for their work. Communication between participants will take place through popular messengers. Additionally, there are plans to consider the possibility of creating a specialized artificial intelligence unit based on a neural network, which will learn investment strategies and capital management from traders.

Keywords: Intelligent system · Neural network · Investment strategies · Cognitive technology

1 Introduction

Investing money in various assets is a fairly common way of accumulating and multiplying capital. This is beneficial both for potential investors and for companies whose assets will be purchased. However, in order to properly manage money, you need to be a sufficiently savvy specialist in the field of finance and the economic system as a whole.

In order for a frequent owner of some capital not to understand the stock market on his own, you can consider the option of transferring funds under the management of investment funds (including hedge funds) or private traders who specialize in this, for a certain percentage of the profit. In this case, it is proposed to implement an intelligent system that automates the process of interaction between investors and traders/funds. At the MVP stage, it is assumed that the system should perform the following operations:

- depositing/withdrawing money from the investor's personal account in the investment fund;
- control of funds and profit distribution;
- drafting contracts and reports;
- other routine operations, completely freeing up traders' time to engage in investments.

To attract more customers, it is proposed to implement several ways of their interaction with the system (from the most to the least important):

- bots in popular messengers;
- web interface available in all browsers;
- Mobile apps for Android and IOS operating systems;
- Desktop applications for Windows and macOS operating systems.

The most preferred way for users to interact with the system is through popular messengers. In this case, users will be able to communicate with the system through a convenient messenger for them. At the first stage, communication is assumed due to a set of predefined commands, and in the future it is possible to add some language model similar to Chat GPT, so that communication is already more convenient and comfortable for users. Also, if you use a messenger, you can worry less about the security of data transmission and there is no need to develop your own user interface. In addition, there are no good solutions for such applications in messengers yet, so you need to have time to occupy this niche.

Paper [1] deals with patterns of cognitive activity in a human vs collaborative robot interactive game. Another paper [2] describes experiments with neural net object detection system. Paper [3] describes convolutional neural networks technologies.

2 Overview of Software Tools

The server part is proposed to be implemented in Java. Among the main advantages of this language over others are the following:

- cross-platform. The language runs under the java JVM virtual machine. JVM distributions are available for all popular operating systems. This feature makes Java convenient for both development and deployment of the system on target servers;
- high level of language proficiency. The performance of the language shows quite high indicators. However, Java is a high-level language that allows you not to waste time solving low-level tasks that C++ programmers face, for example;
- a large number of web development frameworks have been written for this language, which significantly speed up and simplify the development process. It is proposed to use the most popular Spring Boot framework for work.

The PostgreSQL DBMS [4] will be used as the database. It is one of the fastest open source databases.

The following programming languages will be used to implement mobile and desktop applications:

- Kotlin for Android;
- Swift for iOS and macOS;
- C# for Windows.

In this case, the choice is obvious, since these programming languages are tailored for the implementation of applications for these platforms.

To implement the web interface, it is proposed to use Vue.js as the simplest and most functional framework designed to implement the user interface in browsers.

In order to debug the operation of the intelligent system, find errors and incorrect operation, as well as respond in a timely manner to emerging abnormal situations, it is necessary to configure monitoring of the logging status. Grafana/Prometheus tools (for monitoring) and ELK (for logging) will be used to solve this problem.

In addition, for the convenience of launching and managing and the security of both the intelligent system itself and the ecosystem, containerization using the Docker tool will be used.

3 Business Process MVP

At the first stage, the minimum functionality necessary for the operation of the system is implemented. This is necessary in order to assess the demand for a particular functionality, determine the next steps in the development of the system, and also get feedback from users.

For the current stage, the development of a bot based on one of the most popular Telegram messenger in the CIS countries has been started. In addition, this messenger provides a convenient API for creating bots for it.

To identify users, it is proposed to use their unique ID in the telegram system. Each user can have several roles from the list:

- **Admin.** The role of the system administrator. It is necessary to provide the administrator with the ability to manage the system, view and change settings, monitor logs;
- **Investor.** The role of the investor. This role makes it possible to transfer finances under the control of traders, keep records and manage only their own finances;
- **Manager.** The role of the trader. This role allows you to accept money from investors and manage the collected capital on the stock market.

Each investor can create only one chat with a bot and manage through the command menu. To do this, the following functionality is being developed:

- depositing funds to your account in the system. At the first stage, it is assumed that the user will manually transfer money to the account manager and send a check confirming the operation through the bot. The account is in a frozen state until the transaction is confirmed by the manager;
- withdrawal of money. At this stage, it is assumed that the investor requests the withdrawal of a certain amount of money, and the manager transfers money to the specified details manually. Until the confirmation of the receipt of money by the user, the account is in a frozen state;
- view the account status;
- unloading of contracts and statistics on the account.

To implement the described functionality, it is necessary to store information about the user, information about the chat with the user, information about the status of his account, as well as information about the history of account changes. The corresponding implemented data model is shown in Fig. 1.

Data model consists of 6 objects (see Fig. 1): `t_roles_of_user`, `t_finances`, `t_user_history`, `t_chats_of_users`, `t_users`, `t_total_history`.

Each trader must provide daily information on the status of all capital under his management at the time of closing trading on the stock market. Profit is calculated from the difference in the state of capital, the percentage is deducted to the investor’s account. The rest of the funds are distributed to the accounts of investors linked to this manager in proportion to their capital.

To implement this functionality, it is also necessary to keep a history of changes in the entire capital managed by the trader.

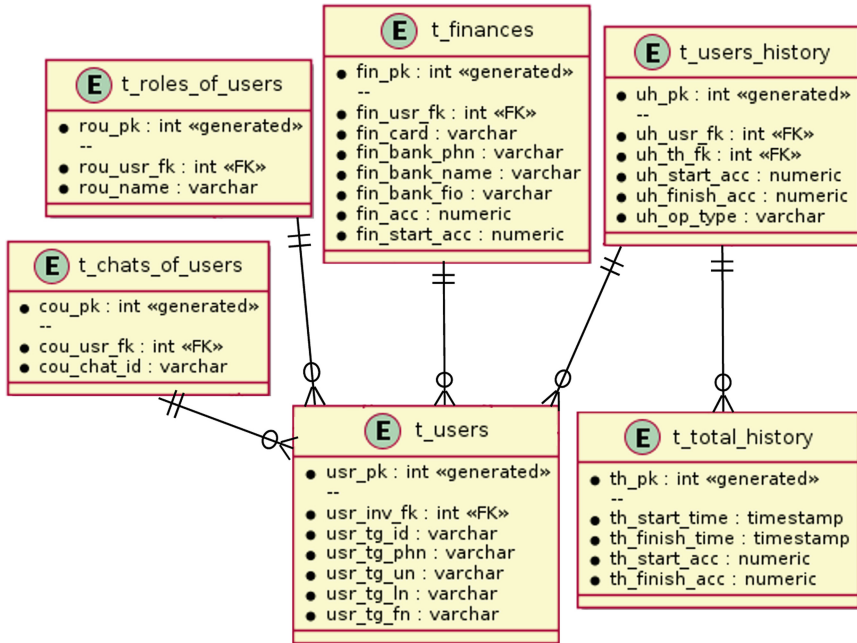


Fig. 1. Data model for MVP.

This means that the results obtained can be used to construct another data model for MVP.

Based on their necessary functionality for MVP, an automated system architecture was proposed, shown in Fig. 2. Services being developed:

- ib-beans is a module with constants, configurations and settings that will be used in all services of the intelligent system;
- ib-dao – database interaction module;
- ib-services – data management module;
- ib-tg-bot is a module for interaction with the Telegram messenger.

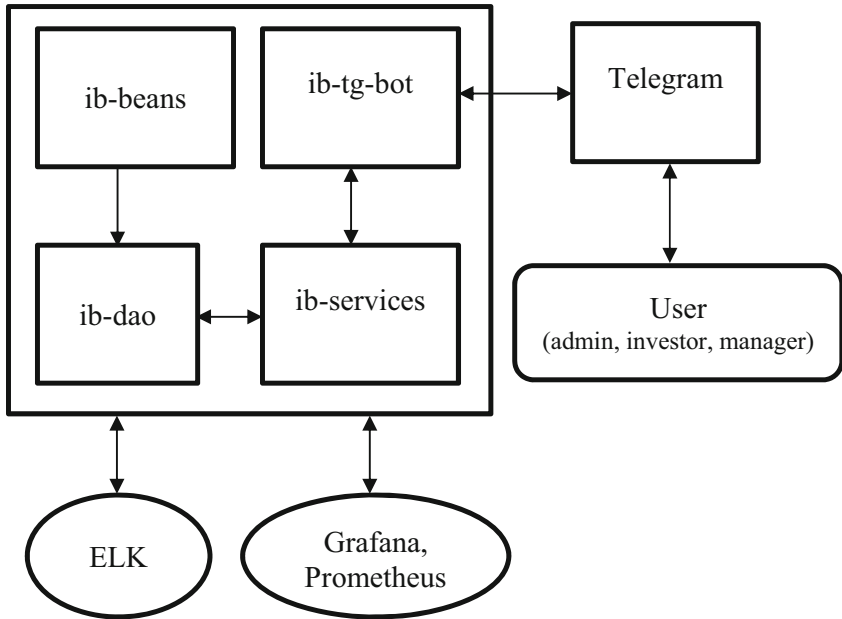


Fig. 2. Architecture for MVP.

4 Further Plans

After the implementation of the MVP, first of all, the described process of depositing investors' funds to their accounts in an automated system will be fully automated. After that, the web user interface will be implemented. In addition to the described functionality, the UI is supposed to add online charts for viewing the dynamics of investors' accounts, the ability to choose the most suitable investment strategies, as well as the ability to collect investment portfolios if desired.

However, from the point of view of traders, the investment process itself can also be a routine and purely mechanical process, for example, when using the following models:

- Investing in securities indices [5];
- Simultaneous opening of long and short positions on different assets;
- Investing in bonds.

To minimize risks and increase potential profits, it is planned to expand the bot with a special intelligent block based on a neural network, which will perform the following operations:

- Analysis of the weighted average correlation of securities markets of different industries and even countries;
- Assessment of the variable cyclical nature of the markets;
- Analysis of the work of companies;
- Analysis of the stock market dynamics in historical terms.

- Based on the obtained statistical models, the most optimal investment strategies will be selected.

It is also assumed that under the guidance of experienced traders, an intelligent block based on a neural network will be trained and supplemented with other, more complex and risky investment models. Excluding the human factor, such a bot will be able to provide greater profitability, freeing up human resources for monitoring and training the neural network.

5 Conclusion

Investing money in various assets seems to be a fairly common and simple way of accumulating and multiplying capital. This is beneficial both for potential investors and for companies whose assets will be purchased. However, not everyone is willing to spend a lot of time studying the stock market in order to get a really significant income from it.

At the same time, there are people who are professionally engaged in this and need more capital to increase it. The system described in this article will help to automate the process of interaction between investors and traders. At the moment, the MVP stage for this system is almost implemented, the authors hope to test the system in real life soon.

We believe that our results are useful for scientists and researchers in the field of cognitive technology [1], projects of the MegaScience class [3], hybridization of intellectual technologies for analytical tasks of decision-making support [6], for mega science class projects [7] and medicine [8], agent technologies for technical information search [9], regular agent technologies [10], forensics technologies [11] neural networks technologies [12], intellectual tutoring systems based on cognitive architectures [13], and more.

Acknowledgments. This work was supported by the MEPhI Program Priority 2030.




References

1. Shirokii, V., Batusov, R., Chubarov, A., Dolenko, S., Samsonovich, A.: Patterns of cognitive activity in a human vs collaborative robot interactive game. *Proc. Comput. Sci.* **145**, 495–499 (2018)
2. Kulik, S.D., Shtanko, A.N.: Experiments with neural net object detection system YOLO on small training datasets for intelligent robotics. In: *Mechanisms and Machine Science*, vol. 80, pp. 157–162. Springer, Cham (2020)
3. Shtanko, A.N., Kulik, S.D.: Scientific personnel training in convolutional neural networks for the implementation of research projects of the MegaScience class. *J. Phys. Conf. Ser.* **1406**(1), 012014 (2019)
4. Documentation for PostgreSQL 9.5.21. The PostgreSQL Global Development Group. Translation into Russian, 2015–2019: Postgres Professional Company (2019)
5. Fin-Plan. Index Investing. <https://fin-plan.org/blog/investitsii/indeksnoe-investirovanie/>. Last accessed 12 Apr 2022
6. Borisov, V.V.: Hybridization of intellectual technologies for analytical tasks of decision-making support. *J. Comput. Eng. Inf.* **2**(1), 11–19 (2014)

7. Kirichenko, A.V.: Acmeological determinants of experts selections for “mega science” class projects. *J. Phys. Conf. Ser.* **1685**, 012016 (2020)
8. Yasnitsky, L.N., Dumler, A.A., Cherepanov, F.M.: Robot-doctor: what can it be? In: *Advanced Technologies in Robotics and Intelligent Systems*, pp. 163–169. Springer, Cham (2020)
9. Ananieva, A., Onykiy, B., Artamonov, A., Ionkina, K., Galin, I., Kshnyakov, D.: Thematic thesauruses in agent technologies for scientific and technical information search. *Proc. Comput. Sci.* **88**, 493–498 (2016)
10. Artamonov, A., et al.: Regular agent technologies for the formation of dynamic profile. *Proc. Comput. Sci.* **88**, 482–486 (2016)
11. Kulik, S., Nikonets, D.: Forensic handwriting examination and human factors: improving the practice through automation and expert training. In: *The Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC2016)*, Proceedings. Moscow, Russia, pp. 221–226, 06–08 July 2016
12. Danilin, S., Shchanikov, S., Zuev, A., Bordanov, I., Korolev, D., Belov, A., Pimashkin, A., Mikhaylov, A., Kazantsev, V.: Design of multilayer perceptron network based on metal-oxide memristive devices. In: *12th International Conference on Developments in eSystems Engineering (DeSE)*, pp. 533–538. IEEE (2019)
13. Samsonovich, A.V., De Jong, K.A., Kitsantas, A., Peters, E.E., Dabbagh, N., Kalbfleisch, M.L. Cognitive constructor: an intelligent tutoring system based on a biologically inspired cognitive architecture (BICA). *Front. Artif. Intell. Appl.* **171**(1), 311–325 (2008). ISSN: 09226389



Testing for Benford’s Law as a Response to the Risks of Material Misstatement Due to Fraud

Viktor M. Sushkov  and Pavel Y. Leonov  

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute),
Moscow, Russia
pyleonov@mephi.ru

Abstract. The paper assesses the possibility of applying Benford’s Law tests as a response to the risks of material misstatement due to fraud under International Standard on Auditing (ISA) 240. The approach has been applied as part of a comprehensive financial statement audit of a leasing company. Its approbation made it possible to identify fraudulent transactions in accounting resulting from management override of controls, while substantially reducing the labor costs of the audit. The confirmation of Benford’s Law efficacy in the auditing process solidifies its status as a reliable and valuable tool for detecting potential fraud and uncovering material misstatements, ultimately enhancing the overall integrity and trustworthiness of financial audits.

Keywords: Benford’s Law · Audit · Fraud · ISA 240

1 Introduction

In accordance with International Standard on Auditing (ISA) 240 “The Auditor’s Responsibilities Relating to Fraud in an Audit of Financial Statements”, the identification and assessment of risks of material misstatement (RMM) due to fraud is a required procedure in an audit of financial statements. At the same time, international practice shows that the effectiveness of fraud detection by auditors tends to be low. According to the Association of Certified Fraud Examiners (ACFE), occupational fraud is detected by external auditors in only 4% of cases, which is 1% lower than disclosure by accident [1]. Given that most cases of occupational fraud have a material impact on the reliability of financial reporting, it is reasonable to assume that audit reports issued without detecting existing fraud may have inaccurate conclusions.

Due to the fact that the audited entity’s accounting data is a large volume array characterized by multiple attributes, manual processing and basic filtering commonly used today do not allow identifying indicators of potential fraud. The use of a statistical method called Benford’s Law is relevant in this regard, allowing the detection of implicit patterns and properties present in the data. Unlike many other methods of fraud detection,

inherent properties and statistical principles of Benford's Law allow for a consistent and objective analysis of numerical data without the need for specific customization or parameterization. This makes Benford's Law an accessible and efficient tool for identifying potential fraud across various industries and accounting systems auditors may encounter.

2 Analytical Part

2.1 Methodology for Identifying Accounting Misstatements Using Benford's Law

In 1881 the American astronomer Simon Newcomb and then in 1938 the American engineer Frank Benford independently identified an unusual pattern inherent in naturally generated data sets. Frequency distribution of the first digits appears to be not equal and decrease with increasing digits. The probability of D_1 being the first digit is calculated using the formula (1).

$$P(D_1 = d_1) = \log_b \left(1 + \frac{1}{d_1} \right) \quad (1)$$

In (1) b is the number system ($b > 2$) and d_1 stands for digits in this number system ($d_1 \in \{1, \dots, b - 1\}$) [2].

For example, for the decimal number system, the probabilities of 1–9 to be in the first digit are shown in Table 1.

Table 1. Expected frequencies of first digits from 1 to 9

Digit	1	2	3	4	5	6	7	8	9
Probability (%)	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6

Benford's Law also applies to financial data such as journal entries, bank transactions, accrued receivables and payables, provisions, etc. Comparison of the distribution in such arrays with the Benford distribution makes it possible to identify anomalies, including those due to fraudulent adjustments to the data.

The greatest contribution to the development of Benford's Law was made by Professor M. Nigrini of West Virginia University, who proposed a methodology for the integrated application of statistical tests based on the Benford distribution to detect financial fraud. The most significant works of M. Nigrini are his Ph.D. thesis "The detection of income tax evasion through an analysis of digital frequencies" [3] and the 2012 edition "Benford's Law: applications for forensic accounting, auditing and fraud detection" [2].

M. Nigrini's methodology includes sequential performance of 8 tests, classified into primary (first digit test, second digit test, first order test), advanced (summation test, second order test) and associated (duplication test, last digit test and distortion coefficient model). Each test identifies irregularities and suspicious patterns in data using specific criteria and statistical analysis. The tests are supplemented with calculation of statistical

characteristics that allow quantify conformity to Benford's Law, including mean absolute deviation (MAD), Z-statistics, Kolmogorov-Smirnov statistics, Chi-square statistics, etc. [2].

Approbation of tests and statistical characteristics on different sets of corporate data proves their effectiveness in identifying non-standard elements in samples of large volume [4, 5]. While the general concept and principles of Benford's Law have been explored and established, there is a lack of comprehensive studies focusing on its specific application and effectiveness in the context of financial auditing. This paper therefore aims to gain a deeper understanding of the nuances, challenges, and best practices associated with utilizing Benford's Law as a tool for fraud detection in audit processes.

2.2 Application of Benford's Law Tests in the Audit

According to ISA 240, required audit procedures responsive to risks related to management override of controls include testing the appropriateness of journal entries recorded in the general ledger and other adjustments made in the preparation of the financial statements (paragraph 32). However, the standard sets out only the framework requirements for this procedure, leaving the auditors an ability to develop their own methodology, including testing of journal records using Benford's Law.

We validated the methodology in a series of real financial audits and chose to present the findings from an audit conducted on a leasing company as it is one of the most representative cases. An assessment of the company's adequacy of internal control over risks of fraud, management and those charged with governance (TCWG) enquiries, analytical procedures and analysis of risk factors identified the high RMM due to fraud, necessitating a detailed analysis at the stage of audit procedures responsive to assessed RMM due to fraud. The data to be analyzed was an array of journal entries for the year under review, containing 27 attributes and 1,562,802 records. The attributes included the accounts, documents, contents, amounts, and other transaction details. To ensure confidentiality, all names and sensitive information have been changed.

The array was then checked against the requirements that need to be met in order for Benford's Law to be used. These include random formation, sufficient volume, absence of categorical data, no minimum and maximum limits, proximity to an exponential distribution, etc. The array has been verified and therefore it was decided to examine the whole set of journal entries over the period consistently using the primary, advanced and associated tests. The confidence level at which the tests were conducted is 95%, which is usually used in Benford's Law analysis.

Graphical representations of the results of some primary, advanced and associated tests are shown in Figs. 1, 2 and 3 respectively.

In addition, statistical characteristics, including Z-statistics, Chi-square statistics, MAD, were calculated for each test. The results of the tests were aggregated. It was found that, firstly, the data does not generally conform to Benford's Law and there is a risk of intentional distortions, excessive rounding and errors in the data. Secondly, the riskiest transaction amounts have certain numerical patterns, as shown in Table 2.

The sampling on the basis of these numerical patterns identified 7 groups of suspicious transactions. In each of the groups there are amounts that may indicate the presence

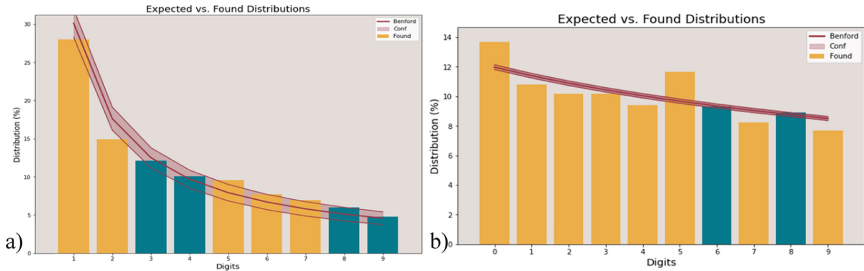


Fig. 1. Results of the a) second digit test; b) first-order test

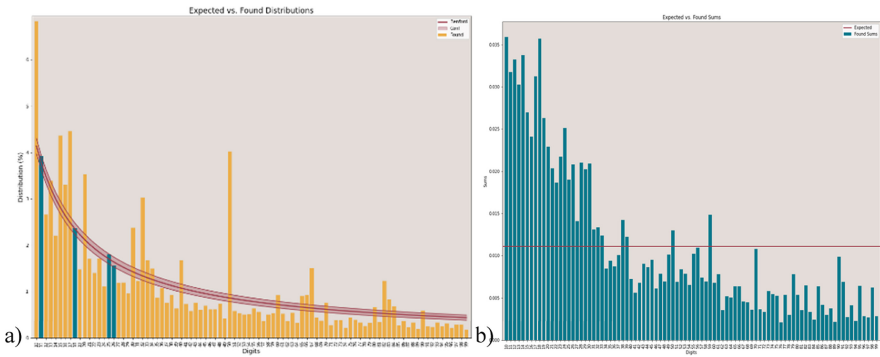


Fig. 2. Results of the a) second-order test; b) summation test (for the first pair of digits)

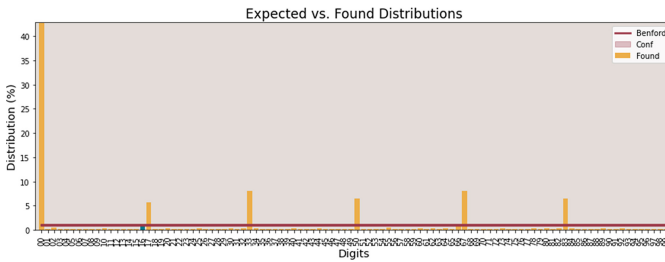


Fig. 3. Results of the last two digits test

of distortions and errors due to nonconformity to Benford's Law. The fraud indicators corresponding to each group of transactions are systematized in Table 3.

In addition, the following inconsistencies were identified in many of the transactions described above:

- the dates of contracts and other documents related to the transactions differ significantly from the transaction period (e.g. the contract is for 2021 or the fine is for 2016, whereas the transaction period is for 2022);

Table 2. Digital templates for the riskiest transaction amounts

Test	Digits
First digits with the highest Z-statistics	2, 5, 7, 1, 6
Second digits with the highest Z-statistics	5, 0, 9, 7, 2, 4
First two digits with the highest Z-statistics	25, 50, 21, 17, 23, 19, 30
First two digits with the largest absolute differences in the sum frequencies	10, 18, 14, 12, 11, 17, 13, 15, 19, 24
Differences from the previous value, whose first two digits have the largest Z-statistics	50, 32, 10, 17, 67, 20, 82, 15, 30, 40
Most frequent	250, 1, 1 500, 1 666.67, 3 000, 10 000, 18 304.02, 500, 15 689.16, 22 697.55
Last two digits with the highest Z-statistics	00, 67, 33, 50, 83, 17

Table 3. Groups of suspicious transactions identified by Benford’s Law testing

Group of transactions	Fraud indicators
1. Payments for services	1.1 Payment of agency fees in instalments, each time in slightly varying amounts. There is a risk of splitting up a large fraudulent operation to avoid attracting attention
	1.2 Contractual payments in small amounts. The counterparty may be trying to conceal the receipt of a large sum of money
2. Offsetting an advance to the suppliers and buyers	2.1 Transactions to offset the buyer’s advance are repeated for the same non-significant amounts
	2.2 Recurring offsetting transactions against advances to suppliers for the same small amounts. These transactions are difficult to control due to their immateriality individually and the likelihood of omissions, although collectively the amounts are significant
3. VAT deductions	3.1 There are more VAT deductions than outgoing VAT accounting transactions
	3.2 VAT deductions are duplicated and registered by the same person. These transactions are an indicator of illegal VAT refunds

(continued)

Table 3. (continued)

Group of transactions	Fraud indicators
4. Accruals and deductions of VAT on prepayment	4.1 Prepayment VAT accruals and deductions are divided into several (often equal) parts for each company and individual entrepreneur. In addition, they are often repeated. This creates a risk of including redundant transactions which are difficult to track
5. Purchase of consumables	5.1 There is a tenfold increase in the expenditure on consumables in 2022 compared to 2021. Such a significant rise may suggest potential misappropriation of funds through overstatement of actual amounts and misappropriation of consumables themselves
6. Purchase of services	6.1 The amounts invoiced for telecommunications and internet services are significantly higher than the company's needs. In addition, most of these invoices are not paid. There is a risk of over-invoicing to avoid taxation
7. Sale of other assets	7.1 Sales of other assets to different companies occur in many transactions for the same large amount of money. In a large sample of similar transactions, there is a potential risk of including fraudulent transactions among the legitimate ones, making it challenging to identify them

- duplication of identical amounts in large numbers: carrying out the same type of repeated transactions, or conversely, different transactions of the same nature for the same amount;
- the amount is not in line with the nature of the transaction;
- the description of the transaction does not correspond to the economic sense of the transaction (e.g. payment of a fine is designated as a contractual service payment).

On the basis of the analysis performed, it can be concluded that there were material misstatements in the designated areas of business operations due to potential fraud, which led to obtaining written representations from management and TCWG, questioning the auditor's ability to continue performing the audit, informing external parties and other procedures stipulated by ISA 240. The sample size was reduced by more than 90% (from 1,562,802 to 97,204 journal records), which significantly reduced the audit workload.

3 Conclusion

The findings demonstrate that Benford's Law as an audit procedure to test the appropriateness of journal entries recorded in the general ledger is an effective audit procedure responsive to assessed RMM due to fraud. The introduction of this methodology into auditing practice is highly recommended and holds significant potential due to the consistent failure of auditors to detect fraudulent practices, as well as the limitations associated with traditional manual detection methods for modern fraudulent schemes.


It is worth noting, however, that current application of Benford's Law in financial auditing remains limited and lacks systematic adaptation to meet the specific requirements of the auditing profession. This highlights the need for further research and development of a theoretical and methodological framework that integrates Benford's Law into the auditing procedures. Without further progress in this area, auditors may continue to struggle in identifying fraudulent practices and face ongoing challenges in the detection of sophisticated fraudulent schemes.

References

1. Occupational Fraud: A Report to the Nations. <https://acfepublic.s3.us-west-2.amazonaws.com/2022+Report+to+the+Nations.pdf> (2022). Last accessed 24 Mar 2023
2. Nigrini, M.J.: *Benford's Law: Applications for Forensic Accounting, Auditing and Fraud Detection*. Wiley, Hoboken, New Jersey (2012)
3. Nigrini, M.J.: *The detection of income tax evasion through an analysis of digital frequencies*. Ph.D. thesis. University of Cincinnati, Cincinnati, OH, USA (1992)
4. Leonov, P.Y., Rychkov, V.A., Ezhova, A.A., Sushkov, V.M., Kuznetsova, N.V., Suits, V.P.: Possibility of Benford's law application for diagnosing inaccuracy of financial statements. In: *Proceedings of the 12th Annual Meeting of the BICA Society. Studies in Computational Intelligence*, vol. 1032, pp. 243–248. Springer, Cham (2022)
5. Leonov, P.Y., Suits, V.P., Norkina, A.N., Sushkov, V.M.: Integrated application of Benford's Law tests to detect corporate fraud. *Proc. Comput. Sci.* **213**, 332–337 (2022)



Greening Telecom: Harnessing the Power of Artificial Intelligence for Sustainable Communications

Anastasiia Suslina¹✉, Konstantin Savin¹, and Irina Suslina² 

¹ IXP Consulting, Kirovogradskaya 32-2-84, 117519 Moscow, Russian Federation
{asuslina, ksavin}@ixp-consulting.com

² National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), 31
Kashirskoye Shosse St., 115409 Moscow, Russian Federation
IVSuslina@mephi.ru

Abstract. The article explores the potential for artificial intelligence (AI) technologies to reduce the environmental impact of the information and communications technology (ICT) industry. The paper notes that the ICT sector is responsible for a significant proportion of global carbon emissions, but that digital solutions can help reduce emissions and promote sustainable practices. The study finds that the use of AI in telecommunications has great potential to reduce the industry's environmental impact by optimizing energy efficiency and reducing carbon emissions. Despite challenges such as data quality, limited availability of data, and ethical and privacy concerns, continued research and development is crucial in realizing the full potential of AI-powered green technologies. The article concludes that by harnessing the power of AI, the telecommunications industry can play a significant role in mitigating the effects of climate change and creating a more sustainable future.

Keywords: Green ICT · Artificial intelligence · Sustainable development · Energy efficiency · Green networks

1 Introduction

According to the latest report by the Intergovernmental Panel on Climate Change (IPCC), the world nowadays is experiencing the effects of climate change and it is expected to worsen in the coming decades if no action is taken [1]. There are several core strategies that exist now to fight climate change: they include reduction of greenhouse gas emissions through the use of renewable energy sources, such as solar and wind power, and the implementation of energy-efficient practices and investing in research and development of new technologies that can help reduce emissions and enhance sustainability [2, 3]. The main international document to combat climate change and reduce greenhouse gases emissions is the Paris Agreement. It is an international treaty that aims to limit global warming to well below 2 °C above pre-industrial levels and to pursue efforts to limit the temperature increase to 1.5 °C [4]. The agreement requires countries to set their own

targets for reducing greenhouse gas emissions and to regularly report on their progress. There are much more national and international agreements and laws, that put reduction of greenhouse gases as the core of the climate strategy [5, 6]. Thus, legal requirements at the international and national levels pose an urgent problem of using ICT to reduce emissions and help achieving sustainability goals, which makes green ICT a very actual research area.

ICT plays a critical role in climate change mitigation and adaptation efforts [7]. Historically, ICT emissions have steadily grown along with global emissions. Several studies up to 2015 have assessed the carbon footprint of ICT sector, they show an increase in the carbon footprint of ICTs over time, even without considering life cycle emissions, with the trend line showing a 40% increase between 2002 and 2012 [8]. The rise in ICT emissions has coincided with a steady rise in our overall global carbon footprint, with global greenhouse gas emissions rising at 1.8% per annum. This indicates that the impact of ICT has likely grown faster than global emissions. Scientists predict that over time, emissions from the ICT sector will continue growing and by 2040 may increase by more than 3 times which contravenes global and regional sustainable aims and makes the area extremely significant to study [8].

The main global body managing development of ICT sector is the International Telecommunication Union (ITU) of the United Nations. One of the key concerns of the ITU is the impact of the ICT sector on the environment, particularly its carbon footprint. The ITU Recommendation calls on ICT companies to reduce their emissions to zero by 2050 [9]. Also, the ITU, in collaboration with the GSM Association (Global System for Mobile Communications, GSMA), the GeSI (Global Enabling Sustainability Initiative) and the SBTi (Science Based Targets initiative), has issued guidance aimed at reducing ICT greenhouse gas emissions. These science-based targets are aimed at reducing carbon emissions made by ICT sector (networks, data centers and end-user devices) by 45% by 2030 and to zero by 2050, in line with limiting global warming to 1.5 °C [10].

AI technologies play a significant role in green telecom by enabling more efficient and sustainable operations, among such technologies are machine learning, Natural Language Processing, computer vision, predictive analytics, optimization algorithms and some others.

This article explores the potential of artificial intelligence (AI) technologies based on cognitive technology solutions. In the course of the analysis, key areas of AI-based ICT were identified, that can help to achieve sustainable development goals and reduce environmental impact. The article provides examples and discusses how AI-based decisions can be used to optimize the energy efficiency and reduce carbon footprint produced by mobile networks, data centers and end-user devices, which have been identified as the most significant contributors to the total carbon footprint, as per the guidelines issued by the ITU, the GSMA, the GESI and the SBTi [10].

2 AI Technologies in Telecommunications

AI technologies have been applied in the telecommunications industry since the early 2000s. They are used to automate processes, optimize network performance, and improve the customer experience. The use of AI technologies in telecommunications is growing

and expected to continue to grow in the coming years, as telecom companies seek to improve network efficiency and customer satisfaction [11]. We have identified and analyzed 3 the most important areas in green telecom, where cognitive AI technologies are applied: mobile networks, data centers and end-user devices.

2.1 Mobile Networks

Increasing Energy Efficiency, Network Optimization and Traffic Prediction

A mobile telecommunication network typically comprises thousands of sites, with each site requiring a significant amount of power. As telecommunication technologies continue to develop and mobile traffic increases exponentially, operators are confronted with the challenge of controlling the energy component of their operating expenditure [12].

Special technologies that are dedicated to controlling and reducing the energy consumption of telecommunication networks are commonly referred to as Energy Saving Features (ESFs). These ESFs are in the form of software that is installed on network equipment across all generations of mobile telecommunication technologies, including 2G, 3G, 4G, and 5G. ESFs operate on several levels, ranging from micro-shutdowns on transmission for a few microseconds to shutting down circuits and transmission/reception branches or frequencies for several hours. The highest level of energy savings is achieved by shutting down transmission branches or frequency bands entirely.

The traffic on each cell of a mobile network is subject to continuous variation and never has an identical profile on two different cells. However, the greatest variability observed by all operators on their networks is the diurnal variation in traffic. The decline in activity during the night is considerable, ubiquitous, and occurs nightly. Sites that are designed to handle high traffic loads during the day are over-dimensioned during the night due to the drastic drop in traffic. In order to meet the constraints of quality of service, the first generation of ESFs was based on manually defined elements that considered the reality and variability of traffic on the networks.

Artificial intelligence has a pivotal role to play in the energy management of telecommunication networks. Intelligent algorithms enable ESFs to be activated on a 24-h basis and adapted to each cell, allowing for an autonomous, reactive system capable of performing automated actions in networks with hundreds of thousands of cells. AI enables the management of ESF requirements during the day at the network level, overcoming the limitations of manual and static programming for ESF activity during the night, often between midnight and 6:00 am.

The intelligent system automatically and continuously analyzes the traffic for each cell to establish a statistical profile and detect repetitive behaviors, which enables it to make predictions about future traffic evolution. Using this forecast, the system can anticipate and with a high degree of accuracy, identify future periods of traffic decrease during the day and night for each cell on the network and for each day of the week. During these periods of traffic decline, the system can activate ESFs to minimize the impact on the quality of service. Outside of these periods, the system blocks the operation of the ESFs to ensure the maximum capacity of the sites is available to users.

Several vendors and operators have demonstrated positive results of implementation AI technologies. For example, iPowerStar Huawei solution employs intelligent algorithms to achieve the goal of “one site, one policy” by managing energy-saving in time, space, frequency, and power domains, while ensuring stable network performance. Additionally, it empowers multi-objective optimization for energy-saving, enabling operators to build networks by striking the optimal balance between performance and energy consumption. In comparison to traditional energy-saving solutions, which provide an average of 15% energy savings, iPowerStar boasts energy savings of 30% [13]. Another example is Telefonica Spain, they have tested Ericsson’s Radio Deep Sleep Mode energy saving functionality based on Artificial Intelligence and Machine Learning algorithms and reported savings of up to 8%, considering the site’s total 24-h consumption, and up to 26% in low traffic hours [14].

Thus, reducing energy consumption and optimizing network performance can help reduce the carbon footprint of the telecommunications industry and traffic prediction can help prevent congestion and also reduce the need for additional hardware and accordingly energy consumption.

Predictive Maintenance

AI technologies can be used to implement predictive maintenance in mobile networks. Predictive maintenance involves using data analytics and machine learning algorithms to predict when equipment is likely to fail and proactively schedule maintenance before the equipment breaks down [15]. This approach is core for achieving circular economy and helps to minimize downtime and reduce the need for costly emergency repairs. It prolongs life-cycle of the network equipment and allows to reduce e-waste, which is now one of the fastest growing types of waste and simultaneously potentially profitable, if managed properly [16]. One of the examples of such existing technologies is the iFault-Care Huawei’s solution, which focuses on network fault scenarios, introduces predictive capabilities to network operation and maintenance (O&M), enabling operators to shift from responsive to proactive O&M. It predicts potential problems and risks through intelligent analysis and modeling, resulting in a 22% decrease in cell service interruption time [13]. iFaultCare precisely identifies faults and locates root causes. In a city in Anhui province, China, iFaultCare is being widely applied. By introducing intelligent fault identification and diagnosis capabilities to the operator’s workflow, iFaultCare reduced work orders by more than 20% [17]. Moreover, the operator’s annual O&M costs are expected to decrease by millions of CNY.

2.2 Data Centers

Data centers are a significant contributor to the overall carbon footprint of the ICT sector [18]. According to estimates, data centers could consume 4 times more electricity by 2030, and they currently account for around 1% of global greenhouse gas emissions [19, 20]. As such, it is essential to find ways to reduce the energy consumption and carbon footprint of data centers. AI technologies can play a significant role in achieving this goal by:

- **Cooling efficiency:** Air conditioning is one of the highest sources of energy consumption in data centers, accounting for up to 40% of total energy consumption [21]. AI

analyzes data center temperature patterns, airflow, and other environmental variables to optimize cooling, reduce energy consumption, and improve cooling efficiency [22].

- Power usage optimization: AI uses data center sensors to monitor energy consumption, power usage, and performance parameters in real-time. AI analyzes the data to identify trends and patterns, enhancing power usage optimization, and enabling data centers to optimize workloads, reduce energy waste, and improve performance [23].
- Predictive Maintenance: AI can potentially predict equipment failure and maintenance requirements in advance, preventing unplanned downtime and reducing energy consumption by limiting the need for ongoing repairs and maintenance [15].
- Load Balancing and Server Optimization: AI monitors data center workloads in real-time, analyzing their performance and predicting potential bottlenecks. AI can then balance the workload across the data center's infrastructure, optimizing power usage, reducing energy consumption, and enhancing performance.
- Energy Sourcing: AI analyzes renewable energy sources, identifies the best time to connect to the grid and when to leverage generated power from renewable sources such as solar and wind [24].

2.3 End-User Devices

The term “end-user device” refers to any device that is used by an end-user to access a telecommunications network or service. This can include devices such as smartphones, tablets, laptops, and desktop computers, as well as specialized devices such as routers and modems. End-user devices are an important part of the telecommunications ecosystem, as they enable individuals and organizations to access and utilize network services. The energy consumption of end-user devices is also be significant, particularly in the case of devices such as smartphones and laptops that are used frequently and require regular charging [25]. As a result, energy efficiency measures in end-user devices are an important strategy for reducing the environmental impact of the telecommunications industry.

AI can be used in several ways to make the usage of mobile phones greener and more sustainable:

- Battery optimization: AI can analyze a user's usage patterns and adjust the phone's power usage accordingly. For example, if the AI detects that a user isn't using their phone, it can put it into a low-power mode, conserving battery life;
- Power-saving modes: Many mobile phones now have power-saving modes that can be activated manually or automatically. AI can optimize and automate these modes, conserving energy whenever the phone is not in use or when its battery is low;
- Smart charging: AI can optimize the charging process by analyzing usage patterns and charging the phone when it's most convenient for the user. AI can also limit the charging duration and slow down the charging rate to conserve energy;
- Energy-efficient apps: AI can be used to optimize the performance of apps on mobile phones to reduce their energy consumption. This can include optimizing the code, reducing unnecessary background processes, and minimizing data usage;
- Virtual assistant: Virtual AI assistants are well-suited for conducting engaged conversations, acquiring and analyzing customer data, playing music, placing online orders, providing recommendations, and setting reminders, among other functions.

Accordingly, they can help users reduce energy usage by automating tasks that would typically use more energy. For example, a user can use a virtual assistant to set a reminder to turn off lights or adjust the thermostat.

Thus, AI can make mobile phone usage greener and more sustainable by reducing energy usage, optimizing power management, recycling and reusing components, and automating energy-saving tasks.

2.4 Challenges Associated with Integrating AI into Green Technologies

Implementing AI for greening ICT sector presents several challenges that must be overcome to realize the full potential of these technologies. We have analyzed them and distinguished into 6 main groups.

Firstly, it's data quality: to achieve optimal results, AI needs accurate and reliable data. In telecom technologies, this can be challenging as the data sets may be incomplete, inconsistent, or outdated. This can lead to inaccurate predictions and suboptimal performance.

Secondly, it's limited availability of data. Some telecom technologies may not have enough data to train AI models. For instance, emerging technologies, such as those using bio-based materials, may not have enough data for AI to make accurate predictions.

The third challenge can be limited computing power. Green technologies such as renewable energy systems may require large amounts of computing power to optimize their performance using AI. Achieving this can be costly and time-consuming, especially for small and medium-sized businesses or organizations.

Another obstacle can be complex modeling. Green technologies are often complicated and can require complex AI models to optimize their performance. These models can be challenging to develop, they require significant expertise in both the technology and AI.

Also, there are ethical and privacy concerns: AI used in green technologies may involve collecting significant amounts of personal data and tracking user behavior. This requires careful consideration of data privacy and ethical concerns.

Finally, it's integration with legacy systems. Many green technologies are integrated with legacy systems, making it difficult to integrate AI-powered solutions seamlessly. The integration process may require significant investment and time to ensure compatibility and avoid disruptions.

Consequently, integrating AI into telecom networks can deliver significant benefits in terms of sustainability and efficiency. However, it is important to address these challenges to maximize the potential of AI-powered green technologies.

3 Conclusions

In conclusion, the use of artificial intelligence (AI) technologies in green telecommunications has shown great potential to reduce the environmental impact of the industry. The article identified how AI can be used to optimize energy efficiency and reduce carbon footprint in networks and data centers, make usage of end-user devices more sustainable

and efficient. The research also includes analysis of the current limitations of AI application in green telecom technologies—challenges that need to be addressed, including data quality, limited availability of data, limited computing power, complex modeling, ethical and privacy concerns, and integration with legacy systems. Despite these challenges, we consider the potential of AI-powered green technologies to be significant, and we see continued research and development to be crucial in realizing the full potential of these technologies. By harnessing the power of AI, the telecommunications industry can play a vital role in mitigating the effects of climate change and creating a more sustainable future.


References

1. Intergovernmental Panel on Climate Change: Synthesis Report of the IPCC Sixth Assessment Report (AR6): Summary for Policymakers. Intergovernmental Panel on Climate Change (2023)
2. Intergovernmental Panel on Climate Change: Mitigation of Climate Change. Working Group III Contribution to the IPCC Sixth Assessment Report (2022)
3. Yousaf, L., Ge, S., Zeeshan, F., Salman, A., Muhammad, A.B.: Do financial development and energy efficiency ensure green environment? Evidence from R.C.E.P. economies. *Econ. Res. Ekon. Istraž.* **36**(1), 51–72 (2023). <https://doi.org/10.1080/1331677X.2022.2066555>
4. Paris Agreement to the United Nations Framework Convention on Climate Change. T.I.A.S. No. 16-1104, 12 Dec 2015
5. Fetting, C.: The European Green Deal ESDN Report. ESDN Office, Vienna (2020)
6. Zhang, X., Wang, Y.: How to reduce household carbon emissions: a review of experience and policy design considerations. *Energy Policy* **102**, 116–124 (2017). <https://doi.org/10.1016/j.enpol.2016.12.010>
7. Imam, N., Hossain, M., Saha, T.: Potentials and Challenges of Using ICT for Climate Change Adaptation: A Study of Vulnerable Community in Riverine Islands of Bangladesh (2017). https://doi.org/10.1007/978-3-319-56523-1_7
8. Freitag, C., Berners-Lee, M., Widdicks, K., Knowles, B., Blair, G.S., Friday, A.: The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. *Patterns* **2**(9), 100340 (2021). <https://doi.org/10.1016/j.patter.2021.100340>
9. International Telecommunication Union: Recommendation ITU-T L.1470 Greenhouse Gas Emissions Trajectories for the Information and Communication Technology Sector Compatible with the UNFCCC Paris Agreement (2020)
10. ITU, GESI, GSMA, SBTi: Guidance for ICT Companies Setting Science Based Targets (2020)
11. Khan, M.: AI-enabled transformations in telecommunications industry. *Telecommun. Syst.. Syst.* **82**, 1–2 (2023). <https://doi.org/10.1007/s11235-022-00989-w>
12. Morley, J., Widdicks, K., Hazas, M.: Digitalisation, energy and data demand: The impact of Internet traffic on overall and peak electricity consumption. *Energy Res. Soc. Sci.* **38**, 128–137 (2018). <https://doi.org/10.1016/j.erss.2018.01.018>
13. Huawei Homepage: <https://www.huawei.com/en/news/2022/10/intelligent-5g-mbbf-2022>. Last accessed 15 June 2023
14. Telefónica Homepage: <https://www.telefonica.com/en/communication-room/telefonica-drives-energy-consumption-optimisation-through-solutions-based-on-artificial-intelligence-and-machine-learning/>. Last accessed 15 June 2023

15. Rojek, I., JasiulewiczKaczmarek, M., Piechowski, M., Mikołajewski, D.: An artificial intelligence approach for improving maintenance to supervise machine failures and support their repair. *Appl. Sci.* **13**, 4971 (2023). <https://doi.org/10.3390/app13084971>
16. United Nations University Homepage: <https://unu.edu/press-release/global-e-waste-surg-ing-21-5-years>. Last accessed 15 June 2023
17. Mobile World Live Homepage: <https://www.mobileworldlive.com/huawei-updates/calvin-zhao-of-huawei-intelligent-bringing-intelligent-network-into-reality/>. Last accessed 15 June 2023
18. Balasooriya, P., Wibowo, S., Wells, M.: Green cloud computing and economics of the cloud. *J. Comput.* **5** (2016)
19. Andrae, A.S.G., Edler, T.: On global electricity usage of communication technology: trends to 2030. *Challenges* **6**(1), 117–157 (2015). <https://doi.org/10.3390/challe6010117>
20. IEA: Data Centres and Data Transmission Networks, IEA, Paris. <https://www.iea.org/reports/data-centres-and-data-transmission-networks>, License: CC BY 4.0 (2022)
21. Han, Z., Sun, X., Wei, H., Ji, Q., Xue, D.: Energy saving analysis of evaporative cooling composite air conditioning system for data centers. *Appl. Therm. Eng.* **186**, 116506 (2021). <https://doi.org/10.1016/j.applthermaleng.2020.116506>
22. Park, B.R., Choi, Y.J., Choi, E.J., Moon, J.W.: Adaptive control algorithm with a retraining technique to predict the optimal amount of chilled water in a data center cooling system. *J. Build. Eng.* **50**, 104167 (2022). <https://doi.org/10.1016/j.job.2022.104167>
23. Conti, G., Jimenez, D., del Rio, A., Castano-Solis, S., Serrano, J., Fraile-Ardanuy, J.: A multi-port hardware energy meter system for data centers and server farms monitoring. *Sensors* **23**(1), 119 (2023). <https://doi.org/10.3390/s23010119>
24. Tanveer, A., et al.: Artificial intelligence in sustainable energy industry: Status Quo, challenges and opportunities. *J. Clean. Prod.* **289**, 125834 (2021). <https://doi.org/10.1016/j.jclepro.2021.125834>
25. Sutton-Parker, J.: Is sufficient carbon footprint information available to make sustainability focused computer procurement strategies meaningful? *Proc. Comput. Sci.* **203**, 280–289 (2022). <https://doi.org/10.1016/j.procs.2022.07.036>



Neuropunk Revolution: Further Results

Max Talanov^{1,2} 

¹ Institute for Artificial Intelligence R&D, Novi Sad, Serbia

² ITIS, KFU, Kazan, Russia

max.talanov@gmail.com

Abstract. We use the term “neuropunk revolution” to identify the set of technologies based on the idea to use neuromorphic computing devices to compensate the pathological work of a patients nervous system. In this paper, I present approach for the distributed architecture of the implants infrastructure and further results of the “neuropunk revolution” technologies. Firstly I present one of the possible solution for the bypass problem to process the inbound data from the motor cortex of paralysed patient by real-time thalamus, brain stem and spinal cord models to orchestrate a spinal cord and muscles stimulation taking into account the feedback loop. Secondly I present the digital components interface to “speak the language” of the nervous system for the implementation of the neurointerface with real-time neurosimulation for spasticity compensation via the antagonist muscle/nerve stimulation. Thirdly I present the memristive implementation of the self-organising reservoir computing spiking neural network replicating the walking pattern recovery of a rat after complete spinal cord injury.

Keywords: Neurosimulation · Neuromorphic computing · Memristor

1 Introduction

This paper describes multidisciplinary approaches that we call “neuropunk revolution” in several domains: brain to computer interface (BCI), neurosimulations and artificial spiking neural networks implemented in memristive devices. Firstly starting from seminal works of Kevin Warwick [1–3] where he determined the main directions of the BCI: (1) integration of two nervous systems, he integrated invasively his nervous system with the nervous system of his wife; (2) extension of the nervous system with digital actuators, he managed the robotic hand with the neuronal activity from his hand through the Atlantic ocean; (3) extension of the nervous system with digital sensors, he plugged in the output of the rage finder into the his hand nerve. In addition want to mention the other work by the Nicolelis and Lebedev dedicated to the development of the specific neuronal circuitry in the monkey brain to control the robotic arm integrated into the sensory cortex, using the feedback loop implemented using electrical stimulation [4–6].

Close to the application domain of the article is the implementation of neurointerfaces and neuromodulation used for the neurorehabilitation of the complete spinal cord (SC) trauma published in works by Igor Lavrov and his team [7–10] and the group of Grégoire Courtine [11, 12].

Secondly huge domain of neurosimulations and it is quickly developing with two gigantic projects: the European – Human Brain Project [13] and the US BRAIN initiative. Both projects include: simulation and brain-inspired technologies. There are several neuronal models used in the neurosimulation field, Hodgkin-Huxley (H-H) [14] is the ubiquitous due to its nature as Lego blocks that provides the option to model a wide range of the neurobiological phenomena like the cell membrane, membrane potential, trans-membrane currents, receptor types conductance, refractory periods, etc. The disadvantage of the above mentioned flexibility in the significant computational burden of the H-H neuron even for the modern clusters. Several light-weight still bio-plausible models were developed: Izhikevich neuron [15, 16] and FitzHugh-Nagumo neuron [17]. Recently we published a new model that we called even simpler real-time neuron (ESRN), that combines real-time processing (here 5 ms of the simulation time must be calculated during 5 ms) [18] and bio-compatibility. I propose this model as the building block for the real-time neurosimulations of the “neuropunk revolution” approach.

Artificial neural networks are well known starting from works of Frank Rosenblatt and Marvin Minsky [19, 20] they exploit the primitive and really effective model of neuron for several traditional domains including computer vision [21] and natural language processing. Recently we see the rise of the scientific interest in the field of spiking neural networks (SNN) [22, 23] and neurohybrid technologies [24] integrating the artificial SNNs and neuromorphic devices like memristive devices with biological neuronal slices and cultures. This interest lies the base-ment for the neuromorphic systems development exploiting hybrid approaches to self-adopt for both traditional and medical domain including the possible integration with biological nervous system [25, 26].

2 SNNs with Real-Time Neurosimulations

I propose to use of the SNNs with real-time neurosimulations as the integration middleware between the biological objects (tissues and organs) and digital sensors and actuators (Fig. 1). This is one of the central ideas of “neuropunk revolution” presented earlier [27, 28]. An implementation of a part of the nervous system as well as neuro-messaging in the real-time provides the option to compensate the damaged or injured part of the nervous system with it’s model. Due to it’s nature the neurosimulation “speaks the language” of the nervous system and has self-organisation/self-learning option to reconstruct and adopt itself to the biological condition of particular patient [25].

The system reads its inputs from the nervous system (Fig. 1) using BCI and digital sensors later it processes the inbound data in the SNN with bio-compatible model whereas the output of the SNN could be implantable or non-invasive neuroprostheses as well as digital actuators, for example servomotors

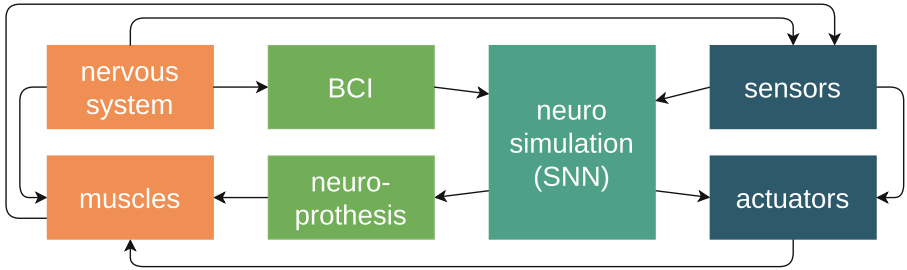


Fig. 1. The example of the real-time SNNs neurosimulation architecture. The SNN is used as middleware between the technologies: BCI and neuroprosthetics oriented to the biological objects and digital world: sensors and actuators. Legend: orange stands for the parts of a biological system, green – the interfaces to a biological system, teal – SNN bio-simulation technologies and blue stands for digital technologies.

of exoskeletons. Neuroprostheses could be implemented as the infrastructure of electrical and chemical stimulators implanted in a body or simply as noninvasive multichannel electrical stimulator.

3 Implanted Infrastructure

The future of the “neuropunk revolution” technologies lies in the integrated intra-body infrastructure and network communicating through ultra-sound and RF with chemical and electrical stimulators orchestrated through the hybrid databus working in real-time feedback loop with self-adopting SNNs. Self-learning as the part of the self-adaptation of biological neurons could be implemented via the memristive devices, this option was demonstrated in works of Victor Erokin as memristive synaptic prosthesis [29], later the the integration of biological neurons with the memristive array was demonstrated in the perspective [24]. I see that further development of memristive devices can provide the energy effective hardware seamlessly integrated with biological nervous system.

The communication and management of the implanted computational infrastructure could be the implemented as the FPGA/ASIC and/or memristive devices communicating with each other organising the implanted and orchestrated network in the body of a patient (Fig. 2). I propose to use several implanted neurosimulation models: (1) brain model that could be used to process the brain data locally and later transfer it to further neurosimulations, (2) distributed SC model as the middleware processing the inbound data from the brain model to translate into SC and muscle electrical activity. Both models are proposed to be implemented as real-time bio-plausible reservoir computing (RC) SNNs in and implemented in FPGA/ASIC/memristive technology.

The complete architecture of the implanted infrastructure is presented in Fig. 2. Firstly implanted sensors read data from the brain, later SNN neurosimulation (brain model) process the inbound data and generates the stimulation

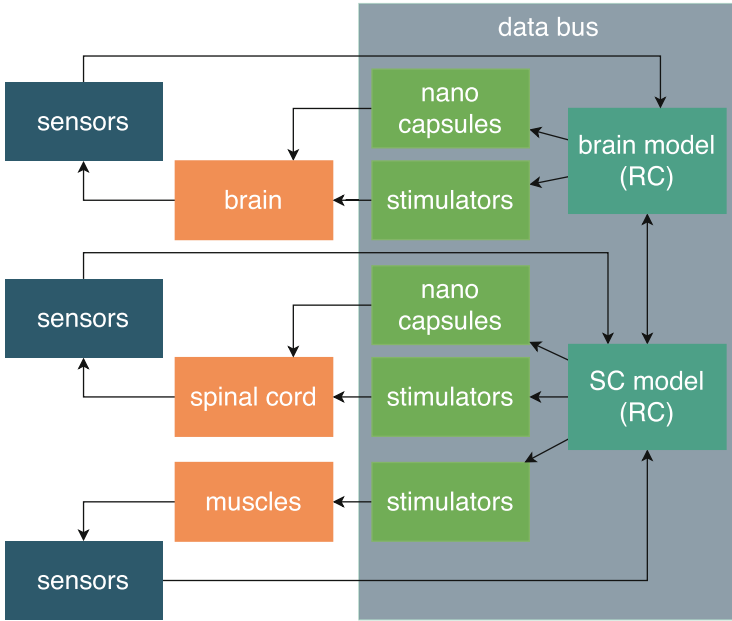


Fig. 2. Implanted infrastructure high-level design. The system reads the electrical activity of the brain using implanted set of sensors; neurosimulation in SNN RC manages brain electrical and chemical (nanocapsules) stimulators in a feedback loop. The brain neurosimulation model is connected via hybrid data bus to the SC model that manages chemical and electrical stimulation of the SC and muscles with the feedback implemented via implanted sensors. Legend: orange stands for the parts of a biological system, green – the interface to a biological system, teal – bio-simulation SNN RC technologies and blue stands for digital technologies.

electrical and chemical profiles applied to electrical stimulators and nanocapsules [30] forming first level of the feedback loop. There are set of the sensors that could be useful in this context including electroencephalography (EEG) and concentration sensors. Secondly the generated data of the brain model, possibly the part of the brain needed for the real-time processing, is transmitted via hybrid data bus (described below) and processed by the SC model for the low level operation, for example, muscles. This model could be implemented in the distributed computational infrastructure of SNNs. Models (teal) and stimulators (green) use the orchestration of the intra-body/transcutaneous (IBT) data bus. The communication of the IBT data bus could be implemented in the hybrid way: (1) using the transcutaneous RF wireless channels or (2) using ultrasound communication channels for intra-body communication. Thirdly the distributed SC model orchestrates the set of chemical (nanocapsules) and electrical stimulators that triggers the neuronal activity in the SC and muscles of a patient. Fourthly sensors mounted close to the spinal cord and muscles sensing concentra-

tion of chemical stimulators and electrical neuronal and electrical activity close the feedback loop of the SC model (second feedback loop).

Due to the extended feedback loops the SNNs with neurosimulation models could exploit reservoir computing approaches with self-organising memristive devices [25, 26].

4 Bypass Problem

One of the possible application of the “neuro punk revolution” approach could be the bypass problem solution presented in Fig. 3.

In the case of spinal cord injury (SCI) there is no signal transmission between the brain and the part of the spinal cord below the trauma. The current medical neuro-rehabilitation approach is using implanted electrical epidural stimulation arrays (EES) [10] that trigger the neuronal activity in the SC below the spinal

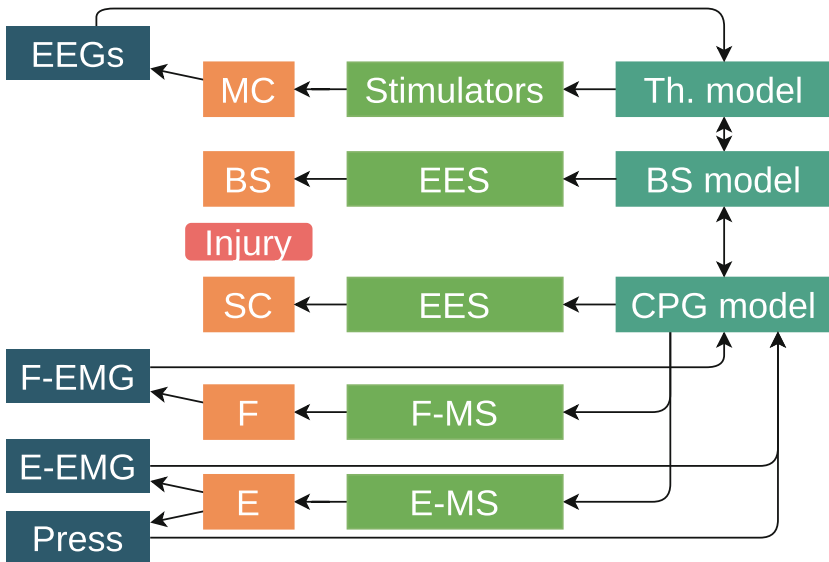


Fig. 3. High level design of the bypass solution. Implanted EEG sensors transmit the data to the thalamus (Th.) + brain stem (BS) models implemented in RC SNNs, the feedback loop closes via EES and network of stimulators of the BS and motor and sensory cortex (MC). Below the SC injury the distributed spinal CPG models near the SC and limb muscle groups process the inbound data of the BS model and orchestrate the epidural electrical stimulation as well as flexor and extensor muscles electrical stimulators. The feedback is organised using the EMG and pressure (Press) as well as joint angles sensors (not shown) to CPG models implemented using memristive technologies. Legend: orange stands for the parts of a biological system, green – the interface to a biological system, teal – bio-simulation SNN RC technologies and blue stands for digital technologies.

cord trauma. The strait-forward approach to read the data of neuronal activity from the motor cortex (MC) and transmit it directly to muscles or motor neurons of the SC does not take into account the spinal neuronal circuitry thus seems to be very limited. Here I propose the application of the SNN real-time neurosimulations of the thalamus (Th.) and brain stem (BS) and spinal central pattern generator (CPG) integrated to implement bio-plausible bypass around SCI. The BS model is responsible for the processing the readouts from the MC and possibly stimulation of the SC above the injury; CPG model is responsible for the low level processing including motor neuronal outputs to muscles via stimulators: flexor (F-MS), and extensor (E-MS). In case of SCI I see important role of the adaptability of the CPG model using STDP and synaptogenesis automatically compensating the damaged part of the SC [25]. The CPG model could be implemented as the reservoir computing model having its inputs from the electro myogram (F-EMG, E-EMG), pressure sensors and outputs to flexor (F) and extensor (E) muscles thus closing smaller feedback loop. The ascending projections from SC through BS and Th. to the brain sensory and MC and later to the model of Th. and the SC, thus closing the bigger feedback loop. Overall adaptability of the proposed model should guarantee the automatic compensation of wide range of injuries and neurodegenerative deceases including: complete SCI, incomplete SCI, palsy, stroke etc.

5 Implemented Cases

5.1 SNN Neural Interface to Compensate Spasticity

In 2022 we published results of implementation of the neurointerface with the real-time SNN models as the middleware [31]. We have implemented the experimental setup with of oscillator motifs (OMs) [32], using the ESRN [18]. As the result of experimental work we indicated the highest motor response of finger displacement using 5 OMs comparing to the 20 Hz and 40 Hz (traditional to neurorehabilitation domain) with lower discomfort rates for the OM generated electrical activity of the neurointerface.

Later in 2023 we presented work using the above described approach for the spasticity compensation device [33]. Experimentally we determined the effect of the spinal cord reciprocal inhibition of the agonist muscle with the use of the electrical stimulation of the antagonist nerve or muscle to compensate for the “spastical” neuronal activity. We used both muscles and ulnar nerve stimulation with different delays (0–20 ms, step 5 ms) to simulate the “spastical” neuronal activity in healthy volunteers. The ulnar nerve innervates the antagonist muscle (flexor) and Ia afferents inhibiting the motor neuronal pool of the agonist “spastical” muscle (extensor). In case of nerve and muscle stimulation we indicated the significant reduction of the “spastical” (extensor) muscle force the strongest decrease from 0.94 ± 0.59 kgf down to 0.19 ± 0.09 kgf was registered in the experiment with a 20 ms delay between stimulations of antagonist muscles, the discomfort rate was 2.4 ± 1.1 out of 10. During the ulnar nerve stimulation, the most significant decrease of the muscle force was registered during 10 ms delay

between the nerve and the extensor muscle stimulations from 1.19 ± 0.23 kgf to 0.31 ± 0.17 kgf, the discomfort rate was 3.4 ± 2.9 out of 10. As the result we indicated the effusiveness of the real-time SNN with bio-plausible model for the inhibition of the spastical neuronal activity.

5.2 Memristive Self-organising Device

In 2023 we have published ground breaking work dedicated to the use of the memristive deices to replicate the neurorehabilitation process. Due to the self-learning properties of the developed memristive neurons we indicated the reconstruction of the electrical activity resembling walking pattern observed in rats during the neurorehabilitation after the complete SCI. During the experiments we used only 4 neurons with 2 of them with memristive synapses. After the training process the resistance of the memristive devices due to the simulated feedback from the insole operated by the simulated muscle we observed the self-organisation of the trapezoidal pattern specific for the recovery of the walking pattern after the complete SCI in rats [25].

6 Conclusion

In this paper we presented the approach to solve the bypass problem using the distributed infrastructure of the SNN RC implants. Further I have presented several examples of the use of the “neuropunk revolution” approach for the medical cases.

First case is the compensation of the spastic muscles activity where we indicated the significant decrease the muscle force via stimulation of the antagonist muscles or ulnar nerve using the bio-plausible pattern generated SNNs.

During the second case we have indicated the self-organisation of the walking pattern in the memristive schematic replicating the recovery of the electrical activity of the SC during the rats neuro-rehabilitation.

Acknowledgement. This paper has been partly supported by the Kazan Federal University Strategic Academic Leadership Program (“PRIORITY-2030”).

References






1. Kevin, W., Dimitris, X., SlawomirJ, N., VictorM, B., MarkW, H., Julia, D., et al.: Controlling a mobile robot with a biological brain. *Def. Sci. J.* **60**(1), 5–14 (2010)
2. Kevin, W., Mark, G., Benjamin, H., Iain, G., Peter, K., Brian, A., et al.: The application of implant technology for cybernetic systems. *Arch. Neurol.* **60**(10), 1369 (2003)
3. Kevin, W.: Superhuman enhancements via implants: beyond the human mind. *Philosophies* **5**(3), 14 (2020)
4. MikhailA, L., MiguelAL, N.: Brain-machine interfaces: From basic science to neuroprostheses and neurorehabilitation. *Physiol. Rev.* **97**(2), 767–837 (2017)

5. Nicolelis, M.A. and Lebedev, M.A.: Principles of neural ensemble physiology underlying the operation of brain-machine interfaces. *Nat. Rev. Neurosci.* **10**(7), 530–540 (2009)
6. Wessberg, J., Stambaugh, C.R., Kralik, J.D., Beck, P.D., Laubach, M., Chapin, J.K., Kim, J., Biggs, S.J., Srinivasan, M.A., Nicolelis, M.A.: Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature* **408**(6810), 361–365 (2000)
7. Lavrov, I., Dy, C.J., Fong, A.J., Gerasimenko, Y., Courtine, G., Zhong, H., et al.: Epidural stimulation induced modulation of spinal locomotor networks in adult spinal rats. *J. Neurosci.* **28**, 6022–6029 (2008)
8. Gill, M., Linde, M., Fautsch, K., Hale, R., Lopez, C., Veith, D., Calvert, J., Beck, L., Garlanger, K., Edgerton, R., Sayenko, D.: Epidural electrical stimulation of the lumbosacral spinal cord improves trunk stability during seated reaching in two humans with severe thoracic spinal cord injury. *Front. Syst. Neurosci.* **14** (2020)
9. Igor, L., YuryP, G., RonaldoM, I., Gregoire, C., Hui, Z., RolandR, R., et al.: Plasticity of spinal cord reflexes after a complete transection in adult rats: relationship to stepping ability. *J. Neurophysiol.* **96**(4), 1699–1710 (2006)
10. Gill, M.L., Grahn, P.J., Calvert, J.S., Linde, M.B., Lavrov, I.A., Strommen, J.A., Beck, L.A., Sayenko, D.G., Van Straaten, M.G., Drubach, D.I. and Veith, D.D.: Neuromodulation of lumbosacral spinal networks enables independent stepping after complete paraplegia. *Nat. Med.* (2018)
11. FabienB, W., JeanBaptiste, M., CamilleGLe, G.-M., Robin, D., Salif, K., Marco, C., et al.: Targeted neurotechnology restores walking in humans with spinal cord injury. *Nature* **563**(7729), 65 (2018)
12. Capogrosso, M., Wenger, N., Raspopovic, S., Musienko, P., Beauparant, J., Bassi Luciani, L., et al.: A computational model for epidural electrical stimulation of spinal sensorimotor circuits. *J. Neurosci.* **33**(49), 19326–19340 (2013)
13. Human Brain Project (2019). Page Version ID: 889932985
14. Hodgkin, A.L., Huxley, A.F.: A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* **117**(4), 500–544 (1952)
15. Izhikevich E.M.: Which model to use for cortical spiking neurons? *IEEE Trans. Neural Netw.* **15**(5) (2004)
16. Eugene, I.: Polychronization: computation with spikes. *Neural Comput.* **18**(2), 245–282 (2006)
17. Richard, F.H.: Impulses and physiological states in theoretical models of nerve membrane. *Biophys. J.* **1**(6), 445–466 (1961)
18. Leukhin, A., Talanov, M., Suleimanova, A., Toshev, A., Lavrov, I.: Even simpler real-time model of neuron. *BioNanoScience* 1–4 (2020)
19. Frank, R.: The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**(6), 386 (1958)
20. Minsky, M.L, Papert, S.A.: *Perceptrons: expanded edition* (1988)
21. Zhang, J., Dong, B., Zhang, H., Ding, J., Heide, F., Yin, B., Yang, X.: Spiking Transformers for event-based single object tracking, pp. 8801–8810 (2022)
22. Sengupta, A., Ye, Y., Wang, R., Liu, C., Roy, K.: Going deeper in spiking neural networks: VGG and residual architectures. *Front. Neurosci.* **13** (2019)
23. Nøkland, A.: Direct feedback alignment provides learning in deep neural networks (2016). [ArXiv:1609.01596](https://arxiv.org/abs/1609.01596)
24. Mikhaylov, A., Pimashkin, A., Pigareva, Y., Gerasimova, S., Gryaznov, E., Shchanikov, S., Zuev, A., Talanov, M., Lavrov, I., Demin, V., Erokhin, V.: Neuro-

- hybrid memristive CMOS-integrated systems for biosensors and neuroprosthetics. *Front. Neurosci.* **14** (2020)
25. Mashev, D.N., Suleimanova, A.A., Prudnikov, N.V., Serenko, M.V., Emelyanov, A.V., Demin, V.A., Lavrov, I.A., Talanov, M.O., Erokhin, V.V.: Memristive circuit-based model of central pattern generator to reproduce spinal neuronal activity in walking pattern. *Front. Neurosci.* **17** (2023)
 26. Matsukatova, A.N., Prudnikov, N.V., Kulagin, V.A., Battistoni, S., Minnekhanov, A.A., Trofimov, A.D., Nesmelov, A.A., Zavyalov, S.A., Malakhova, Y.N., Parmegiani, M., Ballesio, A.: Combination of organic-based reservoir computing and spiking neuromorphic systems for a robust and efficient pattern classification. *Adv. Intell. Syst.*
 27. Max, T., Neuropunk revolution: preliminary results. In: 6th Scientific School Dynamics of Complex Networks and their Applications (DCNA), vol. 2022, pp. 274–277. IEEE (2022)
 28. Talanov, M., Vallverdu, J., Adamatzky, A., Toshev, A., Suleimanova, A., Leukhin, A., Pozdeeva, A., Mikhailova, Y., Rodionova, A., Mikhaylov, A., Serb, A.: Neuropunk revolution. *Hacking Cognitive Systems Towards Cyborgs 3.0*. arXiv preprint (2022)
 29. Juzekava, E., Nasretdinov, A., Battistoni, S., Berzina, T., Iannotta, S., Khazipov, R., Erokhin, V., Mukhtarov, M.: Coupling cortical neurons through electronic memristive synapse. *Adv. Mater. Technol.* (2018)
 30. Svetlana, E., Laura, P., Vladimir, S., Victor, E.: Nanoengineered Polymeric Capsules for Bio-computing. Rhodes, Greece (2015)
 31. Talanov, M., Suleimanova, A., Leukhin, A., Mikhailova, Y., Toshev, A., Milit-skova, A., Lavrov, I., Magid, E.: Neurointerface implemented with oscillator motifs. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4150–4155. IEEE (2021)
 32. Talanov, M., Leukhin, A., Suleimanova, A., Toshev, A., Lavrov, I.: Oscillator motif as design pattern for the spinal cord circuitry reconstruction. *BioNanoScience* (2020)
 33. Mikhailova, Y., Pozdeeva, A., Suleimanova, A., Leukhin, A., Toshev, A., Lukmanov, T., Fatyhova, E., Magid, E., Lavrov, I., Talanov, M.: Neurointerface with oscillator motifs for inhibitory effect over antagonist muscles. *Front. Neurosci.* **17** (2023)



Data Preparation for Advanced Data Analysis on Elastic Stack

M. S. Ulizko^{1,2} , R. R. Tukumbetova¹ , A. A. Artamonov¹ ,
E. V. Antonov¹ , and K. V. Ionkina¹ 

¹ Plekhanov Russian University of Economics, Stremyanny Pereulok, 36, 115093 Moscow, Russia

{mulizko, rrtukumbetova}@kaf65.ru

² National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashira Hwy, 31, 115409 Moscow, Russia

Abstract. This paper presents approaches for preparing different types of data to be loaded into the document-oriented NoSQL Elasticsearch database. The considered database allows not only to store data, but also provides an opportunity to use Kibana, data visualization utility, which is a powerful tool for data analysis. The task of preprocessing is essential, because well-prepared data not only allows you to increase the accuracy of the analysis, but also expand its capabilities. For more coverage, the approaches are described with the use of real cases that have been solved by analysts. The paper presents methodological and practical ways to solve problems both by transforming the data and adding new fields, and by correctly mapping for Elasticsearch indexes. For a clear demonstration of the approaches, their practical application is given on the example of two datasets with bibliographic information on papers and information on funding of scientific and technical projects. The demonstration shows the difference between initial and enriched data, as well as the charts built by working with the data, which enables advanced data analysis.

Keywords: Elasticsearch · Kibana · ETL · Mapping

1 Introduction

Nowadays, a conventional process of data processing consists of ETL-process (Extract, Transform, Load) and subsequent analysis. This paper discusses the problem of data preparation and database customization for more thorough analysis using the No-SQL Elasticsearch database as an example.

Since real data is sometime incomplete, inconvenient and noisy, all data processing pipelines include data preprocessing which makes initial data suitable for analysis. It is vital in AI and NLP tasks where the models' accuracy depends on the quality of data. However, data preprocessing is important for No-SQL databases.

Elasticsearch is an open search and analytics system based on the Apache Lucene high-performance full-text search library [1], based on the term inverted index. Elasticsearch is often considered along with the Logstash data pipeline and the Kibana visualization platform, forming a component called ELK.

Elasticsearch is a NoSQL document-oriented database, in terms of which the database is called an index. In addition, Elasticsearch is a schemaless database, for which it is sufficient to send a JSON object as input data, which will be correctly processed and added to the index [2]. However, you can explicitly set an index schema by defining index settings (number of index parts, backups, text analyzers, etc.) and fields (type, format, analyzers used, etc.).

Elasticsearch is a popular solution for data storage and processing, as evidenced by the 512 works indexed by Scopus. Most papers explain the choice of Elasticsearch [3], and the ELK set is used as a tool, as in [4–6], without defining either the index settings or the fields. If some ELK settings are given in the article, it is mostly related to the architecture [3, 7, 8] or examples [8, 9, p. 114].

This is largely due to the fact that ELK has detailed documentation, including settings, which allows users to deploy this software independently [10]. Among the scientific papers it is worth to mention the following ones, which cover the field settings (so-called mapping). Walter-Tscharf [11] compares the definition of date and text fields in terms of further analysis in Elasticsearch. Rosenberg et al. [12] describe the main advantages of an inverted index and apply them to data storage, dividing text fields into those supporting filtering and parsing (type text) and those not (type keyword) or using tokenization as n-gram, which gives good results in file name processing and reindexing. Kim and Cho [13] also take advantage of Elasticsearch, using Chinese, Japanese and Korean analyzers, improving the accuracy and performance of text processing. The most detailed guidelines for mapping are described in Bajer [1], which can be of use in developing an efficient repository, but these guidelines only describe numeric fields and distinguish between text and keyword types.

Thus, typically Elasticsearch descriptions are usually limited to the architectural solution, being occasionally added with recommendations to field naming. In other words, setting up Elasticsearch indexes is more of engineering work, which exploits the techniques that have been developed in the past. This work fills a niche and provides a set of recommendations that can be useful for researchers to create indexes that solve a wide range of problems.

Let us assume that there is a set of N documents, each of which has an age field, and when you load the data into an Elasticsearch index, the first document's age is written as the string "18", and the second document's age is written as number 23. If a schema for the index is not provided, then as the data is loaded, for the first document Elasticsearch determines that the age field is of type text. However, because the age field in the second document does not match the text type, Elasticsearch will report an error (for a Python client, this is the `BulkIndexError` with the comment "failed to parse field [age] of type [text]...").

Let us consider how to solve this and other problems at the stage of creating an index using two datasets: the bibliographic information of the list of publications of the Joint Institute for Nuclear Research, published in journals indexed in the peer-reviewed Scopus database [14] and the National Institutes of Health projects, published in NIH RePORTER, a repository of research projects funded by NIH [15].

2 Methodology

Data preprocessing is essential part of data process pipeline. It contains many aspects, which can be divided into the following [16–20]:

1. Data cleaning:
 - Filtering noisy data (low-pass filter, Butterworth filter, etc.);
 - Missing values imputation techniques (mean-value imputation, etc.).
2. Dimension-based transformation:
 - Different sampling methods for objects (rows);
 - Changing features (columns): remove redundant features (based on Pearson correlation coefficient) or add features.
3. Data transformation:
 - Transformation of categorical features to numerical (dummy-coding, text vectorization, etc.);
 - Data normalization (scale data into similar ranges: min-max normalization, etc.).
4. Outlier detection/anomaly detection.
5. Data unification (mostly for relational databases).
6. Etc.

Elasticsearch is a schemaless database, for which the field type can be defined by the database itself as the data is loaded. This property reduces the efforts for developing your own repository. However, successful and correct identification of the data type requires thorough data preparation (Transform step in ETL-notation), which is related to the following fact.

Since the stored data are needed to perform statistical analysis rather than to build models with accuracy estimates, preprocessing the data does not require all of the above transformations. To be more precise, the preprocessing in the paper consists of unifying the data and adding new features.

2.1 Case 1. Different Date Formats

Elasticsearch supports different data types, including the date type. If you have fields of this type in your data, it is recommended to add them to mapping for correct indexing, since Elasticsearch allows you to use different time formats.

Most of the time there are no problems with dates, because usually the format of date representation is preserved in the initial dataset. In the case of different formats in the initial dataset, the following situations are possible.

- The data is distinct and clear. For example, the date is represented by the formats “dd-MM-yyyy” and “dd-MM-yyyyTHH:mm:ss”. In this case, to unify data you can use two date formats in mapping, or you can perform preprocessing and date-time conversions to one format.

- The data is ambiguous. For example, the date is represented by the formats “dd-MM-yyyy” and “MM-dd-yyyy”. In this case, the string “01-02-2022” can be either the 2nd of January 2022 or the 1st of February 2022. Different date parsers often use the American representation of dates in the format “MM-dd-yyyy”, so their use should be controlled. A possible way to overcome this problem is to use statistical analysis to determine the more frequent format.

2.2 Case 2. Data Conversion. CSV to JSON

Let us consider a situation where it is necessary to create an index using a dataset in CSV format. The file contains three list fields: keywords, authors’ names and authors’ IDs (see Table 1).

Table 1. Data representation in CSV file

Keywords	Author’s name	Author’s ID
Visualization, Pills, Healthcare	Lue L., Scott W., Willson G	1, 2, 3

Since Elasticsearch uses JSON as the data format, the mentioned list needs to be converted to a dictionary type therefore we need to transform data. For a logically correct and intuitive data representation it is necessary to convert the initial list fields to the following format (see Table 2).

Table 2. Data representation in JSON format

Data
<pre>{ "keywords": ["Visualization", "Pills", "Healthcare"], "authors": [{ "name": "Lue L.", "id": 1 }, { "name": "Scott W.", "id": 2 }, { "name": "Willson G.", "id": 3 }] }</pre>

2.3 Case 3. Geographic Coordinates

One of the features of Elasticsearch and Kibana is the possibility to build maps with objects based on their geographical position. Therefore, it is possible to operate with geographical coordinates. First of all it is important to mention that such fields are specified in mapping, because without that in most cases they are recognized as separate fields of float type, which makes it impossible to build the maps. It is worth noting that geographic coordinates are rarely met in datasets, so it is necessary to add the feature “geolocation”. Let us consider one of the possible ways to extract geographic coordinates.

For instance, let us have a dataset with bibliographic information on articles from Scopus containing the following fields: title, information about the authors, affiliations, abstract, keywords, and list of references. Let us assume that in order to perform the analysis, it is necessary to group the papers by organizations to which authors are affiliated and by the countries of these organizations. In practice, it turns out that authors write the affiliations in multiple different ways, so grouping by the full name of the affiliation is not applicable, so that unification of affiliations is necessary.

Let’s consider the algorithm for determining the unified organization name using the string “Bogoliubov Lab. of Theor. Physics, Joint Institute for Nuclear Research, 141980 Dubna, Moscow Region, Russian Federation”. First of all, let’s distinguish the elements present in all affiliations: official name of the organization (“Joint Institute for Nuclear Research”) and country (“Russian Federation”). Using the official name and country, it is possible to determine a unified name (“Joint Institute for Nuclear Research”), which can be used to determine geographical coordinates using geocoding tools. For example, a possible way to transform the string with author affiliation could be as follows (see Table 3).

Table 3. Data enrichment result

Initial string with affiliation	Enriched data
Bogoliubov Lab. of Theor. Physics, Joint Institute for Nuclear Research, 141980 Dubna, Moscow Region, Russian Federation	<pre> { "name": "Bogoliubov Lab. of Theor. Physics, Joint Institute for Nuclear Research, 141980 Dubna, Moscow Region, Russian Federation", "official_name": "Joint Institute for Nuclear Research", "department": "Bogoliubov Laboratory of Theoretical Physics", "country": "Russia", "country_code": "RU", "location": { "lat": 56.74646, "lon": 37.189408 } } </pre>

It is worth noting that this method may not recognize all affiliations correctly. For example, if the organization in one string is written in English, in another - in the language of the country in which the organization is located. Also, the coordinates for one organization may differ slightly. In this case, before loading the data, it is necessary to make a set of all the organizations with the coordinates and 'glue' all the organizations that are located at a distance of less than 200 m to each other. The distance is conditional, due to the fact that organizations are rarely located closer to each other than the specified distance.

For geographical coordinates, Elasticsearch provides the field type 'geo_point'. It is worth noting that defining a location for organizations not only enriches the initial data set, but also makes grouping by organizations more accurate due to data unification.

2.4 Case 4. Enrichment with Custom Quantitative Fields

While analyzing Scopus papers, let us assume that there is a need to look at the collaborative work of authors from different countries, for example, to understand what countries/organizations a certain country/research team cooperates with, how many papers were written by authors from one, two, three countries, etc. It is quite complicated and laborious to find a solution to these problems using the filtering tools in Kibana. Whereas, data preprocessing makes it possible to add field with the number of unique countries where organizations are located (it's more like metadata than an individual attribute). Adding such a field allows the analyst to determine the number of publications written by authors from the same country, organizations from two different countries that are in close collaboration with each other (see Fig. 1).

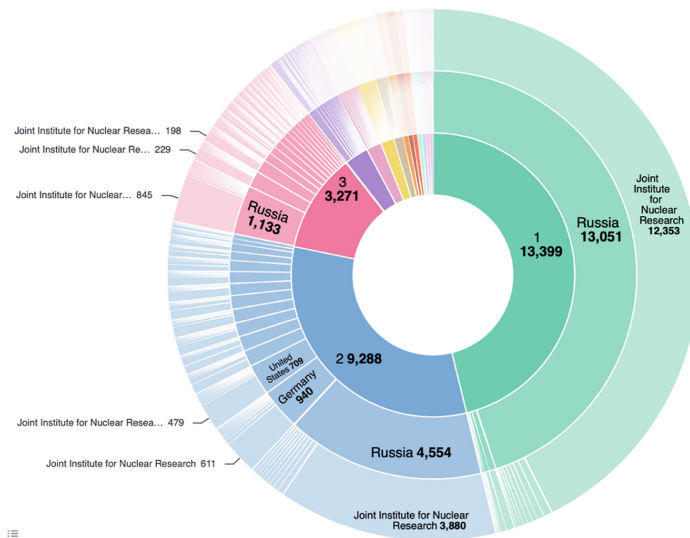


Fig. 1. Donut chart with three layers: distribution of articles by the number of unique countries in which the articles were written; distribution of articles by country, distribution of articles by organization

3 Results and Discussion

The presented methods were put into practice on the bibliographic information of the Joint Institute for Nuclear Research publications list indexed by Scopus and NIH projects published in NIH RePORTER.

In the former case, 36,785 documents in JSON format were formed. The initial document and the processed document in truncated form (without unchanged fields) are presented below (see Table 4).

Table 4. Data enrichment result

Initial document	Processed document
<pre>{ "all_affiliations": [{ "affiliation_name": "Bogolubov Laboratory of Theoretical Physics, Joint Institute for Nuclear Research, Dubna, 141980, Russian Federation", "affiliation_key": "a" }], { "affiliation_name": "Institut de Physique Théorique, Université catholique de Louvain, Louvain-La- Neuve, 1348, Belgium", "affiliation_key": "b" } }] }</pre>	<pre>{ "all_affiliations": [{ "affiliation name": "Bogolubov Laboratory of Theoretical Physics, Joint Institute for Nuclear Research, Dubna, 141980, Russian Federation", "affiliation key": "a", "name": "Joint Institute for Nuclear Research", "country": "Russia", "location": { "lat": 56.74646, "lon": 37.189408 }, "lab": "Bogoliubov Laboratory of Theoretical Physics" }], { "affiliation name": "Institut de Physique Théorique, Université catholique de Louvain, Louvain-La-Neuve, 1348, Belgium", "affiliation key": "b" }], "unique_countries": ["BE","RU"], "unique_countries_fullname": ["Belgium","Russia"], "countries num": 2, }</pre>

It should be noticed that one organization out of two is defined for the above document. This is due to the fact that it's not always possible to determine the organization by affiliation. If completeness rather than accuracy is important, you can use such applications as Google API, Yandex API, ChatGPT, etc. If you use the Google API or the Yandex API to search for organizations, you might get irrelevant results. When using ChatGPT the answers to the same questions often differ, so the use of such chatbots with artificial intelligence is very limited.

To analyze the outlined data, a dashboard in Kibana was built. It consists of 20 visualizations, part of which is presented below (see Fig. 2). Note that 9 of them were built using enriched data. For example, the following statistics and charts are presented: papers, written by authors from different countries, statistics on the number of publications for unified names of organizations, and a map by publications.

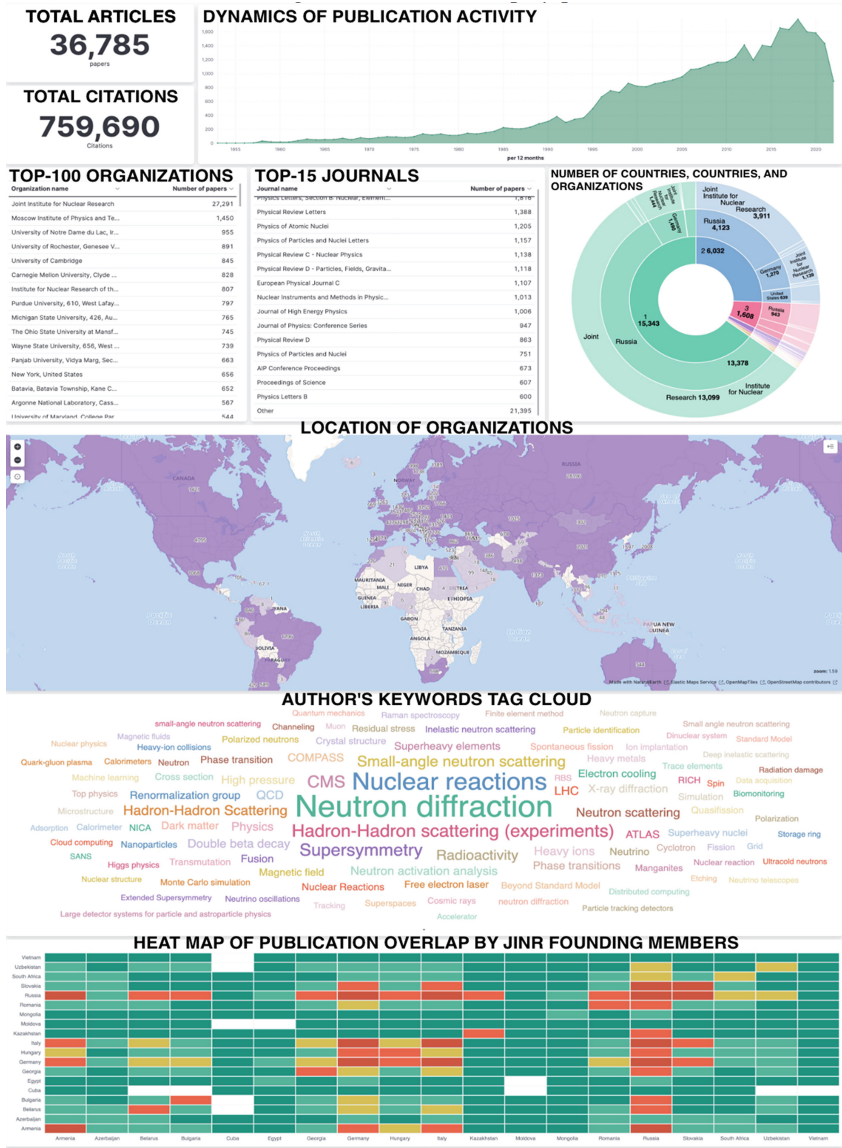


Fig. 2. Dashboard on scientific papers of JINR

The visualizations revealed two problems. Firstly, it is necessary to lowercase keywords. This can be achieved both by preprocessing the data and by using Elasticsearch analyzers. Secondly, a more thorough processing of the affiliations need to be processed more thoroughly by the application of various advanced methods, which is going to be described in another article in detail.

The second dataset is information on NIH-sponsored projects, with a total number of more than 1.5 million projects from 2000 to 2021. The website provides datasets by year in CSV.

The following peculiarities are noticed in these files:

1. Some list fields are written as a string, so they need to be converted to a list.
2. Project lead researchers are represented by two fields: ID and name. As a project may have several researchers, it is necessary to form a dictionary for each one with an ID and a name according to the above scheme.
3. Projects have five fields of date type. One of them has the same format as Elasticsearch's default format. The remaining four fields are of two different formats, and the record format is the same within a year. Therefore, when creating an index, it is necessary to explicitly specify the format used for these fields, and when loading the data, it is also necessary to perform date conversions to the required format.

4 Conclusion

Data preprocessing is an integral and vital step in data analysis. High-quality preprocessing not only eliminates data loading errors, but also widens the analysis capabilities. With Elasticsearch's non-relational database, data preprocessing goes hand-in-hand with mapping, which allows you to create a truly worthwhile database.

In the process of creating indexes there can be different situations: working with dates, peculiarities of CSV to JSON conversion, data unification, data enrichment (defining geographic coordinates, counting quantitative characteristics), etc. The work shows that for Elasticsearch the work on these aspects allows to do the operations which are impossible directly: the construction of the maps or the nested pie charts, etc.

The paper presents situations that have occurred during authors' work, some cases may have been left out of consideration. If such situations are identified, they will be considered in other papers and reports.

Acknowledgements. The study was supported by the Russian Science Foundation grant No. 23-75-30012, <https://rscf.ru/project/23-75-30012/>.






References

1. Bajer, M.: Building an IoT Data Hub with Elasticsearch, Logstash and Kibana. In: 2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), pp. 63–68. IEEE (2017)
2. Talas, A., Pop, F., Neagu, G.: Elastic stack in action for smart cities: making sense of big data. In: 2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 469–476. IEEE (2017)

3. Shah, N., Willick, D., Mago, V.: A framework for social media data analytics using Elasticsearch and Kibana. *Wireless Netw.* **28**(3), 1179–1187 (2018)
4. Lahmadi, F. Beck, Finickel, E., Festor, O.: A platform for the analysis and visualization of network flow data of android environments. In: 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM), pp. 1129–1130. IEEE (2015)
5. Barakhnin, V., Kozhemyakina, O., Mukhamedyev, R., Borzilova, Y., Yakunin, K.: The design of the structure of the software system for processing text document corpus. *Bus. Inform.* **13**(4), 60–72 (2019)
6. Zamfir, V.-A., Carabas, M., Carabas, C., Tapus, N: Systems monitoring and big data analysis using the Elasticsearch system. In: 2019 22nd International Conference on Control Systems and Computer Science (CSCS). IEEE (2019)
7. Haugerud, H., Sobhie, M., Yazidi, A.: Tuning of elasticsearch configuration: parameter optimization through simultaneous perturbation stochastic approximation. *Front. Big Data* **5**, 686416 (2022)
8. Ngo, T.T.T., Sarramia, D., Kang, M.-A., Pinet, F.: A new approach based on ELK stack for the analysis and visualisation of geo-referenced sensor data. *SN Comput. Sci.* **4**(3), 241 (2023)
9. Hunter, T.: *Advanced Microservices: A Hands-on Approach to Microservice Infrastructure and Tooling*. Apress, Berkely, CA, USA (2017)
10. Elastic: <https://www.elastic.co/guide/en/elasticsearch/reference/8.7/mapping.html>. Last accessed: 24 Apr 2023
11. Walter-Tscharf, F.F.W.V.: Indexing, clustering, and search engine for documents utilizing Elasticsearch and Kibana. In: *Mobile Computing and Sustainable Informatics*, pp. 897–910 (2022)
12. Rosenberg, J., Coronel, J.B., Meiring, J., Gray, S., Brown, T.: Leveraging Elasticsearch to improve data discoverability in science gateways. In: *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning)*, vol. 19, pp. 1–5. ACM (2019)
13. Kim, K.-J., Cho, Y.-B.: Improving elasticsearch for Chinese, Japanese, and Korean text search through language detector. *J. Inform. Commun. Converg. Eng.* **18**(1), 33–38 (2020)
14. Scopus Homepage. <https://www.scopus.com>. Last accessed 16 May 2023
15. NIH RePORTER Homepage. <https://reporter.nih.gov>. Last accessed 16 May 2023
16. Agarwal, V.: Research on data preprocessing and categorization technique for smartphone review analysis. *Int. J. Comput. Appl.* **131**(4), 30–36 (2015)
17. García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J.M., Herrera, F.: Big data preprocessing: methods and prospects. *Big Data Analytics* **1**(1), 9 (2016)
18. Fan, C., Chen, M., Wang, X., Wang, J., Huang, B.: A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Front. Energy Res.* **9**, 652801 (2021)
19. Al-Jabery, K.K., Obafemi-Ajayi, T., Olbricht, G.R., Wunsch, D.C., II: Data preprocessing. In: *Computational Learning Approaches to Data Analytics in Biomedical Applications*, pp. 7–27 (2020)
20. Uematsu, H., Nguyen, P., Takeda, H.: Design for data structures: data unification and federation with Wikibase. In: 2022 IEEE International Conference on Big Data, pp. 6169–6178. IEEE (2022)



Brain Neural Network Architectures in Sleep-Wake Cycle

Vadim L. Ushakov^{1,4,5} , Maria L. Khazova² , Polina E. Zhigulina³,
Vyacheslav A. Orlov³ , Denis G. Malakhov³ , and Vladimir B. Dorokhov⁶ 

¹ Institute for Advanced Study of the Brain, Lomonosov Moscow State University, Moscow, Russia

falsetiuq@yandex.ru

² Lomonosov Moscow State University, Moscow, Russia

ml@mkhazova.ru

³ National Research Center “Kurchatov Institute”, Moscow, Russia

⁴ National Research Nuclear University MEPhI, Moscow, Russia

⁵ Mental Health Clinic No. 1 named after N.A. Alexeev of Moscow Health Department, Moscow, Russia

⁶ Institute of Higher Nervous Activity, Neurophysiology of RAS, Moscow, Russia

Abstract. The article presents an attempt to search for neurophysiological mechanisms of maintaining different levels of consciousness by studying the cognitive processes of the sleep-wake transition. To perform this study, a psychomotor test was developed, the monotonous performance of which, during 60 min, causes alternating episodes with the disappearance of consciousness when falling asleep (“microsleep”) and its recovery when waking up (wakefulness). It is shown that the structure of transitions is individual in terms of the spatial localization of neural networks. A common tendency is the extensive activation of cortical-subcortical areas 2-4 TR (4-8 s) before the moment of waking fixation. Functional magnetic resonance imaging (fMRI) signal changes in brain neural networks were recorded during dynamic transitions: thalamus, hippocampus and parahippocampal gyrus, cerebellum, sensorimotor cortex, somato-sensory association cortex, secondary motor area, visual cortex, bilateral inferior parietal cortex, frontal temporal regions, pale globus and putamen area, insular cortex, cuneus, precuneus, anterior and middle cingulate cortex. The results reflect the complex individual dynamic activity of brain neural networks involved in the sleep-wake states, some of which is probably related to the preparation for the realization of the test motor task.

Keywords: Consciousness · Sleep-wake cycle · fMRI · Brain neural networks · Hidden Markov models

1 Introduction

The neurobiological basis of consciousness is the focus of current research. A reliable and rather simple model for investigating the neural correlates of consciousness is the comparison of waking and sleeping states.

In the field of research on functional systems involved in the maintenance of the sleep state it has been shown to be observed a significant increase in the level of fluctuations of the BOLD signal (blood oxygen level dependent) in several cortical areas, with the most significant increase in the visual cortex [23]. Correlations between brain regions involved in the DMN (Default Mode Network) are preserved. There is also a decrease in functional connectivity between the cortex and the thalamus/hypothalamus, which is not usually observed for deeper stages of sleep [22]. During stages N1 and N2, several groups found decreased anti-correlation between DMN and task-positive networks, DAN (Dorsal Attention Network) and FPN (Fronto-Parietal Network). The same decrease in anti-correlation was also observed during wakefulness after partial or complete sleep deprivation, in addition to strong changes in the whole brain functional connectome [16]. Application of the hidden Markov models method to fMRI data of sleep (3 classical stages N1, N2, N3) and wakefulness showed the presence of about 14 states of brain neural networks with different spatial distribution, with some states dominated by one of the stages and others with all 4 stages (three stages of sleep and wakefulness) [21]. Thus, the correspondence between Markovian states and sleep stages is ambiguous - there are more states, and they are not linked to a specific stage, but may be present in several stages at once. Rather than being simply a state of reduced wakefulness, sleep is now understood as a complex state [13] and has characteristic features of neuroelectric and metabolic activity. What is activity during the transitional periods from wakefulness to sleep and from sleep to wakefulness and how is it organized? The answers to these questions are relevant not only for the study of neurophysiological mechanisms of falling asleep and awaking but can also provide insight into the functioning of the brain.

2 Materials and Methods

2.1 Selection

MRI data were obtained for eight healthy subjects (three males and five females, mean age 25 years \pm 3 years, all right-handed with no neurologic symptoms). Informed consent was obtained from each participant. Permission to conduct the experiment was granted by the local ethical committee of NRC Kurchatov Institute (Minutes No. 10 of August 1, 2018).

2.2 Materials and Methods

Methods of the study. A specially designed psychomotor test was used in this work, the monotonous performance of which causes alternating episodes of “microsleep” and awakening for 60 min [6]. When performing this test, the subject with closed eyes counts from 1 to 10 and simultaneously presses a button, alternately with the right and left hands. Spontaneous recovery of the test after an episode of “microsleep” requires activation of consciousness, which is accompanied by conscious performance of the test with verbal counting and simultaneous button pressing. The reproducibility of this test under MRI conditions obtained under normal conditions has been demonstrated. In 10 out of 14 subjects, 3–10 episodes of “microsleep” followed by awakening were recorded

in each subject during a 60-min experiment performed in an MRI chamber. Simultaneously recorded EMG of the finger muscles, motor activity accelerometer, EEG and fMRI to assess the spatiotemporal dynamics of the brain during the psychomotor test. A pneumatic button sensitive to the force of presses, which was previously developed for fMRI studies [6], was used to register presses. Simultaneously with the registration of the mechanogram of button pressing, the electromyogram (EMG) of the short muscle withdrawing the thumb of the right hand (musculus abductor pollicis brevis) was registered.

The scans were performed with a Magnetom Verio 3T MR tomograph (Siemens, Germany) using a 32-channel head MR coil installed in the NBICS-technology complex of NRC Kurchatov Institute. For each subject, a high-resolution anatomical image was acquired based on T1-weighted sequence: TR = 1470 ms, TE = 1.76 ms, 176 slices, voxel size: $1 \times 1 \times 1 \text{ mm}^3$. Functional MR scanning was based on a standard EPI sequence with parameters TR = 2000 ms, TE = 20 ms, 42 slices, voxel size $2 \times 2 \times 2 \text{ mm}^3$, flip angle = 90° , FOV = $192 \times 192 \text{ mm}^2$.

Preprocessing performed using fMRIPrep 22.1.1 [7, 8], which is based on Nipype 1.8.5 [10, 11].

Preprocessing of B0 inhomogeneity mappings. A total of 1 fieldmaps were found available within the input BIDS structure for this particular subject. A B0 nonuniformity map (or fieldmap) was estimated from the phase-drift map(s) measure with two consecutive GRE (gradient-recalled echo) acquisitions. The corresponding phase-map(s) were phase-unwrapped with prelude (FSL 6.0.5.1:57b01774).

Anatomical data preprocessing. A total of 1 T1-weighted (T1w) images were found within the input BIDS dataset. The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection [24], distributed with ANTs 2.3.3 [2], and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast [25]. Brain surfaces were reconstructed using recon-all [5], and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle [15]. Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: ICBM 152 Nonlinear Asymmetrical template version 2009c [9].

Functional data preprocessing. For each of the 1 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt [14]. The estimated fieldmap was then aligned with rigid-registration to the target EPI (echo-planar imaging) reference run.

The field coefficients were mapped on to the reference EPI using the transform. BOLD runs were slice-time corrected to 0.97 s (0.5 of slice acquisition range 0–1.94 s) using 3dTshift from AFNI [4]. The BOLD reference was then co-registered to the T1w reference using `bbregister` (FreeSurfer) which implements boundary-based registration [12]. Co-registration was configured with six degrees of freedom. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD was computed using two formulations following Power [19] and Jenkinson [14]. FD and DVARS are calculated for each functional run, both using their implementations in Nipype [19]. The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction [3]. Principal components are estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 2% variable voxels within the brain mask. For aCompCor, three probabilistic masks (CSF, WM and combined CSF + WM) are generated in anatomical space. The implementation differs from that of Behzadi et al. in that instead of eroding the masks by 2 pixels on BOLD space, a mask of pixels that likely contain a volume fraction of GM is subtracted from the aCompCor masks. This mask is obtained by dilating a GM mask extracted from the FreeSurfer's `aseg` segmentation, and it ensures components are not extracted from voxels containing a minimal fraction of GM. Finally, these masks are resampled into BOLD space and binarized by thresholding at 0.99 (as in the original implementation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each [20]. Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. Additional nuisance timeseries are calculated by means of principal components analysis of the signal found within a thin band (crown) of voxels around the edge of the brain, as proposed by [18]. The BOLD time-series were resampled into standard space, generating a preprocessed BOLD run in MNI152NLin2009cAsym space. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using `antsApplyTransforms` (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels [17]. Non-gridded (surface) resamplings were performed using `mri_vol2surf` (FreeSurfer).

Many internal operations of fMRIPrep use Nilearn 0.9.1 [1], mostly within the functional processing workflow.

Statistics. Fixed effects models were constructed for each subject in a Bayesian approach to data analysis. The number of sleep-wake transitions varied for each subject. The average number of transitions over the entire one-hour experiment was 10 events. The duration of sleep before awakening varied from a few minutes to 10 min. Regressors were generated using the onset and duration of the psychological events of interest and collapsing them using the canonical hemodynamic response function. Specifically, the following events were modeled using the synchronous signal recording method: [1] a time period beginning with the onset of keystrokes after episodes of microsleep lasting 4 s (regressor 1); [2] a time period beginning with sleep 6 s before keystrokes (regressor 2). After model estimation, the following contrasts were assessed in each subject: «regressor 1» relative to 0, «regressor 2» relative to 0, «regressor 2» > «regressor 1». The following type of comparison was used to evaluate the dynamics:

- “regressor 2” was shifted by 1–10 TR toward the “sleep” state with subsequent estimation of each GLM model.

The HMM_MAR package written under Matlab <https://github.com/OHBA-analysis/HMM-MAR> was used to extract from the time series the brain activity networks common to the subjects and to observe their dynamics. Hidden Markov models, a family of models that allow describing time series by a finite number of hidden states, each of which is a probability distribution, were taken as a basis. All distributions belong to the same family and a state is characterized by the parameters of this distribution. Thus, a state describes unique patterns of activity occurring in different parts of the data.

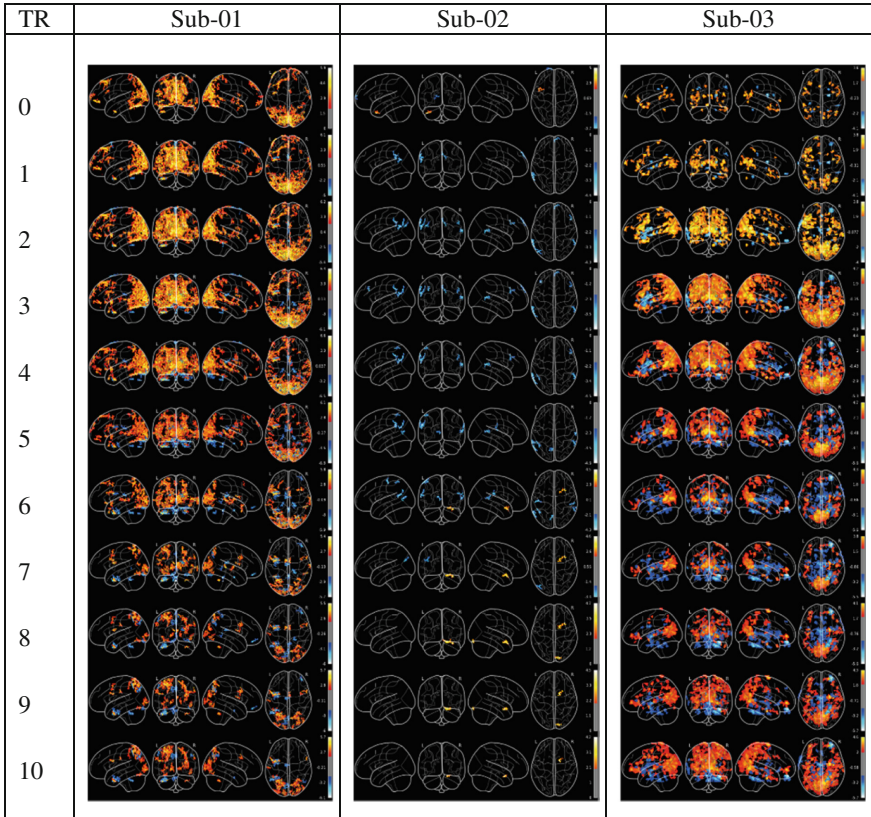
Statistical analysis was performed using Student’s T-statistics ($p = 0.05$ corrected for multiple comparisons (FWE)).

3 Results and Discussion

To observe the dynamics of sleep-wake transitions, the sleep-wake bar was shifted from a fixed state of wakefulness to sleep. The extended network of cortical and subcortical areas is activated during awakening from sleep 2–4 TR before the expected wakefulness (see Table 1) and has an individual character (the figures show examples of individual dynamics for three subjects). In the dynamic transitions, functional magnetic resonance imaging (fMRI) signal changes in brain neural networks were recorded: thalamus, hippocampus and parahippocampal gyrus, cerebellum, sensorimotor cortex, somato-sensory associative cortex, secondary motor area, visual cortex, bilateral inferior parietal cortex, frontal temporal regions, pale globus and putamen, insular cortex, cuneus, precuneus, anterior and middle cingulate cortex. Activation of the cerebellum, sensorimotor cortex, somatosensory associative cortex, and secondary motor area is presumably associated with preparation for a motor task. Different degrees of change in the activity of brain neural networks are presumably associated with different degrees of sleep immersion - so for subjects 1 and 3 the number of falls asleep with a duration of more than 10 s is approximately 1.5–2 times greater than for subject 2.

The transitional stages were marked up using the button press data and the following stages were identified: sleep, falling asleep, left hemisphere activity only, right hemisphere activity only, and wakefulness. When carefully analyzing the obtained markup,

Table 1. Z-maps comparing “regressor 2” (sleep) versus “regressor 1” (wakefulness) when “regressor 1” is fixed and “regressor 2” is shifted between 1 TR and 10 TR toward the sleep state. Red indicates a significant increase in BOLD signal, blue indicates a decrease ($p = 0.05$ adjusted for multiple comparisons (FWE)).



one peculiarity was noticed: activity of only the left hemisphere is much more frequent than activity of only the right hemisphere, which may be related to the fact that right-handed people - people with the leading left hemisphere - are more numerous. Based on this, a hypothesis was put forward that involuntary falling asleep begins with the slave hemisphere and only then the hemispheres can change places. It would be interesting to test this hypothesis in further studies, fixing which of the subjects is right-handed and which is left-handed.

The study of state dynamics was carried out on the basis of hidden Markov models. To select the number of principal components, a graph of the dependence of unexplained variance on the number of components was plotted. In this case, 22 components were chosen, which corresponds to 75% of the explained variance. As a result, the model was run in the space of 22 principal components. Statistical plots were used to determine the optimal number of states in this group: free energy, maximum fractional occupancy and

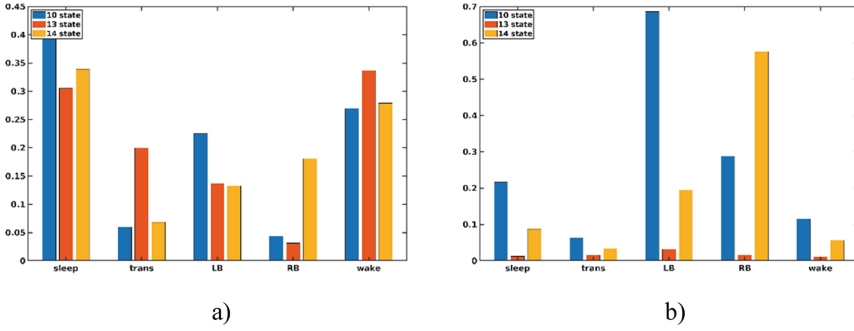


Fig. 1. Distribution of states 10, 13 and 14 in relation to sleep stages. (a) - Sensitivity to sleep stages. (b) - Specificity for sleep stages. 5 stages: sleep (sleep) - no tapping, asleep (trans) - errors in the number of taps, left hemisphere (RB) - tapping with one hand only, right hem-isphere (LB) - tapping with one hand only, awake (wake).

average duration of states. The value of free energy decreased with increasing number of states with reaching “saturation” around the 18th state. Thus, based on the plots of free energy and average duration of states, the value of $K = 18$ was chosen. However, after running the model with $K = 18$, when analyzing the dynamics of these states, 4 states were found that are practically not activated. It was decided to reduce the number of states to 14. Thus, the final model consisted of 14 states: their description and probability of activity over time were also obtained. There were three states common among several subjects - 10, 13, and 14 - and they were taken for further analysis. The spatial localization of these states (5% of the brightest areas and their deviation from the average activity in a given state) was distributed in cortical structures of the visual, temporal and parietal cortex and subcortical structures: Lingual_L, Occipital_Inf_L,R, Caudate L,R Thalamus_L for the first condition, Frontal_Inf_Orb_R, Occipital_Inf_R, Parietal_Sup_R, Parietal_Inf_R, Pallidum_L, Heschl_L for the second condition, and Occipital_Sup_R, Occipital_Mid_R, Thalamus_L, R (L-left, R-right hemisphere) for the third condition. Figures 1 and 2 show the distribution of these states in relation to sleep stages and subjects.

Based on the data obtained, it was concluded that there is no pure state of brain neural networks that characterizes only sleep, but there are networks that work both in sleep and in wakefulness, and during the transition from one to the other, only in different proportions and have an individual character of temporal distribution across subjects. This view of sleep is consistent with previous studies [21].

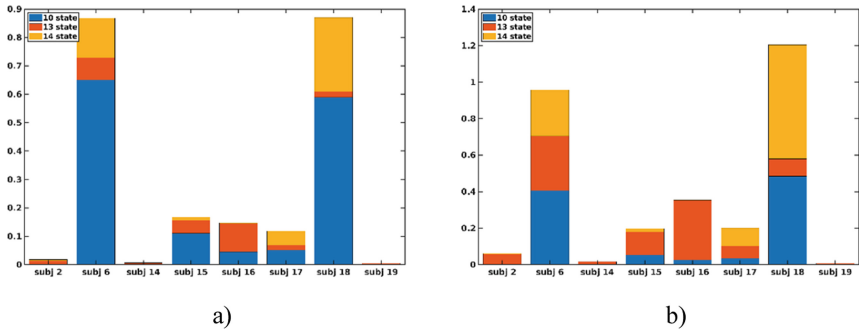


Fig. 2. Distribution of states 10, 13, and 14 with respect to subjects. **(a)**—Sensitivity to subjects. **(b)**—Specificity to subjects

4 Conclusion

The brain mechanisms by which we transition from sleep to wakefulness, i.e., to a conscious state, remain largely unknown, in part because of methodological problems. During the sleep-wake cycle, the brain undergoes profound dynamic changes that are subjectively manifested as transitions between conscious experience and the unconscious. However, neurophysiological signatures that can objectively distinguish between different states of consciousness are few. The results reflect the complex individual dynamic activity of brain neural networks involved in sleep-wake transition states and affecting cortical and subcortical structures.

5 Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements. This study was partially supported by the Russian Foundation for Basic Research grant № 23–78–00010, <https://rscf.ru/en/project/23-78-00010/>

References

1. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G.: Varoquaux: machine learning for neuroimaging with scikit-learn. *Front. Neuroinf.* **8** (2014)
2. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* **12**(1), 26–41 (2008)
3. Behzadi, Y., Restom, K., Liao, J., Liu, T.T.: A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage* **37**(1), 90–101 (2007)
4. Cox, R.W., Hyde, J.S.: Software tools for analysis and visualization of fMRI data. *NMR Biomed.* **10**(4–5), 171–178 (1997)

5. Dale, A.M., Fischl, B., Sereno, M.I.: Cortical surface-based analysis: I segmentation and surface reconstruction. *NeuroImage* **9**(2), 179–194 (1999)
6. Dorokhov, V.B., Malakhov, D.G., Orlov, V.A., Ushakov, V.L.: Experimental model of study of consciousness at the awakening: fMRI, EEG and behavioral methods. *Adv. Intell. Syst. Comput.* **848**, 82–87 (2019)
7. Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., Oya, H.: fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019)
8. Esteban, O., Ross, B., Christopher, J.M., Shoshana, L.B., Craig, M., Feilong, M., Ayse, I.I., et al.: fMRIPrep 22.1.1. Software (2018)
9. Fonov, V.S., Evans, A.C., McKinstry, R.C., Almlí, C.R., Collins, D.L.: Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage* **47**(Supplement 1), S102 (2009)
10. Gorgolewski, K., et al.: Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front. Neuroinform.* **5**, 13 (2011)
11. Gorgolewski, K.J., Esteban, O., Markiewicz, C.J., Ziegler, E., Ellis, D.G., Notter, M.P., Jarecka, D., Johnson, H., Burns, C., Manhães-Savio, A., Hamalainen, C.: Nipype. Software (2018)
12. Greve, D.N., Fischl, B.: Accurate and robust brain image alignment using boundary-based registration. *NeuroImage* **48**(1), 63–72 (2009)
13. Hobson, J.A.: Sleep is of the brain, by the brain and for the brain. *Nature* **437**, 1254–1256 (2005)
14. Jenkinson, M., Bannister, P., Brady, M., Smith, S.: Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**(2), 825–841 (2002)
15. Klein, A., Ghosh, S.S., Bao, F.S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Chaibub Neto, E., Keshavan, A.: Mindboggling morphometry of human brains. *PLOS Comput. Biol.* **13**(2), e1005350 (2017)
16. Krause, A.J., et al.: The sleep-deprived human brain. *Nat. Rev. Neurosci.* **18**(7), 404–418 (2017)
17. Lanczos, C.: Evaluation of noisy data. *J. Soc. Ind. Appl. Math. Ser. B Numer. Anal.* **1**(1), 76–85 (1964)
18. Patriat, R., Reynolds, R.C., Birn, R.M.: An improved model of motion-related signal changes in fMRI. *NeuroImage* **144**:74–82 (2017)
19. Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E.: Petersen: methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage* **84**(Supplement C), 320–41 (2014)
20. Satterthwaite, T.D., Elliott, M.A., Gerraty, R.T., Ruparel, K., Loughhead, J., Calkins, M.E., Eickhoff, S.B., Hakonarson, H., Gur, R.C., Gur, R.E., Wolf, D.H.: An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage* **64**(1), 240–56 (2013)
21. Stevner, A.B.A., et al.: Discovery of key whole-brain transitions and dynamics during human wakefulness and non-REM sleep. *Nat. Commun.* **10**(1), 1035 (2019)
22. Tagliazucchi, E., Laufs, H.: Decoding wakefulness levels from typical fMRI resting-state data reveals reliable drifts between wakefulness and sleep. *Neuron* **82**(3), 695–708 (2014)
23. Tagliazucchi, E., van Someren, E.J.W.: The large-scale functional connectivity correlates of consciousness and arousal during the healthy and pathological human sleep cycle. *Neuroimage* **160**, 55–72 (2017)

24. Tustison, N.J., et al.: N4itk: improved N3 bias correction. *IEEE Trans. Med. Imaging* **29**(6), 1310–1320 (2010)
25. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* **20**(1), 45–57 (2001)



Speech Recognition from MEG Data Using Covariance Filters

Vitaly Verkhlyutov¹(✉), Victor Vvedensky², Konstantin Gurtovoy²,
Evgenii Burlakov³, and Olga Martynova¹

¹ Institute of Higher Nervous Activity and Neurophysiology of RAS, Moscow, Russia
verkhliutov@ihna.ru

² RNC Kurchatov Institute, Moscow, Russia

³ Derzhavin Tambov State University, Tambov, Russia

Abstract. Speech recognition from EEG and MEG data is the first step in the development of BCI and AI systems for further use in the decoding of inner speech. Great achievements in this direction have been made with the use of ECoG and stereo-EEG. At the same time, there are few works on this topic on the analysis of data obtained by noninvasive methods of brain activity registration. Our approach is based on the evaluation of connections in the sensor space with the extraction of the MEG connectivity pattern specific to a given segment of speech. We tested our method on 7 subjects. In all cases, our processing pipeline was sufficiently reliable and worked either without recognition errors or with few errors. After “training” the algorithm is able to recognize a fragment of spoken speech in a single presentation. For recognition, we used MEG recording segments of 50–1200 ms from the beginning of the word. A segment of at least 600 ms was required for high-quality recognition. Intervals longer than 1200 ms degraded the quality of recognition. Band-pass filtering of MEG showed that the quality of recognition is higher when using the gamma frequency range compared to the low-frequency range of the analyzed signal.

Keywords: Speech decoding · MEG · EEG · Connectivity in sensor space · Semantic systems · Theta-rhythm · Alpha-rhythm · Gamma-rhythm · BCI · AI

1 Introduction

Decoding of inner speech and speech stimuli from brain activity data is an urgent task for theoretical and applied purposes of modern neurophysiology. Within this direction, researchers are trying to solve the problem of compensation for lost functions in various types of disorders of speech reproduction and perception at the cortical level, which has direct relevance to BCI. At the same time, the study of this issue helps to move towards the improvement of AI systems. Significant progress has been made with intracranial ECoG recordings [1] and stereo EEG

[2]. However, invasive techniques have a limited range of applications. Recent studies have shown that decoding macroscopic fMRI data using a trained language model can quite accurately decode internal speech based on semantic information [3].

Other non-invasive recording methods, such as EEG and MEG, have proven that speech perception and playback affect rhythmic [4–6] and evoked [7] brain electrical activity. Thus, there are all prerequisites for speech decoding based on MEG and EEG data. However, for the analysis of brain activity in this case [8] neural network technologies are used, the results of which are difficult to interpret. For these purposes, we propose to use a simpler technique to investigate the connectivity of MEG in the sensor space, which is based on observations showing a remarkable similarity of the current MEG activity on lead clusters when listening to words, as well as the dynamic reorganization of these clusters when recognizing the semantic meaning of a speech stimulus [4].

2 Measurements

2.1 Subjects

Seven volunteer subjects (four men and three women) participated in the pilot study aimed at testing the technique. One of the subjects was left-handed at the age of 23. The average age of the young right-handed subjects was 23.8 ± 0.5 years. The elderly right-handed subject was 67 years old. All subjects had no history of neurological or psychiatric disorders. The study was carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki) for experiments involving humans, and the Ethical committee of the Institute of Higher Nervous Activity and Neurophysiology of the RAS approved the research protocols (Protocol No. 5 of January 15, 2020). The studies took place from 12 to 15 o'clock.

2.2 Stimuli

The subject was presented with three series of speech stimuli in the form of Russian adjectives. Each series included eight original words, which were repeated five times. All forty words were randomly mixed. Before each series, three words from the same series of words were presented to adapt the subject, but the recorded data from these presentations were not considered for analysis. The series of words differed in sound duration and were 600, 800, and 900 ms, respectively. The loudness of the sound was selected for each subject and ranged from 40 to 50 dB. The frequency of the digitized words as an audio file did not exceed 22 kHz. After a word was presented, the subject had to press the hand-held manipulator button if he understood the meaning of the presented word. Pressing the button after 500 ± 100 ms (randomized) was followed by the next stimulus, but no later than 2000 ms after the previous presentation.

2.3 Experimental Procedure

Before the experiment, the coordinates of the anatomical reference points (left and right preauricular points and the bridge of the nose) were determined using the FASTRAK 3D digitizer (Polhemus, USA), as well as indicator inductance coils attached to the subject's scalp surface in the upper part of the forehead and behind the auricles. During the experiment, the subject was in a magnetically shielded multilayer permalloy chamber (AK3b, Vacuumschmelze GmbH, Germany) and his head was placed in a fiberglass helmet, which is part of a fiberglass Dewar vessel with a sensor array immersed in liquid helium. The test subject was seated so that the surface of the head was as close as possible to the sensors. To avoid artifacts, sound stimuli were delivered through a pneumatic system delivering sound from a standard audio stimulator. The stimulator was programmed using Presentation software (USA, Neurobehavioral Systems, Inc). The subject was asked to relax and close his eyes. His right hand touched a console with buttons. He had to press one key with his index finger after recognizing the heard word. At the end of the series of presentations, the subject was allowed to rest for 1–2 minutes.

2.4 Registration

Registration parameters are described in the Zenodo service. The data (MEG and MRI) are also available there [9].

3 Data Analysis

We did not use any additional signal processing methods except MaxFilter and bandpass filtering to determine the contribution of the delta-gamma of the MEG frequency range to the correctness of word recognition. Actual segments of the MEG were identified using the marks of the beginning of the sounding of words. These segments were used to build covariance matrices, calculating the Pearson correlation of each registration channel with each and thus forming a covariance vector for the original word

$$C_{nk} = \text{cov}(\overline{\mathbf{M}}_{nk}), \quad (1)$$

where $\overline{\mathbf{M}}_{nk}$ is the vector of the MEG data for the k -th repeat of the n -th word ($n = 1, \dots, N$, $k = 1, \dots, K$). The covariation matrices were averaged for all words by subtracting from them the averaged matrix for each word with replacing the principal diagonal elements by zeros and setting to zero all elements that less than 0.7 in the resulting matrices:

$$\mathbf{F}_n = \left(\frac{1}{K} \sum_{k=1}^K C_{nk} - \mathcal{O}_{0.7} \left(\frac{1}{KN} \sum_{n=1}^N \sum_{k=1}^K C_{nk} \right) \right) \quad (2)$$

where the operator \mathcal{O}_x replaces the principal diagonal elements by zeros and setting to zero all elements that less than x . The resulting patterns were used

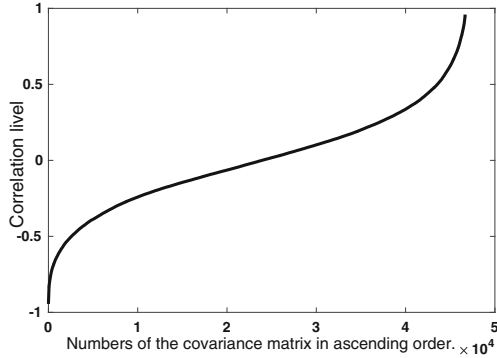


Fig. 1. Correlation coefficients when comparing each sensor to each from a one-second segment of MEG at the time of word sound. Mirror elements have been removed.

to calculate the weights of newly presented words. The weight w of a word with the covariation matrix C can be assessed with respect to the \hat{n} -th word filter \mathbf{F}_n as

$$w = \text{Sum}(C \circ H\mathbf{F}_n), \quad (3)$$

where the functional Sum translates any matrix to the sum of its elements, the binary operation \circ represents the elementwise product of two matrices (of the same dimension), and H is the elementwise Heaviside function. If the weight of a recognised word exceeded the weight of all others, then the word was considered recognised. Thus, the number of the recognised word can be found from the relation

$$\text{word number} = \underset{n=1, \dots, N}{\operatorname{argmax}} w(\mathbf{F}_n). \quad (4)$$

Weights were calculated for all 40 words. The 5 maximum weight values belonged to the recognized word. The recognition error was considered to be weight reduction below the maximum weight for all presentations of all the remaining 35 other words. Thus, the system could make a maximum of 5 errors when recognizing one word and 40 errors when recognizing 8 original words. At the same time, we could evaluate the success rate of recognition as a percentage. At 100% recognition, all 5 identical words were identified in a sequence of 40 words. One error reduced the recognition success score by 2.5%. The described algorithm is implemented by Matlab [10].

4 Results

Correlation analysis of the one-second segments of MEG showed the range of the correlation coefficients from $r < 0.9$ to $r > -0.9$ (Fig. 1). However, we only used r values that are greater than 0.7 for the analysis.

Weights were obtained for all 24 words in three series (Fig. 2) for each subject. It was not always possible to successfully recognize a target word with a weight

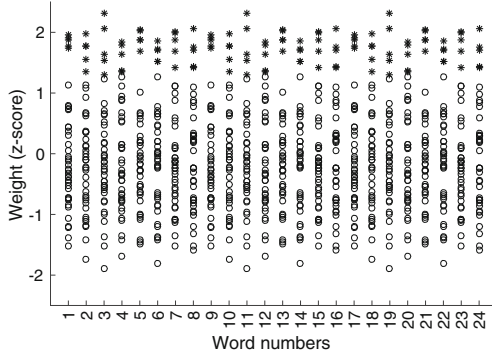


Fig. 2. Normalized weights for 24 original words (5 presentations for each word) for 3 series of presentations of subject V4. Each column contains 40 weights. Asterisks show 5 recognized words, circles show background words. If there are less than 5 stars, then there is layering, and the weights have very close values.

close to or less than one of the background words or words of choice. In this case, recognition was considered inaccurate. An attempt to select the optimal length of the MEG segment for recognition was made (see Fig. 3). Reducing the analyzed MEG segment had little effect on the quality of recognition until the segment of 850–1000 ms. Figure 3 shows the percentage of errors for subject V1 for three series of presentations separately. At intervals from 200–1000 ms to 750–1000 ms, a zero error rate is observed in the individual series (the circle lies on the zero line). By increasing the same analyzed MEG interval from 0–50 ms to 0–1200 ms, the quality of recognition began to increase with durations from 0 to 600 ms (Fig. 4). Allocating 100 ms segments for analysis caused a deterioration of word recognition quality, especially at segments from 200 to 300 ms and from 1000 to 1100 ms (Fig. 5).

We evaluated the quality of recognition in all subjects (Table 1). We did not find any tendency depending on age, gender, and dominant hand (subject V3 was elderly and subject V5 was left-handed).

5 Discussion

An important factor for the behavior of neuronal populations is their synchronization, which allows many neurons to work in parallel and process many properties of the input signal simultaneously, establishing their multiple connections with other mental objects and their properties [11].

In our experiments, we observe that some populations are active in the perception of any word, and some are specific to a particular word. A speech recognition system can be based on the specificity property. In doing so, we investigate only amplitude connectivity, which is due to distant connections [13] as opposed to phase connectivity, which in turn is provided by local interactions. The presence of phase synchronization in our experiments proves the presence of

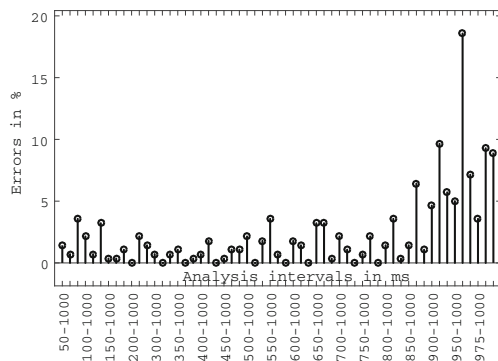


Fig. 3. The effect of misrecognition, when the analyzed MEG segment is reduced from 50–1000 ms to 975–1000 ms from the beginning of the word in the subject V1. The unit stem denotes the percentage of errors in the recognition of one set of 8 original words. Intervals are indicated only for the first set of three.

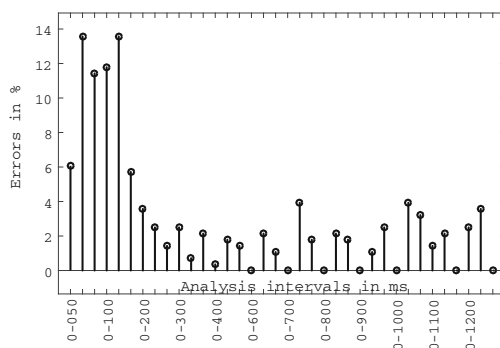


Fig. 4. The effect of misrecognition, when the analyzed MEG segment is increased from 0–50 ms to 0–1200 ms from the beginning of the word in subject V1. The unit stem denotes the percentage of errors in the recognition of one set of 8 original words. Intervals are indicated only for the first set of three.

Table 1. Percentage of correct recognition in 7 subjects using a 0–1000 ms interval of unfiltered MEG from word onset for three sets of words.

Subject	Sex	Age	Set1 (0–1 s) %	Set2 (0–1 s) %	Set3 (0–1 s) %
V1	Male	24	100	97,5	95
V2	Female	24	100	97,5	92,5
V3	Male	67	97,5	100	97,5
V4	Male	27	100	100	100
V5	Male(L)	23	100	100	97,5
V6	Female	20	85	90	100
V7	Female	24	95	100	100

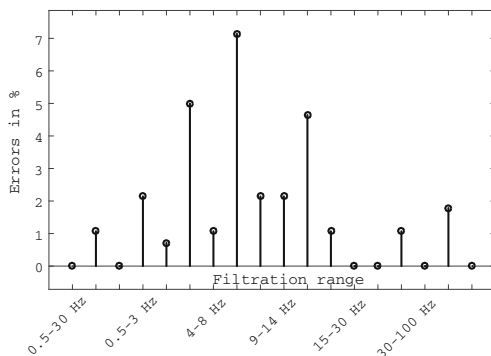


Fig. 5. The effect of misrecognition, when bandpass filtering the interval 200–1000 ms from the beginning of the word in the subject V1 in the ranges 0.5–30 Hz, 0.5–3 Hz, 4–8 Hz, 9–14 Hz, 15–30 Hz, 30–100 Hz. The single stem denotes the percentage of errors in the recognition of one set of 8 original words. Intervals are marked only for the first set of three.

both positively and negatively correlated data. Phase coupling indicates possible effects associated with the rotation of current dipoles, which are due to cortical traveling waves [12]. The effectiveness of individual segments for decoding is possibly related to implicit perception, which allows the brain to view part of a phrase as having complete meaning [14]. Bandpass filtering has shown that high-frequency components are most effective for decoding, though along with other frequency bands, due to long-range interaction (through myelinated fibers) between electrical brain sources [15, 16].

Acknowledgments. This study was supported by the Russian Science Foundation - Grant no. 23-78-00011. The authors are grateful to Chernyshev B. V. and Prokofyev A. O. from the Center for Neurocognitive Research (MEG Center) at the Moscow City University of Psychology and Education.



References

1. Anumanchipalli, G.K., Chartier, J., Chang, E.F.: Speech synthesis from neural decoding of spoken sentences. *Nature* **568**(7753), 493–498 (2019). <https://doi.org/10.1038/s41586-019-1119-1>
2. Norman-Haignere, S.V., Long, L.K., Devinsky, O., Doyle, W., Irobunda, I., Merriks, E.M., Mesgarani, N.: Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nat. Hum. Behav.* **6**(3), 455–469 (2022). <https://doi.org/10.1038/s41562-021-01261-y>
3. Tang, J., LeBel, A., Jain, S., Huth, A.G.: Semantic reconstruction of continuous language from non-invasive brain recordings. *Nat. Neurosci.* (2023). <https://doi.org/10.1038/s41593-023-01304-9>
4. Vvedensky, V., Filatov, I., Gurtovoy, K., Sokolov, M.: Alpha rhythm dynamics during spoken word recognition. *Stud. Comput. Intell.* **1064**, 65–70 (2023). https://doi.org/10.1007/978-3-031-19032-2_7

5. Lizarazu, M., Carreiras, M., Molinaro, N.: Theta-gamma phase-amplitude coupling in auditory cortex is modulated by language proficiency. *Hum. Brain Mapp.* **44**(7), 2862–2872 (2023). <https://doi.org/10.1002/hbm.26250>
6. Neymotin, S.A., Tal, I., Barczak, A., O’Connell, M.N., McGinnis, T., Markowitz, N., Lakatos, P.: Detecting spontaneous neural oscillation events in primate auditory cortex. *Eneuro* **9**(4), ENEURO.0281–21 (2022). <https://doi.org/10.1523/ENEURO.0281-21.2022>
7. Anurova, I., Vetchinnikova, S., Dobrego, A., Williams, N., Mikusova, N., Suni, A., Palva, S.: Event-related responses reflect chunk boundaries in natural speech. *NeuroImage* **255**, 119203 (2022). <https://doi.org/10.1016/j.neuroimage.2022.119203>
8. Dash, D., Ferrari, P., Wang, J.: Decoding imagined and spoken phrases from non-invasive neural (MEG) signals. *Front. Neurosci.* **14**, 290 (2020). <https://doi.org/10.3389/fnins.2020.00290>
9. Verkhlyutov, V.: MEG data during the presentation of Gabor patterns and word sets. ZENODO, 7458233 (2022). <https://zenodo.org/record/7458233>
10. <https://github.com/BrainTravelingWaves/22SpeechRecognition>
11. Defosse, A., Caucheteux, C., Rapin, J., Kabeli, O., King, J.-R. Decoding speech from non-invasive brain recordings. *ArXiv*, 2208.12266, 1-15. (2022). <http://arxiv.org/abs/2208.12266>
12. Sato, N.: Cortical traveling waves reflect state-dependent hierarchical sequencing of local regions in the human connectome network. *Sci. Rep.* **12**(1), 334 (2022). <https://doi.org/10.1038/s41598-021-04169-9>
13. Rolls, E.T., Deco, G., Huang, C.-C., Feng, J.: The human language effective connectome. *NeuroImage* **258**, 119352 (2022). <https://doi.org/10.1016/j.neuroimage.2022.119352>
14. Liaukovich, K., Ukraintseva, Y., Martynova, O.: Implicit auditory perception of local and global irregularities in passive listening condition. *Neuropsychologia* **165**(July 2020), 108129 (2022). <https://doi.org/10.1016/j.neuropsychologia.2021.108129>
15. Proix, T., Delgado Saa, J., Christen, A., Martin, S., Pasley, B.N., Knight, R.T., Giraud, A.-L.: Imagined speech can be decoded from low- and cross-frequency intracranial EEG features. *Nat. Commun.* **13**(1), 48 (2022). <https://doi.org/10.1038/s41467-021-27725-3>
16. Arnulfo, G., Wang, S.H., Myrov, V., Toselli, B., Hirvonen, J., Fato, M.M., Palva, J.M.: Long-range phase synchronization of high-frequency oscillations in human cortex. *Nat. Commun.* **11**(1), 5363 (2020). <https://doi.org/10.1038/s41467-020-18975-8>



Prediction of the Correct Firing Position with a Pistol Based on a MANFIS Model

David Alberto Vique Almeida^(✉) , Luis Armando Chicaiza Conteron , José Luis Carrillo Medina , and Edison Gonzalo Espinosa Gallardo 

Universidad de las Fuerzas Armadas ESPE, Latacunga, Ecuador
davique@espe.edu.ec

Abstract. Simulator systems are intended to facilitate access to practice in different test environments through a closed environment using computer vision. In this paper, a MANFIS model is implemented to improve the position of a trainee through positioning practice, which is responsible for detecting and displaying the errors that the trainee has when adopting the shooting position in real time. This environment qualifies, evaluates the pistol shooting position in a closed environment, analyzes, compares, and discusses the results obtained to determine the best option for evaluating the shooting position where the integrity and safety of the practitioner are paramount. The programs are adjusted to the process of identification and evaluation by Computer Vision using algorithms and processing methods. The evaluation reached an overall efficiency $89.59 \pm 3.36\%$, for 12 participants, determining the virtual simulator as adequate for training practices of the correct firing position.

Keywords: Computer vision · MANFIS · Virtual simulator · Firing position · Pose estimation

1 Introduction

The posture a shooter adopts on the practice range becomes critical because it can affect the accuracy and speed of the shot, where the moving parts of the body, muscles, and bones (such as the legs, back, and arms) are involved [1, 2]. Proper shooter's posture helps to maintain balance, if the shooter does not control his posture during practice, it will increase body fatigue, which affects his physical condition, concentration, and shooting efficiency [3–5].

In the teaching-learning process, constant practice and correction of errors help to improve posture, accuracy, and performance on the shooting range through a continuous evaluation whose results allow a significant improvement in the functional capacity of the practitioner [6, 7]. Once this stage has been completed, technical and postural corrections are difficult to make again, an aspect that is evident in the sample studied in this work, as part of a preliminary diagnosis of the level of influence of the position of the body in front of the target. With the use of weaponry (see Sect. 2), the technical focus of the preparation lies in identifying specific alterations that may affect the shooter's postural control and consequently his performance [8, 9].

This article proposes the integration of technologies oriented to the use of immersive virtual reality, using an adaptive neural-fuzzy inference system for the prediction of the correct shooting position. The proposed model has been tested and compared with real cases of correct position evaluation in shooting ranges. The evaluation allows the student to develop his skills, abilities, and professional training. At the end of the practical exercises, a joints analysis of the results obtained on the position simulator is presented for experienced and inexperienced trainees.

2 System Design

Virtual reality now allows for the visualization of biomechanical agents in shooting position simulators during conflict situations [10], offering potential training enhancements [11, 12]. Figure 1 displays the system's design, highlighting key components: (a) trainee positioning capture, (b) recording and playback, (c) a projection showcasing trainee silhouettes, and (d) 3D representations in the projection area, supported by (e) equipment rack, (i) computer with connections, and (ii) the projection area.

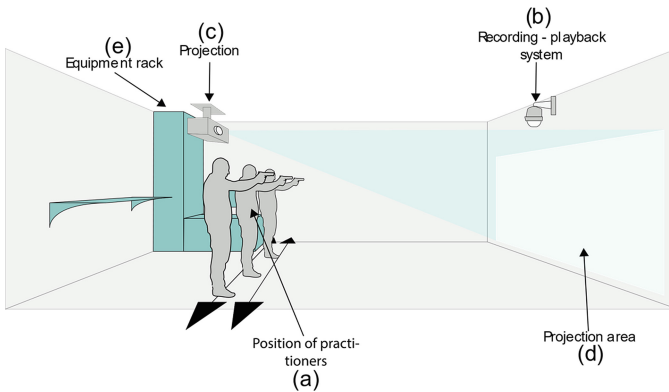


Fig. 1. Main components of system.

The virtual simulator uses a Computer Vision technique present in the Kinect sensor to detect the position of the practitioner [10, 11], and 4 Adaptive Neural-Fuzzy Inference System of Mamdani (MANFIS) techniques to evaluate the correct shooting position [13, 14]. To validate the results, we propose the creation of a scenario that visualizes the position with a 3D representation and the silhouette of the biomechanical agents of the practitioner (see Sect. 4.1), which facilitates the control of the evaluation [15].

3 Functional Scheme of the Firing Range Simulator

Using the Unity3D platform, a simulator was developed to teach the correct firing position by analyzing biomechanical agents. The collected data is stored adaptively [15]. Figure 2 shows the functional schematic of the simulator with two modules: (1) User

interaction, which places practitioners in a shooting position, and (2) Dialogue flow, which processes the practitioner's position and provides feedback. The reasoning is applied through the instructor's Knowledge Base (see Sect. 4). The final result shows both the firing posture and the percentage of positioning achieved using a 2D silhouette.

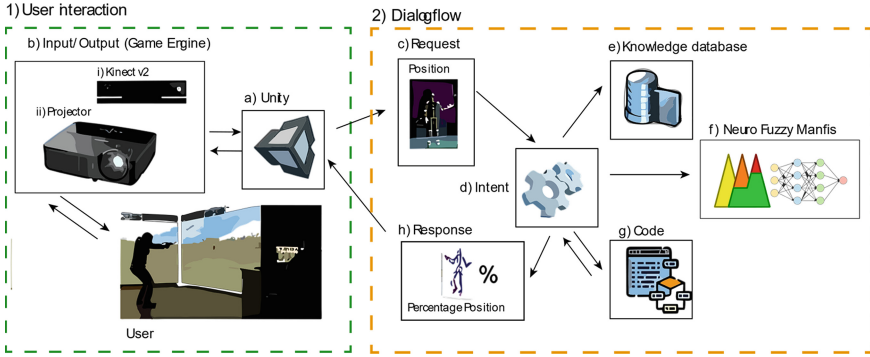


Fig. 2. Functional scheme of the proposed application: (1) User Interaction for communication with the simulator, (2) Dialog Flow receives inputs, processes them, and returns them as outputs.

The graphics engine, (a) Unity 3D, was used to build the virtual and simulation environments (Unity SDK, Kinect 2 SDK). Blender and multimedia resources from the Unity Asset Store. (b) The inputs and outputs of the system are realized by operating (i) the Kinect 2 integrated motion sensor to obtain (c) request or position information to be processed by (f) MANFIS (see Sect. 4), and (e) the knowledge database (see Table 2) and (g) code return the percentage of correct firing position where the data visualization is obtained through the (ii) projection of the scenario.

4 Adaptive Neural-Fuzzy Inference System of Mamdani (MANFIS)

Biomechanics applies mechanical principles to study human body movement, especially in shooting sports [7, 16]. Despite extensive research on licensed weapons and shooting psychology, there's a gap in technical and mechanical studies [17]. Biomechanical indicators (parts of a body) from shooters are pivotal for technical control [18, 19]. The Kinect 2 sensor enhances this analysis by (a) tracking the human body and identifying the joints being studied by (b) shooter position using parameters and shooter biomechanics to obtain (c) Euclidean distances and angles to be used as input data in (d) the Mamdani neuro-fuzzy system, uses membership functions and a neural network for defuzzification, with fully connected backpropagation learning (see Fig. 3).

The tracking of (a) human body skeleton is provided by the Kinect 2 through the integrated SDK, which uses human motion capture based on the Shotton algorithm using Support Vector Machine (SVM) for body part classification based on depth and RGB data and Randomized Decision Forests for precise joints position prediction, revolutionizing human motion capture [18, 19]. With this technology, the human motion of up to 6 people is tracked simultaneously, identifying 25 joints [19]. Figure 3 shows the tracking

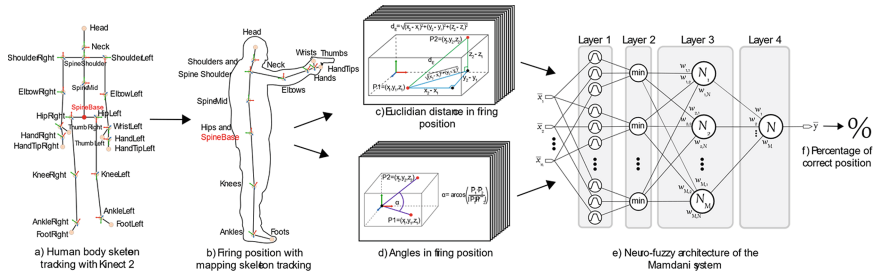


Fig. 3. Human body biomechanical agents tracking with Kinect 2 and correct position evaluation with a basic neuro-fuzzy architecture of the Mamdani system [17]

of the human body to capture its motion, where (b) the shooter’s position given by Sect. 4.1. The biomechanics of the shooter is used to obtain (c) the distances and (d) the angles of its biomechanical agents using parameters of the membership functions. These parameters are the input of the neural network that will be introduced in the fuzzy logic system, to be integrated into (e) the MANFIS in charge of predicting (f) the percentage of the correct firing position.

4.1 Biomechanics of the Shooter

Biomechanics studies the movement of the human body using mechanical principles [7, 20]. In shooting sports, a general technique is integrated with other scientific disciplines [2, 6]. Some factors can influence the assessment of athlete performance [21, 22]. During training, instructors evaluate posture and weapon grip [6, 7]. Initially, shooters are taught correct positioning, but post-training technical corrections are rare. This neglect becomes evident in posture studies [8, 9]. Improper posture can cause increased fatigue, affecting a shooter’s concentration and performance. Proper postural control is vital for maintaining balance and shooting efficiency [3–5].

The standard position for firing a weapon, as shown in Fig. 3, section (b), has the arm extended, the sights oriented approximately 90° towards the feet and shoulders, to counterbalance the weight of the weapon [10]. However, this is not universal; factors such as posture, body composition, and height influence arm angle. Shorter shooters may raise their arms more, while taller shooters may aim lower [11]. The standard shooting position involves facing the target with feet shoulder-width apart, knees slightly bent, and aligned hips, back, shoulders, and head.

The body should lean forward slightly, forming an “isosceles triangle” with extended arms and straight shoulders. This stance ensures optimal balance, minimizes muscle strain, and offers comfort [11] (refer to Fig. 3). Angles refer to the degrees of rotation or flexion at specific joints, while distances indicate the spatial separation between body parts (see Table 1). Their measurement helps to correct the firing position. This study aims to analyze biomechanical indicators in firearm action by quantifying physical traits and comparing results from various practices.

4.2 Fuzzy Logic and Training

A Mamdani-type fuzzy system [21] can be represented by a multilayer architecture similar to a neural network [23]. This neuro-fuzzy architecture has as its first layer the calculation of the angle between two vectors in a 3D plane (see Eq. 1), Fig. 3 and the Euclidean distances between two points (see Eq. 2), of the biomechanical agents of the shooter in a 3D environment, according to Sect. 4 and Table 1, evaluated in the membership functions, which gives me an idea to improve the firing position.

$$\alpha = \arccos\left(\frac{P_1 \cdot P_2}{|P_1| \cdot |P_2|}\right) \quad (1)$$

$$d_E(P_1 \cdot P_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \quad (2)$$

In the second layer, the Cartesian product or minimum function is performed, where there are N results given by the number of IF-THEN fuzzy rules. The last two layers are responsible for defuzzification by the centroid method, constrained by fuzzy inference. However, this can be represented by a neural network [23, 24], so that various defuzzification formulas can be substituted as a function of the training samples, which constitute a learning sequence [25–28]. This process represents a fuzzy system modeling approach [29] for detecting the correct firing position with defuzzification using a neural network with backpropagation learning. In the following, the terms of the fuzzy system are defined:

- The Universe of Discourse: for the angles is between -180 to $+180$ and in distance between 0 to 2 m.
- The Crisp set: the correct position of the practitioner is determined by comparing the angles between the centre of mass and the limbs of the body. This is considered accurate when it aligns with the shooter's biomechanics (see Sect. 4.1).
- The Fuzzy set: the Head and Trunk angles are 90° and the left and right arms 40° according to [20, 24]; However, these elements belong to a classical set therefore, a representation with adjustment to a fuzzy set given by Table 1 is proposed taking into account the instructor's indications and the noise generated by the Kinect 2 sensor.
- The Membership function: appropriate ranks in the fuzzy set are identified using specific membership functions [29]. These functions use Gamma (Γ) (see Eq. 3), Lambda (Λ) (see Eq. 4) for triangular shapes, and Pi (Π) (see Eq. 5) for trapezoidal shapes themselves to represent fuzzy sets Table 1 designates a membership function for each biomechanical agent within a category based fuzzy set.

$$\mu(x) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{x-a}{m-a} & \text{if } a < x < m \\ 1 & \text{if } x \geq m \end{cases} \quad (3)$$

$$\mu(x) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{x-a}{m-a} & \text{if } a < x < m \\ \frac{b-x}{b-m} & \text{if } m < x < b \\ 0 & \text{if } x \geq b \end{cases} \quad (4)$$

$$\mu(x) = \begin{cases} six \leq a \\ 0 & sia < x \leq b \\ \frac{x-a}{b-a} & \cdot \\ 1 & sia < x \leq c \\ \frac{d-x}{b-c} & sic < x \leq d \\ 0 & \cdot \\ six > d \end{cases} \quad (5)$$

Angle measures include head (H), trunk (T), left arm (LA), right arm (RA), left knee (LK), and right knee (RK). On the other hand, distance measurements comprise the arms (DAR), knees (DK) and ankles (DAN). These abbreviations simplify the identification and categorization of measurements, as angles represent joints positions and distances quantify spatial relationships.

Table 1. Membership functions on the fuzzy sets of the pistol shooting position.

Element	Membership function in the fuzzy set														
	Bad					Regular					Good				
	a	b	c	d	F	a	b	c	d	F	a	b	c	d	F
H (°)	105	120	-	-	Γ	115	130	145	155	Π	150	180	-	-	Λ
T (°)	105	120	-	-	Γ	115	130	145	155	Π	150	180	-	-	Λ
LA (°)	0	40	-	-	Λ	30	35	40	55	Π	40	65	85	105	Π
	130	180	-	-	Γ	95	105	120	135	Π					
RA (°)	0	40	-	-	Λ	30	35	40	55	Π	40	65	85	105	Π
	130	180	-	-	Γ	95	105	120	135	Π					
LK (°)	85	100	-	-	Γ	95	110	125	135	Π	130	180	-	-	Λ
RK (°)	85	100	-	-	Γ	95	110	125	135	Π	130	180	-	-	Λ
DAR (cm)	50	100	-	-	Γ	20	40	50	70	Π	0	25	-	-	Λ
DK (cm)	75	150	-	-	Γ	40	55	65	80	Π	0	50	-	-	Λ
DAN (cm)	30	100	-	-	Γ	25	30	40	55	Π	0	40	-	-	Λ

Table 1 provides an overview of the membership functions for a fuzzy set, where each element represents a biomechanical agent, complete with units of measurement. The descriptors “Bad”, “Regular”, and “Good” describe the acceptable features of the position, which are associated with upper and lower bounds, represented in columns a, b, c, and d. The Gamma, Lambda, and Pi functions assign these limits, denoted by x, and are described in Sect. 4.1 using specific letter representations. Essentially, Table 1 categorizes body measurements, such as the angle of the head or the distance between arms, into fuzzy sets using these membership functions. These parameters shape the functions and offer a structured framework for understanding and categorizing biomechanical measurements.

The defuzzification method employs a backpropagation neural network for training, using various inputs and their classifications [30]. This network takes its input from the results of the Mamdani table-based fuzzy rules [31] in the second layer of the MANFIS system. Table 2 shows an approach for modeling a backpropagation neural network, belonging to a MANFIS system, showing each biomechanical agent (see Sect. 4.1) in their respective units (degrees or centimeters) along with their percentages of correct position (see Fig. 3).

Table 2. The training data set used in this study for the correct position of fire obtained by sensor Kinect 2 and instructor.

#	0	1	2	3	4	...	6	...	299	300
H (°)	167	116	159	151	122	...	174	...	141	114
T (°)	160	106	168	154	149	...	173	...	140	119
LA (°)	43	16	99	42	47	...	104	...	120	142
RA (°)	72	27	48	39	36	...	101	...	54	33
LK (°)	166	119	169	146	150	...	156	...	146	112
RK (°)	157	111	151	132	135	...	154	...	137	112
DAR (cm)	9	66	5	30	32	...	22	...	21	126
DK (cm)	49	81	44	54	78	...	20	...	77	100
DAN (cm)	28	107	37	45	29	...	31	...	26	141
Percentage	71	27	75	73	51	...	96	...	55	21

Table 2 is a structured data set for training a neural network. The inputs are the biomechanical agents, while the percentage is the output data indicating the quality of the posture (rows), and the number of records used is 300 (columns) [25–28]. Using this data set in the training of a neural network, a model capable of predicting the percentage quality of the pistol shooting stance from the input values was developed. This predictive capability can be used in various contexts, such as postural control applications, virtual exercise assistants, or postural correction systems [30]. With the use of this dataset and a trained neural network model, a system capable of assessing and providing information about the quality of a person’s posture can be obtained, which can contribute to improving health and prevent possible injuries related to incorrect posture.

5 Execution of the Practice

The simulator displays a 3D virtual environment in which the correct position of a character is displayed. To improve training results, practitioners must address and rectify positional faults. The simulation system evaluates the user’s biometrics in real time, scoring them from 0 to 100 based on the distances and angles of the biomechanical agents (see Table 1). A score above 70 indicates optimal positioning, which directly influences the accuracy of the shot [31, 32].

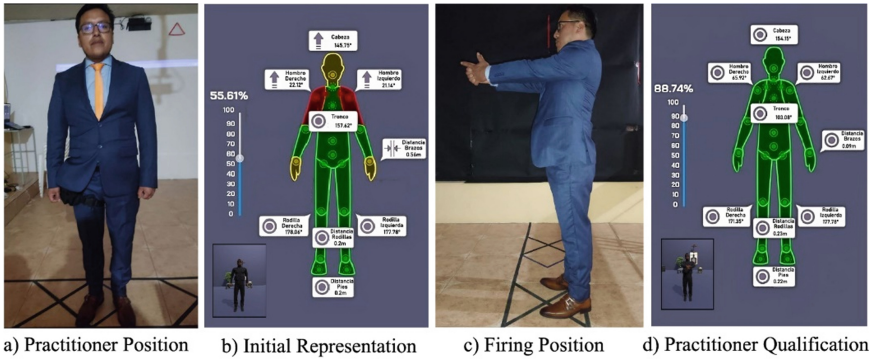


Fig. 4. Scene 1, position training practice

The scene is designed to detect the biomechanical components of the practitioner. It starts with (a) the practitioner's initial position in front of the Kinect, leading to (b) a real-time representation of the practitioner's silhouette. The white boxes represent biomechanical angle inputs. The colors (red, yellow, and green) indicate the good position of each joint. The box at the bottom right projects the practitioner's movements in real-time, the process continues with (c) the adoption of the shooting position and concludes with (d) the evaluation of the practitioner's correct position, represented by a 0–100 score bar on the right side of the figure.

6 Simulation Cases and Results

This section presents the results of 12 simulation cases performed by trainee security officers on the range simulator. This function makes it possible to evaluate and obtain real-time information on the shooter's position. By providing immediate data on the shooting position, the system allows for quick analysis and adjustment, improving training and performance evaluation. The methodology of this research focuses on applied research with a descriptive scope and its design is quasi-experimental. The data are not manipulated, but the situation of the trainee's position at a given time is evaluated. The target population is the students of the private security training center Taurhus CIA. LTDA, between 18 and 48 years of age. As the population is a group of students made up of 9 males and 3 females, the study will be applied to the entire universe.

In this type of simulation, a brief induction on the scenario operation is carried out (see Fig. 4), where the final objective is to determine the relevance of adopting a good firing position. The rating metric establishes that a rating above 70% is determined as good and a rating above 90% is considered as a correct shooting position.

Table 3 shows the positional results of 5 interactions in 25 s and ending with an individualized average, for 12 pupils. The instructors of the CIA Taurhus. LTDA. Instructors consider that this time is sufficient to adopt a correct shooting position. After an introduction to the virtual simulation system, the trainees adjust their positions with real-time feedback. During the exercise, it is observed that in each interaction, the trainees

Table 3. Positioning practices on the virtual simulator

Practitioner	Interaction					
	First	Second	Third	Fourth	Fifth	Average
1	87,23	94,12	78,34	83,23	93,12	87,208
2	73,12	71,23	85,23	87,34	89,23	81,23
3	89,04	87	91,4	87,34	91,5	89,31
4	78.6	89,4	78,23	91	89,23	86,965
5	89,45	93,12	93,5	89,23	95	92,06
6	92	94,23	93,45	93,6	95,34	93,724
7	87.56	89,2	94,23	98,23	92,45	93,5275
8	85,23	88	89,34	93,65	91,23	89,49
9	89,23	93,23	94,2	89,56	96,2	92,484
10	89	92,23	94,23	87,2	94,2	91,372
11	88,3	90	94,2	87,23	89,34	89,814
12	84,23	85,34	87	93,2	90	87,954

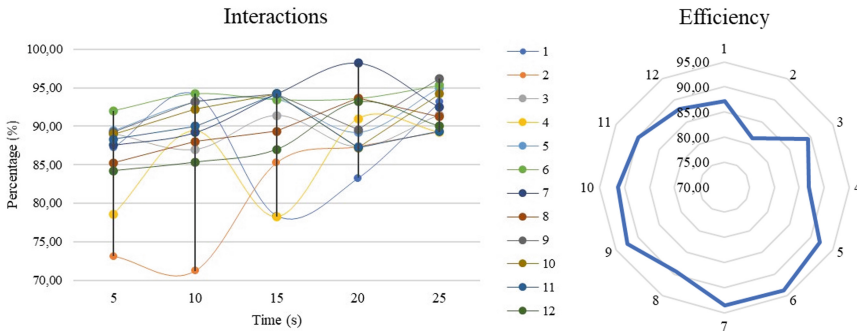


Fig. 5. Comparison of results of five interaction and efficiency by practitioner

obtain good scores thanks to the real-time evaluation of the system, which proves its effectiveness and is approved by the instructors of CIA Taurhus. LTDA.

Figure 5 Interactions: shows that the score of the group of participants is above 70 out of 100 points. Each participant takes the shooting position for 5 s and at the end of the cycle, a new practice begins, thus obtaining individual results that are used for a more efficient analysis. For 25 s, spread over 5 interactions, it is observed that the overall trajectory decreases with each practice. With repeated use of the system, the results become more linear. Efficiency is calculated from 0 to 100, showing that constant interaction with the system leads to an improvement that can be visualized from the first practice to the last. In the first iteration, the practitioners' correct position score ranged from a minimum of 73.12% to a maximum of 92%. In the last iteration, the score ranged

from a minimum of 89.23% to a maximum of 96.20%, indicating an alleged progressive improvement see Fig. 5. First and last interaction. Real-time evaluation allows immediate adjustment of the score without the need for constant intervention by the instructor.

7 Conclusions and Future Work

The research highlighted the effectiveness of the MANFIS model in the assessment of shooting posture as a valuable tool for handling complex data. The virtual simulator of this research met the expectations and the participants showed great interest. Through this simulator, trainees observe their virtualization in 3D and 2D formats (see Fig. 3). This immersive experience, coupled with real-time feedback, facilitates the adoption of an optimal shooting posture. Trainees experience a reduction in body fatigue (see Sect. 1) and an increase in concentration, improving shooter efficiency. The importance of correct posture for safe shooting was emphasized. Despite research in the field of shooting, biomechanics, particularly in MANFIS, is an under-explored area. This study helps to fill this gap and suggests that MANFIS should be incorporated into training programs. Finally, the success of the model suggests its potential use in other areas of shooting and related sports.

This simulator plays a crucial role in helping shooters perfect their shooting position. With 300 postural records obtained by the instructors (see Table 2), the system's effectiveness is constantly monitored. The evaluation is carried out with inexperienced participants on a shooting range. The most remarkable result is an efficacy rate of $89.59 \pm 3.36\%$, thanks to the proposed tool that guides the shooters to perfect their posture and achieve an appropriate position. This drastically improves shooting accuracy and efficiency, providing a competitive advantage.

Shooters can view real-time scores on the simulator, allowing instant adjustments and improved techniques (see Fig. 4). While accuracy in virtual simulators can vary based on various factors, experienced shooters usually achieve around 90% accuracy in physical practice [20]. Tests with the proposed simulator, mainly with new shooters, indicate its potential for high performance and alignment with traditional training.

Future research will broaden the study of shooting positions to cover all body biomechanics and adapt to various weapons. There's a plan to develop virtual shooting training programs that evaluate a practitioner's shooting traits and accuracy. Based on the results, a proficiency report will be produced, confirming the effectiveness of the shooting simulator training.

Acknowledgments. This study is in addition to the research project entitled "Development of video surveillance modes and/or algorithms for the Re-Identification of people based on Soft-Biometrics features in closed-circuit cameras using computer vision and machine learning techniques". We thank all the research assistants of the project, as well as the Intelligent Systems, Data Analysis, Robotics, Cybersecurity, and Software (SIARCyS) Research Group for their support in the development of this work.

References

1. Vučković, G., Jovanović, A., Dopsaj, M.: Povezanost između takmičarske efikasnosti gađanja pištoljem na 20 metara i mehaničkih karakteristika sile različitih mišićnih grupa **10**, 194–201 (2001)
2. Vučković, G., Dopsaj, M., Blagojević, M.: The relationship between 10 m. distance pistol shooting efficiency and indicators of muscle force regulation mechanisms at different groups **28**, 301–302 (2001)
3. Muñoz Gil, L.H.: Propuesta de guía metodológica para la enseñanza de los fundamentos técnicos básicos del tiro deportivo con pistolas neumáticas en deportistas juveniles (2014)
4. Aalto, H., Pyykkö, I., Ilmarinen, R., Kähkönen, E., Starck, J.J.O.: Postural stability in shooters **52**(4), 232–238 (1990)
5. Mon, D., Zakyntinaki, M.S., Calero, S.: Connection between performance and body sway/morphology in juvenile Olympic shooters (2019)
6. Mon, D., Zakyntinaki, M.S., Cordente, C.A., Antón, A.J.M., Rodríguez, B.R., Jiménez, D.L.: Finger flexor force influences performance in senior male air pistol olympic shooting **10**(6), e0129862 (2015)
7. Benavides, M.A.B., Villalba, T.F.R., Saavedra, R.L.Y., Apolo, E.G.C.: Estudio biomecánico del lanzamiento de granada entre deportistas principiantes y de alto rendimiento **36**(2), 228–238 (2017)
8. Lourenço, C.P., Silva, A.L.: Controle postural e sistema vestibulo-oculomotor em atletas de tiro esportivo da modalidade pistola **19**, 313–316 (2013)
9. Mon, D., Zakyntinaki, M., Cordente, C., Barriopedro, M., Sampedro, J.: Body sway and performance at competition in male pistol and rifle Olympic shooters **6**(1) (2014)
10. Lele, A.: Artificial Intelligence (AI): Disruptive Technologies for the Militaries and Security, pp. 139–154. Springer (2019)
11. Lee, B., Kim, J., Shin, K., Kim, D., Lee, W., Kim, N.: A study on the actual precision shooting training based on virtual reality **18**(4), 62–71 (2018)
12. Castro, C., Aguilar, W.J.S.: Ecuador: Universidad de las Fuerzas Armadas ESPE, Desarrollo de un sistema de calificación para un polígono virtual de tiro basado en visión por computador (2018)
13. Rutkowski, L., Cpalka, K.: Flexible neuro-fuzzy systems **14**(3), 554–574 (2003)
14. Gomathi, V., Ramar, K., Jeevakumar, A.S., Engineering, I.: Human facial expression recognition using MANFIS model **3**(2), 288–292 (2009)
15. Mayer, I., et al. (2014). The research and evaluation of serious games: toward a comprehensive methodology **45**(3), 502–527
16. Blazeovich, A., Blazeovich, A.J.: Sports Biomechanics: The Basics: Optimising Human Performance. Bloomsbury Publishing (2017)
17. Rutkowska, D.: Neuro-Fuzzy Architectures Based on the Mamdani Approach. In: Neuro-Fuzzy Architectures and Hybrid Learning, pp. 105–126. Springer (2002)
18. Corti, A., Giancola, S., Mainetti, G., Sala, R.J.R., Systems, A.: A metrological characterization of the Kinect V2 time-of-flight camera **75**, 584–594 (2016)
19. Shotton, J., et al.: Real-time human pose recognition in parts from single depth images. In: CVPR 2011, pp. 1297–1304. IEEE (2011)
20. Viñachi Guerrón, G.F.: Biomecánica del tiro de pistola de aire calibre 22 en deportistas de ambos sexos de pichincha **38**(2), 195–209 (2019)
21. Ihalainen, S., Mononen, K., Linnamo, V., Kuitunen, S.: Which technical factors explain competition performance in air rifle shooting? *Int. J. Sports Sci. Coach.* **13**(1), 78–85 (2018)
22. Kim, M.-S.: The kinematic factors of physical motions during air pistol shooting **26**(2), 197–204 (2016)

23. Wang, L.: *Adaptive Fuzzy Systems and Control*. PTR Prentice Hall, ed: Englewood Cliffs New Jersey (1994)
24. González de Garibay Barba, A.: *Optometría Deportiva. Tiro Olímpico con Arma Corta* (2016)
25. Lee, K.-M., Kwang, D.-H., Wang, H.L.: A fuzzy neural network model for fuzzy inference and rule tuning. *Internat. J. Uncertain. Fuzziness Knowl. Based Syst.* **2**(03), 265–277 (1994)
26. Lee, K.M., Kwak, D.H., Lee-Kwang, H.: Fuzzy inference neural network for fuzzy model tuning. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **26**(4), 637–645 (1996)
27. Rutkowska, D.: *Intelligent Computational Systems: Genetic Algorithms and Neural Networks in Fuzzy Systems*, ed: PLJ, Warszawa (1997)
28. Jang, J.-S., Sun, C.-T.: Neuro-fuzzy modeling and control **83**(3), 378–406 (1995)
29. Kosko, B., Isaka, S.J.S.A.: Fuzzy logic **269**(1), 76–81 (1993)
30. Nadeem, A., Jalal, A., Kim, K.: Human actions tracking and recognition based on body parts detection via Artificial neural network. In: *2020 3rd International Conference on Advancements in Computational Sciences (ICACS)*, pp. 1–6. IEEE (2020)
31. Lucero Urresta, E.K.: Sistema de entrenamiento de tiro de precisión mediante realidad aumentada para el Club Deportivo Especializado Formativo Polygono, Universidad Técnica de Ambato. Facultad de Ingeniería en Sistemas Electrónica e Industrial. Carrera Ingeniería Electrónica y Comunicaciones) (2020)
32. Clements, J.M., et al.: Neurophysiology of visual-motor learning during a simulated marksmanship task in immersive virtual reality. In: *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 451–458. IEEE (2018)



Designing a Neural Network Cascade for Object Detection in Drawing and Graphical Documentation Processing

Kirill Vitko^(✉) and Anna Tikhomirova

National Research Nuclear University MEPhI, Kashirskoe Shosse, 31, Moscow 115409,
Russian Federation
vitko.1998@mail.ru

Abstract. The article is devoted to the problem of recognizing images of different types of rolled metal products during the analysis of the drawing and graphic documentation, which is the basis for the production enterprise to manufacture metal structures. Determining the rolled product type and calculating its consumption solves the problem of automation of keeping stock records and product costing, when it comes to the use of non-unit items, which consumption is measured in meters. Two processing methods were chosen to solve the problem - automatic parsing and neural network cascade, the assembly of which made it possible to obtain a high quality of object search.

Keywords: Automation · Metal structure drawings · Analysis · Computer vision · Neural networks · Detection · Text recognition

1 Introduction

Making customized products carries great risks and costs for the manufacturer, including keeping stock records – how many materials and resources are currently in stock and how much will be spent on each specific order. This is mostly handled by a person in Excel, and automation of this process is not possible due to the large differences in the business processes in each production facility. However, with a detailed analysis, it is possible to create a mechanism that automates the cost accounting process. For this purpose, it is possible and, in most cases, necessary to use various practices of machine learning methods.

2 Analysis of Preprocessing Methods of Drawing and Graphic Documentation

Various application software packages are used to create drawings in the production of metal structures. AutoCAD is quite often used for these purposes. It is a modern automated design system developed by Autodesk for creating drawings and three-dimensional models, the most accurate and productive due to specialized functions aimed

at creating projects for mechanical engineering, architecture, electrical engineering and other areas [1]. AutoCAD stores drawings in the binary dwg format, closed to the external work except for plugins and AutoCAD API. AutoCAD plugins cannot be built into web applications, and AutoCAD APIs are expensive, therefore, it was decided to use an open format for exchanging and working with drawings, DXF. This format was created by Autodesk for AutoCAD and is supported by almost all PC-based CAD systems. Exporting dwg to dxf files with the ODA File Converter utility and parsing the dxf file with the ezdxf library will allow to correctly calculate in most cases, but errors may occur between the formats because of improper export. To fix them, it is necessary to analyze the image stored in the source file's metadata using artificial intelligence methods.

As in many other machine learning tasks, a class of artificial intelligence models - neural networks - is effective, however, even they are not able to show high quality in some tasks [2]. Therefore, each specific task requires some preprocessing of the original image, depending on the application, the task, and the images themselves. They increase the efficiency and quality of extraction and recognition of searched or studied objects. Preprocessing methods depend on research tasks, are quite diverse and may include, for example, extraction of the most informative fragments, their magnification, obtaining 3D images, color mapping, realization of high spatial resolution, increase of contrast resolution, improvement of image quality, etc.

First, it is necessary to analyze the original images and select objects on them, which need to be highlighted and classified. The objects of interest are profiles, corners, squares, sheets and strips. In addition to determining the type of object, it is also important to correctly determine its size, since each of them is a separate item in the warehouse accounting. For this purpose, in addition to detecting the object itself, it is necessary to detect its size marked with text using AutoCAD objects like Text, Multileader and Rotated Dimension. Because of the labor-intensive nature of this task and the problem of matching the size with the object itself, one should try to select objects whose size determination by pixels will have a small error. Such objects are squares, sheets and stripes, while calculation of size of profiles and corners from an image is a very time-consuming task due to complexity of form of these objects, which shown in Fig. 1.

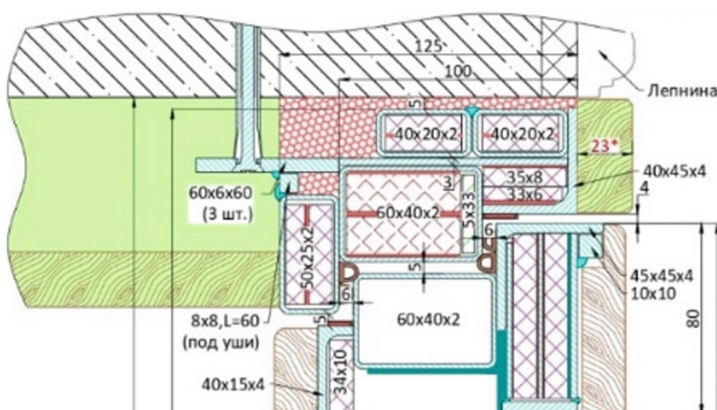


Fig. 1. An example of drawing

Thus, it is necessary to solve two tasks of computer vision, which are detection and classification of objects, and text recognition. All objects have different shapes, and the searched groups of objects (for example, rolled steel, which objects we are trying to find, and bullet-proof steel) have different colors, including the objects we detect. This means that it is possible to use color filters to simplify the original image [3]. Drawing and graphic documentation is characterized by many simple lines, which in turn are the elements of the objects considered – letters. So, preprocessing should help to separate hatchings and letters.

The following methods will be used in this work: grayscale conversion, Gaussian blur, thresholding, binary morphology.

Grayscale images store values for levels of brightness as opposed to color information. A typical grayscale image is 256 shades of gray ranging from 0 (black) to 255 (white). When converting an RGB color image to grayscale all the color is replaced with shades of gray.

Gaussian blur is a low-pass filter that smoothes uneven pixel values in an image by cutting out the extreme outliers. Both grayscale and color images can contain a lot of noise, or random variation in brightness or hue among pixels. The pixels in these images have a high standard deviation, which just means there's a lot of variation within groups of pixels. Because a photograph is two-dimensional, Gaussian blur uses two mathematical functions (one for the x-axis and one for the y) to create a third function, also known as a convolution. This third function creates a normal distribution of those pixel values, smoothing out some of the randomness.

Thresholding separates objects from the surrounding background when the brightness of the object and background pixels are concentrated near the two prevailing values. Typically, a filter is used to produce a bi-level (binary, i.e., black and white) image from a grayscale image, or to remove noise.

When applied to binary images, morphology is used to describe any properties of that image. And operations of mathematical morphology are used for transformations of sets. In this regard, it is convenient to use them to process binary images with subsequent analysis. In order to carry out the operation of binary morphology, we need the original binary image and a structuring element. A structuring element is a description of an area of some form. This area can have any size and shape, which can be represented as a binary image. Structuring elements are some kind of mask, which defines the area of the image, over which the binary morphology operation will be performed. Structuring elements also have a point of origin. It is usually located in the center of the mask, but may be in an arbitrary place. The origin of the structuring element must coincide with the current pixel of the binary image in order to make transformations over it.

Applying these filtering steps will help prepare the image for subsequent processing by object detection models and textual information recognition.

When developing complex models, the first thing to do is to turn to already existing solutions and use them as a lower bound on the quality of the model's performance. For example, the most popular one for text recognition is Tesseract. Tesseract is a free OCR software that was developed by Hewlett-Packard from the mid-1980s to the mid-1990s, and then shelved for 10 years. In August 2006 Google bought it and opened the source code under the Apache 2.0 license for further development.

OCR (optical character recognition) uses neural networks to search for and recognize text in images. Tesseract looks for patterns in pixels, letters, words, and sentences. Tesseract uses a two-step approach called adaptive recognition. It takes one pass through the data to recognize characters, then a second pass to fill in any letters it wasn't sure of with letters that are likely to match a given word or sentence context...

In computer vision tasks training from scratch takes quite a long time due to the specifics of the domain (each image is a set of a huge number of pixels, each of which is perceived by the neural network as a feature). Therefore, neural networks pre-trained on huge image bases, which can be retrained to perform a specific task, or ready-made solutions that require only fine-tuning, are often used. For example, when recognizing text from documents, Adobe Acrobat or a trained neural network running in the shell of the product and in the public domain are typically used [4].

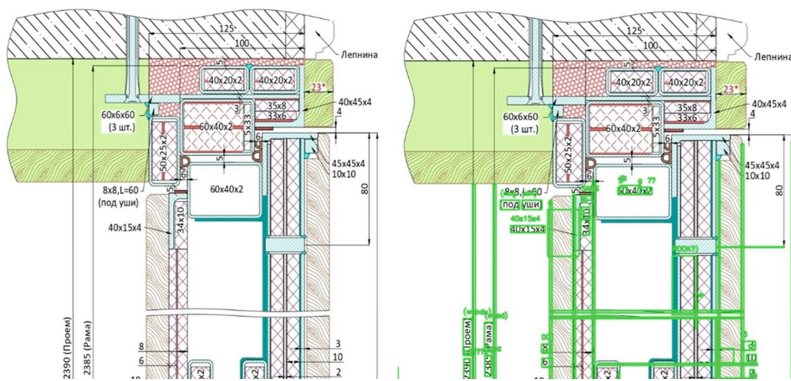


Fig. 2. An example of using the pytesseract library

As shown in Fig. 2, the quality is quite low. This is primarily associated with the fact that, despite attempts to unify the use of these methods, they are developed to recognize text from images similar to scanned pages, photographs of books or at least images of natural nature without text noisiness. Thus, it is necessary to develop a proprietary neural network capable of correctly recognizing textual information on drawing and graphic documentation. The results of the considered solutions can be taken as a lower bound on the quality of the model under development.

3 Designing an Assembly of Preprocessing Methods and Models for Object Detection and Textual Information Recognition

Each of the directions of reducing the number of restrictive frames under consideration has its own leading solutions and approaches. There are two-stage methods (R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN), one-stage methods (YOLO, SSD) and a specific method of Anchor boxes [5].

Two-stage methods, instead of using fixed-size sliding windows to search for images, in the first stage identify a subset of regions, rectangular frames of different sizes, that

might contain an object. This ensures faster and more efficient finding of objects regardless of object size, camera distance, or angle of view. The total number of regions for each image generated in the first step is approximately equal to two thousand.

The YOLO algorithm first divides the image into $N \times N$ grids [6]. If the center of the object falls inside the cell coordinates, then this cell is considered responsible for determining the object location parameters. Each cell describes several variants of bounding box locations for the same object. Each of these options is characterized by five values—the coordinates of the center of the bounding box, its width and height, and the degree of certainty that the bounding box contains the object. It is also necessary to determine for each pair of object class and cell the probability that the cell contains an object of that class. Thus, the last layer of the network that makes the final decision about bounding boxes and object classification works with the dimension tensor

$$N * N * (5 * B + C) * N * N * (5 * B + C) \quad (1)$$

where B is the number of predicted bounding boxes for the cell, C is the number of object classes defined initially [7].

The YOLO algorithm works faster than the algorithms of the R-CNN family because it supports splitting into a constant number of cells instead of suggesting regions and calculating a solution for each region separately. However, poor recognition of objects with complex shapes or groups of small objects due to the limited number of candidates for bounding boxes is indicated as one of YOLO disadvantages.

The SSD model applies the idea of using a pyramidal hierarchy of convolutional network outputs to efficiently detect objects of different sizes. The image is sequentially transmitted to the layers of the convolutional network, which are reduced in size. The output from the last layer of each dimensionality is involved in the decision to detect objects, thus adding up the pyramidal characteristic of the image. This makes it possible to detect objects of different scales, since the dimensionality of the outputs of the first layers strongly correlates with the bounding boxes for large objects, and the last ones for small objects [8].

To get all positions and their correct location, the following steps must be followed:

- apply filters that will reduce the amount of input data;
- apply a neural network that detects objects on drawing and graphic documentation;
- apply a neural network that recognizes text and stores its coordinates;
- compare the size of the object found as text and the classified object.

The first part of the assembly is a consistent application of filters, while the second and third require special attention [9].

The first part of the assembly requires any method to select the ranges of color characteristics of the pixels of the sought objects [10]. In addition, the first part of the assembly identifies potential text, cuts out the part of the image corresponding to it, and matches each such segment with the coordinates of the overall image from which the fragment was removed, and the second part of the model recognizes text in the fragments of images and, in case of detection, matches the coordinates with the already found text.

The processed image is then transferred to the input of the network, and the obtained and classified objects are highlighted in the original image.

The result of the detection and recognition of textual information is shown in Figs. 3 and 4.

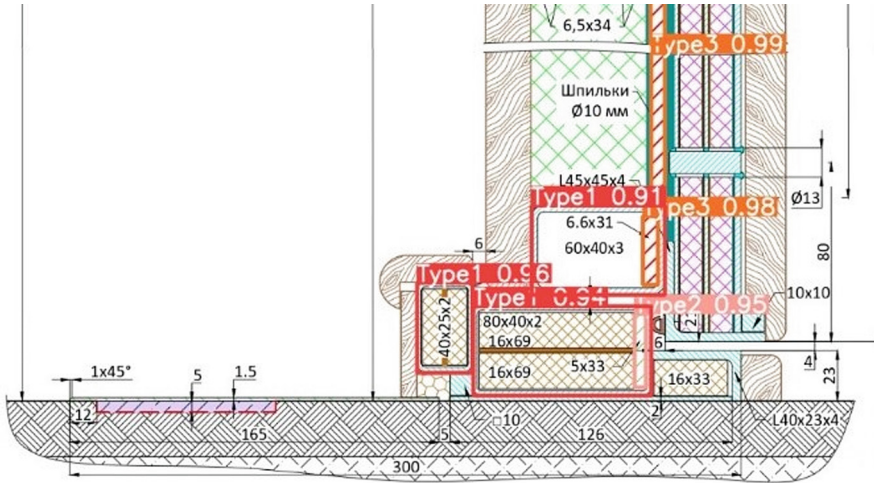
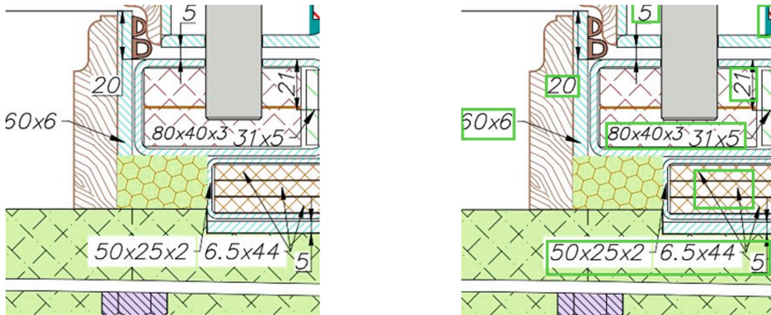


Fig. 3. An example of drawing object detection



50x6 → {'60x6': (0, 107),
 '20': (102, 64),
 '50x25x2 6.5x44 15': (104, 212),
 '80x40x3 31x5': (163, 145),
 '5': (184, 0),
 '': [(205, 174), (319, 0)],
 '21': (291, 62),}

Fig. 4. An example of text recognition in a drawing

The production technologists marked 100 different fragments of various drawings. They identified both objects and text using a special program. In total, 1272 objects were found, including 824 text blocks and 448 objects. Each object and label was considered as a separate entity for recognition, resulting in the test results shown in Table 1.

Table 1. Recognition errors

	True for text	False for text	True for object	False for object
Reject	793	22	439	42
Not reject	31	-	9	–

As shown in the table, the quality of recognizing the test and objects is quite high, however, in the case of objects, the model often finds incorrect areas and recognizes objects on them. To achieve high accuracy in production, this part needs to be improved.

The use of this approach allowed to obtain a high quality of position detection in the drawing, which improves the already implemented dfx-document parser.

4 Conclusion

For automation of the routine calculation and material accounting process for customized products, a cascade was implemented, which included a set of preprocessing methods and two neural networks. The first one solves the problem of object detection, and the second one – the recognition of textual information.

All three approaches described above were implemented. The OpenCV library was used for various transformations over the whole image, such as convolution, changing color saturation or image tones. The pytorch library was used for the implementation of neural networks.

The most efficient of the object detection methods by speed proved to be the YOLO method – about 20 s. Therefore, given the rather high quality of the predictions of all approaches, preference was given to it (more than 90% accuracy and completeness).

According to the results of the development, it was possible to save about 1 man-day in production. In the future, it is planned to expand the functionality by adding the ability to account for scraps and pieces of rolled metal.

Acknowledgements. This work was supported by Competitiveness Growth Program of the Federal Autonomous Educational Institution of Higher Professional Education National Research Nuclear University MEPhI (Moscow Engineering Physics Institute).

References

1. Official website of autodesk. <https://www.autodesk.ru/>
2. Samsonovich Preface, A.: *Advances in Intelligent Systems and Computing*, vol. 848, Q3 (2019)
3. Bradski, G., Kaehler, A.: *Learning OpenCV: OReilly* (2013)
4. Xu, C., Prince, J.L.: *Gradient Vector Flow Deformable Models* (2000)
5. Tensorflow Freezegrph Tool Source Code. https://github.com/tensorflow/tensorflow/blob/master/tensorflow/python/tools/freeze_graph.py
6. Moreno-García, C.F., Elyan, E. and Jayne, C.: New trends on digitisation of complex engineering drawings. *John Macintyre Neural Comput. Appl.* **31**, 1695–1712 (2019)

7. Nurminen, J.K., Rainio, K., Numminen, J.-P.: Object detection in design diagrams with machine learning. *Janusz Kacprzyk Adv. Intell. Syst. Comput.* **27–36**, 2020 (2019)
8. Elyan, E., Jamieson, L., Ali-Gombe, A.: Deep learning for symbols detection and classification in engineering drawings. Vasile Palade, Danilo Mandic, Ariel Ruiz-Garcia. *Deep Neural Netw. Represent. Generative Adversarial Learn.* **129**, 91–102 (2020)
9. Mani, S., Haddad, M.A., Constantini, D.: Automatic digitization of engineering diagrams using deep learning and graph search. In: *CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 673–679 (2020)
10. Samsonovich, A.: A conceptually different approach to the empirical test of Alan Turing. *Procedia Comput. Sci.* **123**, **Q2**, 512–521 (2018)



Crowdsourcing-Based Appropriation of Communicative Behaviour Elements on the F-2 Robot: Perception Peculiarities According to Respondents

Liliya Volkova^{2,3} , Artemy Kotov^{1,3,4} , and Andrey Ignatev² 

¹ National Research Center «Kurchatov Institute», pl. Akademika Kurchatova 1, 123182
Moscow, Russia

kotov_aa@nrcki.ru

² Bauman Moscow State Technical University, 2-ya Baumanskaya ul. 5-1, 105005 Moscow,
Russia

liliya@bmstu.ru

³ Russian State University for the Humanities, Miusskaya pl. 6, 125993 Moscow, Russia

⁴ Moscow State Linguistic University, ul. Ostozhenka 38 str.1, 119034 Moscow, Russia

Abstract. This article is dedicated to F-2 the companion robot and to interpretations of respondents' estimations of designed communicative multimodal behaviour. The affective robot is described: it represents a platform for implementing and verifying various individual behavioural traits for robots. F-2 interprets multimodal input: text, face orientation and tactile signals; it translates the input into facts, which are seeds for further affective behaviour. Facts trigger behavioural patterns for reacting—concurrent scenarios with their activation degrees varying over time. The most activated scenario is implemented via one reaction out of a pool of corresponding scenario reactions. Each reaction is multimodal and includes one or several components: speech, gestures, gazes. Robot behaviour is estimated by human assessors during conducted experiments on communication. Several notable effects were observed during perception of implemented communicative behaviour of F-2. These effects are discussed, they are presumed to be evidence to common human expectations transfer from human-human interaction to human-robot interaction.

Keywords: Human-robot interaction · Companion robots · Affective interfaces · Multimodal communication · Human-machine interaction

1 Introduction

In various languages “to respond mechanically” means to answer without involvement, without affection or taking the interlocutor’s feelings into account. This is a marker of technologic insufficiency of most existing dialogue systems, as well as the fact that most people avoid chat-bots in online services, preferring communication and interaction with a living person. What prevents people from using most question answering

machines is not only their limitedness of scenarios and question types, but importantly a lack of personal touch and affection. People tend to treat robots as social actors, to subconsciously attribute personality traits to them [1] and to transfer psychological aspects of communication from human-human interaction onto human-robot interaction (HRI). In general, people are not only or not always interested in information, they crave for (affective) communication *per se*; this has been recently proved during worldwide lockdowns. It is in general use and of importance to show caring attitude in order to satisfy the addressee's needs in conversation. Applied psychology is highly concerned with enciphering and deciphering additional information from multimodal behavioural patterns, all in intention to increase the communication efficiency and to decrease the psychological discomfort brought up with inappropriate means of expressing one's attitude, of delivering the intended sense in cues. In HRI relative problems can be observed. Question-answering systems along with assistants in smart phones and houses are a global trend nowadays, and regarding the aforesaid, adding some affection into their mechanistic thinking is a problem to solve.

One particular field of dialogue systems is development of affective companions. Given a body, virtual or physical, such a companion would have an ability of nonverbal communication. On one hand, nonverbal communicative behaviour isn't the main channel for communication and is less controlled by the speaker as compared to speech. On the other hand, it is necessary for personal communication, as its elements can highlight the speaker's emotional state or one's attitude towards the topic, as well as it can give additional information about pragmatics of the message [2]. The nonverbal communicative behaviour can express up to 60% of the information [3] and gives a personal touch. The core of an affective reaction is associated with a communicative function (i.e. an intention to express some pragmatics, e.g. hesitation, negation, joy etc., referred to a particular stimulus by the addressee). Pragmatics of particular multimodal expressions can be investigated basing on a corpus of recorded multimodal emotional reactions, annotated with implementations of communicative functions. As to choice, corpora recorded with actors show pretty clear portraits of emotions; still, these are not fully applicable within a task of simulating communicative behaviour where communicative functions and their expressions can be blended. A more promising approach is investigating natural dialogue recordings, such as TV shows records in EmoTV [4], cinema in Multimedia Russian Corpus [5], oral exams and happy people recordings in Russian Emotional Corpus [2]. Affective reactions can be adapted onto companion robots and 3D avatars of dialogue systems.

Modern companion robots interact with people in different modes: via speech, gestures and gazes. Although the tactile channel is rarely the preferred mode of interaction for social robots, it becomes important within the two major areas. First, children tend to establish tactile contact with attractive robots, for they are used to tactile interaction with their toys. People contacting with robots at home or at a private space also tend to establish tactile communication with appropriate robots; that observation has helped to create a growing segment of tangible companion robots [6, 7]. Tactile feedback can be also of a great importance, if robots are used for therapy of people with health deficits or people passing medical treatment at a hospital [8–11]. Secondly, those who have just encountered a robot for the first time, try to evaluate its adequacy by checking some basic

cognitive responses: people wave with their hand in order to attract the robot's attention and to check its gaze response, they say hello or ask a simple question (e.g. "What is your name?") to check its speech competence, and in many cases they slightly touch the robot to get its behavioural response to that. So, the social touch, although not frequent in modern social interaction space, becomes an important feature for companion robots. In this work we design a tactile input processing subsystem for the companion robot F-2, we extend its basic responsive features with those aimed at tactile stimuli, and organize the crowdsourcing procedure to check the perception of behavioural cues. Beyond this, we discuss a number of observations given by our respondents: the more the variety of opinions and of perception peculiarities is accounted, the more aspects of perception of robot's behaviour are discovered. F-2's behavioural diversity in social interaction is often source for hypotheses and sometimes for a more narrow experimental investigation of psychological aspects, transferred from human-human interaction onto human-robot interaction.

2 Overview

That's the way the emotional intellect affects the robot's behavior. The same event can affect us in different ways, depending on our mood or on circumstances.

"Eva" by Sergi Belbel, Cristina Clemente, Martí Roca, Aintza Serra

AI classics gave birth to a dream of a different kind of robots, exceeding the limitation of pure automation of heavy and sometimes dangerous labor. Alan Turing test has been finally passed (on the simulated ground of interlocutor's youth), the Loebner Prize has been awarded multiple times, which proves that dreams of humanity made their first steps in real life. Dialogue systems started resembling people in their mechanical minds and in the way they think (SIR showed simple logic for inference and solving straightforward logical problems [12]) and conduct conversation: ELIZA [13]—via rule-based reformulation of input, A.L.I.C.E. [14]—via heuristic matching of an input phrase to samples in its knowledge base, PARRY [15]—via dialogue strategies. SHRDLU [16] modeled commands understanding and performing, particularly it could state a lacuna and ask what would the missing term mean. CYC [17, 18] performed human-like reasoning and adapted to novel situations.

Affective robots are the next step on the stairway of robotic companions: we're looking forward to *a future sort of friendship, of a robot-human kind*, according to Cynthia Breazeal. Simulating surface emotional phenomena [19] is perceived in HRI as pleasant, it contributes to establishing closer contact and affection towards robots. The first robotic toy which shortly became object of attachment was Tamagotchi, and the corresponding effect of affection was named after it. As to toys, one particular problem is their limitedness of reactions: people get used to that and lose their interest. Hence, high variability of reactions scenarios is of big importance.

Several breakthroughs were made in the classic field of dialogue systems, inspired by AI pioneers. These are affective robots and agents based on text semantics processing and simulating surface emotional phenomena: IBUG [20], SEMAINE [21], Greta [22], Kismet [19], Max [23, 24]. The latter is interesting for changing its emotional state over time:

it represents emotions via an activation axis and a valency axis. Each emotion is a point in this space, e.g. two ‘happiness’ points may differ: one having a high activation level, the other one having a low one—a passive sort of happiness. These projects model more natural reactions with behavioural elements which enable simulating surface emotional phenomena (referred to as affective behaviour), and often infer displaying emotional mimics or gestures adapted from human ones.

In communication, speech and gestures often cooccur and coexpress the speakers’ message as a composite signal [25–27]. Hence, multimodal communication should combine elements of behaviour targeted at several layers of perception. The BML (behavior markup language, sort of a lingua franca for behaviourists in robotics [28]) format is handy: it is flexible and allows multiple layers in a behaviour frame, in particular speech, gestures, gazes, mimics.

Implementation of personality for social robots is a prospective HRI field. First, there is such an opportunity due to dialogue systems worldwide spread via smartphones. Second, while robots (and this general tendency covers AI as well, beyond fundamental science) are considered useful tools rather than social function carriers [29], the latter function is getting more and more actual for social robotics. It is stated that “personality is a key element for creating socially interactive robots”, and that “studies on this dimension will facilitate enhanced human–robot interaction” [30]. This is due to providing users with better affordance, which makes it intuitive and natural for the users to understand the robot’s behaviours [31]. Personality “represents those characteristics of the person that account for consistent patterns of feelings, thinking, and behaving” [32]. Particular traits of personality can distinguish one robot from others and impress people more, which leads again to personal touch [33].

3 F-2: The Platform for Implementing Affective Behaviour Elements and Personal Traits

3.1 A Communicative Agent Inside a Robot

—*Max, what’s your emotional level?*

—*Standard eighth.*

—*We’re not accustomed to that much emotional robots. Turn it down to 5.*

“*Eva*” by *Sergi Belbel, Cristina Clemente, Martí Roca, Aintza Serra*

F-2 first was developed as an agent comprehending text and expressing its artificial attitude to the input. Classic text analysis stages are executed. Morphological analysis operates with a morphological dictionary of 100,000 lemmas, resulting in a stack with annotated tokens. Next, syntax and semantics are extracted to form a dependency tree. The stack head is reduced when possible with any of the existing 850 syntactic rules: each rule can reduce a list of tokens in its right-hand side to the left-hand side head (e.g. to form a predicate group out of a predicate and its object). Several trees can exist simultaneously (in case different rules can be applied), unsuccessful trees are rejected. In the resulting tree, nodes are annotated with semantic roles (valencies) and with semantic features (semantic markers). Facts are to be extracted afterwards according to a set of

templates, e.g., a subtree with following semantic valencies in nodes: {*agens, predicate, patiens, instrumentative*} (subject, verb, object, instrument). It is similar to triads [34, 35] and frames [36], but we store more than one feature per valency in most cases, see example on Fig. 1.

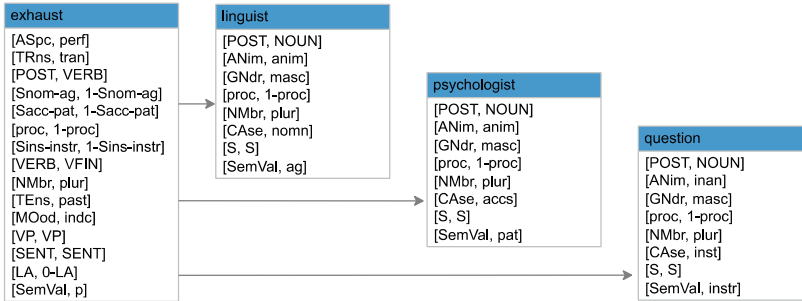


Fig. 1. A sample syntactic dependency subtree corresponding to a semantic fact template.

Facts are source for forming and expressing an attitude. We adopted the approach of concurrent states, the dominating one taking control in order to produce a reaction to the input: each input fact is matched to communicative goals (intentions to implement, standing for communicative strategies, e.g. to express happiness, to show negation, to draw attention, to hesitate etc.). The most activated goal implies selection of a multimodal response out of a pool of possible reactions corresponding to this goal. While the selected reaction is produced, all of the activated goals remain for further processing. Once the current reaction is complete, the corresponding goal is deactivated, and the next topmost goal is selected from those having non-null activation degree. There is a background goal implementing the standby mode via a small constant activation degree. In order to look awake when unoccupied, F-2 slightly moves all the time instead of freezing as if turned off: in between stimuli it looks at its hands or watches the ceiling, it moves its gaze as if thoughtful. All of the implemented elements of behaviour are selected and translated into behavioural elements for F-2 from our REC corpus (Russian Emotional Corpus) [2], and then appropriated.

F-2 is taught to receive multimodal input as facts: from text (we use our own syntactic and semantic parser along with Yandex SpeechKit [37] for speech-to-text translation and for further speech synthesis) and from video (human face orientation is perceived [38]). The reaction is formed as a BML frame, its components are played by different modules of our robot control subsystem [38], performing speech synthesis, gesticulating, gazes. F-2 has 6 Dynamixel AX-12 motors (2 per each hand and 2 for its neck, enabling F-2 to nod and to turn the head left and right). An LCD monitor stands for the face. The robot is connected to a PC.

Research based on REC showed that dominant channels of nonverbal communication are hands (20% reactions), head (37%) and mimics (~ 30%), while body is used relatively rarely: the average is 6,6% cases per communicative function, except for 3 functions—absence/impossibility (35%), separation (97%), inspiration (34%) [2].

Hence, body movements are not obligatory for affective companions, as to Russian (different cultures have different gesticulation peculiarities and usage frequencies). This is the reason why F-2 gesticulates only with its hands and head.

3.2 Perception of Tactile Signals

People tend to develop excessive expectations of social robots due to generalization from behaviour and mental models of human kind [39]. The team and our respondents thought of F-2 only as of a talking head with mimics and hands, until one particular series of experiments. When making acquaintance with F-2, youngsters happened to involve the tactile channel of communication, trying to scratch or to tap the robot—and got upset for their touches were ignored. Thus an expectation gap [39] was discovered: our respondents expected F-2 to perceive touch as well, as pets do.

In order to fill this perception gap we developed a module perceiving tactile signals. Basing on an overview conducted in [40], we selected flexible resistive tactile sensors of appropriate size to place in palms, on cheeks and on top of head (2 on every cheek, 3 on top)—round sensors \varnothing 1.2 cm from the Interlink FSR 400 series [41]. We also placed 2 square sensors of the same series on F-2's belly. Since these sensors don't have an interface for connecting to a computer nor to a controller, they were connected using a third-party programmable controller, via an Arduino UNO board.

A software module was created for analysis of tactile signals: every sensor states the force of pressure. A training set was recorded; the overall data amount is 2.5 k recordings. The decision tree classifier showed the topmost precision, overwhelming logistic regression, Bayesian and support vector machine. The developed module classifies tactile input with precision from 75% to 100% for 5 touch types: tap (or simple touch), smoothing over, scratching, handshake, hit. Smoothing over and scratching are registered for head, cheeks and belly zones. Hit is registered for head and cheeks. Tap is allowed for all of the zones, including hands. Following four target communicative goals are implemented: to focus on the touched zone, to show discomfort, to express pleasure, to express sadness and pain. These communicative goals are part of F-2's set of goals; they cover five touch types and four touch zones.

4 Results

4.1 Evaluation of Communicative Strategies Implementation

Since we bridge the affective divide that parts human interlocutors from most AIs, human estimations of F-2's affective reactions are what matters in the first place. The most important and the main criteria for estimating designed behaviour in whole and in part are assessing and interpreting impressions of human interlocutors. We conduct experiments with assessors conversating with F-2 (or two F-2 robots which show slightly different behaviour). Assessors are asked to interact with F-2 in particular settings and to evaluate the robot's reactions in terms of "human-like", "realistic", "nice". Beyond the approbation, priceless remarks sometimes give us food for thought. We aggregate feedback on different stages of F-2's evolution. This is crowdsourcing: all of the respondents are volunteers invited to interact with the robot.

In one series of experiments we check one particular function of F-2’s, reducing its functionality—of maintaining dialogue, recognizing people, reacting to gazes—in order to evaluate the function subject to research. Particular results on emotional expressions evaluation can be found in [42], F-2’s gaze behaviour is analyzed in [38], its use of politeness strategies—in [43]: F-2 helps learning words in Latin, it gives a word in Russian and expects the answer in cycle. When one tries to guess the word and fails, this is an emotional situation, a threat to one’s social face [44]. We essay at working with this embarrassment: in its multimodal response (a) F-2 points at the error (“Incorrect!”), (b) it ignores it, giving a hint (“it’s similar to...”) or (c) uses a politeness formula (a hedge: “no, *a bit* incorrect”), mitigating the face loss. In the reverse setting, F-2 loses its social face and uses multimodal hedges to compensate this: hedges contribute not only to expressing politeness, but to expressing emotional and cognitive states as well—the robot is perceived as nervous and hesitating [43].

4.2 Assessing Perception of Reactions on Tactile Input

Feedback from respondents is the most important factor for developing the robot. 24 respondents were invited to evaluate F-2’s reactions to tactile input. The feedback was systematized and interpreted; results are given in Table 1. Overall expressed attitude is positive. Respondents express positive opinions to the new functionality of F-2.

Table 1. Feedback from respondents: attitude grouped by touch type.

Touch type	Positive	Negative	Neutral	Number of responses
Smoothing over	90%	1%	9%	68
Tap	53%	28%	19%	78
Hit	82%	11%	7%	44
Scratching	68%	6%	26%	31
Handshake	52.5%	27.5%	20%	40
Overall	71%	13%	16%	249

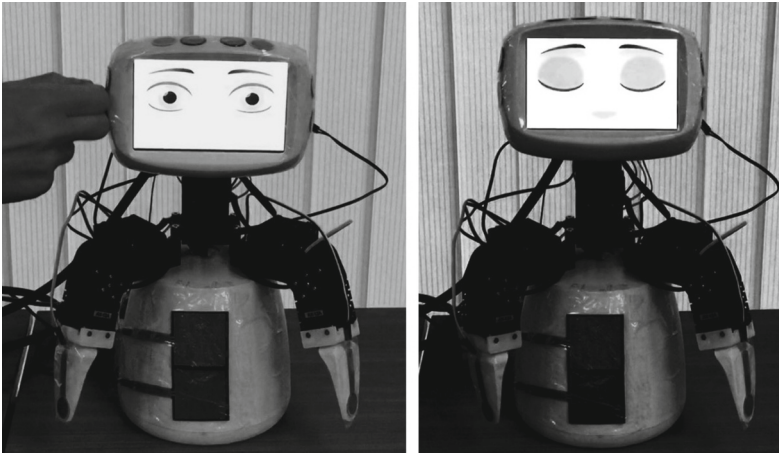
During experiments, our respondents introduced 2 new touch types: tickling the belly (associated with scratching) and holding or shaking both hands (handshake touch group). In Table 1 these are included in associated basic groups, in Table 2 both are stated separately and marked with asterisks, thus basic types being presented clearer.

In comparison of handshake in Table 1 and handshake subdivided in Table 2, basic handshake positive estimations changed from 52.5% cases to 60%. This difference is due to neutral or negative perception of F-2’s reactions to an unforeseen touch type “hold hands”, which was classified as one of predefined touch types. Produced reactions didn’t meet respondents’ expectations—an expectation gap again. On the other hand, separating “tickling the belly” from basic scratch makes it evident that this new type seems to fit existing patterns of reacting. Obviously, it is prospective to highlight new touch types and to develop corresponding reactions patterns.

Table 2. Feedback, attitude by touch type: basic and 2 new types introduced by respondents.

Touch type	Positive	Negative	Neutral	Number of responses
Smoothing over, tap, hit	72.6%	14.7%	12.6%	190
Scratching, basic	75%	8%	17%	24
Tickling the belly *	43%	0%	57%	7
Handshake, basic	60%	29%	11%	35
Hold both hands *	0%	20%	80%	5
Overall, basic types	73.5%	16.5%	10%	237
Overall, new types *	25%	8%	67%	12

Figure 2 shows an example: F-2 expresses pleasure (with lifting its head, closing its eyes and saying “cool” with slight raising hands) as reaction to smoothing over.

**Fig. 2.** F-2 modification equipped with touch sensors reacts to smoothing over its cheek.

5 Discussion

5.1 Perception of Positive Reactions in Case of Low Variability of Reactions

—Is there anything I can help you?

—Yes. Turn your emotional level back to 8.

“Eva” by Sergi Belbel, Cristina Clemente, Martí Roca, Aintza Serra.

During approbation of the tactile input detecting subsystem, our experiment wasn’t aimed at detailed conversation, thus respondents weren’t asked to give cues for further

discussion with F-2. So the robot didn't have the opportunity to be talkative: it was mostly touched and sometimes hit (no robot was harmed). F-2's reactions spectrum was limited by the context to reactions associated with the predefined communicative goals enlisted in Sect. 3.2. In this setting, we collected comments from assessors who came in small groups and actually saw the robot communicating with other respondents for a while before or after taking one's turn. Their prolonged observations resulted in a few particular surprising comments, which can be summed as follows: the robot mostly shows positive reactions, it is "too happy", which was estimated by two people as "unnatural, for life isn't mostly positive".

Keeping in mind that the reactions observed were limited by target goals within the experiment under discussion, the F-2 team concluded the following. First, assessors tend to lose interest in case of prevailing homogeneous reactions in context of less informative communication (as compared to the context of applying core F-2 functions of conducting conversation and commenting on cues with a touch of personal touch [42, 45]). As this less talkative setting is possible during exhibitions when people first meet F-2 and gather first impressions (contrary to sharing thoughts during conversation), in such a case the robot's behaviour could be balanced with at least two more communicative strategies: to make acquaintance (a face recognition module along with a long-term faces memory is in progress) and to present oneself to unfamiliar people who showed interest via tactile actions. Second, a goal of expressing the need in distancing (which was previously considered as possible, but not primary for implementation) is prospective, perhaps with a shade of irony. E.g. the "tap" touch on the head or on the belly might result in a cue "I'm not THAT social" with shaking its head, or in a cue "Stay away from wild robots!", which is a not rude variation of "keep your hands off me" (the latter harsh one is absolutely not to be implemented).

Our respondents were asked to punch the robot in different zones, but most of them stated they didn't want to or weren't willing to. Those who tried punching can be subdivided into two groups: (1) parameterizing the punch force in order to find the threshold separating punch from simple touch, and (2) writing afterwards that they had feared to hurt the robot. The 2nd subgroup outnumbered cold-minded naturalists (and partly intersected it). A possible research direction arises: as we developed a communicative robot intended to be a pal, a friend, it is of interest to estimate general expectations of assessors in such sort of HRI (the approach described in [39]).

5.2 Righty or Lefty?

Assessors sometimes point out particular observed behaviour elements and draw their conclusions, which surprise us. During approbation of F-2's reactions to tactile stimuli, one assessor concluded the robot was right-handed: most reactions were shown with the right hand. Waving or pointing at something are performed by the right hand in the gestures database, which seems to correspond to the following consideration: people most frequently perform gestures with the leading hand. On the other hand, it is only one of possible points of view, depicting only part of the whole. Gestures of the second hand can be associated with the emotional intelligence, with enhancing metaphor explanation [46]. Particular investigations show that semantics partially determines hand choice for gesture production. According to [47], spatial aspects of a message determine the choice

of the right or left hand for gesturing (e.g., use of left hand to gesturally depict an object moving in the relative left position). Speakers tend to use their dominant hand to represent messages with positive connotations in political debates [48]; an assumption was made that emotional valence (positive–negative) matched to the way right- and left-handers represent valence (e.g., the dominant side, either left or right, is positive) may determine hand choice for gesturing [49].

In future research, a robot can be specified as a lefty or as a righty, or balanced towards ambidexterity. Non-dominant hand usage for expressing rather emotional gestures (as contrary to strictly informative) is prospective as well, and it is subject to investigate in context of approbation with invited respondents and psychologists.

5.3 Gender Perception

According to [50], many of the gender-related perceptions and expectations formed in human-human interactions may be inadvertently and unreasonably transferred to interactions with social robots. Human perception is affected by gender-related expectations when judging both humans and robots with minimal gender markers, such as voice or even a name (which is the case of F-2: it recently got a higher voice as compared to the previous, both synthesized). We didn't conduct intentional experiments on gender perception yet, but our respondents already provided us with first results.

An experiment was carried out with two F-2 robots in order to compare perception of gazes-based communicative behaviour. When human face orientation changed, the video analysis software module detected one's gaze direction. If it was towards one of robots, the corresponding robot performed its reaction. In our simulation the left robot turned towards the interlocutor and looked straight at one's face, while the right one looked away after eye contact. The interesting fact is, the person with the highest emotional intellect estimation among present respondents (according to a test we conducted; emotional intellect is briefly defined as the ability to perceive, understand, and manage emotions [51–54]) assumed the left robot to be a boy, and the right one to be a girl. The motivation is as follows: the robot which doesn't look away demonstrates traits of "courageous and masculine behaviour", compared to the "shyness" of the right robot, associated by the interlocutor with rather feminine behaviour. Such perception peculiarities depend on a particular person's experience, for the cited traits may be associated not only with gender, but also with context. In any case, this difference might be one of important traits for developing different personalities for robots.

During the experiment dedicated to robot's reactions on tactile signals, our respondents either didn't state F-2's gender at all, or asked if it was a girl. This happened after we re-designed the robot's eyes: they were surrounded with gradient grey color for a more expressive look. That new feature turned out to look like cosmetics when viewed from the side, and was interpreted as a feminine feature by most respondents among that third part of those who paid attention to the gender aspect.

5.4 Effect of Enlightenment of Mutual Understanding

Last but not least, every human being starves for being accepted and understood, and people tend to transfer their communicative expectations from human-human to human-robot interactions [50]. F-2 was created as such a character showing its understanding and replying affectively (as taught to simulate it): it doesn't perform commands, but it can hold conversation via expressing its attitude to human interlocutor cues. We conduct a series of experiments on storytelling: F-2 comments on phrases, expressing its artificial attitude, and/or roves its gaze, which is often interpreted as thinking on the story [38], as a perceptual ability (though not necessarily staring at the interlocutor, as in [55]). During one experiment, the robot gave its cues as usual, this time including "Hmm" and "I understand" repeatedly. One respondent gave a feedback of awe, pointing at being understood. Thus, one can see the evidence to success of the adopted approach of imitation modelling [56], which is transferring human behaviour elements associated with emotions in human-human interaction onto the robot, and simulating affective reactions driven by communicative goals. Successfully selected and adopted communicative behavioural strategies meet human expectations in HRI, which is proved experimentally, according to feedback given by respondents.

6 Conclusion

F-2 is designed as a communicative robot, which comments on interlocutor's cues and multimodal behaviour and expresses its opinion with artificial affect. In this article we presented the introduction of a tactile input channel along with approbation results. Human-machine communication was set up with respondents who gave their opinions on what they experienced. We discuss selected observations grounded on human estimations of implemented behavioural strategies of F-2, as this crowdsourced feedback is the most important quality metric for this sort of HRI projects. Effects of transfer of expectations and perception from human-human interaction to human-machine interaction are highlighted. Presence of these effects enables us to conclude that F-2 is a successful project as a robot-companion: not only its behaviour is perceived as human-like and estimated mostly positively, but our respondents also treat it like a personality (while we fill their expectation gaps when discovered).

Due to the F-2 architecture, more behavioural strategies can be implemented along with personality traits. In our architecture, input facts trigger several communicative goals, which is quite universal, as this mechanism enables researchers to balance existing strategies via varying activation functions decrease principle: slowly fading activation for relatively important events, and no fading for background processes (e.g. the stand-by mode). Furthermore, tempers can be simulated on a robot basing on implementing different sets of activation functions, e.g. following an approach stated in [57]: 4 basic tempers could be modeled via activation and deceleration. With varying goals importance via initial activation values and fading rules (i.e. the activation function form), different tempers can be tuned. These are considered as prospective research directions along with conducting comparative analysis of perception of various culture-related, gender-related, social role-related behavioural patterns.

Acknowledgements. The reported study was supported by the grant of the Russian Science Foundation № 19–18-00547, <https://rscf.ru/project/19-18-00547/>.

The F-2 team wishes to express gratitude to all of our respondents, including students of the Power Engineering department of BMSTU. All of the crowdsourced feedback is precious for us as source of knowledge and as seeds of future research.

References




1. De Graaf, M. M.A., Ben Allouch, S.: Expectation setting and personality attribution in HRI. In: HRI, ACM/IEEE International Conference on Human-Robot Interaction, pp. 144–145. IEEE, Piscataway (2014)
2. Kotov, A.A., Zinina, A.A.: Functional analysis of non-verbal communicative behavior (in Russian). In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue”, 14(21) vol. 1, pp. 308–320. RSUH, Moscow (2015)
3. Engleberg, I.N., Wynn, D.R.: Working in Groups: Communication Principles and Strategies, My Communication Kit Series, p. 133. Allyn & Bacon, Boston (2006)
4. Martin, J.-C., Devillers, L.: A multimodal corpus approach for the study of spontaneous emotions. In: Affective Information Processing, pp 267–291. Springer-Verlag, Heidelberg (2009)
5. Grishina, E.A.: Gestures and grammatical features of speech act. In: Multimodal Communication: Theoretic and Empiric Investigations (in Russian), pp. 25–47. Buki Vedi, Moscow (2014)
6. Gansohr, C., Emmerich, K., Masuch, M.: Hold me tight: a tangible interface for mediating closeness to overcome physical separation. In: Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, vol. 178, pp. 74–85. Springer, Heidelberg (2017)
7. Paiva, A., Chaves, R., Piedade, M., Bullock, A., Andersson, G., Höök, K.: SenToy: a tangible interface to control the emotions of a synthetic character. In: Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 1088–1089. ACM, New York (2003)
8. Shibata, T.: Therapeutic seal robot as biofeedback medical device: qualitative and quantitative evaluations of robot therapy in dementia care. *Proc. IEEE* **100**, 2527–2538 (2012)
9. Shibata, T., Kawaguchi, Y., Wada, K.: Investigation on people living with seal robot at home. *Int. J. Soc. Robot.* **4**, 53–63 (2012)
10. Inoue, K., Wada, K., Shibata, T.: Exploring the applicability of the robotic seal PARO to support caring for older persons with dementia within the home context. *Palliat. Care Soc. Pract.* **15** (2021). <https://doi.org/10.1177/26323524211030285>
11. Takayanagi, K., Kirita, T., Shibata, T.: Comparison of verbal and emotional responses of elderly people with mild/moderate dementia and those with severe dementia in responses to seal robot, PARO. *Front. Aging Neurosci.* **6**(SEP), 257 (2014)
12. Raphael, B.: SIR, a computer program for semantic information retrieval. In: Minsky, M. (ed.) *Semantic Information Processing*, pp. 33–144. M.I.T. Press, Cambridge (1968)
13. Weizenbaum, J.: ELIZA. *Commun. ACM* **9**, 36–45 (1966)
14. Wallace, R.S.: The Anatomy of A.L.I.C.E. In: Epstein, R., Roberts, G., Beber, G. (eds.) *Parsing the Turing test*, pp. 181–210. Springer Science+Business Media, London (2009)
15. Colby, C.M.: *Artificial Paranoia: A Computer Simulation of Paranoid Processes*. Elsevier Science Inc., New York (1975)
16. Winograd, T.: *Understanding Natural Language*. Academic Press, New York (1972)

17. Lenat, D.B., Borning, A., McDonald, D., Taylor, C., Weyer, S.: Knoesphere: building expert systems with encyclopedic knowledge. In: Proceedings of the Eight International Journal of Conference on Artificial Intelligence, IJCAI'83, vol. 1, pp. 167–169. ACM, New York (1983)
18. Lenat, D., Prakash, M., Shepherd, M.: CYC: using common sense knowledge to overcome brittleness and knowledge acquisition [sic] bottlenecks. *AI Mag.* **6**(4), 65–85 (1986)
19. Breazeal, C.: *Designing Sociable Robots*. MIT Press, Cambridge (2002)
20. i bug. <http://ibug.doc.ic.ac.uk/>. Last Accessed 18 May 2023
21. Shröder, M.: The SEMAINE API: towards a standards-based framework for building emotion-oriented systems. *Adv. Hum. Comput. Interact.* **2010**, 319406 (2010)
22. Pelachaud, C.: Greta: a conversing socio-emotional agent. In: Proceedings of the 1st ACM Sigchi International Workshop on Investigating Social Interactions with Artificial Agents, pp. 9–10. ACM, New York (2017)
23. Max. <http://cycling74.com/products/max/>. Last Accessed 18 May 2023
24. Becker, C., Kopp, S., Wachsmuth, I.: Simulating the emotion dynamics of a multimodal conversational agent. In: ADS 2004, LNAI, vol. 3068, pp. 154–165. Springer, Heidelberg (2004)
25. Engle, R.A.: Not channels but composite signals: speech, gesture, diagrams and object demonstrations are integrated in multimodal explanations. In: Gernsbacher, M.A., Derry, S.J. (eds.) Proceedings of the Twentieth Annual Conference of the Cognitive Science Society, pp. 321–326. Psychology Press, London (1998)
26. Kelly, S.D., Ozyürek, A., Maris, E.: Two sides of the same coin: speech and gesture mutually interact to enhance comprehension. *Psychol. Sci.* **21**, 260–267 (2010)
27. Kendon, A.: *Gesture: Visible Action as Utterance*. Cambridge University Press, Cambridge (2004)
28. Vilhjálmsón, H., et al.: The behavior markup language: recent developments and challenges. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007, LNCS, vol. 4722, pp. 99–111. Springer, Heidelberg (2007)
29. Severinson-Eklundh, K., Green, A., Huttenrauch, H.: Social and collaborative aspects of interaction with a service robot. *Robot. Auton. Syst.* **42**, 223–234 (2003)
30. Miwa, H., Takanishi, A., Takanobu, H.: Experimental study on robot personality for humanoid head robot. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, vol. 2, pp. 1183–1188. IEEE, Piscataway (2001)
31. Hara, F., Kobayashi, H.: Use of face robot for human-computer communication. In: IEEE International Conference on Systems, Man and Cybernetics, Oct. 1995, pp. 1515–1520. IEEE, Piscataway (1995)
32. Pervin, L.A., John, O.P.: *Personality Theory and Research*. Wiley, New York (1997)
33. John, O.P., Srivastava, S.: The Big-Five trait taxonomy: history, measurement, and theoretical perspectives. In: Pervin, L.A., John, O.P. (eds.) *Handbook of Personality: Theory and Research* (2nd edn.), pp. 102–138. Guilford Press, New York (1999)
34. Simmons, R.F.: Storage and retrieval of aspects of meaning in directed graph structures. *Commun. ACM* **9**, 211–214 (1966)
35. Simmons, R.F., Burger, J.F., Long, R.E.: An approach towards answering English questions from text. In: Proceedings on fall Joint Computer Conference, pp. 357–363. Spartan, New York (1966)
36. Minsky, M.: A framework for representing knowledge. In: Winston, P.H. (ed.) *The Psychology of Computer Vision*. McGraw-Hill, New York (1975)
37. Yandex SpeechKit. <https://cloud.yandex.com/en/services/speechkit/>. Last Accessed 18 May 2023
38. Velichkovsky, B.M., Kotov, A., Arinkin, N., Zaidelman, L., Zinina, A., Kivva, K.: From social gaze to indirect speech constructions: how to induce the impression that your companion robot is a conscious creature. *Appl. Sci.* **11**(21), 10255 (2021)

39. Kwon, M., Jung, M., Knepper, R.: Human expectations of social robots. In: 11th ACM/IEEE International Conference on Human-Robot Interaction, pp. 463–464. IEEE, Piscataway (2016)
40. Ignatev, A., Volkova, L.: On equipping F-2 the affective robot with tactile sensors. In: Uvaysov S.U., Ivanov I.A. (eds.) Information Innovative Technologies: Materials of the International Scientific–Practical Conference, pp. 123–127. Association of Graduates and Employees of AFEA Named After Prof. Zhukovsky, Moscow (2021)
41. InterLink Electronics. FSR 400 Series. <https://www.interlinkelectronics.com/fsr-400-series>. Last Accessed 18 May 2023
42. Zinina, A., Zaidelman, L., Kotov, A., Arinkin, N.: The perception of robot’s emotional gestures and speech by children solving a spatial puzzle. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue”, vol. 19 (26), pp. 811–826. RSUH, Moscow (2020)
43. Malkina, M., Zinina, A., Arinkin, N., Kotov, A.: Multimodal hedges for companion robots: a politeness strategy or an emotional expression? In: Selegey, V.P., et al. (eds.) Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, Issue 22, pp. 319–326. RSUH, Moscow (2023)
44. Brown, P., Levinson, S.C.: *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge (1987)
45. Kotov, A.: D-script model for synthesis and analysis of emotional speech. In: SPIIRAS (ed.) Proceedings on SPECOM’2004: 9th Conference “Speech and Computer”, Saint-Petersburg, Russia, September 20–22, 2004, pp. 579–585. ISCA Archive (2004). http://www.isca-speech.org/archive/specom_04. Last Accessed 18 May 2023
46. Argyriou, P., Mohr, C., Kita, S.: Hand matters: left-hand gestures enhance metaphor explanation. *J. Exp. Psychol. Learn. Mem. Cogn.* **43**(6), 874–886 (2017)
47. Lausberg, H., Kita, S.: The content of the message influences the hand choice in co-speech gestures and in gesturing without speaking. *Brain Lang.* **86**, 57–69 (2003)
48. Casasanto, D., Jasmin, K.: Good and bad in the hands of politicians: spontaneous gestures during positive and negative speech. *PLoS ONE* **5**(7), e11805 (2003)
49. Park, W., Choi, W., Jo, H., Lee, G., Kim, J.: Analysis of control characteristics between dominant and non-dominant hands by transient responses of circular tracking movements in 3D virtual reality space. *Sensors* **20**(12), 3477 (2020)
50. Chita-Tegmark, M., Lohani, M., Scheutz, M.: Gender effects in perceptions of robots and humans with varying emotional intelligence. In: 14th ACM/IEEE International Conference on Human Robot Interaction, pp. 230–238. IEEE, Piscataway (2019)
51. Salovey, P., Mayer, J.D.: Emotional intelligence. *Imagin. Cogn. Pers.* **9**(3), 185–211 (1990)
52. Bar-On, R.: *The emotional Intelligence Inventory (EQ-i): Technical Manual*. Multi-Health Systems, Toronto (1997)
53. Bar-On, R.: The Bar-On Model of Emotional-Social Intelligence. *Psicothema* **18**(Suppl.), 13–25 (2006)
54. Ciarrochi, J., Forgas, J.P., Mayer, J.D.: *Emotional Intelligence in Everyday Life: A Scientific Inquiry*. Psychology Press, Philadelphia (2001)
55. Iwasaki, M., et al.: That robot stared back at me! Demonstrating perceptual ability is key to successful human-robot interactions. *Front. Robot. AI* **6**, 85 (2019)
56. Rudakov, I.V., Paschenkova, A.V.: A hierarchical method for verification of software algorithms via hierarchical Petri nets. *Eng. J. Sci. Innov.* **2**(14) (in Russian). BMSTU Press, Moscow (2013)
57. Karpov, V.E., Karpova, I.P., Kulinich, A.A.: *Social Communities of Robots*. URSS, Moscow (2019). (in Russian)



A Wavelet-Based Method for Morphing Audio Recordings of Interjections from One Voice to Another

Liliya Volkova^{1,2}(✉) , Arina Untilova¹ , and Maksim Kozlov¹ 

¹ Bauman Moscow State Technical University, 2-ya Baumanskaya ul. 5-1, 105005 Moscow, Russia

liliya@bmstu.ru

² Russian State University for the Humanities, Miususkaya pl., 6, 125993 Moscow, Russia

Abstract. This article is dedicated to a new method for morphing audio recordings of interjections: the original speaker's voice is transformed to the target one within a given voice pair, with preserving the original intonation. In speech synthesis solutions for Russian, interjections are poorly covered: the intonation set is limited, the contrary to human voice richness. Introducing a variety of interjections for communicative robots and voice assistants is of interest, as this would enable them to appear more natural in human-machine interaction, as well as to overcome existing constraints of dialogue systems and voice interfaces which lack affection. We designed an interjections synthesizing method aimed at fulfilling the F-2 communicative robot's reactions database. For verifying the applicability of synthesized recordings, we matched interjections-based reactions to input interjections, in the manner of Campbell's classic examples of dialogues limited to interjections, typical in colloquial human speech. We invited respondents to interact with F-2, to give cues and subsequently estimate answers. Collected evaluation data are discussed; most synthesized interjections are perceived as natural. Thus our research procedure enabled us to enrich the variety of our robot's communication means, and to check the appropriateness of a number of reacting scenarios: several reactions are dropped out as inappropriate, according to our respondents' opinions. The introduced wavelet-based method is applicable for morphing interjections from various speakers to the target voice, including the cases of multimodal corpora utilized for developing human-like reactions for affective communication solutions.

Keywords: Speech synthesis · Interjections · Human-machine interaction · Affective robots · Dialogue systems · Wavelets · Digital signal processing

1 Introduction

This research paper lies in the field of human-machine interaction (HMI). Interfaces for HMI have evolved from textual to multimodal in intention to bring in more natural input means [1]. Still, computer capacities are not used in full, and much remains to be done to simplify interfaces for their use by untrained common people [2]. Natural language

processing (NLP) along with speech processing methods is the answer, enabling software developers to create interfaces for fluent HMI. Methods for computer vision and tactile input detection followed, and multimodal communication became possible in a rather human way: its facilitation effect may contribute to the ease of conversational interaction [1] in order to achieve successful message comprehension and response preparation under the tight temporal constraints of conversation [3, 4].

Several research projects conjugate multimodal interfaces with text semantics processing; they simulate surface emotional phenomena, basing on studies in psychology and linguistics: *i bug* [5], *SEMAINE* [6], *Greta* [7], *Kismet* [8], *Herme* (based on the *Nao* robot) [9], *Max* [10, 11], *F-2* [12, 13]. These projects form one important applied field of HMI, the one of sociable and communicative agents and robots; they succeed at contribution to trustability of HMIs and robots, to their appearance as likeable and pleasant to respondents, even to attribution of personality traits to them [14, 15].

Bringing in the naturalness of interaction keeps a human interlocutor involved in communication with a robot, beyond the straightforward task of commands completing and efficient processing of factual information [9, 16]. Communicative robots and agents can also perform an important social function in health care applications [17–19]. Beyond listening informationally and critically, the goal of empathic listening is to build a relationship or help the speaker solve a problem [20–22]. Particular experiments in the setting of storytelling with a listening robot [14] show the effect of enlightenment of mutual understanding: while *F-2* the robot commented a story with several uses of “*Hmm*” and “*I understand*”, one respondent gave a feedback of awe, pointing at being understood. Hedges are another means of establishing personal contact and making communication more comfortable, namely by defusing the situation of social face threatening act [23]. Probes in conversation (e.g. interjections, brief remarks) can open up or direct the interlocutor and uncover useful information [22].

Campbell states that in NLP there has long been an implicit assumption that a text transcription adequately embodies all the relevant parts of a spoken message. Now this approach is challenged, and the affective colouration of a message is now being considered as an essential component for a successful interpretation of the speaker’s intended message [24]. One particular speech mode is interjections. They are most often intonationally rich and quite poorly transcribed in written text. As to speech synthesis, existing solutions are often not competent at interjections, however good they are at voicing simple texts with a regular intonation. For dialogue systems and communicative robots which utilize speech synthesis methods, the task of acquiring a database of affective reactions filled with voiced interjections is of interest, as it is aimed at enriching the spectrum of available modes of communicative reactions.

2 Interjections in Communication

No class of words has better claims to universality than interjections, and no category has more variable content than this one [25]. Campbell [24] shows that in natural communication 49% of utterances can be non-lexemic, namely interjections. He gives a classic example of the brevity of natural interjection-based human-human interaction as

compared to lexically loaded dialogue that predominates in systems currently considered by linguists and engineers; he places more emphasis on tone-of-voice and prosodic inflections [24], in contrast to the solely text processing.

This article focuses on interjections—monolexicemic linguistic items that typically function as standalone utterances, namely those formed of vowels. Some scientists point out semantic criteria linking interjections to the expression of feelings or mental states [26–28], but this approach excludes a class of one-word utterances less clearly aimed at expressing affect [25]. Nevertheless, interjections are commonly understood as public emissions of private emotions, not necessarily intentional, but often including a specific appeal-to-the-listener-function [29] from the communicative point of view at the dialogue conducted. In our research, we imposed the interjections class limitation: we selected only one interjection of the class of secondary interjections that originate from words (“da”, i.e. “yes” in Russian) and are used as interjections, and concentrated on the opposite class of primary interjections which do not [30].

Conversations are not only scenario-driven, and the common tendency for dialogue systems is following the Minsky’s approach: cognitive units take control over the body of an artificial robot or agent [31]. The underlying processor analyzes multimodal input data, extracts the meaning, compares it to some model state or behavior scenario, and then produces responses to input stimuli. The affective component of behavior is modelled dynamically, via reactions to the input. These reactions should vary for each one communicative goal (what an agent or a robot ought to express, e.g. happiness, sadness, compensation) so that the robot’s processor could choose one reaction out of a pool of synonyms for a given goal. Interjections are a stand-alone segment of voiced responses. Existing software for Russian often shows unsatisfactory results for interjections, see Table 1 (an asterisk marks a limit on the available set).

Table 1. Comparison of existing speech synthesis solutions for Russian.

Criteria	Yandex SpeechKit	Acapela	Robot Talk	L&H
Speech intelligibility	Good	Good	Good	Satisfactory
Results sound similar to natural speech	Good, sometimes unsatisfactory	Good, sometimes unsatisfactory	Satisfactory	Satisfactory
Option of changing voice characteristics	A number of predefined options	No	No	Yes
Option of intonations	Yes *	No	No	Yes *
Synthesizing interjections	Yes *	No	No	Yes *, with a foreign accent

Hence, in Sect. 4 we introduce a method for synthesizing interjections which would fill affective robots’ and agents’ reactions databases; in Sect. 5 we verify the synthesized

recordings applicability via matching interjections-based reactions to the robot's current communicative intention and via crowdsourcing-based estimation of implemented scenarios adequacy and naturalness, basing on respondents' opinions. Our conversational robot analyzes interlocutor's speech and forms reactions on the basis of triggered communicative goals, which stand for script-based expressing affective and/or rational attitude to what's been said. F-2 listens to the input interjection, selects the reaction scenario automatically and sounds the reaction. Only the interjection reaction channel is used in this setting, as in the Campbell's example [24].

3 Related Work: The Phoneme-Based Approach

Intonation refers to the way in which a voice changes in pitch to convey meaning [32] and to the structured variation in pitch which is not determined by lexical distinctions as in tone languages [33]. The nature of phonetic implementation is context-sensitive and language-specific [34]. Existing software succeed in synthesizing utterances with either monotonic or monotonically rising or falling intonation. Markov chains and Petri nets account for probabilistic rules of changing intonation in phrases, while particular research works describe intonational grammars for particular languages [35].

Yandex SpeechKit, the best speech synthesizing solution for Russian, performs following steps: (1) preprocessing text via decoding abbreviations and numerals, (2) text subdivision into phrases with the same intonation, with punctuation marks and stereotypic linguistic constructions taken into account, (3) determining words phonetic transcription with appropriate stresses via (a) manually composed dictionaries, (b) rule-based transcription for words absent in dictionaries, (c) corpora-based statistical rules utilized in case when usual rules are insufficient [36]. Such phoneme-wise speech synthesis approach is applicable to texts. But when it comes to interjections, most often there is one phoneme of variable length. Yandex SpeechKit reduces input interjection transcriptions to short ones ('a-a-ah' → 'ah'), which is its current limitation.

The limitation of pitch transforming methods of voice morphing [37–39] is that compared to segmental mapping for phrases, for intonationally rich monophonemic utterances with complex pitch patterns it is a non-trivial task to join separate segments with raising and falling pitch; special techniques are required for smoothing artifacts in the resulting signal [39]. This implies the need for a method which would morph such short utterance without segmentation. It should preserve the intonation that appears to be unique, as a speaker never utters the same sound pattern twice [39]: pitch tracks differ even for one interjection [32]. Evaluating mean pitch values is hardly applicable to complex intonations. Thus it is of interest to utilize some sort of signal derivatives. For this purpose we selected the discrete wavelet transform (DWT). Not only does it store data on the contribution of frequencies in the signal, but it also localizes this contribution at each level of decomposition [40]. One stage of linear signal decomposition produces 2 vectors: approximating and detailing coefficients [41].

4 The Interjection Morphing Method

Given a source signal (SS) and the target interjection voice sample (TS), we preprocess both wav-files with interjections: (a) we normalize signals by amplitude, so that absolute maximum value is 1; and (b) we use linear interpolation to elongate the shorter signal; we recommend selecting target signal of pretty much the same length. Then for both signals we perform DWT until the 3rd level, we put aside 3 vectors of detailing coefficients and we calculate envelopes separately for positive and for negative approximating DWT coefficients of the 3rd level (AC). The following step is morphing the shape of AC envelopes for TS (*Source*) so that each of the two envelopes resembles the corresponding AC envelope for SS (*Target*) with preserving the sign. For positive values the resulting $Value_i$ is equal to the small value ε if $Source_i \leq \varepsilon$, otherwise $Value_i = Source_i \cdot Target_i / Max_i (Source)$; an analogous formula is used for negative values. Two *Value* vectors serve source for reconstructing the signal along with detailing coefficients which were previously put aside, these detailing coefficients preserving what's personal in the voice. The last step is correcting the length of the synthesized signal to that of TS, and consequent multiplying the result by maximum amplitude of TS. Figure 1 shows an example: 2 envelopes for positive SS AC and TS AC values result in the morphed of positive AC values, and 2 envelopes for negative AC values result in the morphed negative AC values envelope.

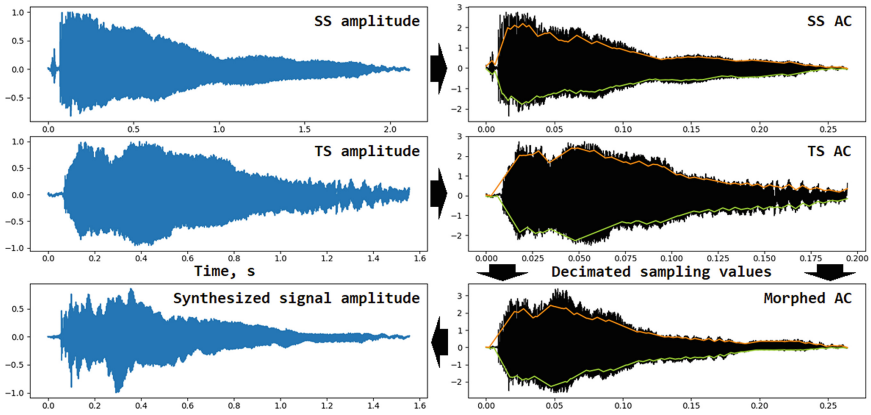


Fig. 1. An example: SS converted to SS AC, TS converted to TS AC, both ACs resulting into the morphed AC, and the reconstructed signal. Two envelopes are depicted for each AC.

5 Results

We recorded 700 interjection audios for 5 communicative goals (which are to demonstrate a certain emotion). Table 2 shows the implementation of emotions in our dataset with different lexemes in Russian, given as interjections transcription.

Studio recordings contain no sighs and no extraneous noise. There are 63 recordings of 9 male voices and 77 audios of 11 female voices. As interjections are characterized

Table 2. Implementation of emotions with different lexemes in the source dataset.

Emotion	‘ah’	‘ekh’	‘oho’	‘okh’	‘oh’	‘oo’	‘wow’	‘da’ (‘yes’)
E—enlightenment	+		+		+			+
S—sadness	+	+	+	+	+	+		+
J—joy	+	+	+	+	+	+	+	+
I—indignation	+	+	+	+	+	+		+
A—astonishment	+		+		+	+		+

with monolexemicity and conventional utterancehood [25], they are short (most often up to 5 s), this limitation being associated with the outward breath duration.

In tasks dedicated to speech synthesis human ear and brain are supreme judges: this is a test without any standard answer, and generated signals are intended to sound naturally to assessors. In order to evaluate the quality of synthesized signal, we asked 20 respondents to assign one of 3 emotion labels to interjections recordings, as perceived. Three classes of recordings were used: human-recorded, synthesized with an intended label (basing on ‘ah’, ‘oh’, ‘oo’ sounds), and extended recordings with an extra word accompanying each interjection (“oh, great!”, “ah, pity”). Percentage of given estimations are stated in Table 3, with values for matching labels marked in bold. The results vary slightly for the first two classes: precision of human recognition of intended emotions is 73%, of synthesized interjections—66%. Respondents tend to confuse joy and astonishment in all classes of recordings. Restrained sadness was mistakenly identified as astonishment as well. Hence, we recommend that joyful surprise and sad surprise should be clearly separated. Last three columns show that interjections are better recognized along with additional words, with precision over 80%.

Table 3. Labels that respondents assigned to interjections, grouped by recordings class.

Intended	Assigned								
	Human-recorded			Synthesized			Synthesized + words		
	J	S	A	J	S	A	J	S	A
J—joy	60%	20%	20%	60%	10%	30%	80%	5%	15%
S—sadness	10%	85%	5%	15%	80%	5%	5%	85%	10%
A—astonishment	10%	15%	75%	30%	15%	55%	15%	0%	85%

We also asked our respondents to evaluate, how natural did the synthesized interjections sound and how appropriate they sounded within 11 reactions scenarios comprising an input emotion (IE) paired with a communicative goal, or the target emotion (TE) to express in the reply. We recognized the input label via a random forest classifier with 31 trees of height 6, processing three feature vectors which represent the sound tone as adapted to human ear perception of sound—a chromagram [42–45], Mel-frequency

cepstral coefficients [45, 46] and a Mel-spectrogram [47, 48]. The classifier results in accuracy values 0.85 for female voices and 0.7 for male voices. Respondents assessed the synthesized reactions generally positively. According to results presented in Table 4 for 11 proposed scenarios $IE \rightarrow TE$, most of the generated interjections are evaluated positively ('+' : 51%), while there are 40% neutral reactions ('~', acceptable), and 9% negative ('-'). These data enabled us to identify 3 scenarios that were not assessed as appropriate (marked with bold on the last line). The most successful scenarios are highlighted in the 2nd line. These results are to be taken into account during final selection of scenarios for the F-2 robot's behaviour.

Table 4. Fractions of assessment values for scenarios $IE \rightarrow TE$, according to respondents

Grade	I → S	I → A	E → J	E → A	S → S	S → A	A → A	A → J	A → S	J → A	J → J
+	0.40	1.00	0.50	0.45	0.75	0.35	0.85	0.40	0.15	0.80	0.45
~	0.60	0.00	0.40	0.45	0.25	0.50	0.15	0.30	0.60	0.15	0.55
-	0.00	0.00	0.10	0.20	0.00	0.15	0.00	0.30	0.25	0.05	0.00

We also compared our results to assessments of the same scenarios implemented with basic interjections generated with Yandex SpeechKit [49]. Joy was selected as TE for the reason of lack of other 4 TE in SpeechKit synthesis settings. Results are pretty much the same, slightly worse in 2 cases out of the following three: $E \rightarrow J$ 0.4 '+' / 0.3 '~' / 0.3 '-', $A \rightarrow J$ 0.4 '+' / 0.4 '~' / 0.2 '-', $J \rightarrow J$ 0.3 '+' / 0.5 '~' / 0.2 '-'. The drawback is the limitedness of intonations and lexemes: all interjections are reduced to short forms and sounded with one of 3 basic intonations (regular, question, exclamation, the latter used for joy TE). The lack of complex intonation patterns, which are typical for people, makes it pretty impossible to solve the given task of introducing intonationally variative interjections for dialogue systems and communicative robots. Moreover, our team assessed SpeechKit-generated intonations as 'weird' multiple times, regarding that for synthesizing basic, non-affective speech it's the best solution for Russian.

6 Conclusion

As the conducted survey shows that no solution for Russian is able to synthesize trustable interjections, we designed a method for morphing interjections with changing individual speaker's voice characteristics while retaining the intonation pattern. Evaluation of the naturalness and appropriateness of synthesized interjections used as reactions in robot-human dialogue proves that we overcame constraints of the phonemes-based approach typical for text-to-speech solutions: we developed an approach adapted to monophonemic cues, as our task is concerned with intonations rather than phonemes. The synthesized data are ready to use in dialogue systems and companion robots which implement affective reactions.

Affective robots are prospective for production due to credible long-term human interest towards it, ascertained by a richer reactions spectrum [50]. Provided with highly variable reactions, they don't bore people as opposed to first robotic toys with a limited

set of reactions. In order to synthesize various reactions, our approach is applicable, for it allows recording a bulk of voiced interjections and to consequently transfer them onto the selected speaker voice, preserving the intonation.

The described procedure involving respondents to assess the relevance of the proposed reaction scenarios is the necessary means of verifying particular reactions applicability; it allows accepting positively rated behaviour elements as well as rejecting tested scenarios which were rated low and marked as inappropriate. The described means of checking the reactions adequacy enables researchers to make a decision on admitting particular interjection-based reaction scenarios into the reactions database.

The abovementioned allows us to conclude the applicability of the proposed method of interjection synthesis and of a subset of synthesized interjections which sound natural and appropriate to respondent ears. With adding this sort of human-like behavioral patterns of spoken reactions in dialogue, researchers in HMI are enabled to take one more step to more natural interfaces, which show unimodal and multimodal reactions and are estimated as friendly, comfortable, likeable [12, 14, 51–54].

Acknowledgements. This research is supported by the grant of the Russian Science Foundation № 19–18–00547, <https://rscf.ru/project/19-18-00547/>.

We would like to express gratitude to our volunteer speakers who helped us recording a dataset, and to respondents whose opinions are priceless to our sort of projects, when it comes to assessment of behavioural patterns developed for use in HMI.

References

1. Drijvers, L., Holler, J.: The multimodal facilitation effect in human communication. *Psychon. Bull. Rev.* **30**(2), 792–801 (2023)
2. Ronzhin, A.L., Karpov, A.A., Lee, I.V.: *Speech and Multimodal Interfaces*. Nauka, Moscow (2006). (in Russian)
3. Holler, J., Kendrick, K.H., Levinson, S.C.: Processing language in face-to-face conversation: questions with gestures get faster responses. *Psychon. Bull. Rev.* **25**(5), 1900–1908 (2018)
4. Levinson, S.C.: Turn-taking in human communication—origins and implications for language processing. *Trends Cogn. Sci.* **20**(1), 6–14 (2016)
5. i bug. <http://ibug.doc.ic.ac.uk/>. Last Accessed 18 May 2023
6. Shróder, M.: The SEMAINE API: towards a standards-based framework for building emotion-oriented systems. *Adv. Hum. Comput. Interact.* **2010**, 319406 (2010)
7. Pelachaud, C.: Greta: a conversing socio-emotional agent. In: *Proceedings of the 1st ACM Sigchi International Workshop on Investigating Social Interactions with Artificial Agents*, pp. 9–10. ACM, New York (2017)
8. Breazeal, C.: *Designing Sociable Robots*. MIT Press, Cambridge (2002)
9. Han, J.G., Campbell, N., Jokinen, K., Wilcock, G.: Investigating the use of non-verbal cues in human-robot interaction with a Nao robot. In: *Proc. 3rd IEEE Int. Conf. on Cognitive Infocommunications (CogInfoCom 2012)*, Kosice, Slovakia, pp. 679–683. IEEE, Piscataway (2012)
10. Becker, C., Kopp, S., Wachsmuth, I.: Simulating the emotion dynamics of a multimodal conversational agent. In: *ADS 2004, LNAI*, vol. 3068, pp. 154–165. Springer, Heidelberg (2004)
11. Max. <http://cycling74.com/products/max/>. Last Accessed 18 May 2023

12. Zinina, A., Zaidelman, L., Arinkin, N., Kotov, A.: Non-verbal behavior of the robot companion: a contribution to the likeability. *Procedia Comput. Sci.* **169**, 800–806 (2020)
13. Kotov, A.A., Zinina, A.A.: Functional analysis of non-verbal communicative behavior (in Russian). In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue”*, **14**(21) vol. 1, pp. 308–320. RSUH, Moscow (2015)
14. Velichkovsky, B.M., Kotov, A., Arinkin, N., Zaidelman, L., Zinina, A., Kivva, K.: From social gaze to indirect speech constructions: how to induce the impression that your companion robot is a conscious creature. *Appl. Sci.* **11**(21), 10255 (2021)
15. De Graaf, M. M. A., Ben Allouch, S.: Expectation setting and personality attribution in HRI. In: *HRI, ACM/IEEE International Conference on Human-Robot Interaction*, pp. 144–145. IEEE, Piscataway (2014)
16. Jokinen, K., Wilcock, G.: Modelling user experience in human-robot interactions. In: *MA3HMI 2014 Workshop, LNAI*, vol. 8757, pp. 45–56. Springer, Heidelberg (2014)
17. Shibata, T.: Therapeutic seal robot as biofeedback medical device: qualitative and quantitative evaluations of robot therapy in dementia care. *Proc. IEEE* **100**, 2527–2538 (2012)
18. Takayanagi, K., Kirita, T., Shibata, T.: Comparison of verbal and emotional responses of elderly people with mild/moderate dementia and those with severe dementia in responses to seal robot, PARO. *Front. Aging Neurosci.* **6**(SEP), 257 (2014)
19. Inoue, K., Wada, K., Shibata, T.: Exploring the applicability of the robotic seal PARO to support caring for older persons with dementia within the home context. *Palliat. Care Soc. Pract.* **15** (2021). <https://doi.org/10.1177/26323524211030285>
20. Spacapan, S., Oskamp, S.: *Helping and Being Helped: Naturalistic Studies*. Sage, Newbury Park (1992)
21. Weaver, J.B., Kirtley, M.D.: Listening styles and empathy. *South Commun. J.* **60**, 131–140 (1995)
22. Adler, R.B., Rodman, G.: *Understanding Human Communication*, 9th edn. Oxford University Press, New York (2006)
23. Malkina, M., Zinina, A., Arinkin, N., Kotov, A.: Multimodal hedges for companion robots: a politeness strategy or an emotional expression? In: Selegey, V.P., et al. (eds.) *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”*, Issue 22, pp. 319–326. RSUH, Moscow (2023)
24. Campbell, N.: Extra-semantic protocols; input requirements for the synthesis of dialogue speech. In: *Proceedings of Affective Dialogue Systems, Tutorial and Research Workshop, ADS 2004, LNCS 3068*, pp. 221–228. Springer, Heidelberg (2004)
25. Dingemans, M.: Interjections (preprint). In: Eva van Lier (ed.) *The Oxford Handbook of Word Classes*. Oxford University Press, Oxford (2021)
26. Goffman, E.: Response cries. *Language* **54**(4), 787–815 (1978)
27. Wierzbicka, A.: The semantics of interjection. *J. Pragmatics* **18**(2–3), 159–192 (1992)
28. Wharton, T.: *Pragmatics and Non-verbal Communication*. Cambridge University Press, Oxford (2009)
29. Elffers, E.: Interjections and the language functions debate. *Asia Pac. J. Hum. Resour.* **50**(1), 17–29 (2008)
30. Bloomfield, L.: *An Introduction to The Study of Language*. Holt, New York (1914)
31. Minsky, M.L.: *The Society of Mind*. Touchstone Book, New York (1988)
32. Dingemans, M., Torreira, F., Enfield, N.J.: Is “Huh?” a universal word? Conversational infrastructure and the convergent evolution of linguistic items. *PLoS ONE* **8**(11), e78273 (2013). <https://doi.org/10.1371/journal.pone.0078273>
33. Gussenhoven, C.: *The Phonology of Tone and Intonation*. Cambridge University Press, Cambridge (2004)
34. Pierrehumbert, J.B.: Phonological and phonetic representations. *J. Phon.* **18**, 375–394 (1990)

35. Dehé, N.: An intonational grammar for Icelandic. *Nordic J. Linguist.* **32**, 5–34 (2009)
36. How does it work? The speech synthesis / Yandex blog (in Russian), <https://yandex.ru/blog/company/kak-eto-rabotaet-sintez-rechi>. Last Accessed 16 Mar 2023
37. Arslan, L.M., Talkin, D.: Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum. In: Proceedings of 5th European Conference on Speech Communication and Technology (Eurospeech 1997), pp. 1347–1350. ISCA (1997)
38. Gillett, B., King, S.: Transforming F0 contours. In: Proceedings of 8th European Conference on Speech Communication and Technology (Eurospeech 2003), pp. 101–104. ISCA (2003)
39. Banerjee, A., Pandey, S., Khushboo, K.M.: Voice intonation transformation using segmental linear mapping of pitch contours. In: 2018 IEEE 4th International Conference on Computer and Communications (ICCC), pp. 1278–1282. IEEE, Piscataway (2018)
40. Stark, H.G.: *Wavelets for Signal Processing: An Application-Based Introduction*. Springer, Berlin (2005)
41. Mallat, S.G.: A theory of multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Machine Intell.* **11**, 674–693 (1989)
42. Shepard, R.N.: Circularity in judgements of relative pitch. *J. Acoust. Soc. Am.* **36**(12), 2346–2353 (1964)
43. Ruckmick, C.C.: A new classification of tonal qualities. *Psychol. Rev.* **36**(2), 172–180 (1929)
44. Bartsch, M.A., Wakefield, G.A.: Audio thumbnailing of popular music using chroma-based representations. *IEEE Trans. Multimedia* **7**(1), 96–104 (2005)
45. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* **28**(4), 357–366 (1980)
46. Ustubioglu, A., Ustubioglu, B., Ulutas, G.: Mel spectrogram-based audio forgery detection using CNN. *Sig. Image Video Process.*, 1–9 (2022)
47. König, W.: A new frequency scale for acoustic measurements. *Bell Telephone Lab. Rec.* **27**, 299–301 (1949)
48. Devi, J.S., Srinivas, Y., Nandyala, S.: Speaker emotion recognition based on speech features and classification techniques. *Int. J. Comput. Netw. Inf. Secur.* **7**, 61–77 (2014)
49. Yandex SpeechKit. <https://cloud.yandex.ru/services/speechkit>. Last Accessed 16 Mar 2023
50. Volkova, L., Kotov, A., Klyshinsky, E., Arinkin, N.: A Robot Commenting Texts in an Emotional Way. In: CCIS, vol. 754, pp. 256–266. Springer, Heidelberg (2017)
51. Kamide, H., Mae, Y., Takubo, T., Ohara, K., Arai, T.: Development of a scale of perception to humanoid robots: PERNOD. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5830–5835 (2010)
52. Rosenthal-von der Pütten, A.M., Krämer, N.C., Hoffmann, L., Sobieraj, S., Eimler, S.C. An experimental study on emotional reactions towards a robot. *Int. J. Soc. Robot.* **5**(1), 17–34 (2013)
53. Waytz, A., Heafner, J., Epley, N.: The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. *J. Exp. Soc. Psychol.* **52**(1), 113–117 (2014)
54. Aremyr, E., Jönsson, M., Strömberg, H.: Anthropomorphism: an investigation of its effect on trust in human-machine interfaces for highly automated vehicles. In: Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018). IEA 2018. *Advances in Intelligent Systems dsx vand Computing*, p. 823 (2019)



Extended and Distant Cortical Areas Coordinate Their Oscillations Approaching the Instant of Decision Making During Recognition of Words

Victor Vvedensky¹(✉), Vitaly Verkhlyutov², and Konstantin Gurtovoy¹

¹ RNC Kurchatov Institute, Moscow, Russia
VictorLvo@yandex.ru

² Institute of Higher Nervous Activity and Neurophysiology of RAS, Moscow, Russia

Abstract. By measuring magnetic signals of the brain during recognition of spoken words, we have found that several large neural populations in the cortex start to oscillate in concert before the subject makes a decision about which word was heard. This coordination takes place for several cycles of basic oscillatory activity and the events differ in the number, relative amplitude and shape of the cycles and their duration. We observe nine functional groups which are active during 40 words long recognition task. However, only some of them produce high amplitude signals when a single word is recognized. For every new word, a new combination of neural clusters is formed. Such measurements imply that nearly the whole rear part of the brain cortex is involved in the word recognition process.

Keywords: Word recognition · Decision making · Cortical oscillations · Magnetic encephalography

1 Introduction

The cognitive systems emerging with artificial intelligence are starting to control different processes and have to make decisions in different circumstances. The operation of these biologically inspired intelligent systems should be perpetually compared to how people solve the same problem. This can probably add a new dimension to the development of artificial control devices. Though, we know little about what happens in the brain before decision making, studies are under way [1, 2]. We measured magnetic signals of the brain in the period preceding the decision regarding which word the person heard. The EEG version of these experiments was reported earlier [3].

2 Measurements

Seven volunteers took part in the study. All subjects signed an informed consent. The study was approved by the local ethics committee of the Institute of Higher Nervous Activity and Neurophysiology of Russian Academy of Science and was conducted following the ethical principles regarding human experimentation (Helsinki Declaration).

The measurements were made in the Center for Neurocognitive Research (MEG Center) at the Moscow City University of Psychology and Education.

Magnetic brain responses were recorded with the sampling rate of 1000 Hz using a helmet-shaped whole head magnetometer (ElektaNeuromag MEG system, 306 channels). Native hardware filters with band pass 0.1–330 Hz were used and no additional filtering was applied to avoid distortion of the signal shape. General data processing was done using the Brainstorm software [4].

During the measurement the participant was sitting in a magnetically shielded room with their eyes closed to avoid suppression of the alpha rhythm. The stimuli were words spoken by a human and reproduced by earphones. The subjects were instructed to recognize the word they heard and to confirm recognition by a keystroke. Playback of the next word started 1 s after the button was pressed. In a single experiment every subject heard 40 words of the same duration. This measurement lasted for about 80 s. 8 different Russian adjectives with close meaning were presented in this sequence, 5 times each in random order. Three different sets of 8 adjectives with different word duration (680, 830, 910 ms) were presented to each subject. We analyzed offline time segments of the recorded signals corresponding to the word recognition process.

3 Results

In this paper we focus attention on the brain signals recorded just before the subject makes a decision on the meaning of the word and presses the button, confirming the choice. Figure 1 shows the magnetic signals from several sensors within approximately a second before the keystroke. The behavior of magnetic signals is similar for all subjects in all recognition events. In each event one can always find several groups of sensors (most often 2 or 3) where the pattern of the signals becomes similar approaching the instant of the keystroke. Usually this is a couple of oscillation cycles with a characteristic shape and duration, with amplitude differing sometimes. Before this time interval, the signals in different sensors were uncorrelated. The sensors measure a gradient of the magnetic field and hence are mostly sensitive to the current sources in the neural tissue located just under the sensor. The colors of sensors in Fig. 1 indicate three territories in the brain under the helmet where neural populations start to oscillate in concert before the word is recognized. The rhomboid arrow shows the sensitivity axis of the sensor picking up the signal from active neurons – it corresponds to direction of current flowing in the brain. There are two sensors on each square chip in the helmet surrounding the head, which measure two orthogonal gradients.

The measurements reveal formations of several large, nearly independent, neural populations during decision making in word recognition. They occupy considerable portions of cortical surface in the rear part of the brain. For example, the “blue” group in Fig. 1 extends in a linear fashion from the head vertex down to theinion.

Other groups also tend to form chains of sensors which pick up correlated signals from coordinated activity of distant cortical areas. A new set of groups of sensors is formed during every new process of recognition. Representative examples of these groups of sensors are given in Fig. 2. We combined several sensors into a common group if their signal course contained a congruent time segment before the end of the

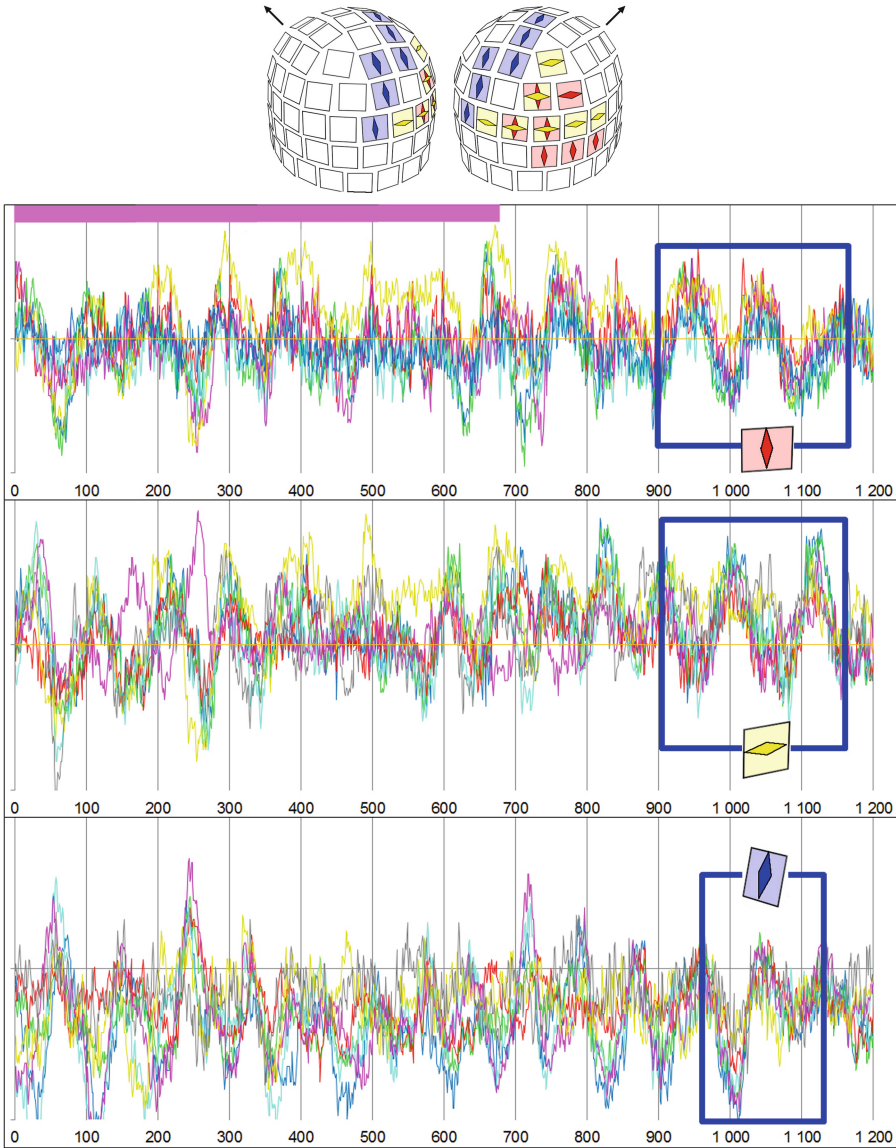


Fig. 1. Signals recorded by three groups of magnetic sensors marked by red, yellow and blue rhomboid arrows. Positions of the square sensors are shown on the helmet surrounding the subject's head. Arrows indicate direction of the subject's gaze. Signals in each sensor of each group are presented in different contrast colors and overlap. Magenta strip indicates the time when the word *kudryavy-curly* sounds. Time scale is in milliseconds. In the end of the recording, the subject recognized the word and pressed a button. Curves that have not matched for most of the recording time become almost identical as they approach the moment the key is pressed.

recognition process. We individually scrutinized 40 signal curves of each experimental session and identified 9 different clusters of sensors, which are shown in Fig. 2 in different colors. These are data for one subject, others display similar assortment of clusters though with individual variations.

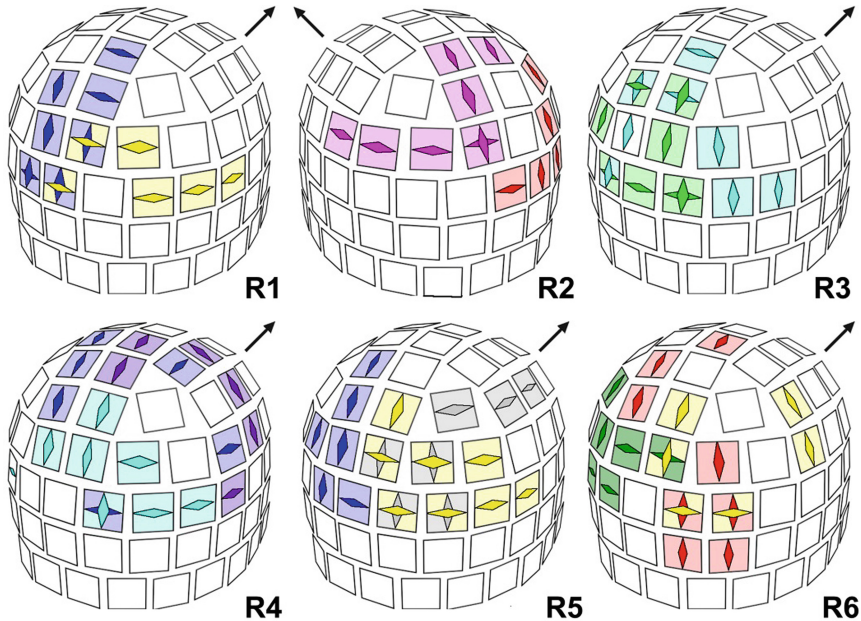


Fig. 2. Groups of sensors, where magnetic signals start to coordinate their time course just before the word recognition. These are distributions of coordinated activity during six representative events of one subject who have heard 40 words in a row. The images of the helmet are the same as in Fig. 1. Different colors correspond to nine groups of sensors. Each group picks up signals from a certain cortical area, where neural currents tend to be generated in one of two orthogonal directions. These groups nearly completely cover the rear part of the sensor array. The array monitors neural activity in the rear part of the brain, which seems to be totally involved in the word recognition process.

Our data provide information on the general behavior of cortical oscillatory activity during a series of word recognition events. Before terminating running recognition, several large territories of cortex start to oscillate in concert. Individual features of this coordination are depicted in Fig. 2. The helmet views correspond to recognition of words: *zavitoy-frizzy* R1, *kurchavy-curly* R2, *kurchavy-curly* R3, *volnisty-wavy* R4, *petlyayuschy-winding* R5, *petlyayuschy-winding* R6. During individual recognition of a word only some (1–5) of the clusters pick up high amplitude oscillations, while others are nearly silent. The next time another combination of clusters comes into play. Each cluster has its own territory and preferred orientation, though from time to time it “invades” neighbors and temporarily includes some “foreign” sensors. The example is shown in Fig. 2-R2, where the magenta cluster in this particular recognition run includes a vertical row of sensors (nearly orthogonal to the basic horizontal row), which usually belongs

to the blue cluster. This can be treated as the manifestation of interaction between large independent neural populations during the decision making process. The blue cluster is most frequently active during the whole 40 word recognition task, and is present three times in Fig. 2.

Most often, coherent signals are seen in clusters which occupy a compact area on the helmet, as do the yellow clusters in plots R1 and R5. However, quite coherent signals can be seen in distant and orthogonally oriented sensors shown in plot R6. For this particular event we attribute these scattered areas to a common yellow cluster since they oscillate coherently. The same holds true for the red cluster in plot R6 where upper sensors usually “belong” to the blue cluster. Similar split of cyan cluster can be seen in plot R4, where the neighboring parts of the cluster pick up signals which are generated by the current sources oriented orthogonally to each other.

We come to conclusion that we observe a smart system which presumably controls word recognition process. Virtually all cortical territories in the rear part of the brain are involved in the decision making during word recognition process.

4 Discussion

We summarize our results in the following way. Each of 204 sensors, measuring gradient of magnetic field, monitors mostly its own cortical patch. Often a pair of neighboring sensors records the signals with quite a different time course. This means that the areas under these sensors tend to oscillate independently. On the other hand, we frequently see that several sensors record signals of nearly congruent shape, which means that large cortical areas in this case oscillate coherently. When the recognition process approaches the instant of decision making on which word was heard, the cortical areas tend to coordinate their oscillations. Most often, we see a couple of congruent waves before the subject presses the button, though sometimes train of up to four oscillations appears, or just a single cycle. These cortical events can be treated as stereotyped episodes, which are permanently observed when encephalographic signals are recorded [5, 6]. During the whole task, 40 words long, we observe about a dozen of centers in the rear part of the brain, which agglomerate other cortical territories into a common pool in a course of each word recognition event. Each center is located under a certain cluster of magnetic sensors and has preferred orientation of electric current it generates in the neural tissue. Only two or three (sometimes 1–5) of these centers are active during a particular recognition event. Usually they “switch off” when the next word is presented. The cortical oscillations during perception of this word coordinate itself around a new set of activity centers.

The observed behavior of activity centers allows us to make assumptions about the mechanisms that support this coordination. In order to force other cortical territories (sometimes distant) to imitate the oscillations in the center, the latter has to broadcast a message on its own state, which presumably needs resources. The previous recognition depleted these resources, and a new set of centers, which were idle in a previous run, will take part in the processing of the next word. One should experimentally find what influences the selection of neural populations that will be involved in the next recognition. Most probably the accumulated amount of the resource, needed for recruiting cortical territories into coherent oscillation, is a decisive factor.

Specific features of the magnetic signals, generated in the brain, should be taken into account, when analyzing links between different centers in the brain. One of the important parameters, which should be transmitted into an area, which will follow the oscillation in the center, is the instant when the slope of the signal curve changes considerably, most often on the tip of usually sharp wave. These sharp tips one can easily see in Fig. 1 despite of the present noise. Monitoring these sequences of sharp events in the cortex, one can clear up the interrelations of the centers, which perform an effective teamwork when the brain recognizes words.

This study was supported by the Russian Science Foundation - Grant no. 23-78-00011. The authors are grateful to Chernyshev B.V. and Prokofyev A.O. from the Center for Neurocognitive Research (MEG Center) at the Moscow City University of Psychology and Education and Martynova O.V. from the Institute of Higher Nervous Activity and Neurophysiology for their help in the experiment.

References

1. Peixoto, D., et al.: Decoding and perturbing decision states in real time. *Nature* **591**, 604–609 (2021). <https://doi.org/10.1038/s41586-020-03181-9>
2. Keuken, M.C., et al.: Brain networks of perceptual decision-making: an fMRI ALE meta-analysis. *Front. Hum. Neurosci.* **19**(8), 445 (2014). <https://doi.org/10.3389/fnhum.2014.00445>
3. Vvedensky, V., Filatov, I., Gurtovoy, K., Sokolov, M.: Alpha rhythm dynamics during spoken word recognition. In: Kryzhanovsky, B., Dunin-Barkowski, W., Redko, V., Tiumentsev, Y. (eds.) *Advances in Neural Computation, Machine Learning, and Cognitive Research VI. Neuroinformatics 2022. Studies in Computational Intelligence*, vol. 1064, pp. 65–70. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-19032-2_7
4. Tadel, F., Baillet, S., Mosher, J.C., Pantazis, D., Leahy, R.M.: Brainstorm: a user-friendly application for MEG/EEG analysis. *Comput. Intell. Neurosci.* **2011**, ID 879716 (2011). <https://doi.org/10.1155/2011/879716>
5. Neymotin, S.A., et al.: Detecting spontaneous neural oscillation events in primate auditory cortex. *eNeuro* **9**(4) (2022). <https://doi.org/10.1523/ENEURO.0281-21.2022>
6. Kaplan, A.Y., et al.: Nonstationary nature of the brain activity as revealed by EEG/MEG: methodological, practical and conceptual challenges. *Signal Process.* **85**, 2190–2212 (2005). <https://doi.org/10.1016/j.sigpro.2005.07.010>



Laser Detection of Surface Quality of Electrical Contacts Based on Ensemble Learning

Chao Wang^(✉) and Cheng Jun Guo

College of Science and Technology, Ningbo 315000, ZJ, China
490895161@qq.com

Abstract. The quality of silver-based contacts directly affects the overall performance and service life of electrical connectors. There are many factors affecting the surface quality: production process, equipment accuracy, transportation, vibration. In the assembly process of electrical connection equipment, it is necessary to first screen out contacts with unqualified surface quality. But the traditional testing process requires the participation of professional workers, which is difficult to meet the demand for high-speed and efficient testing under the trend of output expansion. Fortunately, in recent years, the research of machine learning has created the possibility of online identification and classification of fine surface defects of silver-based contacts. In this paper, based on machine learning, laser ranging, automatic classification and other methods for surface quality detection, contact automatic detection and sorting. The results show that the detection time can be greatly reduced with the help of computer, and the detection data can be integrated to provide data support for the subsequent life estimation of electrical connectors.

Keywords: Ensemble model · Silver base contact · Surface quality · Machine learning · Automatic sorting

1 Introduction

Electrical equipment, power systems are widely used switching appliances [1], switch through the contact to control the working state of the electrical connector. The contact of the electrical connector should not only resist the current surge impact caused by microdynamic jump during repeated breaking. At the same time, it is necessary to suppress the strong temperature rise caused by the arc action. And only under the working conditions to ensure the correct parting and closing of the contact can the electrical connector play its due performance, so the core component of the switching appliance is the contact material [2].

In the recent years, the research of silver-based environmental materials replacing traditional cadmium materials is becoming more and more perfect, but there is still a lot of room for improvement in production process optimization. Due to the limitations of the existing process and the contact material is too small, the finished contact is always doped with a large number of defective parts. And common quality problems

such as uneven coating, peeling surface, scratches, bumps and dents [3–6]. As shown in Fig. 1, these surface defects directly or indirectly lead to increased contact resistance, abnormal temperature rise, contact viscosity and other potential factors affecting the life of electrical connectors.

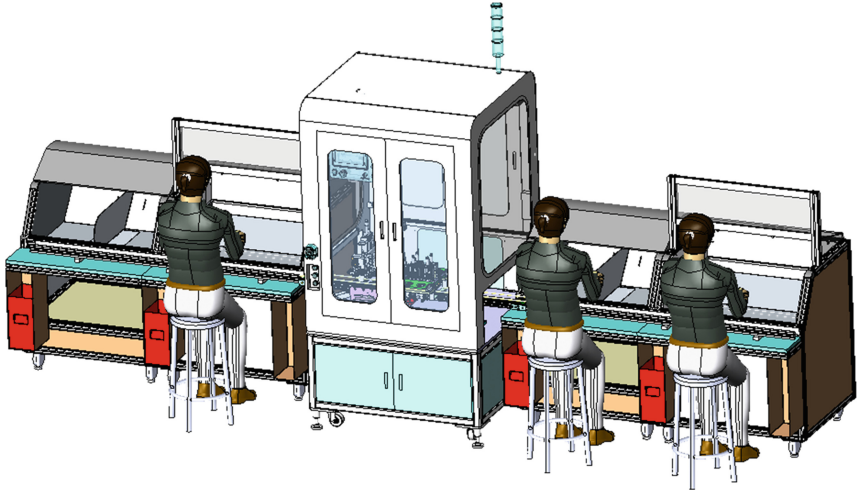


Fig. 1. Manual detection of silver-based contacts

The manual method is very dependent on the professional ability and experience of the operator, the different workers can not guarantee the similar yield, and the manufacturing volume of the contact is gradually increasing, the manual can not bear the long-term, repeated detection work, and can not meet the efficient and high-speed production mode. Automated contact detection system has become a major research direction.

In recent years, there have been many studies using industrial cameras to collect information about contact surface problems [7, 8]. The feature set established by this method is characterized by multi-dimension, large redundancy, convenient feature extraction, etc. The most intuitive advantage of this method is that the physical feature set of the tested object can be saved and inherited to the detection of subsequent assembly links, such as size detection, gloss detection, electrical life detection, etc. However, it is worth noting that the camera can obtain almost all the texture features and geometric information at once, but it is vulnerable to the stability of the light source arrangement and the multiple effects of ambient noise. Therefore, in order to obtain the best detection robustness, it is necessary to regularly maintain and adjust the light source equipment and camera state, which increases the cost of the detection system to a certain extent. In addition, the size of the supporting equipment is also an important factor limiting the full performance of the optical camera, in the batch contact detection pipeline, there is no more space to allow the installation of more than one identification camera and its auxiliary light source system, which will greatly affect the detection system's recognition rate and detection efficiency. Can the acquisition means of the detection system be optimized to meet the purpose of screening good contacts at a lower cost?

In the detection before the contact assembly link, the primary purpose is to screen out the qualified electrical performance of the contact components in the shortest time, its roughness and surface defects are the key indicators affecting the electrical performance, and these two physical parameters only need to be easily obtained through the laser rangefinder. This paper presents a detection method that relies on laser rangefinder to collect surface texture features of contact. The laser acquisition instrument has low cost, simple installation, convenient volume control of supporting equipment, and long maintenance cycle, which is very consistent with the indicators of long-term automated detection. And because the characteristic dimension of laser acquisition is highly condensed, there is no need to remove excess characteristic information, simply from the data processing capacity, the information collected by the laser rangefinder is one order of magnitude smaller than that of the optical photo, which means that the sorting speed will be faster.

The only problem is that the identification of defects at the micro level is more complex, there are more potential types of defects, and the machine cannot guarantee the accuracy and robustness when classifying uneven information. This requires the establishment of a suitable classification model through machine learning to improve the accuracy of the detection system.

2 Contact Defect

2.1 A Subsection Sample

As shown in Fig. 2, the working contact area of the contact component is the top end of the contact. The contactor realizes the change of the working state of the electrical appliance by controlling the opening and closing of the contact and the metal plate or between the contact and the contact. The surface quality of the top contact surface directly affects the working reliability and even the service life of the connector. Therefore, the detection system designed in this paper will detect, sort and analyze the surface fine defects in this area.

From production to transportation to welding and riveting assembly, each link will inevitably make the working contact surface of the contact suffer from impact, scratching or small amplitude regular periodic vibration. These factors may have little impact on large metal parts, but may cause fatal damage at the microscopic level. For example, silver-based material surface cracking, surface peeling, metal coating delamination, oxide migration and other permanent damage [9].

In Fig. 3, surface cracking, surface scratch and layer layering are shown respectively.

The occurrence of mechanical damage cannot be solved by simply improving the wear resistance of the metal coating, even if the coating is prepared into cemented carbide by adding a variety of alloying elements. It is still impossible to completely eradicate the wear rate during transportation, and even the comprehensive performance of the electrical connector will only do more harm than good. In order to improve the wear resistance of the contact and prolong the life, J song adds cobalt, iron and nickel to the material to improve the hardness and wear resistance of the gold plate [10]. However, the wear test results show that the improvement of the wear rate in the contact area is always limited, and this method will reduce the conductivity between the contact and

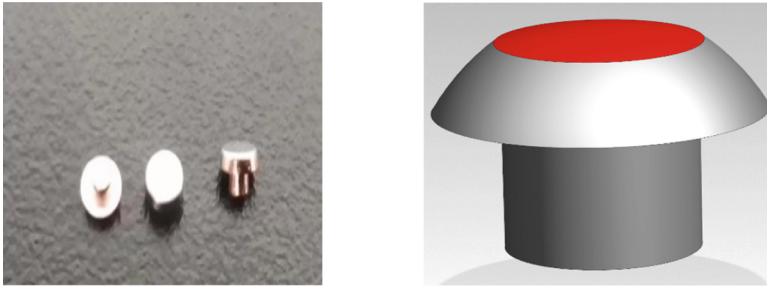


Fig. 2. Silver contact and its effective contact area

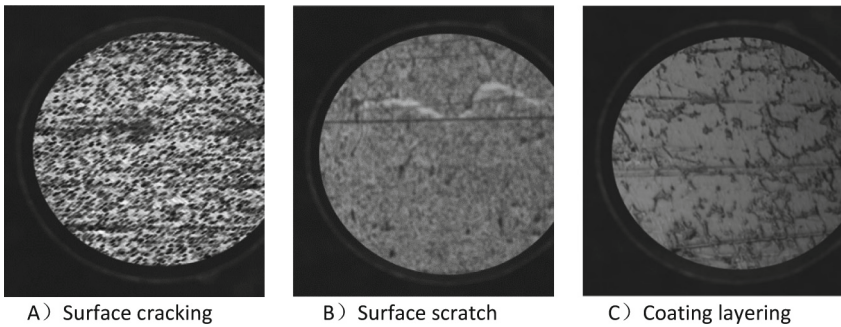


Fig. 3. Contact surface coating failure form

the metal plate, affecting the sensitivity of the corrosion resistance of the working face. Since mechanical damage in transportation is inevitable, it is the most economical and fast way to sort out the defective parts.

Traditional silver-based contacts are mostly silver cadmium oxide, with excellent electrical properties and long-term use, its manufacturing process is relatively mature, but cadmium has a certain toxicity, manufacturing process pollution, gradually replaced by emerging environmentally friendly materials. However, the manufacturing process of environmentally friendly silver-based materials is relatively immature, and in many cases, the control of oxide size cannot be taken into account in order to obtain better electrical properties, and the inconsistency of oxides in the material may lead to bumps, dents and even the migration of oxides on the surface of the contact, as shown in Fig. 4. The defects caused by insufficient manufacturing process will be directly reflected in the electrical properties, and the bumps and depressions will reduce the effective contact area.

In addition, the more complex the surface structure, the higher the sensitivity to fretting corrosion [11], wear debris and dust in the air will gradually erode the surface of the contact, causing secondary damage, and the contact resistance will continue to increase until the component fails.

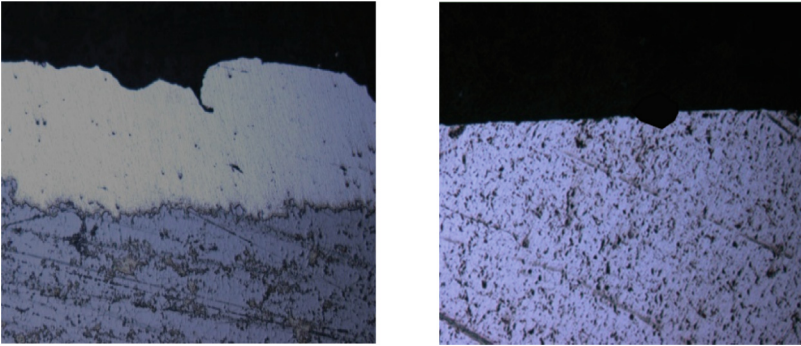


Fig. 4. Failure mode of contact matrix

In this paper, Crack, Scratch, Exfoliation and Delamination were mainly detected and identified for surface coating defects, while Matrix Swelling and Matrix Depression were mainly detected for contact matrix.

3 Detection System

3.1 Acquisition System

The size of the contact electrical components is relatively small, and the non-contact detection device with strong anti-interference ability is used to collect the texture characteristics of the contact surface, which can not only achieve non-destructive online detection, but also ensure the correctness and recognition efficiency of the final sorting results. On the market, the detection of metal surface texture and roughness is mostly measured by contact methods, which can only be sampled offline, and friction with the contact coating will cause surface damage, which is difficult to be used in the detection system of contact surface quality [12].

Contact surface quality detection before assembly only needs to meet the requirements of single surface texture feature collection, and does not need to collect multiple feature information in an all-round way. Therefore, this paper adopts laser speckle method to collect contact surface texture information. Using the speckle caused by the interference of the reflected light phase difference, the texture features of surface quality can be extracted based on the wavelet model [13]. As shown in Fig. 5, using the main light source to illuminate the surface of the measured material at a specified phase Angle, the light detector and light intensity detector first analyze the deviation and intensity of the test laser to determine whether the optical collector needs to suppress or promote the acquisition intensity. After the collection of speckle information, the optical collector can preprocess the original image and extract texture through the built-in switch, or directly transmit the image to the PC for more precise defect detection and classification.

Laser measurement can eliminate the interference caused by the metallic luster of the contact, omit the adjustment of the auxiliary light source, and further improve the intensity of the detection system; The distance between the detection contact and the detection instrument is called the potential object distance. During the long-term service

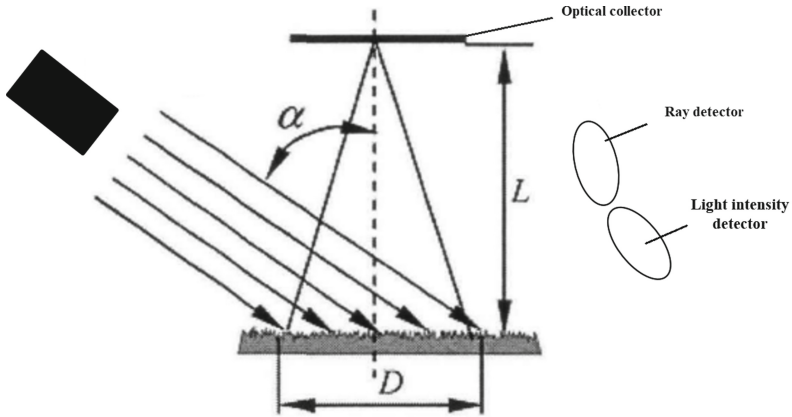


Fig. 5. Schematic diagram of laser acquisition system

of the detection system, the distance deviation will be caused by a variety of external factors (vibration, positioning error, aging of laser components). The laser rangefinder can continue to work as long as the error compensation is introduced in the software level after calibration, and the maintenance cost is relatively low [14].

3.2 Sorting System

The detection system collects the texture features of the contact surface. The smaller the incidence Angle of the laser, the more obvious the scattering effect of the light through the uneven surface, but too small incidence Angle will also cause the speckle image distortion. The contact surface of the contact is curved, while the image in the training set is two-dimensional.

$$\frac{\lambda}{\gamma} \cos^{-1} \frac{H|R - r/2|}{MAXR' - MINR'} \quad (1)$$

In the formula, α is the corrected compensation Angle, λ is the slope of the surface of the contact, γ is the Angle of laser injection, H is the height of the contact, R is the overall radius of the contact, r is the radius of the effective contact area, $MAXR'$ and $MINR'$ are the actual measured maximum diameter and minimum diameter of the contact.

The scanned speckle image is simulated to synthesize numerical image, which is convenient for subsequent surface defect recognition and feature extraction. Due to installation errors, measurement errors, fixture errors and other factors in the actual detection process, data size will be lost when the real contact size is mapped into a two-dimensional image, especially in the subtle scale deviation will be amplified. Remove the influence of noise on the recognition results, and establish the transfer function of speckle contrast and surface roughness:

$$V = \frac{(\Delta I)^{\frac{1}{2}}}{I} = \frac{\sqrt{\Delta I^2 - i^2}}{I}, \quad (2)$$

where ΔI is the average difference between random point area light intensity and speckle light intensity, I is the light intensity of random point area on the speckle projector surface, and i is the average light intensity of the speckle projector surface.

The fitted image data is sent to the computer, and the first step is to judge whether the contact is qualified. The unqualified texture features are included in the feature data set training. In order to improve the efficiency of computer recognition of surface damage types, the contact image is grayed to minimize the information interference of normal surfaces and enhance the image information of damaged parts. As shown in Fig. 6.

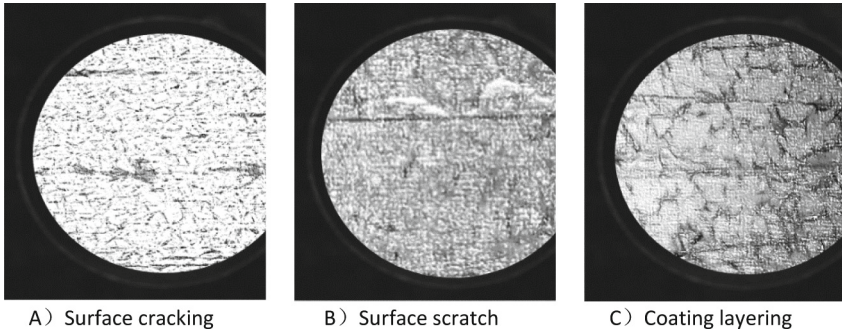


Fig. 6. Enhanced image information

The information of the intact part of the contact is very messy, which will affect the accuracy of the result prediction during deconstruction. The image thinning algorithm is used to eliminate the areas that do not need to be identified, highlight the texture characteristics of the damaged area, and suppress the interference of redundant information. The basic idea of the thinning algorithm is to divide and isolate each region to form a square matrix. The region in the middle is the region to be judged, and the region adjacent to the region to be judged and close to the diagonal is called the first-order symbiotic region, as shown in Fig. 7.

P2	P3	P4
P5	P1	P6
P7	P8	P9

Fig. 7. The region to be judged and the first-order symbiosis region

Suppose that the roughness of the first-order symbiotic region is 1 when the roughness is abnormal, 0 when the roughness is normal, and the number of abnormal symbiotic regions is Numb-P. To determine whether the region to be judged should be eliminated,

all the following conditions must be met at the same time:

$$\begin{cases} (1) P1 \neq 1 \\ (2) P2 * P9 \neq 0 \text{ or } P4 * P7 \neq 0 \\ (3) P3 * P8 \neq 0 \text{ or } P5 * P6 \neq 0 \\ (4) Numb - P < 6 \end{cases} \quad (3)$$

After the image refinement, for example Fig. 8, only the texture features of the defects on the surface of the contact are left, so it is easier to identify the types of defects and train the integrated model.

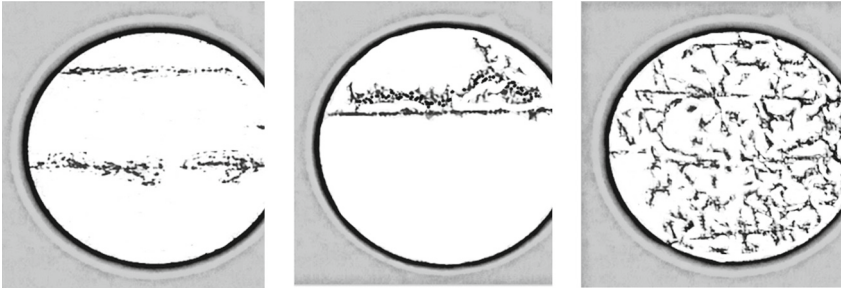


Fig. 8. Defect refinement image

4 Integrated Model Training

Machine learning is to collect a large number of quantifiable raw data for a specific application environment, divide the data set with a reasonable algorithm, and train the computer how to predict the results, and improve the efficiency and accuracy of the prediction results in continuous training. A single model maps the qualified data into the corresponding data set through a preset classification function. At present, the most popular single classification models are decision tree (DT), SVM classifier and K-nearest neighbor method (KNN).

The refined image is extracted for surface texture features, and the above three models are used for classification and judgment respectively. The types of contact defects are judged from the two dimensions of microscopic surface roughness and surface coating peeling degree. The definition of classification labels is summarized as Table 1.

Table 1. Contact defect label

Label	CD1	CD2	CD3	CD4	CD5	CD6
Categories	Crack	Scratch	Exfoliation	Delamination	Matrix swelling	Matrix depression

A single classification model is easily affected by the quality of the data set, and the recognition accuracy cannot be improved due to the lack of anti-interference ability. Moreover, each single model has its own limitations. In the face of the delivered data set, the performance of a single model is shown as Fig. 9.

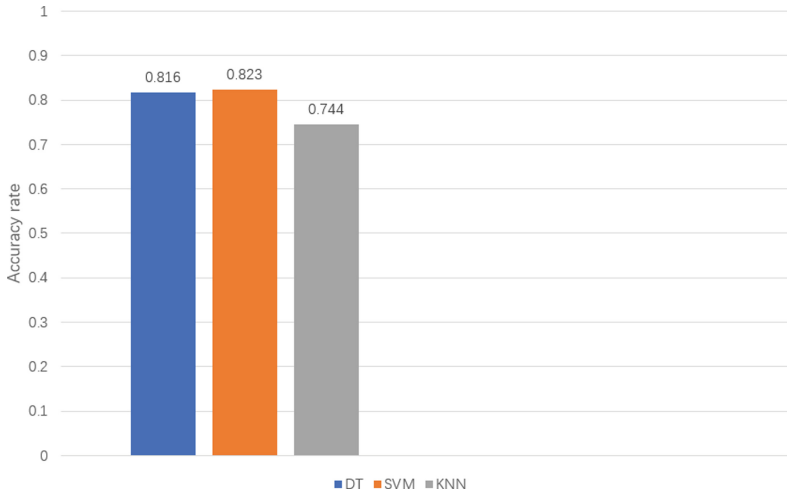


Fig. 9. Accuracy of single classification model

The performance of a single model is not optimistic, and its low accuracy can not meet the general production requirements, especially the KNN model, which relies on the retrieval of adjacent samples in the feature space to optimize the target parameters. If there is no good consistency in the sample set, the feature parameters will be fuzzy, and a large number of samples can not ensure the accuracy and can only be selectively forgotten.

The accuracy and generalization ability of a single model is always limited. In order to further improve the performance of the training model, the most convenient way is to use an integrated algorithm to integrate a single model as a base model into a diversified large model. Parallel classification is carried out among each base model, and independence is maintained between each prediction result. The final result of the integrated model is the average of multiple classification results. In continuous cycle training, it can be found which base model can update the weight of this part of base model according to the test data, so as to achieve a more ideal training model. This paper selects the mainstream integrated models of Random Forest (RF) and XGBoost for training. Figure 10 compares the performance of the integrated model with that of a single model.

In the continuous recognition of a single model, the average speed of individual sample judgment is only 0.67 s, and the average speed of integrated model recognition is about 0.83 s, while the manual screening method takes dozens of times more time. The laser detection system not only has the advantage of shortening time, but also quantifies

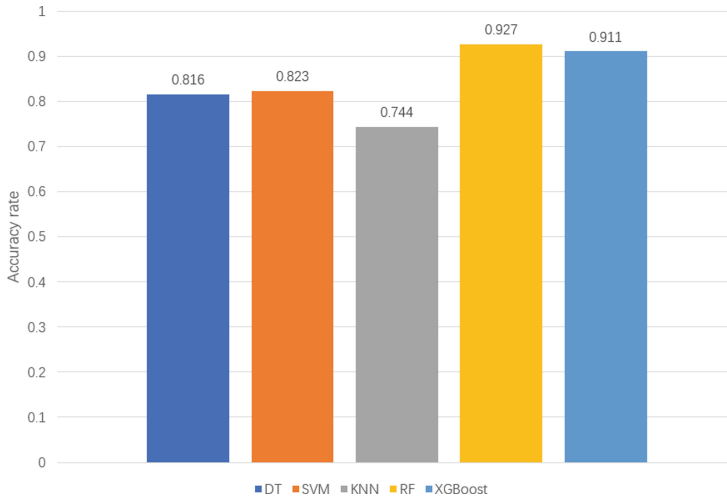


Fig. 10. Accuracy of Ensemble classification model

surface texture features while identifying defects. It provides reference and basis for the life evaluation of Electric contactor.

5 Conclusion

In this paper, a simple online detection system for microscopic defects on the surface of contact is designed by means of laser detection, which realizes the rapid identification of surface texture and defect types. The accuracy of a single model for texture recognition is low, while the integrated model shows high accuracy after training. However, in the case of continuous recognition, the classification efficiency of the training model will gradually deteriorate. Therefore, further research on the recognition of surface defects by convolutional neural network will be very meaningful.

References

1. Xie, M., Wang, S.: Overview of AgSnO₂ electrical contact materials. *Electr. Mater.* **2013**(2), 36–39 (2013)
2. Zhou, X.L., Xiong, A.H., Liu, M.M., et al.: Electrical contact properties of AgSnO₂NiO electrical contact material. *Rare Met. Mater. Eng.* **48**(9), 2885–2892 (2019)
3. Wang, J., Duan, C., Chen, S., et al.: Arc erosion behaviors and surface characteristics of SnO₂ nanofiber/particle reinforced Ag-based composite. *J. Mater. Eng. Perform.* **2023**, 1–11(2023)
4. Liu, M., Chen, J., Cui, H., et al.: Temperature-driven deintercalation and structure evolution of Ag/Ti₃AlC₂ composites. *Ceram. Int.* (2018)
5. Allen, S.E., Streicher, E.: The effect of microstructure on the electrical performance of Ag/Wc/C contact materials. In: *Proceedings of the 44th IEEE Holm Conference on Electrical Contacts*, pp. 276–285 (1998)

6. Swingler, J., McBride, J.W.: The erosion and arc characteristics of Ag/CdO and Ag/SnO₂ contact materials under DC break conditions. *IEEE Trans. Compon. Packag. Manuf. Technol. Part A* **19**(3), 404 (1996)
7. Sun, P.Z.: *Research on Online Machine Vision Technology for Relay Contact and Riveting Point*. Zhejiang University (2019)
8. Mu, D.Y.: *Study on Surface Defect Detection Technology for Contact Materials*. Harbin Institute of Technology
9. Meng, F.B., Lu, J.G.: Failure characteristics of Ag-SnO₂ contact material in AC level. *Rare Met.* **27**(1), 18 (2008)
10. Song, J., Koch, C.: Wear patterns and lifetime of electric contacts. In: *Proceedings of the 54th IEEE Holm Conference on Electrical Contacts, 2008*, pp. 238–244. IEEE (2008)
11. Mashimo, K., Ishimaru, Y.: Computational modeling and analysis of a contact pair for the prediction of fitting dependent electrical contact resistance. In: *2011 IEEE 57th Holm Conference on Electrical Contacts (Holm)*, pp. 1–6. IEEE (2011)
12. White, D.J.: Stylus contact method for surface metrology in the ascendancy. *Meas. Control* **31**(2), 48–50 (1998)
13. Dainty, J.C.: *Laser Speckle and Related Phenomena*. Springer-Verlag, Berlin (1975)
14. Kumar, R., Kulashekar, P., Dhanasekar, B., Ramamoorthy, B.: Application of digital image magnification for surface roughness evaluation using machine vision. *Int. J. Mach. Tools Manuf.* **45**(2), 228–234 (2005). <https://doi.org/10.1016/j.ijmachtools.2004.07.001>



When and Where Conceptual Maths Equals to Conceptual Modeling: Reasons for Using in Cognitive Modeling

Viacheslav Wolfengagen¹(✉), Larisa Ismailova¹, and Sergey Kosikov²

¹ National Research Nuclear University “Moscow Engineering Physics Institute”,
Moscow 115409, Russian Federation

jir.vew@gmail.com , jir.lyui@gmail.com

² NAO “JurInfoR”, Moscow 119435, Russian Federation
kosikov.sv@gmail.com

Abstract. The means and constructions of conceptual mathematics that are most closely related to conceptual modeling are selected. This selection is made in full accordance with the doctrine of computational thinking. An example of a construction representing the processability of individuals and concepts is given. It is shown that functors and natural transformations can serve as a basis for the subsequent construction of non-standard conceptual modeling based on variable domains. On the basis of conceptual mathematics, a conceptual framework of cognitive modeling has been laid, which allows taking into account the stages of knowledge and the transitions between them. It turns out that using conceptual mathematics, one can express cognitive maps.

Keywords: AGI · Cognitive modeling · Information process · Channeling · Computational thinking · Semantic information processing · Computational model · Variable domains

1 Introduction

Why not start right away with a discussion of the epistemological status of category theory? This theory begins to play an increasingly prominent role in applied research, making its way from purely abstract reasoning towards cognitive research and a full-scale system of knowledge representation – conceptual mathematics in the sense of Lawvere. Let us dwell on some points related to the possibilities of solving problems, paying attention to what kind of problems are solved in this case.

Category theory alone does not solve difficult problems in topology or algebra. It clears out a confusing set of individually trivial problems. This puts difficult problems in clear garb and makes their solution possible.

A category in its full generality is nothing more than a generalization of a labeled directed multigraph – a class of objects and a class of arrows (also known

as morphisms) between them. Category theory is used in various branches of mathematics, both to unify some of the 'natural' definitions and to the tools it can help develop. Its origins lie in the algebraic topology of the mid-20s, but in less than a century since its development, the subject has become widespread.

In particular, category means are used in homological and cohomological algebra, algebraic geometry, and algebraic topology. They appear as knot invariants, abstract vector spaces, and more. They also used outside of pure mathematics: in mathematical physics, biology, and (especially) computer science, where categories provide a useful language for functional programming.

It is known that the most important difference between different areas of mathematics are the types of functions that are considered there are linear transformations, homeomorphisms, automorphisms, and so on. A better understanding is achieved after one begins to study category theory. The maxim to remember is that *an object is defined by its relations to other objects* – succinctly, 'it's all about maps', or, in other terminology, mappings.

For uniformity and ease of perception, the conceptual mathematics of Lawvere [13] will be used. Where possible, the variable set construction [11, 12] developed in his papers over a number of years will be used. The task set is quite complex and peculiar: to bring such an abstract discipline as category theory [17] to the possibility of constructing structures that can be found useful in computer science [24] and cognitive research in the development of cognitive models.

Individual concepts are understood in the sense of Carnap-Bar-Hillel [18], 'legitimate representatives' of concepts are domains, understood as variable sets in the sense of Lawvere. Individuals have been considered by many authors, but, apparently, Scott went deeper, understanding individuals as processes [19–21]. It is not so much about physical individuals and their variability, but about the 'legitimate representatives' of individuals, which allows us to build their theory.

Then the plan for researching the problem domain – subject area, – is as follows.

1. Individual objects are distinguished, which are subdivided into *real*, *possible* and *virtual*. This can be done, for example, by using individualizing functions, requiring sufficient selectivity from them. The individuals are named.
2. Relationships are established between individuals, and these relationships are named.

All this refers to the *decomposition* phase, if we apply the *computational thinking* doctrine of Denning [1, 2].

In the *recognition* phase, we collect the individuals in the aggregate – into Lawvere's *variable domains*. Dependencies – *conceptual relationships* are established between domains. This can be considered as some version of the conceptual model, with the difference that the concepts are now represented by variable sets. Immediately, we note that this is a non-standard conceptual model.

At the *abstraction* phase, a *representation* is chosen to be the *functorial category*. On this *comprehension* of the problem domain is still far from complete. If standard conceptual modeling leads to a semantic network, now we have to look

for answers to the acute questions of artificial intelligence from other positions: what's in a link? what's in a node? These new answers are simple in appearance: there is a natural transformation on the link, and a functor is located in the node.

At the algorithmization phase, it remains to arrange one or another model of *information processes*, which are the 'legitimate representatives' of individuals and individual concepts.

The whole thing, if desired, can be seen as an example of the application of computational thinking.

What is the role and place given to *cognitive modeling*?

The defended thesis is that computational thinking is considered as one of the skills in solving problems of cognitive modeling, and quite useful.

The functor category turns out to be quite representative, functors are associated with 'stages of knowledge', natural transformations model transitions between states. The presence of stages of knowledge is characteristic of the whole variety of cognitive models.

2 Related Works

This paper touches upon the issues of conceptual modeling to some extent. Conceptual process-based modeling introduces certain cognitive difficulties [25].

Cognitive Modeling (CM) finds application in [10] use cases, for concept acquisition [16], formation of cognitive context [15]. The question arises, to which there is still no convincing answer, how to use the notion of a function in the conditions of variability of its domain of definition and range in the process of defining this function. In addition, the function definition itself and its argument can, in turn, *be generated by processes*.

The interaction of a function with an argument is also not at all instantaneous, but has a *process interpretation*. The above questions, as it turns out, are of a certain fundamental nature. The following works of the group of authors are devoted to clarifying the answers to them, and, as it turned out, not all answers are unambiguous. Nevertheless, a number of them find full or partial explanation in the context of *conceptual mathematics*.

An analysis of the variability of characters is given in [3]. The results using continuity of functions allow one to obtain a commutative diagram of concepts given in [5]. An important result is the constructed 'conceptual hanger' for domains/concepts.

Ismailova et al. [7] explores the conceptualization/individuation that is important for the development of non-standard conceptual modeling. It is shown that the applicative approach forms a metatheoretical basis for cognitive systems. The paper [9] shows how to use composition and the commutativity to generate information channels. There is a shift in the understanding of computing as a natural science [26]. On other hand, [27] investigated one of the conditions for safe interaction of semantic processes, including cognitive interference. The process of non-standard investigation of the subject area was considered in [8],

and in [4] conceptual mathematics was applied to build a computational model. The paper [6] deals with an applicative model for semantic analysis, and [23] deals with the problem of generating words in a context-sensitive language for the needs of semantic analysis.

3 Elements of Conceptual Mathematics

Let \mathcal{C} is a category and \mathcal{S} is a category of sets and arbitrary maps. The definitions below can be found in any textbook on category theory, e.g. [17]. Nevertheless, we follow the book ‘Conceptual mathematics’ [13] with the elements and notations from [14, 21, 22].

The details are straightforward and this was known from a very early chapter in topos theory; but we use it in other context and understanding. What is a contravariant functor? It is a mapping $F : \mathcal{C} \rightarrow \mathcal{S}$ that associates to every domain A of \mathcal{C} a set $F(A)$ of \mathcal{S} and to every map $f : B \rightarrow A$ of \mathcal{C} a function $F(f) : F(A) \rightarrow F(B)$ (and note the change of order!) So that:

$$\begin{aligned} F(1_A) &= 1_{F(A)} \\ F(f \circ g) &= F(g) \circ F(f), \end{aligned} \tag{1}$$

provided $f : B \rightarrow A$ and $g : C \rightarrow B$ in \mathcal{C} . It was one of the major early insights of category theory to see that the functors form a category in themselves, what is needed is a definition of transformation between functors. We call such maps $\nu : F \rightarrow G$ ‘natural transformation’ for reasons explained in category theory books. What is a natural transformation $\nu : F \rightarrow G$? It is a association with every domain A in \mathcal{C} of a function

$$\nu_A : F(A) \rightarrow G(A) \tag{2}$$

so that whenever

$$f : B \rightarrow A \tag{3}$$

in \mathcal{C} , then the following diagram commutes in \mathcal{S} and is shown in Fig 1. This means

$$\nu_B \circ F(f) = G(f) \circ \nu_A. \tag{4}$$

An example will help explain this though looking slightly compressed.

For each C of \mathcal{C} , let

$$H_C(A) = \{h|h : A \rightarrow C\} \tag{5}$$

end if $f : B \rightarrow A$ in \mathcal{C} , let $H_C(f)$ be the map taking $h \in H_C(A)$ into $h \circ f \in H_C(B)$. It is easy to show H_C is a (contravariant) functor. It is often called the *representable functor* (corresponding to \mathcal{C}), and we shall see that it is very ‘representative’.

Now let

$$g : C \rightarrow D \tag{6}$$

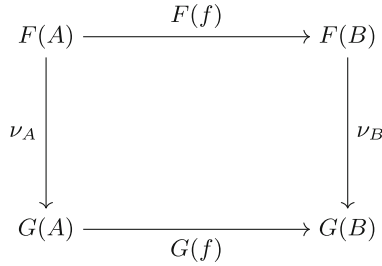


Fig. 1. Commutative diagram for natural transformations

in \mathcal{C} . There is a natural transformation $H_g : H_C \rightarrow H_D$ because, for each $h \in H_C(A)$, we can map it to $g \circ h \in H_D(A)$, naturally.

The composite map for $f : B \rightarrow A$ takes

$$h \in H_C(A) \tag{7}$$

into

$$g \circ h \circ f \in H_D(B), \tag{8}$$

and there are two equal ways to calculate it owing to the associativity of composition (in \mathcal{C}); that is why the necessary diagram commutes.

Not only are the H_C pleasant functors with cooperative natural transformations between them, but then by now classic Yoneda Lemma proves for us that the only natural transformations $\nu : H_C \rightarrow H_D$ are those of the form H_g for some $g : C \rightarrow D$.

4 Conclusion and Future Research

The means and constructions of conceptual mathematics that are most closely related to conceptual modeling are selected. This selection is made in full accordance with the doctrine of computational thinking. An example of a construction representing the processability of individuals and concepts is given.

- It is shown that functors and natural transformations can serve as a basis for the subsequent construction of non-standard conceptual modeling based on variable domains.
- On the basis of conceptual mathematics, the conceptual framework of cognitive modeling is laid, which allows taking into account the stages of knowledge and the transitions between them. It turns out that using conceptual mathematics it is possible to express cognitive maps, but a more detailed example is left for the future.
- It seems possible to build a property change model. This is left for the future, but the path is approximately clear: categories allow and imply an analysis of the existence and formation of objects, which is important for cognitive modeling.

References

1. Carnegie Mellon University Center for Computational Thinking. www.cs.cmu.edu/CompThink. Accessed: 11 Sep 2021
2. Denning, P.J., Tedre, M.: Computational Thinking. The MIT Press (2019)
3. Ismailova, L., Wolfengagen, V., Kosikov, S.: A mathematical model of the feature variability. *Procedia Comput. Sci.* **190**, 312–316 (2021). <https://doi.org/10.1016/j.procs.2021.06.041>. www.sciencedirect.com/science/article/pii/S1877050921012850. 2020 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: Eleventh Annual Meeting of the BICA Society
4. Ismailova, L., Wolfengagen, V., Kosikov, S.: Applicative approach to construe a computational model of concepts and individuals. *Procedia Comput. Sci.* **213**, 463–470 (2022). <https://doi.org/10.1016/j.procs.2022.11.092>. www.sciencedirect.com/science/article/pii/S1877050922017835. 2022 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: The 13th Annual Meeting of the BICA Society
5. Ismailova, L., Wolfengagen, V., Kosikov, S.: Conceptual hanger for displaced concepts: a framework for information processes variability. *Procedia Comput. Sci.* **213**, 588–595 (2022). <https://doi.org/10.1016/j.procs.2022.11.109>. www.sciencedirect.com/science/article/pii/S1877050922018087. 2022 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: The 13th Annual Meeting of the BICA Society
6. Ismailova, L., Wolfengagen, V., Kosikov, S.: Elements of semantic analysis based on lambda-calculus. *Procedia Comput. Sci.* **213**, 471–476 (2022). <https://doi.org/10.1016/j.procs.2022.11.093>. www.sciencedirect.com/science/article/pii/S1877050922017847. 2022 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: The 13th Annual Meeting of the BICA Society
7. Ismailova, L., Wolfengagen, V., Kosikov, S.: Lambda-calculus, combinators and applicative computational technologies. *Cogn. Syst. Res.* **76**, 93–100 (2022). <https://doi.org/10.1016/j.cogsys.2022.10.002>. www.sciencedirect.com/science/article/pii/S1389041722000468
8. Ismailova, L., Wolfengagen, V., Kosikov, S.: A prototype system for supporting a network of information graphs with the ability to assess the nature of the subject's knowledge. *Procedia Computer Science* **213**, 16–20 (2022). <https://doi.org/10.1016/j.procs.2022.11.033>. www.sciencedirect.com/science/article/pii/S1877050922017240. 2022 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: The 13th Annual Meeting of the BICA Society
9. Ismailova, L., Wolfengagen, V., Kosikov, S., Andronov, S.I.: The applicative approach to the synthesis of a data structure with the given combinatory characteristic. *Cogn. Syst. Res.* **77**, 88–93 (2023). <https://doi.org/10.1016/j.cogsys.2022.10.010>. www.sciencedirect.com/science/article/pii/S1389041722000602
10. Kim, J.H., Lee, S.M.: Case-based cognitive modeling: a student modeling methodology for an intelligent tutoring system. Ph.D. thesis (1993). AAI9415973
11. Lavwere, F.W.: Variable quantities and variable structures in topoi. In: Heller, A., Tierney, M. (eds.) *Algebra, Topology, and Category Theory*, pp. 101–131. Academic Press (1976). <https://doi.org/10.1016/B978-0-12-339050-9.50014-6>

12. Lawvere, F.W.: Continuously variable sets: Algebraic geometry = geometric logic. In: Rose, H., Shepherdson, J. (eds.) *Logic Colloquium '73*, Studies in Logic and the Foundations of Mathematics, vol. 80, pp. 135–156. Elsevier (1975). [https://doi.org/10.1016/S0049-237X\(08\)71947-5](https://doi.org/10.1016/S0049-237X(08)71947-5)
13. Lawvere, F.W., Schanuel, S.J.: *Conceptual Mathematics: A First Introduction to Categories*. Cambridge University Press, Cambridge (1997)
14. Mac Lane, S.: Logic, Meaning and Computation: Essays in Memory of Alonzo Church, chap. The Lambda Calculus and Adjoint Functors, pp. 181–184. Springer, Netherlands, Dordrecht (2001). https://doi.org/10.1007/978-94-010-0526-5_7
15. Mei, L.: Cognitive context elicitation and modeling. Ph.D. thesis, CAN (2011). AAINR78288
16. Miller, C.S.: Modeling concept acquisition in the context of a unified theory of cognition. Ph.D. thesis, USA (1993). UMI Order No. GAX93-32133
17. Schubert, H.: *Categories*. Springer, Berlin Heidelberg, Berlin, Heidelberg (1972)
18. Scott, D.: Advice on modal logic. In: K. Lambert (ed.) *Philosophical Problems in Logic: Some Recent Developments*, pp. 143–173. Springer, Netherlands, Dordrecht (1970). https://doi.org/10.1007/978-94-010-3272-8_7
19. Scott, D.: The lattice of flow diagrams. In: Engeler, E. (ed.) *Symposium on Semantics of Algorithmic Languages*, pp. 311–366. Springer, Berlin Heidelberg, Berlin, Heidelberg (1971)
20. Scott, D.: Identity and existence in intuitionistic logic, pp. 660–696. Springer, Berlin Heidelberg, Berlin, Heidelberg (1979). <https://doi.org/10.1007/BFb0061839>
21. Scott, D.: Relating theories of the λ -calculus. In: Hindley, J., Seldin, J., Curry, H.B. (eds.) *Essays on Combinatory Logic, Lambda-Calculus and Formalism*, pp. 403–450. Academic Press, Berlin (1980)
22. Scott, P.J.: Functorial polymorphism and semantic parametricity. *Diagrammes* **22**, 77–90 (1989). www.numdam.org/item/DIA_1989__22__77_0
23. Slietsov, I.O., Wolfengagen, V.E., Kosikov, S.V.: Constructing generator of words of context-sensitive language on example of typed λ -calculus. *Procedia Computer Science* **213**, 556–562 (2022). <https://doi.org/10.1016/j.procs.2022.11.104>; 2022 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: The 13th Annual Meeting of the BICA Society. www.sciencedirect.com/science/article/pii/S1877050922017951
24. Spivak, D.I.: *Category Theory for the Sciences*. The MIT Press (2014)
25. Tolk, A., Turnitsa, C.: Conceptual modeling with processes. In: *Proceedings of the Winter Simulation Conference, WSC '12*. Winter Simulation Conference (2012)
26. Wolfengagen, V., Ismailova, L., Kosikov, S.: Cognitive technology to capture deep computational concepts with combinators. *Cogn. Syst. Res.* **71**, 9–23 (2022). <https://doi.org/10.1016/j.cogsys.2021.10.001>. www.sciencedirect.com/science/article/pii/S1389041721000747
27. Wolfengagen, V., Ismailova, L., Kosikov, S., Babushkin, D.: Modeling spread, interlace and interchange of information processes with variable domains. *Cogn. Syst. Res.* **66**, 21–29 (2021). <https://doi.org/10.1016/j.cogsys.2020.10.016>. www.sciencedirect.com/science/article/pii/S1389041720300905



A Prediction Model for the Equivalent Parameters of an Acoustic Transducer Based on DPSD and LSTM Neural Network

Yuhui Xue^{1,2}, Zhidi Jiang^{1,2}(✉), and Mudan Yu²

¹ Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo Zhejiang 315211, China

² College of Science & Technology, Ningbo University, Ningbo Zhejiang 315300, China

jiangzhidi@nbu.edu.cn

Abstract. This article proposes a novel model for predicting the equivalent parameters of underwater acoustic transducers. To address this challenging task, the model utilizes Digital Phase-Sensitive Detection(DPSD) on the detection signal to obtain the amplitude and phase difference. Subsequently, the model applies Long Short-Term Memory(LSTM) and Gated Recurrent Unit(GRU) neural networks for classification and regression predictions of the equivalent models and parameters of the transducer. Through simulation experiments using MATLAB, the experimental results prove that the LSTM prediction model has higher prediction accuracy and robustness than the GRU prediction model. This provides a new method for modeling equivalent parameters of underwater acoustic transducers.

Keywords: Acoustic transducers · Equivalent parameter prediction model · Digital phase-sensitive detection · Long short-term memory · Gated recurrent unit

1 Introduction

The hydrophone is a sensor that can convert sound wave signals into electrical signals or vice versa. To better understand and apply hydrophones, researchers often use equivalent parameter modeling. However, due to the complexity of the hydrophone, its equivalent model and parameters are often difficult to predict accurately. This paper proposes a hydrophone equivalent parameter prediction model based on DPSD and LSTM neural networks. The model aims to improve the accuracy and efficiency of predictions. Firstly, the DPSP technology based on the cross-correlation principle can effectively extract the amplitude and phase information of the tested hydrophone at different frequencies from noise, which reflects the characteristics of the hydrophone.

Belagoune et al. used LSTM neural network to detect, diagnose and locate transmission line faults in large-scale multi-machine power systems [1]. Sharma et al. proposed a real-time structural damage assessment method based on LSTM network. Both studies achieved high accuracy and robustness in fault recognition and damage recognition by using regression and classification to detect and locate faults and damage in [2]. Rana proposed a method for emotion classification from noisy speech based on GRU, and the experimental results showed that using GRU performed well in the classification task [3]. LSTM and GRU are commonly used recurrent neural network models, where LSTM is more suitable for processing long sequence data and has shown better performance in various application fields than GRU. Therefore, this paper selects LSTM to predict the equivalent model type and parameters of the tested hydrophone through classification prediction and regression prediction. Five possible hydrophone RLC equivalent models were constructed in this paper, and a large amount of hydrophone data with different RLC parameters were generated through MATLAB simulation for training and testing LSTM neural network models.

2 Research Methods

2.1 Digital Phase-Sensitive Detection

Digital Phase-Sensitive Detection is a technique for measuring the amplitude and phase of a signal. It can be used to extract weak signals from noisy backgrounds by filtering out components that are not coherent with the reference signal [6]. After sampling with an analog-to-digital converter, DPSD converts the continuous signal into a discrete sequence, resulting in the following discrete sequence.

$$x[n] = s[n] + w[n] = A_0 \sin\left(2\pi \frac{n}{N} + \phi\right) + w[n] \quad (1)$$

$$p[n] = \sin\left(2\pi \frac{n}{N}\right) \quad (2)$$

$$q[n] = \cos\left(2\pi \frac{n}{N}\right) \quad (3)$$

where $N = \frac{f_s}{f}$, f_s is the sampling frequency, f is the signal frequency, A_0 is the signal amplitude, ϕ is the signal phase, $w[n]$ is random white noise. $p[n]$ and $q[n]$ are reference signals with the same frequency and amplitude and a phase difference of $\pi/2$.

The digital phase-sensitive detection method uses the correlation relationship between the effective sine wave in the detection signal and the reference signal to obtain the amplitude and phase difference. Since system noise is random and has no correlation with the reference signal, it can be effectively suppressed by

cross-correlation calculation. The detection signal is cross-correlated with $p[n]$ and $q[n]$ respectively, as shown in formula (4) and (5).

$$R_{xp}[k] = \sum_{n=0}^{N-1} x[n]p[n - k] \tag{4}$$

$$R_{xq}[k] = \sum_{n=0}^{N-1} x[n]q[n - k] \tag{5}$$

Substituting $x[n]$, $p[n]$, and $q[n]$ into the above formula yields.

$$R_{xp}[k] = \sum_{n=0}^{N-1} \left(A_0 \sin \left(2\pi \frac{n}{N} + \phi \right) + w[n] \right) \sin \left(2\pi \frac{n - k}{N} \right) = R_{sp}[k] + R_{wp}[k] \tag{6}$$

$$R_{xq}[k] = \sum_{n=0}^{N-1} \left(A_0 \sin \left(2\pi \frac{n}{N} + \phi \right) + w[n] \right) \cos \left(2\pi \frac{n - k}{N} \right) = R_{sq}[k] + R_{wq}[k] \tag{7}$$

In the equation, $R_{xp}[k]$ and $R_{xq}[k]$ represent the correlation relationship between the effective components in the detection signal and the reference signal, which is also the key to obtaining the amplitude and phase difference [7]. $R_{wp}[k]$ and $R_{wq}[k]$ represent the correlation relationship between noise signals and reference signals. Since noise is random and has no correlation with reference signals, it satisfies formulas (8) and (9).

$$R_{wp}[k] = 0 \tag{8}$$

$$R_{wq}[k] = 0 \tag{9}$$

Let $k = 0$, then we have.

$$R_{xp}[0] = R_{sp}[0] = \frac{A_0}{2} \cos \phi \tag{10}$$

$$R_{xq}[0] = R_{sq}[0] = \frac{A_0}{2} \sin \phi \tag{11}$$

The amplitude and phase of the detection signal can be calculated through formulas (12) and (13).

$$A = 2\sqrt{R_{xp}^2[0] + R_{xq}^2[0]} \tag{12}$$

$$\phi = \arctan \frac{R_{xq}[0]}{R_{xp}[0]} \tag{13}$$

After the cross-correlation operation, it is found that the DPSD algorithm can remove unrelated noise in the original signal, extract phase and amplitude information, and achieve a filtering effect. However, traditional cross-correlation methods require a large number of multiplication and addition operations, which are relatively time-consuming in calculation. By referring to [8], it is known that the Fast Fourier Transform (FFT) algorithm can quickly calculate the frequency spectrum of the signal due to its high precision, high efficiency, and parallelism. Therefore, it can perform cross-correlation operations more quickly to improve the calculation accuracy and operation efficiency of digital phase-sensitive detection.

To use FFT to calculate cross-correlation, it is necessary to zero-pad the detection signal and reference signal so that their lengths are:

$$L = 2M - 1 \tag{14}$$

where M is the length of the detection signal and reference signal. Perform FFT on the zero-padded sequence to obtain $X[k]$, $P[k]$, and $Q[k]$. Then perform conjugate multiplication to obtain formulas (15) and (16).

$$R_{XP}[k] = X[k]P^*[k] \tag{15}$$

$$R_{XQ}[k] = X[k]Q^*[k] \tag{16}$$

Perform IFFT on $R_{XP}[k]$ and $R_{XQ}[k]$ to obtain the cross-correlation functions $R_{xp}[k]$ and $R_{xq}[k]$. Then use formulas (12) and (13) to solve the amplitude and phase of the detection signal. The improved digital phase-sensitive detection principle using FFT to calculate cross-correlation is shown in Fig. 1.

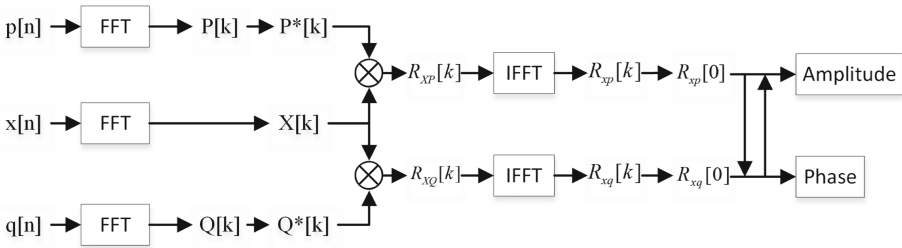


Fig. 1. Improve the schematic diagram of the digital lock-in detection algorithm

Field-Programmable Gate Array(FPGA) is a programmable logic device that can be programmed to perform various digital signal processing functions. Due

to its high-speed parallel processing capabilities, FPGA can achieve high-speed and high-precision digital filtering. Considering that the subsequent work will use FPGA for actual collection and processing of signals through underwater acoustic transducers, this application scenario has high requirements for real-time performance and accuracy. Therefore, this article chooses to use the improved digital lock-in detection algorithm [9].

2.2 Construction of the Equivalent Model for Transducers

According to the different physical mechanisms of transducers from [10], equivalent RLC models can be established as shown in Fig. 2.

2.3 Long Short-term Memory(LSTM)

LSTM is a model based on Recurrent Neural Networks (RNN) and is mainly used to solve sequence data problems. By referring to [11], it is known that the LSTM can automatically learn the characteristics and rules of input data, which performs well in regression prediction, but it faces problems such as overfitting and imbalanced data. To achieve optimal performance, careful design of model structure and adjustment of hyperparameters are required for different tasks and data.

The basic structure of LSTM consists of multiple gate units, each of which has three inputs [12]: input x_t , the previous hidden state h_{t-1} , and memory state C_{t-1} , and outputs the current hidden state h_t and memory state C_t . Among them, memory state C_t is the key to LSTM, allowing the network to retain information for a long time.

A standard LSTM gate unit usually includes the following three parts:

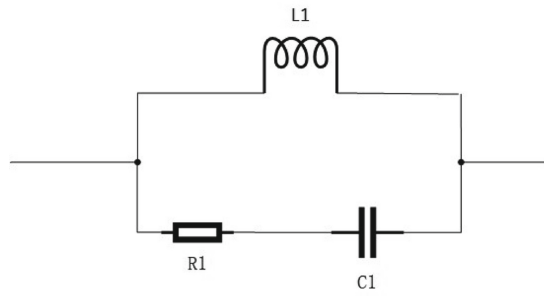
- (1) Input Gate: determines which new information will be added to the memory state. The input gate controls the flow from input x_t and the previous hidden state h_{t-1} to memory state C_{t-1} .
- (2) Forget Gate: determines which old information will be retained in the memory state. The forget gate controls the flow from the previous memory state C_{t-1} to the current memory state C_t .
- (3) Output Gate: determines what kind of hidden state will be passed to the next time step. The output gate controls the flow from the current memory state C_t to the currently hidden state h_t .

The input and output of each gate unit are in vector form, and their calculation method is as follows [13].

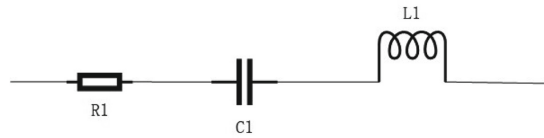
Input Gate:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (17)$$

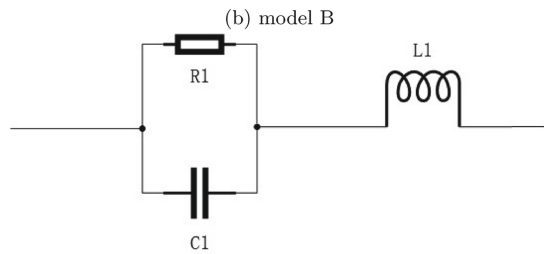
$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (18)$$



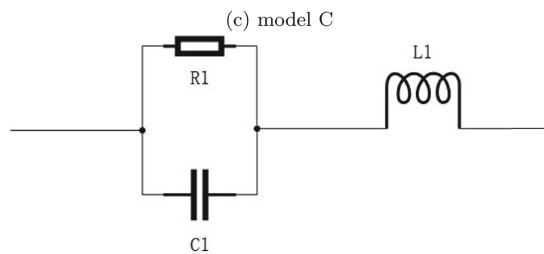
(a) model A



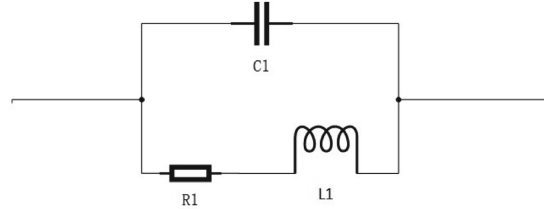
(b) model B



(c) model C



(d) model D



(e) model E

Fig. 2. The equivalent model for transducers

Forget Gate:

$$f_t = \sigma(W_f [h_{t-1}, x_t] + b_f) \quad (19)$$

Memory State Update:

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (20)$$

Output Gate:

$$O_t = \sigma(W_o [h_{t-1}, x_t] + b_o) \quad (21)$$

$$h_t = O_t \times \tanh(C_t) \quad (22)$$

where σ is the sigmoid function, \times represents element-wise multiplication, W_i represents the weight of the previous time step's hidden layer output value for the input gate, W_c represents the weight of the previous time step's hidden layer output value for the memory cell, W_f represents the weight of the previous time step's hidden layer output value for the forget gate, W_o represents the weight of the previous time step's hidden layer output value for the output gate, b_i represents the bias of the input gate, b_c represents the bias of the memory cell, b_f represents the bias of the forget gate, and b_o represents the bias of the output gate.

2.4 Gated Recurrent Unit (GRU)

The Gated Recurrent Unit (GRU) neural network is an improved structure for the RNN. By adding gate mechanisms, the network can better capture the dependency between input sequences and avoid the gradient vanishing problem. Compared to traditional RNN structures, GRU neural networks have stronger memory capabilities and higher training efficiency.

GRU neural networks are basically composed of two important gate units: the update gate and the reset gate [4]. The update gate controls the degree of retention of the memory state generated at the previous time step in the current time step, while the reset gate determines the degree to which the previous memory state should be ignored and the memory state should be recalculated based on the current input. In addition, GRU also introduces a new state vector, the candidate memory state, which replaces the traditional hidden state. This is done to reduce information loss and increase the capacity of the model.

GRU neural networks can be used for various classification and prediction tasks. In practice, due to its smaller number of parameters and higher training speed, it is often used as an alternative to RNN.

The formula derivation process of the GRU neural network is as follows [5]: Assuming that the input at time t is $x_t \in R^d$, the hidden state at the previous time step is $h_{t-1} \in R^n$, the current hidden state is $h_t \in R^n$, and the candidate

memory state is $\tilde{h}_t \in R^n$. The update gate is defined as $z_t \in [0, 1]^n$, and the reset gate is defined as $r_t \in [0, 1]^n$. First, the candidate memory state is calculated based on the input and the hidden state at the previous time step.

$$\tilde{h}_t = \tanh(W_x x_t + r_t \cdot W_h h_{t-1}) \quad (23)$$

whereas $W_x \in R^{n \times d}$ and $W_h \in R^{n \times n}$ are trainable weight matrices.

Calculate two gate vectors: update gate and reset gate. Their calculation methods are as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad (24)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \quad (25)$$

where $W_z, W_r \in R^{n \times d}$ and $U_r \in R^{n \times n}$ are learnable weight matrices, σ represents the sigmoid function.

Finally, we can calculate the hidden state h of the current time step:

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t \quad (26)$$

Formula (26) is the core of the GRU neural network. It adds the previous hidden state and the current candidate memory state according to the proportion of the update gate and reset gate, and generates the current hidden state. In this formula, if the update gate is 0, it means that the previous hidden state is completely discarded; if the update gate is 1, it means that the current candidate memory state is completely accepted. When the update gate is between 0 and 1, it means that the previous hidden state and the current candidate memory state are combined in some way.

3 Simulation Experiment and Result Analysis

3.1 Simulation Implementation Process of Predictive Model for Hydrophone Parameter Using Simulation

The entire simulation implementation of the prediction model is divided into two steps.

In the first step, a discrete sine sequence with a frequency range of 100 KHz–1 MHz and a frequency interval of 100 KHz is generated using MATLAB. Gaussian noise voltage signal sequence is added to the discrete sine sequence. At the same time, two reference signal sequences with the same frequency and mutually orthogonal are generated. Then, the sine sequence with added noise is passed through the transducer under test to obtain a detection signal. The corresponding Direct Current(DC) component is obtained by performing cross-correlation operations on the detection signal and reference signal. The amplitude and phase of the detection signal can be calculated using the estimated DC component value.

Based on the general properties and design principles of underwater acoustic transducers, the equivalent capacitance of an underwater acoustic transducer is usually between several hundred picofarads to several nanofarads, the equivalent inductance is usually between tens to hundreds of microhenries, and the equivalent resistance is usually between tens to hundreds of ohms [14]. For each possible transducer model, build 50,000 test transducers with different R, L, and C values within a reasonable range. After digital lock-in detection, obtain amplitude and phase data for each test transducer at ten different frequency points. Use these 250,000 sets of amplitude and phase data as the input features for the dataset, and use the category of the test transducer and its corresponding RLC parameters as the output features for the dataset.

In the second step, LSTM and GRU neural network model is used to classify and predict the equivalent model of the transducer under test. After determining the equivalent model, R, L, and C parameters in the model are predicted. The flowchart of the entire process is shown in Fig. 3.

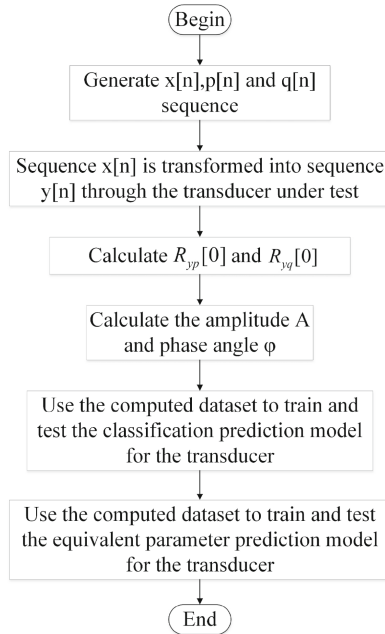


Fig. 3. Flowchart of the simulation implementation process of predictive model for hydrophone parameter.

3.2 MATLAB Simulation of Digital Synchronous Detection

MATLAB software was used to simulate and compare the results of the digital synchronous detection algorithm under two different computation methods and evaluate their specific performance.

The comparison of the calculation results of the digital synchronous detection before and after improvement is shown in Table 1. According to Table 1, the FFT-based digital synchronous detection generally has higher calculation accuracy than before the improvement. When we use the time-domain-based method for cross-correlation calculation, we have to calculate the product at each time point and add all the results to obtain the final result. This requires a very large amount of calculation, especially for longer signal sequences. Using the FFT method, however, we can transform a signal into the frequency domain and perform calculations in the frequency domain. This means that we only need to perform one FFT to obtain the cross-correlation results at all time points.

Table 1. Comparison of transducer phase and amplitude measurement results

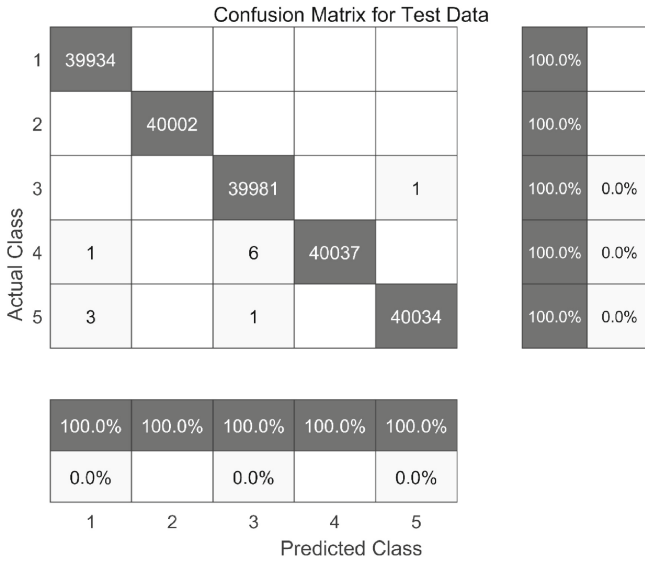
Deviceundertest	Theoreticalphasedifference	Originalphasedifference	Improvedphasedifference	Theoreticalamplitude	Originalamplitude	Improvedamplitude	Testfrequency
A	3.59	3.72	3.63	10.02	10.08	10.07	100 KHz
B	12.52	12.68	12.58	158.87	155.46	160.02	100 KHz
A	10.67	10.85	10.74	10.18	9.94	10.19	300 KHz
B	- 25.38	- 25.45	- 25.52	365.64	367.13	364.23	300 KHz
A	17.44	17.70	17.55	10.92	10.72	11.06	500 KHz
B	- 77.72	- 77.94	- 77.88	111.75	108.40	110.62	500 KHz
A	32.14	32.76	32.24	11.81	12.32	11.88	1 MHz
B	- 88.72	- 90.40	- 89.29	70.41	70.09	70.24	1 MHz

Table 2. Comparison of transducer model and parameter prediction results

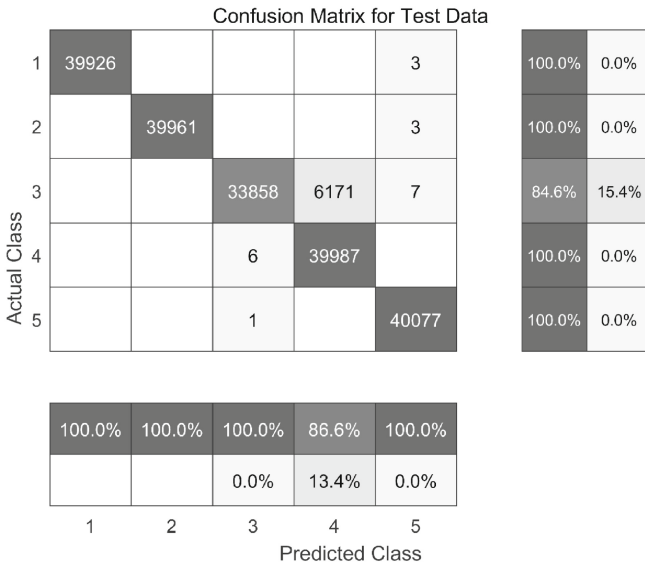
TM	PM	R*TV(Ω)	LSTM R*PV(Ω)	GRU R*PV(Ω)	LTV(10 ⁻⁴ H)	LSTM L*PV(10 ⁻⁴ H)	GRU L*PV(10 ⁻⁴ H)	C*TV(10 ⁻⁹ F)	LSTM C*PV(10 ⁻⁹ F)	GRU C*PV(10 ⁻⁹ F)
A	A	333.30	332.16	336.12	2.39	2.45	2.49	1.16	1.19	1.24
A	A	63.42	64.22	60.22	2.21	2.26	2.32	19.7	20.0	18.9
B	B	271.74	274.18	278.25	1.54	1.57	1.42	21.2	21.9	23.4
C	C	79.39	80.45	77.92	8.13	8.34	8.63	21.4	22.2	23.5
D	D	343.56	349.68	348.52	3.80	3.91	4.35	27.0	27.3	27.9
D	D	237.92	238.05	240.34	1.20	1.22	1.31	4.86	4.91	4.62
E	E	112.27	114.20	108.76	0.590	0.603	0.642	26.2	26.4	25.6

3.3 Analysis of Classification and Prediction Results for Equivalent Models of Transducers

When classifying the equivalent model of the transducer, the input is the amplitude and phase data of 10 different frequency points. The LSTM and GRU neural network models are used to classify and predict the equivalent model of the transducer. The results of classifying 200,000 test samples are shown in Fig. 4.



(a) LSTM classification prediction result Figure



(b) GRU classification prediction result Figure

Fig. 4. LSTM and GRU classification prediction result figure

From the classification prediction results, it can be seen that the GRU neural network model has a prediction accuracy of 96.9%, while the LSTM neural network has a prediction accuracy of 99.9%. The LSTM neural network model has a higher accuracy in classifying and predicting equivalent models such as heat exchangers.

3.4 Analysis of the Prediction Results for Equivalent Parameters of Transducers

To validate the performance of the equivalent parameter prediction model for transducers proposed in this article, three performance evaluation indicators were used, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination, where smaller values of RMSE and MAE indicate more accurate model predictions. The closer R^2 is to 1, the greater the degree of fit and the better the model prediction performance.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y'_i - y_i)^2} \tag{27}$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |y'_i - y_i| \tag{28}$$

$$R^2 = \frac{\sum_{i=1}^N (y'_i - y)^2}{\sum_{i=1}^N (y_i - y)^2} \tag{29}$$

In the formula, N represents the number of predicted samples, y'_i represents the predicted value of the component parameters, y_i represents the true value of the component parameters, and y represents the mean value of the true value.

When predicting the equivalent parameters of a transducer, the input consists of amplitude and phase data at 10 different frequency points. The evaluation metrics for predicting R, L, and C parameters on a test set of 10,000 samples using LSTM and GRU neural network models are as follows.

When using LSTM for parameter prediction, $R^2 = 0.99794$, $RMSE = 6.4\ 007$, and $MAE = 6.0696$; when predicting L, $R^2 = 0.99887$, $RMSE = 3.8862 \times 10^{-6}$, and $MAE = 3.543 \times 10^{-6}$; when predicting C, $R^2 = 0.99805$, $RMSE = 3.8291 \times 10^{-10}$, and $MAE = 3.5889 \times 10^{-10}$.

When using GRU for parameter prediction, $R^2 = 0.98352$, $RMSE = 10.5\ 863$, and $MAE = 9.3745$; when predicting L, $R^2 = 0.95452$, $RMSE = 7.7644 \times 10^{-6}$, and $MAE = 7.368 \times 10^{-6}$; when predicting C, $R^2 = 0.96562$, $RMSE = 7.0235 \times 10^{-10}$, and $MAE = 8.3241 \times 10^{-10}$.

Table 2 shows a comparison between some theoretical and predicted values of the transducer. In Table 2, “TV” represents theoretical values, “PV” represents

predicted values, “TM” represents theoretical models, and “PM” represents prediction models.

The partial data given in the evaluation index and Table 2 can prove that the prediction accuracy of the LSTM neural network model is higher than that of the GRU neural network model when predicting parameters for the heat exchanger.

To validate the parameter prediction model for the transducer proposed in this article, the results of model classification and parameter prediction for two different RLC devices are presented below. For the first device used in the experiment, the equivalent model A of the transducer was employed. The resistance value of the component was 200Ω , the capacitance value was 3nF , and the inductance value was $100\mu\text{H}$. The amplitude and phase of the signal passing through the device were measured using a signal acquisition system based on FPGA. Subsequently, the acquired amplitude and phase were used as test inputs for LSTM and GRU neural network models. Firstly, the equivalent model of device one was correctly identified as A, and then the parameters were predicted. The results are shown in Table 3.

Table 3. Predicted results of parameters for device one

Neural network	R(Ω)	L(μH)	C(nF)
LSTM	198.56	104	3.04
GRU	205.34	107	3.16

When using LSTM for parameter prediction, the prediction errors of R, L, and C are 0.72%, 4%, and 1.33%, respectively. When using GRU for parameter prediction, the prediction errors of R, L, and C are 2.67%, 7%, and 5.3%, respectively.

The second device is modeled as a transducer equivalent model D, with a resistance value of 400Ω , a capacitance value of 20nF , and an inductance value of $200\mu\text{H}$. The collected amplitude and phase are used as test inputs for LSTM and GRU neural network models. First, the equivalent model of device two is correctly identified as D, and then the parameters are predicted, as shown in Table 4.

Table 4. Predicted results of parameters for device two

Neural network	R(Ω)	L(μH)	C(nF)
LSTM	403.35	205	20.56
GRU	390.84	211	19.24

When using LSTM for parameter prediction, the prediction errors of R, L, and C are 0.84%, 2.5%, and 2.8%, respectively. When using GRU for parameter

prediction, the prediction errors of R, L, and C are 2.29%, 5.5%, and 3.8%, respectively.

By testing two RLC devices, the accuracy of the predictive model has been demonstrated. Accurate measurements of the amplitude and phase of the electrical signal after passing through the hydroacoustic transducer are crucial for the equivalent model and parameter prediction of the hydroacoustic transducer.

4 Conclusion

This article proposes a method for predicting the parameters of a hydrophone using an improved digital phase-sensitive detection algorithm combined with a LSTM neural network model. The method utilizes the piezoelectric effect of the hydrophone and the principle that amplitude and phase reflect the circuit characteristics of the hydrophone. By improving the digital phase-sensitive detection, simulated data can be used to train and test the neural network, and only amplitude and phase data from 10 frequency points are required to complete the prediction. The performance of LSTM and GRU neural networks in hydrophone equivalent model classification and parameter prediction was compared, and the results showed that the LSTM prediction model has higher prediction accuracy, stability, and generalization ability. This article provides a new tool for modeling equivalent parameters of hydrophones.

Acknowledgement. This research was funded by the Ningbo Natural Science Foundation(2022J138).

References

1. Andraka, R.: Hybrid floating point technique yields 1.2 gigasample per second 32 to 2048 point floating point FFT in a single FPGA (2022)
2. Belagoune, S., Bali, N., Bakdi, A., Baadji, B., Atif, K.: Deep learning through LSTM classification and regression for transmission line fault detection, diagnosis and location in large-scale multi-machine power systems. *Measurement*, **3**, 109330 (2021)
3. Rana, R.: Gated recurrent unit (GRU) for emotion classification from noisy speech. arXiv preprint [arXiv:1612.07778](https://arxiv.org/abs/1612.07778) (2016)
4. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint [arXiv:1412.3555](https://arxiv.org/abs/1412.3555) (2014)
5. Dey, R., Salem, F.M.: Gate-variants of gated recurrent unit (GRU) neural networks. In: 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), pp. 1597–1600. IEEE (2017)
6. Cai, Z., Duan, L.M., Wu, C.: Phase-sensitive detection for unconventional Bose-Einstein condensation. *Phys. Rev. A* **86**(5), 1–5 (2012)
7. Chen, Y., Zhou, L., Guo, X., He, T., Zhang, J.: Modelling, measurement and optimization of self-noise of hydrophone with preamplifier. *MATEC Web Conf.* **283**, 05004 (2019)
8. Xiao Er-Liang, N.I.: Zhen-Zhen, and Zhai Wan-Li. Research of FFT cross-correlation algorithm in boiler flame dual detection. *Inf. Technol.* (2013)

9. Graves, A., Jaitly, N., Mohamed, A.R.: Hybrid speech recognition with deep bidirectional LSTM. In: Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on (2013)
10. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. *Comput. Sci.* (2015)
11. de Moura, B.F., da Mata, A.M., Martins, M.F., Palma, F.H., Ramos, R.: Implementation of a phase-sensitive detector with CORDIC algorithm in microcontrollers for low-cost EIT demodulation procedure. In: 6th Multiphase Flow Journeys (2021)
12. Sharma, S., Sen, S.: Real-time structural damage assessment using LSTM networks: regression and classification approaches. *Neural Comput. Appl.* **35**(1), 557–572 (2023)
13. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. *Adv. Neural Inf. Process. Syst.* (2014)
14. Li, H., Deng, Z.D., Carlson, T.J.: Piezoelectric materials used in underwater acoustic transducers. *Sens. Lett.* **10**(3–4), 679–697 (2012)



Unraveling the Elements of Effective Altruistic Appeals Through Machine Learning and Natural Language Processing

Sourav Yadav¹(✉), Sankalp Arora², Akash Kumar³, and Kaveri Verma⁴

¹ Illinois Institute of Technology, Chicago, USA
syadav12@hawk.iit.edu

² Vellore Institute of Technology, Vellore, India
sankalparora5@gmail.com

³ SRM Institute of Science and Technology, Chennai, India
akash0697@gmail.com

⁴ Manipal University, Jaipur, India
vkaveri10@gmail.com

Abstract. In today's world, online platforms such as social media, philanthropic communities, and Q&A websites provide opportunities for people to be altruistic by donating money or answering questions without expecting anything in return. The r/Random Acts Of Pizza subreddit on Reddit is one such online community where users can post requests for free pizza while explaining their current situation, and the outcome of each request is either successful or unsuccessful. This study seeks to explore the determinants that impact the outcome of such selfless appeals. To achieve this, we propose a new model architecture that combines two models, one that deals with sparse text vectors and the other that analyzes dense features from previous works to predict the success of a request. The study reveals that the probability estimated from the first model and the number of comments on the request post is crucial in predicting the outcome of a request.

Keywords: Natural language processing · Machine learning · Feature engineering

1 Introduction

The amount of information shared in the form of text ever since the dawn of the internet has been appalling. In contemporary times, written information is a pivotal, if not the paramount, reservoir of knowledge. Efforts to get information out of text (or speech) have been underway since the 1950s when in the Georgetown-IBM Experiment, scientists successfully translated 60 Russian sentences into English and made some progress toward machine translation [1].

Until the 1980s, the research work under Natural Language Processing was based on sets of hand-written rules, which became obsolete after the introduction of machine learning algorithms to textual data. The beginning of statistical NLP gave way to the advancement of many different subfields like machine translation, text classification, speech recognition, etc. Further, the demand for advanced maneuvers in many natural language processing applications gave rise to what is today known as Neural Natural Language Processing or Neural NLP [2]. While utilizing deep learning methods, Neural NLP has become increasingly essential for sectors like healthcare and medicine.

Recently, a wave of use cases in Computational Social Science has surfaced, requiring state-of-the-art NLP tools to solve them [3]. Be it determining biased news, identifying emotion behind tweets, or assessing the persuasiveness of an argument in a debate, the use of NLP tools for social causes has clearly increased. With the popularity of social media and Natural Language Processing in a social setting, the rise of philanthropic communities on the internet is notable. Websites like Kickstarter, Ketto, Reddit, and Kiva have made their presence immense and facilitated many people worldwide to get out of unpleasant situations via crowdfunding. For example, the Indian online crowdfunding platform, Ketto, has been responsible for raising over Rs. 2000 crores across 3.2 Lakh online fundraisers. This new era of online fundraising has undoubtedly helped people around the world and will certainly continue to do so. Our project focuses on a similar community forum, under the Reddit website, known as r/Random Acts Of Pizza [4]. The forum was created in the year 2010 with the aim of getting pizza to those who need it. Requesters could explain their situation in a post, giving information about why they need a pizza and how it would help them. It would then depend on the other users to donate a pizza to the requester.

Our study relied on information mentioned on the CS Stanford website, and the foundation for this project has been the work done by Althoff et al. [4]. The dataset contains 5671 requests collected from the Reddit community between December 8, 2010, and September 29, 2013. There are 33 features present in the dataset, including a target variable containing information about whether the requester received the pizza or not. The project proposes an architecture of a two-stack binary classification model to predict the success of an altruistic request. The first model deals with the textual data by utilizing textual feature extraction techniques to convert the raw text data into vector form and then classify the vector data against the target variable. The finished model is then used to estimate the probabilities of the request being successful by the text. This probability estimate is then used as a dense feature among other features, like account age of the requester, the number of likes and dislikes, etc., to predict request outcomes. We have then used the final model to determine the importance of each feature and how they contribute towards predicting the target variable.

2 Literature Review

In recent years, along with the increase in the research work under NLP in a social setting, there have been multiple attempts to decode linguistic factors in textual data and how they affect the success of requests. With the objective of interpreting what makes a favor successful, Althoff et al. [5] discuss high-level social factors, like the status of the requester and the similarity between the requester and the donor, textual factors, like evidence of need and the sentiment of text, and, finally, temporal factors like the age of the account and the time the request was made. A logistic regression model then binds these factors together to classify whether the request was successful or not. The model also allows the researchers to reason out the significance of independent variables in predicting the dependent variable. Similar to the base paper, Jacek and Sabrina et al. [6] critically evaluate different algorithms over the RAOP dataset. In addition to the social and linguistic factors, such as the requester's identity and how the requester is asking, the researchers also aim to extend the work by investigating the sentimental factors satisfying a donor.

Hsieh et al. [7] extend the base paper by introducing additional features like Topic, Role, and Centrality. By employing Bag of Words & N-Gram, along with considering the requester's interaction with other users, these characteristics effectively addressed the underlying subtext of the text. Furthermore, a groundbreaking model called the Graph-based Predictor for Request Success was utilized in this context. Utilizing a Request Graph and a Propagation-based Optimization algorithm, this approach captures the interrelatedness of features among requests within the dataset. By learning feature weights, it calculates the probabilities of requests being successful or unsuccessful. Specifically, requests with higher similarity scores are assigned the same label, further enhancing the accuracy of the model. Ahmed et al. [8] discuss the same method, focusing more on the Topic and Role as additional features.

Durmus and Cardie et al. 's [9] work is based on debate.org, which provides a platform for people to express their opinions and beliefs. It takes on the task of judging which debater will be more successful. This research paper introduces a novel dataset to explore the impact of linguistic usage versus pre-existing beliefs on persuasive communication. It also proposes a controlled framework that incorporates the political and religious ideologies of the readers, thereby providing a structured environment for analysis. It inculcates user-based features like ideology and opinion similarity and linguistic features like text length, sentiment, and features gathered using TF-IDF. Similarly, Yang et al. [10,11] also revolved around constructing a framework for modeling persuasive tactics. This involved the utilization of a semi-supervised hierarchical neural network that incorporates attention mechanisms at both the word-level and sentence-level. The aim was to assess the degree of persuasiveness achieved quantitatively.

3 Proposed architecture and steps

This project utilizes the crucial features gathered from the pre-existing dataset as well as the previous works in the literature review while framing our architecture consisting of a stack of two models to deal with the textual and numerical data separately. In this section, we will be stating the architecture developed and the process followed in creating the two models.

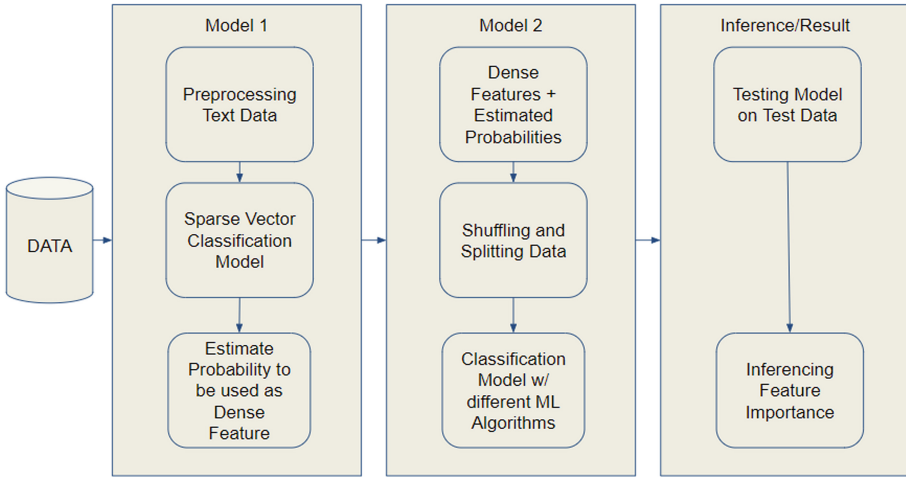


Fig. 1. Model architecture

3.1 Data Exploration

Data exploration allows for a better understanding of the dataset, making it easier for us to navigate the data better and select methods suitable for that data. After a simple exploration of the RAOP dataset, we notice that there are 5671 requests with an average success rate of 24.63%, making it an imbalanced dataset. This will allow us to choose a Stratified K-Fold instead of the standard K-Fold Cross Validation technique while training our models. Stratified K-Fold Cross Validation partitions the dataset so that the validation set will have an equal number of instances of the dependent variable [12]. This ensures that no instance will overpower the other in the validation set.

3.2 Text Preprocessing

Lowercasing, Removal of Punctuation & Stop-words are a few of the most common preprocessing steps that allow us to reduce the redundancy between words.

Similarly, Punctuation in a text can sometimes create difficulty in comprehending the text or extracting tokens. Therefore, it becomes necessary in some use cases to remove the punctuation in the text. Stop-words are widely used irrespective of the language and are usually removed from the text as they lack constructiveness to the text. Therefore, our study employed NLTK [13] list of English stop-words in our project to remove them from the request.

Regular Expression [14] has proven powerful in searching and manipulating text strings and has been an essential tool in text analytics. We can use it to find patterns in the text that we want to remove or edit. For example, in the project, we used the "re" python library to find and expand the contractions such as "don't" and "can't" "do not" and "can not".

Tokenization is the fundamental task of the NLP preprocessing pipeline in which we break a text sample into tinier components. These units might be words, subwords, or characters. We have used the "regex tokenize" function from the NLTK [13] library to tokenize the cleaned text.

Lemmatization is the most crucial text preprocessing step that stems the word while ensuring it does not lose its meaning. The project uses the Wordnet Lemmatizer present inside the NLTK [13] package. Wordnet [15] is a large lexical database that aims to establish structured semantic relationships between words. We utilize the NLTK [13] interface of this database using the function `lemmatize()` of the instance `WordNetLemmatizer()`.

3.3 Feature Engineernig

We split the numeric and textual data and separately worked on it for the two different models.

Numeric Data: Based on the previous literature and knowledge, we manipulated the dataset in order to get the features. We use two features to understand how any posted request is doing in the community by subtracting the number of downvotes from the number of upvotes at retrieval as well as getting the number of comments at retrieval. We also add the length of the text as a feature to check if longer texts lead to the request being successful or not. An "Evidence" feature is calculated by setting a value of 1 if there is a presence of any image link and 0 if there is no link present [5]. For other features, we consider the account age of the requester and whether the requester has posted before in the RAOP subreddit. These features give us some information about the requester.

Text Data: We combine the title of the request and the main text of the request to get the final text which can then be converted into vectors using One-Hot Encoding [16], TF-IDF [17], Word2Vec [18], and Doc2Vec [19] techniques.

One-Hot Encoding or Count Vectorizing is the method of representing words by creating a vector having as many dimensions as the unique words in the vocabulary. Inside the vector, if the text data features that word, we put a "1" in that dimension; otherwise, we put "0". This gives us a huge sparse vector representation of all the text requests in the data. However, one of the disadvantages of

the Count Vectorizing technique is its inability to represent the semantic and statistical relationship between the data.

TF-IDF vectors feature statistical representation of text regarding the word's importance or relevance in a document. This representation is constructive in information retrieval systems as it allows the system to look for more relevant words in the search query. TF-IDF [17] measure is the product of Term Frequency (TF) [20] and Inverse Document Frequency (IDF) [21], where TF looks at how many instances a word is repeated, while IDF indicates the relevance of the word. Therefore, it gives weightage to each word based on its frequency in the corpus.

Word2Vec frequently used embedding techniques which aim at learning embeddings for every word in the corpus [22]. Word2Vec helps in capturing the semantic representation of the text. It uses neural networks to derive the word's meaning from its context. Therefore, Word2Vec projects the meaning of the word in a high-dimensional vector space and clusters the word with similar meanings together. We used the pre-trained Word2Vec model over the Google News Corpus for this project [18]. In order to find the sentence vector by using individual word vector representations, we calculate the average of the vectors of words in the text. Another method that has given good results in the past is calculating the minimum and maximum from the list of vectors of words in the text and concatenating it to form a more extensive vector representation [23].

Doc2Vec [19] is very similar to the Word2Vec [18] model in terms of its general architecture. In addition to learning the word vectors, it also learns the paragraph vector associated with the full text. There are two architectures of the Doc2Vec model but this project uses the Distributed Memory (DM) architecture of the Doc2Vec Model packaged under the Gensim library.

3.4 Modelling

This section briefly describes the stack of two models present in our classification architecture. The architecture consists of two models, one that deals with the high-dimensional sparse vector converted from the textual data and the other model that deals with the dense features pre-existing in the dataset as well as the ones engineered by previous literature.

Text Model: The request title and the request text are combined with the target variable into a separate data frame. This data frame will be helpful in converting each row of the text data into a vector using different text embedding techniques mentioned in Sect. 3.3. The resulting vectors from the different embedding techniques are then fed into different machine learning algorithms like Logistic Regression [24], Gaussian Naive Bayes [25], and Random Forest [26]. (Note that we do not use K Nearest Neighbors [27] or Support Vector Machines [28] as they do not give a probability estimate needed for our second model.) Based on the accuracy of the different models created by the combinations of different text embedding techniques and machine learning algorithms, we select

the models to estimate the probability score for the success of the request and feed it to the final model.

Final Model: The final model consists of the numeric data present in the dataset, the derived features based on previous works, and the probability estimates of the text from the previously mentioned text models performing well. We use Logistic Regression [24], Gaussian Naive Bayes [25], and Random Forest [26] algorithms to create a classification model using the target variable of receiving a pizza as the dependent variable and the other dense features as the independent variable. Our model handled unseen data without overfitting issues because we implemented a Stratified K-Fold cross-validation method [12]. Our findings with regard to the following models are given under Sect. 4 with Accuracy [29], Precision [29], Recall [29], and F1-Score [29] as the metrics.

3.5 Parameter Optimization

Parameter Optimization or tuning is the name given to the process of selecting parameter values for a learning algorithm that leads to better performance of the algorithm. In this project, we decided to tune the hyperparameters of the Random Forest [26] algorithm, as there are a lot of parametric values that can affect the algorithm's performance. However, the parameters are not critical enough to need tuning for other algorithms, like Logistic Regression [24] and Naive Bayes [25]. This paper uses Randomized Search [30] to optimize hyperparameters by randomly selecting points assigned inside a domain. Following this, Grid Search [30] is applied to fine-tune the hyperparameters found by forming a grid of values based on the best values provided by the Randomized Search [30].

3.6 Feature Importance

Feature Importance is the technique of calculating a score for all the input features used in the model. This score refers to the importance of the variable in predicting the target variable. We use the built-in Feature Importance technique in the Random Forest algorithm implemented in scikit-learn using Gini Importance [31]. Gini Importance [31] averages the decrease in node impurity for each feature over all trees in the ensemble to get the feature importance. We also use the SHAP [32] interpretation to find the feature importance from the different models. In order to get this estimate of feature contribution to the prediction, it uses Shapley [32] from game theory which helps us fairly distribute the contribution among the features. We apply both methods to all the final models with the self-created dense features as well as the probability estimates from the text models.

4 Results and Inference

To measure the performance and understand the feature importance in the parts mentioned before, we decided to divide the dataset of 5671 into 4537 training rows and 1134 testing rows. This helps expose the classifier to training on the given data and measure its performance over the unseen data.

4.1 Model Results

For the performance of the text models, we can infer from the above plot, developed using matplotlib, that out of the three machine learning algorithms used, Random Forest [26] performs equally well for each text embedding technique, while the performance under Logistic Regression [24] improves with the use of a more complex and advanced text embedding technique. However, Gaussian Naive Bayes [25] Classifier fails to achieve a reasonable threshold with any embedding techniques. Therefore, based on the inference from the graph, we choose the Random Forest version of all the embeddings. Now to measure the performance of the final model, the test set is separated from the primary data is first converted into vectors using the different embedding methods and then fed into the final model along with other numeric data. We make use of the metrics like Accuracy, Precision, Recall, and F-Score [29] in order to understand how the final model is performing on the test set.

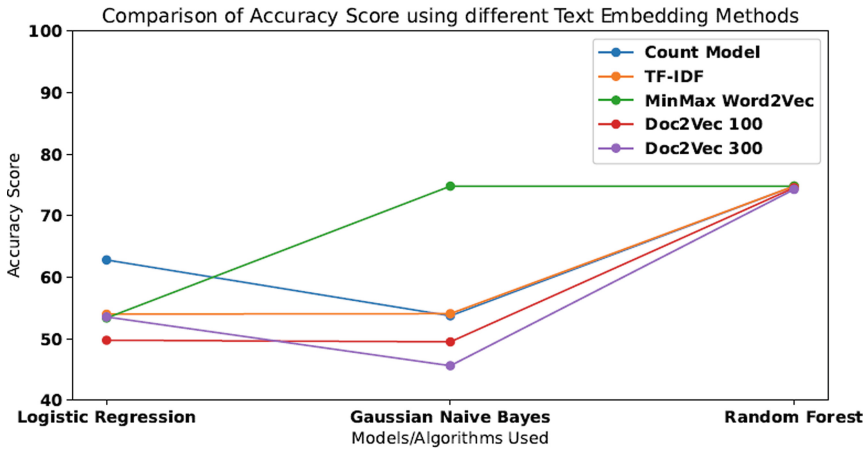


Fig. 2. Comparison of accuracy score using different text embedding methods

Figure 3. displays metrics of different machine learning algorithms applied over the original numeric data combined with the estimated probability received from the Random Forest versions of the embedded text data. We see that the model’s accuracy remains highest at around 75 when we use Random Forest to

model the final data regardless of the embedding method used. However, for our use case, we are predicting whether a user will receive a pizza or not, which is a rare outcome with a chance of 24.63. From the graph, we infer that both the Doc2Vec embeddings, i.e., 100- and 300- dimensions, fed into a logistic model perform much better than other models in predicting the requesters that will receive a pizza. However, we also notice that the same models are less precise in predicting who actually receives the pizza and who does not.

On the other hand, the Logistic Model of the final data having Count and TFIDF embeddings have a comparatively lower but appropriate recall score as well as a good precision score. Therefore, we attempt to get a balance between recall and precision by measuring the F-Score of the models. We see that, on average, all the models perform relatively close to each other in terms of F-score, with the Logistic Regression version of Min-Max Word2Vec & TF-IDF and Random Forest version of 300-dimension of Doc2Vec & Min-Max Word2Vec have high F-scores. This suggests that the mentioned models maintain a good balance between precision and recall while predicting the outcome of a requester receiving a pizza or not.

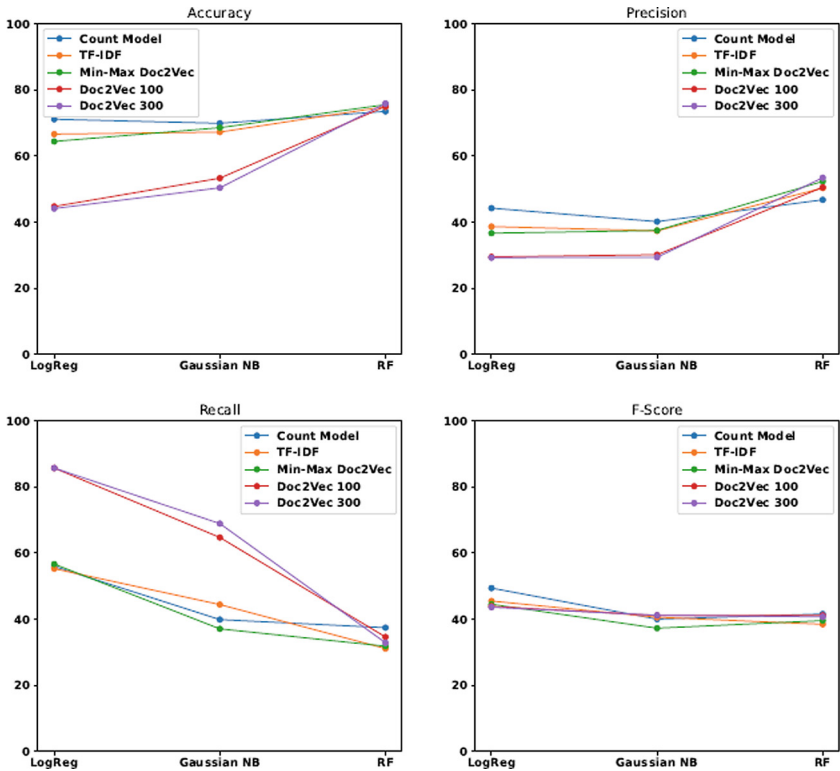


Fig. 3. Performance of the final model in terms of accuracy, precision, recall and F-score

4.2 Model Inference

SHAP [32] values from the classifiers provide the mean SHAP [32] values of each feature utilized to predict the target variable. This mean-SHAP [32] value refers to the impact of a feature on the model. Therefore, we see that the probability estimate generated from the text has a significant role in predicting whether a requester will receive their pizza or not. The number of comments on the request post, the upvotes minus downvotes, and the length of the text also have some part to play in predicting the outcome. However, from the models, we see that it hardly matters whether the requester has posted before in the subreddit or has attached an evidenced link in the text as their SHAP [32] values are very low.

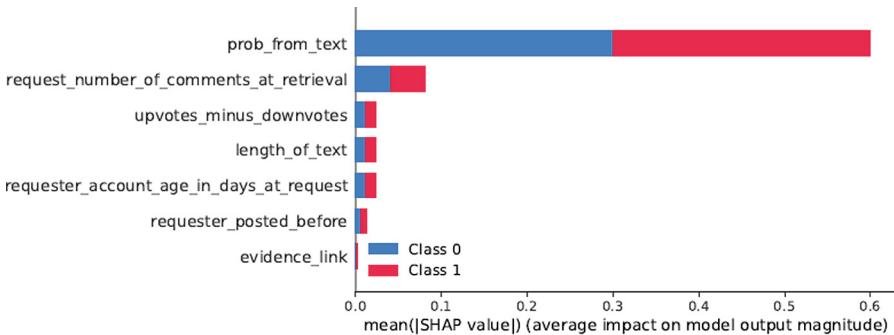


Fig. 4. Ranking of feature importance in terms of SHAP values

5 Conclusion

This work presents a stack of two models that separately deal with the text and numeric data. This allows us to reduce higher dimensional vectors generated from the text data to a probability estimate and retain the importance of the features present in the original dataset. The experimental findings validate the efficacy of our approach in terms of the performance metrics discussed in the Sect. 4. Therefore, given a requester posting a request, our model can predict whether it will stimulate any altruistic behavior from the other Redditors. Inferring the model also gives us an understanding of how useful other variables in the dataset were in predicting the outcome. In order to improve our work in the future, we can work around framing and combining different embedding techniques to get a vector representation of text, leading to better probability estimates and more satisfactory performance. Finally, this work can also be extended by using more sophisticated machine learning and deep learning models, as well as adding more data in order to develop a more robust model.

References

1. Hutchins, W.J.: The Georgetown-IBM experiment demonstrated in January 1954. In: Frederking, R.E., Taylor, K.B. (eds) *Machine Translation: From Real Users to Research*. AMTA 2004. Lecture Notes in Computer Science, vol. 3265. Springer, Berlin, Heidelberg (2004)
2. Young, T., Hazarika, D., Poria, S., Wang, Z.: Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **13**(3), 55–75 (2018)
3. Bamman, D., Doğruöz, A. S., Eisenstein, J., Hovy, D., Jurgens, D., O'Connor, B., Oh, A., Tsur, O., Volkova, S.: Proceedings of the first workshop on nlp and computational social science. In: *Workshop on Natural Language Processing and Computational Social Science (EMNLP 2016)*. Association for Computational Linguistics (ACL) (2016)
4. Althoff, T., Salehi, N., Nguyen, T.: Random acts of pizza: success factors of online requests (2013)
5. Althoff, T., Salehi, N., Nguyen, T.: Random acts of pizza: success factors of online requests (2013)
6. Filipczuk, J., Pesce, E., Senatore, S.: Sentiment detection for predicting altruistic behaviors in Social Web: a case study. In: *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* pp. 004377–004382. IEEE (2016)
7. Hsieh, HP., Yan, R., Li, CT.: Will I win your favor? Predicting the success of altruistic requests. In: Bailey, J., Khan, L., Washio, T., Dobbie, G., Huang, J., Wang, R. (eds) *Advances in Knowledge Discovery and Data Mining. PAKDD 2016*. Lecture Notes in Computer Science, vol. 9651. Springer, Cham (2016)
8. Ahmad, A., Ahmad, T., Bhatt, A.: A novel approach for predicting the outcome of request in RAOP dataset. In: Jain, L., Balas, E.V., Johri, P. (eds.) *Data and Communication Networks. Advances in Intelligent Systems and Computing*, vol. 847. Springer, Singapore (2019)
9. Durmus, E., Cardie, C.: Exploring the role of prior beliefs for argument persuasion. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035-1045, New Orleans, Louisiana. Association for Computational Linguistics (2018)
10. Yang, D., Chen, J., Yang, Z., Jurafsky, D., Hovy, E.: Let's make your request more persuasive: modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3620-3630, Minneapolis, Minnesota. Association for Computational Linguistics (2019)
11. Chen, J., Yang, D.: Weakly-supervised hierarchical models for predicting persuasive strategies in good-faith textual requests. *Proc. AAAI Conf. Artif. Intell.* **35**(14), 12648–12656 (2021)
12. Raschka, S.: Model evaluation, model selection, and algorithm selection in machine learning. *arXiv (Cornell University)* (2018)
13. Loper, E., Bird, S.: NLTK: The Natural Language Toolkit. *arXiv (Cornell University)* (2002)
14. Erwig, M., Gopinath, R.: Explanations for regular expressions. In: de Lara, J., Zisman, A. (eds.) *Fundamental Approaches to Software Engineering. FASE 2012*. Lecture Notes in Computer Science, vol. 7212. Springer, Berlin, Heidelberg (2012)

15. Kilgarriff, A., Fellbaum, C.: WordNet: an electronic lexical database. *Language* **76**(3), 706 (2000)
16. Harris, D.R., Harris, S.: Digital design and computer architecture. In: Elsevier eBooks (2007)
17. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval (2008)
18. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint [arXiv:1301.3781](https://arxiv.org/abs/1301.3781) (2013)
19. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning, pp. 1188–1196 PMLR (2014)
20. Luhn, H.P.: A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.* **1**(4), 309–317 (1957)
21. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. *J. Doc.* **28**(1), 11–21 (1972)
22. Vajjala, S., Majumder, B., Gupta, A., Surana, H.: Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems. O'Reilly Media (2020)
23. De Boom, C., Van Canneyt, S., Demeester, T., Dhoedt, B.: Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recogn. Lett.* **80**, 150–156 (2016)
24. Cramer, J.: The origins of logistic Regression. Social Science Research Network (2003)
25. Rish, I.: An empirical study of the naive Bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, pp. 41–46 (2001)
26. Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001)
27. Cunningham, P., Delany, S.J.: k-nearest neighbour classifiers—a tutorial. *ACM Comput. Surv.* **54**(6), 1–25 (2021)
28. Evgeniou, T., Pontil, M.: Support vector machines: theory and applications. In: Advanced Course on Artificial Intelligence, pp. 249–257. Springer Berlin Heidelberg, Berlin, Heidelberg (1999)
29. Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: European Conference on Information Retrieval, pp. 345–359. Springer Berlin Heidelberg, Berlin, Heidelberg (2005)
30. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**(2), 281–305 (2012)
31. Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., Hamprecht, F.A.: A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinf.* **10**(1) (2009)
32. Lundberg, S., Lee, S.: A unified approach to interpreting model predictions. *Neural Inf. Process. Syst.* **30**, 4768–4777 (2017)



A Novel Feature Selection Method Based on Slime Mold Network Formation Behavior

Chenyang Yan^(✉)

Ningbo City College of Vocational Technology, Ningbo 310000, Zhejiang, China
yanchenyang@nbcc.edu.cn

Abstract. Slime mold (*Physarum polycephalum*) is a remarkable organism that can solve complex problems such as mazes, shortest path networks, and optimal transport networks. Inspired by the adaptive network formation behavior of slime mold, we propose a novel feature selection algorithm (SMFS). The SMFS converts the feature selection into an optimal subgraph problem and employs a slime mold network formation inspired strategy to guide the sub-graph search procedure. We evaluate the performance of SMFS against 4 well-known meta-heuristic feature selection methods using different classifiers (support vector machine and naive Bayes). The experimental results on several benchmark datasets demonstrate the efficiency and effectiveness of the SMFS method and its superiority over previous related methods.

Keywords: Feature selection · Slime mold · Adaptive networks

1 Introduction

1.1 A Subsection Sample

Some non-neuronal organisms, such as plants, bacteria, fungi, or protists, have been shown to make complex decisions under challenging environments, despite having no intelligence [1].

We are particularly fascinated by the work of Tero and Nakagaki [2], who showed the remarkable ability of *Physarum* to design adaptive networks. They set up an experiment with a plate containing 36 food sources that corresponded to the geographic locations of cities in the Tokyo area. They simulated the geographic features by varying the intensity of light, since *Physarum* tends to avoid bright light. They placed a small plasmodium of *Physarum* at the location of Tokyo and observed its behavior. Initially, the *Physarum* occupied most of the available space on the plate, but then it refined its network by thinning out some connections and leaving behind a subset network. This network not only resembled the Tokyo rail system visually, but also had similar efficiency, fault tolerance, and cost as the networks designed by human engineers.

Inspired by Tero and Nakagaki's work, we transform the feature selection problem into an optimal subgraph problem and construct an algorithm based on the rules of the slime mold (*P. polycephalum*) to solve the problem. In a nutshell, the candidate features

are considered vertices in a complete undirected graph. These vertices are also the virtual source. The edges of the graph are the initial tubes. The virtual slime plasmodia flows and forages through these tubes according to specific rules. In the end, the morphology of the slime tubes will converge to form a complete subgraph. All vertices on the subgraph are the selected features (Fig. 1). We call this model SMFS, a Slime Mold inspired Feature Selection algorithm.

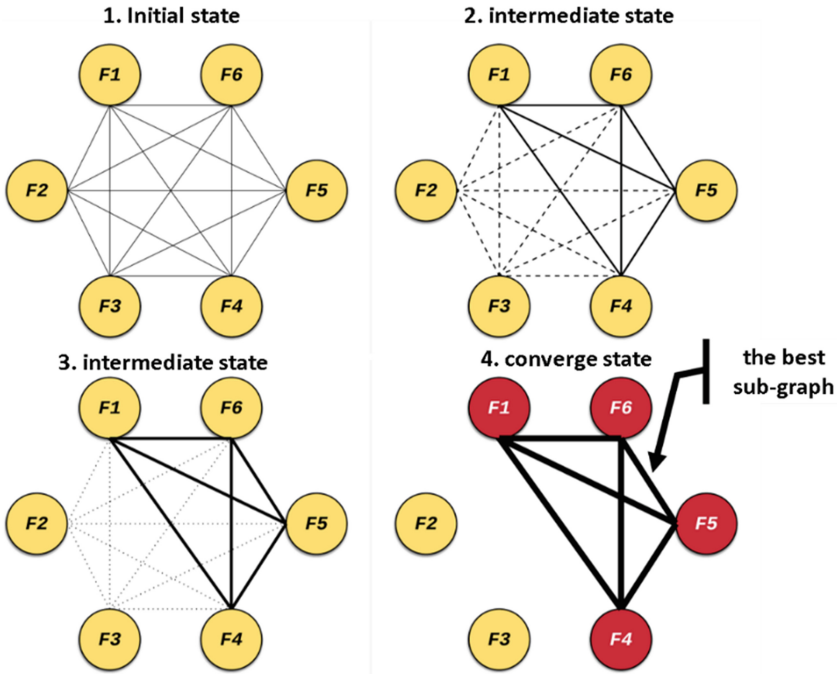


Fig. 1. Imitate the network formation behavior of slime mold to solve the feature selection problem. Each vertex represents a candidate feature, which is also where the food is. The set of all vertices is the candidate feature set. The virtual slime mold forages for food along the edges of the graph. If the network formation rules are designed reasonably, it can be expected that the slime mold will eventually converge into a subgraph, and all the vertices on the subgraph are the selected features.

The paper is structured as follows. Section 2 describes the implementation SMFS. Section 3 presents the experiments and analyzes the results. Section 4 concludes the paper and suggests future directions.

2 The SMFS Method

2.1 Problem Definition

We represent each text as a vector of length N , where N is the number of words (features) in the corpus. Each element in the vector corresponds to a word (feature), and its value is the L2-regularized tf-idf score [3] of the word if it appears in the text, or 0 otherwise.

The sample input E consists of M text vectors and their sentiment polarity tags, so E is an $M \times (N + 1)$ matrix. The word set S of E is the candidate feature set. Most texts only use a small subset of the words in S . Therefore, E is likely to be a sparse matrix (with more than 99% of the elements being 0). The feature set may contain many irrelevant features that affect the classification performance. The feature selection method needs to remove these irrelevant features.

We can formulate the feature selection problem as follows: given the candidate feature set S , find a subset S_{best}^{sub} that maximizes the classification performance of classifier C on sample dataset E based on S_{best}^{sub} .

2.2 Framework

The SMFS uses a feature graph G , where each vertex $V_i (0 < i < |S|)$ represents a feature in S . G is a complete graph. The slime mold can flow for food along the edges $E(G)$. Each vertex V_i has a reward weight R_i , indicating the current short-term reward for the slime mold to flow through the vertex. Each edge E_{ij} has a weight D_{ij} , called conductivity, indicating how easily the slime mold can flow from V_i to V_j .

The SMFS simulates the slime mold flow on G according to specific rules that combine the reward weights and the conductivities, to obtain the maximum reward. The variable Q_{ij} represents the protoplasmic flux on the edge E_{ij} .

When the flow process converges, the slime mold on each edge forms a complete subgraph G^{sub} . The set of all vertices in the subgraph $V(G^{sub})$ represents the features selected by the current plasmodium.

The SMFS algorithm also uses a slime colony with memory [4] to forage in parallel. There are n plasmodia in the colony. The colony has a collective memory and can remember the subgraph G_{ib}^{sub} that has achieved the best reward so far. $f(Q_{ij})$ is the flux increment of slime mold on all edges E_{ij} in G_{ib}^{sub} . $f(Q_{ij})$ is proportional to the optimal reward $Rwd(G_{ib}^{sub})$.

The flux on the edges not in G_{ib}^{sub} will decay. Therefore, as the algorithm iterates, some edges of G will lose their flux and disappear. Eventually, all the fluxes on the graph will converge into a subgraph G_{gb}^{sub} of G . The set $V(G_{gb}^{sub})$ of all vertices in the subgraph is the solution of the algorithm.

The SMFS follows the framework outlined in Table 1. Each iteration of the REPEAT-UNTIL loop consists of three key procedures: initialization, flow and renew. We will discuss them in detail in the next subsections.

2.3 Initialization

This refers to the initialization of the conductivities on each edge, i.e., D . The SMFS uses random initialization, i.e., it generates a random subset of vertices (features) S_{rnd}^{sub} and sets the conductivities on all edges to $D_0 = Rwd(S_{rnd}^{sub})$, which is the classification performance of the classifier trained on the dataset E based on S_{rnd}^{sub} . Since the SMFS solves a binary problem, we use the harmonic mean of precision and recall as the performance measure.

Table 1. SMFS framework.

 SMFS framework.

procedure SMFS (Feature_Set(**S**), Classifier(**C**), Examples(**E**), mold_size)

D := initial (**S**, **C**, **E**)

i := 0, $G_{gb}^{sub} := 0$
repeat
j := 0

while(*j* ++ < *n*)

 $G_j^{sub} := \mathbf{flow}(\mathbf{V}, \mathbf{D})$
end while
 $G_{ib}^{sub} := \operatorname{argmax}_{0 < j < n} (Rwd(G_j^{sub}))$
 $G_{gb}^{sub} := \operatorname{argmax} (Rwd(G_{ib}^{sub}), Rwd(G_{gb}^{sub}))$
D := renew (**D**, G_{gb}^{sub})

until (*i* ++ > max_iteration or converge_criteria)

Return G_{gb}^{sub}
end procedure

2.4 Flow Procedure

To simplify the SMFS, we introduce a virtual vertex V_0 , which represents the starting point of plasmodia network formation. The procedure $\mathbf{flow}(\mathbf{V}, \mathbf{D})$ consists of the following steps: Let $\mathcal{P}^k(t)$ be the set of vertices that plasmodium k has flowed through at time t , and $\mathcal{I}^k(t) = S - \mathcal{P}^k(t)$ be the set of remaining vertices. Let D_{uv} be the conductivity from vertex V_u to vertex V_v . Then the SMFS selects the next vertex V_w with probability P according to Eq. (1):

$$V_w = \operatorname{argmax}_{v \in \mathcal{I}^k(t)} \left(R_v(\mathcal{P}^k(t)) \cdot \sum_{u \in \mathcal{P}^k(t)} D_{uv}(t) \right) \quad (1)$$

Alternatively, with probability $1-P$, the next vertex V_w is selected probabilistically according to Eq. (2), where m is a parameter that specifies the maximum number of vertices that plasmodium can flow through. If $m = |S|$, then it means there is no limit on the number of vertices (features) to be selected.

$$P_w^k(t) = \begin{cases} \frac{\left(\sum_{u \in \mathcal{P}^k(t)} D_{uw}(t) \right)^\alpha \cdot R_v(\mathcal{P}^k(t))}{\sum_{u \in \mathcal{P}^k(t)} \left[\left(\sum_{v \in \mathcal{I}^k(t)} [D_{uv}(t)]^\alpha \cdot R_v(\mathcal{P}^k(t)) \right) \right]} & \text{if } w \in \mathcal{I}^k(t) \wedge |\mathcal{P}^k(t)| < m \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

A natural way to define the reward $R_v(\mathcal{P}^k(t))$ of feature V_v is to use its L2 regularized tf-idf score. SMFS computes all tf-idf scores in preprocessing.

SMFS also introduces a more sophisticated and accurate feature reward measure: the feature importance of V_v , which is the improvement in prediction performance obtained

by adding feature V_v to the feature set $\mathcal{P}^k(t)$. Due to its high computational complexity, it is not feasible to compute all features' importance in preprocessing. Therefore, SMFS calculates the feature importance on-the-fly and caches it for later use.

The conductivity $D_{uv}(t)$ will be discussed in the next section. The parameter α in Eq. (2) is a weight that adjusts the relative importance of conductivity and reward.

The flow procedure of plasmodium k terminates if the reward of the features selected by it drops below the threshold ϵ for ten consecutive times, or if all edges connected to each candidate vertex have zero conductivity.

2.5 Renew Procedure

Following the model proposed by Tero [5], the conductivity and flux between vertex V_i and vertex V_j are related by Eq. (3), where $f(Q_{ij})$ is an increasing function with $f(0) = 0$, and r is a parameter that controls the decay rate, $0 < r < 1$.

$$\frac{dD_{ij}}{dt} = f(Q_{ij}) - rD_{ij} \quad (3)$$

We adopt this assumption in our model. Since our model is discrete, we obtain Eq. (4):

$$D_{ij}(t) = f(Q_{ij}(t-1)) - D_{ij}(t-1)e^{-r} \quad (4)$$

The function $f(Q_{ij})$ is defined as in Eq. (5):

$$f(Q_{ij}(t)) = \begin{cases} \lambda \cdot Rwd(V(G_{gb}^{sub})) + \frac{\mu \cdot (m - |V(G_{gb}^{sub})|)}{m} & \text{if } i \in G_{gb}^{sub} \wedge j \in G_{gb}^{sub} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

The first line of Eq. (5) has two terms: the first term represents the classification performance, where λ is a weight ($0 < \lambda < 1$), and $Rwd(S)$ is the training result obtained by the classifier on data set E with feature set S .

The second term reflects the SMFS's preference for simple solutions, where μ is a weight that indicates the preference for concise solutions, $\mu = 1 - \lambda$. $|V(G_{gb}^{sub})|$ denotes the size of the feature set obtained by the current plasmodium.

3 Experiments

3.1 Dataset

The original corpus consists of user reviews from seven online open courses. SMFS cleans the original corpus using tools like HTML Parser and regular expressions. The cleaned data set has 24608 comments, and some of them are randomly selected as test data. The test data are manually labeled with sentiment, segmented into words, filtered for stop words, and extracted for features (after fine-tuning, the feature extraction weight filter threshold t is set to 0.6), which eliminates about 9/10 of the low-weight features. Table 2 shows the details of the test data set.

Table 2. Test data set.

Categories	Comments number	Feature per comment	Extracted feature number
Positive	4096	12.2	3201
Negative	4096	8.4	1203
Total	8192	10.3	4404

3.2 Comparing Methods

SMFS will be compared with the following three feature selection methods:

- A GA-based feature selection method that uses feature IG as heuristic information (EWGA) [6].
- A multi-swarm particle swarm feature selection algorithm (MSPSO) [7].
- A feature selection method based on ant colony optimization (ACO) [8].
- An artificial bee colony-based feature selection algorithm (ABC) [9].

Meanwhile, SVM and NB methods without feature selection are used as comparison benchmarks.

3.3 Parameters Settings

In the experiment, the parameters of SMFS are set as follows:

- The maximum number of iterations I : $I = 100$.
- The number of plasmodia n : $n = 20$. After the number of plasmodia reaches a certain threshold, its growth will not improve the algorithm's performance but will cause an increased computational overhead.
- Deterministic flow probability P : The probability that the plasmodia will flow deterministically to the vertex can obtain the maximum short-term return, $P = 0.2$.
- Flow return threshold ϵ : The threshold for the colony to stop flowing. If the colony's short-term reward for ten consecutive flows is less than ϵ , the colony flow will end. If the short-term reward is the L2 regularized tf-idf mean of the selected feature, then $\epsilon = 0.65$. If the short-term return is the importance of the feature to the category under the selected feature set, then $\epsilon = 0.001$.
- Flux decay rate r : After a colony flows, the flux on each edge will change according to the current flow. The flux on the side where the colony does not flow will decay exponentially. r is the adjustment parameter of flux decay, $r = 0.5$
- The flux weight α , $\alpha = 2$. α is a parameter that adjusts the relative weight of flux and short-term return in the flow process. The larger the α , the greater the relative weight of flux, and vice versa.
- Return weight λ : $\lambda = 0.4$. λ is the relative weight that adjusts the return value and simplicity of the current subgraph in the renew procedure. The larger the λ , the greater the relative weight of the final return, and vice versa.

It should be noted that the experiment in this article did not precisely tune the parameters of the algorithm but only determined the optimal value based on the experience of the results of several runs.

The parameter settings of the three comparison algorithms refer to the settings in the respective literature. All methods use L2-regularized L2-loss SVC and Multinomial Naive Bayes as classifiers. The Regularization parameter C of SVC is 0.8. The feature sets to be selected are all feature sets after preprocessing. All methods will perform 10-fold cross-validation on the test data set.

3.4 Results and Analysis

In this part, the experimental results will be given, and a brief analysis will be made.

Table 3 shows the average run time of each algorithm (using SVC as a classifier) and the average features subset size (expressed as a percentage of the candidate feature set size) in 10 cross-validation tests (ten runs). SMFS trades time performance for the ability to select better feature sets. In practice, feature selection is often part of data preprocessing, and final classification performance is often more important than runtime, so the performance of SMFS can be considered to be superior to MSPOS to some extent.

Combined with the above test results, SMFS is superior to the benchmark algorithm, EWGA, and ACO algorithm and slightly better than the MSPSO algorithm in the application scenario that is not sensitive to the algorithm run time (Figs. 2 and 3).

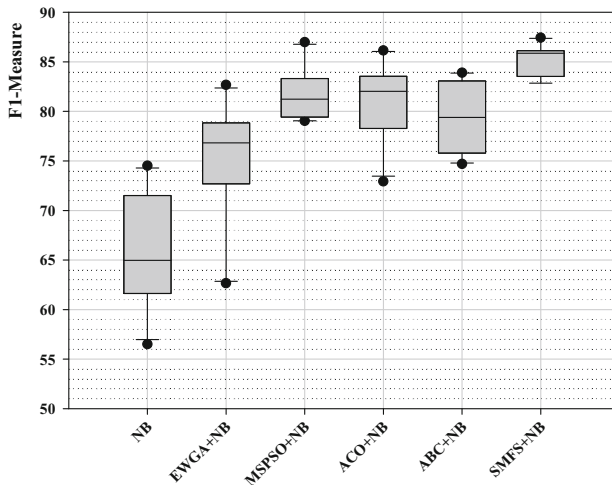


Fig. 2. F1-measure scores of different algorithms in 10-fold cross-validation. The SMFS (feature importance/Multinomial Naive Bayes) is our proposed method, the Multinomial Naive Bayes is the benchmark algorithm, and the EWGA, MSPSO, ABC and ACO are the comparison algorithms. The solid line shows the mean of the scores.

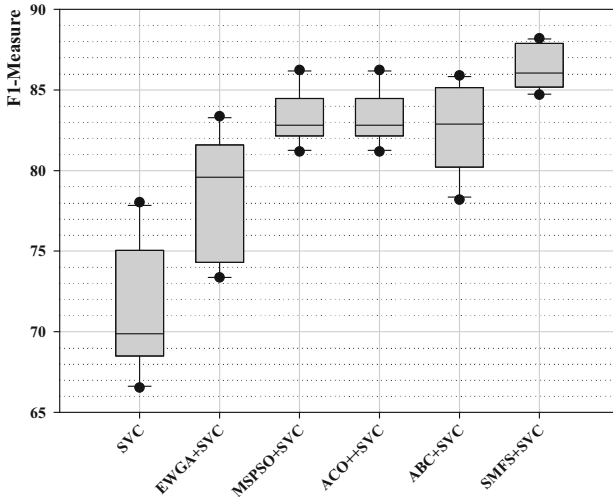


Fig. 3. F1-measure scores obtained by different algorithms in 10-fold cross-validation. The SMFS (feature importance/SVC) is our proposed method, the SVC is the benchmark algorithm, and the EWGA, MSPSO, ABC and ACO are the comparison algorithms. The solid line represents the mean of the scores.

Table 3. The average run time of each algorithm (using SVC as a classifier) and the average selected feature subset size in 10 cross-validation tests (ten runs).

Methods	SVC	EWGA	MSPSO	ACO	SMFS (imp/SVC)
Runtime (s)	850.2	3768.2	4488.7	4670.9	8211.3
Average feature subset size	N/A	31.43%	25.49%	25.44%	25.44%

4 Conclusion

This article briefly reviewed the decision-making mechanism of slime mold network formation and discussed how to use it to solve the feature selection problem.

On this basis, this article proposed a slime mold network formation inspired feature selection algorithm (SMFS) and compared it with the EWGA, MSPSO, ABC, and ACO-based feature selection algorithms on the sentiment recognition problem of online course comments. The results show that the SMFS algorithm can select a proper subset of features and obtain good classification performance on the test dataset.

SMFS is just a preliminary exploration of the application of slime mold network formation mechanism. SMFS's availability and reliability cannot be compared with the state-of-the-art feature selection algorithms. In the follow-up research, on the one hand, SMFS will be improved, and on the other hand, the algorithm inspired by the network formation behavior of slime bacteria will be further expanded so that it can be applied to other machine learning problems.

Acknowledgments. This work was supported in part by the Research Foundation of Education Department of Zhejiang Province [Y201327646]; the Educational Technology Research and Planning Project of Zhejiang Province [JB117], National (China) Educational Technology Research and Planning Project [166243001], and Zhejiang Province's 14th 5-Year Plan for Higher Vocational Education Teaching Reform Project [jg20230321].

References

1. Reid, C.R., Beekman, M.: Solving the Towers of Hanoi - how an amoeboid organism efficiently constructs transport networks. *J. Exp. Biol.* **216**(9) (2013)
2. Nakagaki, T., Yamada, H., Tóth, Á.: Maze-solving by an amoeboid organism. *Nature* **407**(6803), 470 (2000)
3. Ramos, J.: Using tf-idf to determine word relevance in document queries. Paper Presented at the Proceedings of the First Instructional Conference on Machine Learning (2003)
4. Reid, C.R., Latty, T., Dussutour, A., Beekman, M.: Slime mold uses an externalized spatial 'memory' to navigate in complex environments. Paper Presented at the Proceedings of the National Academy of Sciences of the United States of America (2012)
5. Tero, A., Kobayashi, R., Nakagaki, T.: A mathematical model for adaptive transport network in path finding by true slime mold. *J. Theor. Biol.* **244**(4), 553–564 (2007)
6. Uğuz, H.: A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowl.-Based Syst.* **24**(7), 1024–1032 (2011)
7. Zahran, B.M., Kanaan, G.: Text feature selection using particle swarm optimization algorithm. *World Appl. Sci. J.* **7**, 69–74 (2009)
8. Aghdam, M.H., Ghasem-Aghaee, N., Basiri, M.E.: Application of ant colony optimization for feature selection in text categorization. Paper Presented at the 2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence) (2008)
9. Sarac Essiz, E., Oturakci, M.: Artificial bee colony-based feature selection algorithm for cyberbullying. *Comput. J.* **64**(3), 305–313 (2021)



The Role of Social Stress in the Development of Mental Disorders

Shuya Yang^(✉)

Wuhan Britain China School, Gutian 4th Road, Wuhan 430000, Hubei, China
annieyang0908@163.com

Abstract. The primary risk factor for the onset of depression is social stress. In a clinical study, experience to daily social stress has been connected to the occurrence of major depressive disorder. To investigate the underlying mechanisms, animal depressive models have been proposed. In this study, we comprehensively summarized the different methodologies used to induce social-related stress in animals, including the most commonly used resident-intruder model, housing condition, social isolation, early life social event as well as social hierarchy model. Meanwhile, we investigated the behavioral readout adopted in this field. Moreover, we systematically summarized the molecular changes that have been associated with behavior observation. We found that most of the studies were performed on male animals. However, in humans, depression happens 2–3 times greater in females than males, thus we discussed the potential model that has been adopted to study social stress in females as well. We hope this study could provide a good summary for the field to have a comprehensive view of the social stress-related depression model as a foundation for the development of corresponding therapeutic approaches.

Keywords: Social stress · Major depressive disorder · Animal model · Sexual dimorphism

1 General Introduction of Depression and Social Stress

Depression is a multicausal disorder and has been associated with many severe diseases [1]. Several methodologies or animal models have been proposed to measure the mechanisms that lead to depression. For example, chronic mild stress (CMS) could be used to induce depression, and the measure most commonly used to track the effects is a decrease in consumption of a palatable sweet solution [2].

Animal models have been adopted to test the pathology of depression [3], we acknowledge that there are promises and pitfalls in using animal models to study psychiatric illnesses [4]. However, we shall not dismiss their value in promoting our understanding of depression pathology. For example, studies on anti-depressants in animals may help us to know how these medications work and how the mechanisms influence these disorders [5].

The social stressor is an essential risk factor in the formation of depression and are more relevant to human. Clinical studies have suggested a strong relationship between

stressful events and major depression [6, 7]. The study of the effect of using social stress on animals began decades ago. It's been proposed that one of the most powerful stressors as measured by neuroendocrine activation is social rejection [8]. It has been shown that a single social defeat can be long-lasting and makes the stressed animal sensitive to subsequent minor stressors [8]. Moreover, some internal state such as aggression level is strongly associated with the coping behavior type in the social stress model. High aggression is associated with the high level of defense behavior. A low level of aggression is associated with more acceptance of the situation [8].

In this study, we will systematically review the animal models that have been used in social stress, the behavioral and molecular readout of the depressive-related symptoms, as well as sex dimorphism.

2 Animal Models of Social Stress-Induced Depression

2.1 Social Defeat Model

Depression, as one of the most disabling medical conditions, could be investigated utilizing the social defeat animal model [9]. The social defeat model can induce some depression-related phenotypes, for example, an increase in HPA axis activity, increase of immobility, a decrease in sucrose water preference, a reduction of hippocampal volume, and a decrease of cell proliferation in the dentate gyrus [10].

Social stress models let people recognize the ways in which social defeat may have long-lasting psychological, neuroendocrine impacts, particularly on social withdrawal and anhedonia [11]. We conclude that animal models of social stress possess construct validity within many practical applications predicated on the principle attributes of the resident and intruder experiment, colony models, and broad overviews of the behavioral, neuroendocrinological, physiological, neurochemical, and immunological responses that are observed with the aforementioned models [12]. Though studies on social stress may differ in humans and studies on social defeat in animals, with different terminologies, such as bullying in school and rodent resident and intruder paradigm. A reciprocal theory and methodological interaction between the two domains can be beneficial [13].

Experimental rats have behavioral alterations in the resident-intruder paradigm that may be correlated with the depression symptoms, such as anhedonia and motivational impairments [14]. A standardized technique to assess offensive aggression and defensive behavior in a semi-natural environment is the resident-intruder paradigm in rats [15]. For five weeks, rats underwent exposure to daily subjugation stress and exhibit decreased motility, decrease exploratory activity, and reduce sucrose preference. Indicating that this model might mimic the anhedonia and motivational deficit in human depression [16].

Numerous neuropsychiatric disorders, including severe depression, that are associated with genetic, environmental, and epigenetic roots, have been related to social stress [17]. Several methods and different duration of social stress has been used to establish the social defeat model have been proposed:

Using a 5 weeks resident-intruder paradigm to establish a social defeat animal model that may be more relevant to the human daily situations. These animals show three

types of behaviors, the deficit in motivation, behavior despair, and anhedonia [16]. To be more specific, in the 5 weeks of social stress-exposed rats, reduced locomotor and exploratory activity, diminished preference for sucrose solution, lowered body mass gain, elevated weights of the adrenal glands, and reduced testosterone plasma levels have been reported. These effects could be rescued by fluoxetine [14]. In another study, people used the 4 weeks social defeat rats, hyperactivity of HPA, prolonged inactivity time during the forced swimming test, lesser body weight and sweet water intake, and diminished hippocampus volume linked to a reduced rate of dentate gyrus, increased cortical CCK release, have been reported. These effects could be rescued by CCK2 receptor antagonist CI-988 [10].

Acute stress illnesses may result from animals' semi-natural social groups' protection mechanisms failing. And a one significant stressor encounter may lead to long-term effects that last for hours or days [8]. Then social defeat may be the most severe stressor among laboratory stimuli, and a single social failure would cause a stress response that might leave anywhere from four hours to days even weeks. For the long haul, the animals become sensitized to subsequent minor stressors. Also, individual differences should be considered during the testing social stress of animals [8].

In order to test social stress, researchers have identified four behavioral categories in animal models: Resident Dominant (RD), Resident Subordinates (RS), Intruder Dominants (InD), and Intruder Subordinates (InS). These new models of chronic psychosocial stress in male mice could help us to understand the differences between individuals and the effects of susceptibility variation [18]. And they investigate some factors of rats, like their metabolic functions. To sum up, social standing and land ownership were aspects that influenced a person's susceptibility to stress exposure. And these models could also distinguish the dominants and subordinates [19].

Animal models using social defeat, have been demonstrated to make rodents fearful and socially avoidant. Also, using behavioral paradigms, including the social interaction test, they assessed fear and social avoidance in rodents. These paradigms show that social anxiety order (SAD) is strongly accompanied by anxiety and affective disorders [20].

Long experience of daily societal defeat with permanently living with aggressive males change the male C57BL/6J mice's traits of submissive behavior. We discussed the depression's emergence in the subordinate C57BL/6J male rate caused by chronic unavoidable social stress. (Kudryavtseva, 1991) [22]. In the experiment, C57BL/6J mice that experienced social defeat by a larger and more aggressive CD-1 mouse on a regular basis had a marked depressive-like illness. The experiment contains three important stages, selection of aggressive CD-1 mouse, social conflicts between CD-1 and C57BL/6J mice, which put an end to the C57BL/6J mice's avoidance of social situation. (Golden, 2011) (Sam A. Golden et al. 2011) Further experiments measure the implications of singular and dual social conflict on different rat behaviors and body mass. Three experiments were conducted to examine the effects of social defeat. Two days after the most recent conflict, the first experiment was conducted in an open area. In a subsequent experiment, animals were exposed to a strange, gentle, mild mice two days following one particular or dual dispute. (Meerlo, 1996) (Meerlo et al. 1996) [40].

Researchers also looked at how DBA/2 male rat's territorial urine marking and ultrasonic courting vocalizations were affected by acute social stress. Male mice who

suffered social destitution in the first experiment displayed a protracted suppression of territorial urine marking. Four weeks after a social defeat, they saw less possessive marking in the second experiment [22].

It has been suggested that social conflict models are an effective tool to look at fundamental concerns about how social experience affects the brain function as well as the behavior. Examiners take a variety of some animal models, such as social defeat, to assess whether they can help us to understand how experience can shape an animal's behavior and cause physical change in the body [23]. The research demonstrates how a social quarrel in rats can have different long-lasting effects depending on how the subjects handle it. For one hour, vicious related individuals were placed in enclosures with experimental rats (C57BL/6J), and the variations in the rhythm magnitude revealed a definite negative association with the hostility of the testing rats themselves [24]. In Prosolt's test and the open-field test, obedient mice reported a decline in mobility and a spike in inactivity time, but they never expressed aggression toward other animals. In subordinate rats, weight loss and greater gastric mucosa injury were observed, along with a drop in plasma testosterone levels and immunological resistance. Additionally, persistent, unnecessary interpersonal pressure is thought to be a carcinogenic element that causes mice to acquire melancholy abnormalities and anxiety disorders [25].

2.2 Housing Condition Model

The behavioral and physiological reactions of individuals are influenced by their residential environment. Alteration of housing conditions has also been proposed to induce depression. For example, one study aims to analyze how isolated versus communal living affects male mice's susceptibility to the lingering impacts of inevitable social defeat. Single exposure to an abrasive, strange male mice was done with group-housed mice for three weeks, resulting in social defeat. Findings indicated that after suffering social failure, the caged rats underwent long-term, detrimental behavioral and biological alterations. Rats' body development slowed between 7 to 14 days. At the twenty-first day, when social interaction with other rats took place again and again, the individual rat was remarkably motionless. In the open field test, tested mice took much greater time to leave their house cage and were considerably less versatile as well. And mice were more anxious when they were put into elevated plus-maze. When tested at 21 days, the DEX test and CRF text showed that HPA activity was higher, also, ACTH was higher. And adrenals were higher and thymus and seminal vesicles were smaller. In conclusion, the isolated mice after experiencing social defeat showed long-lasting, adverse behavioral and physiological changes [24]. Another study compared the impact of congested dwelling circumstances and individuals on corticosterone and the biochemical stress index. By adjusting the residence's structural and population features independently combining fifty male and fifty female mice, the researchers further expanded their discoveries. And the results show that examiners have to take the residence circumstances into account as an intervening variable that will affect the consequences of the experiment [26].

2.3 Social Isolation Induced Behavioral Changes

Several species of animals, especially rats, have been shown to exhibit intense aggressive tendency when they are isolated [27]. Animal models using social isolation have been shown to induce social avoidance and fear in rodents. Also, using behavioral paradigms, including the social interaction test, we assessed the terror in rats and the avoidance in social situation. These theoretical frameworks show that the anxiety and emotional disorders are closely related to social anxiety disorders (SAD) [20].

Early life social event model. Early life paradigms encompass a range of factors, such as early handling, separation, and deprivation protocols, as well as the presence of impoverished environments, which may contribute as risk factors in the etiology of depression. The manifestation of depressive-like behavior in rodent subjects can be subject to modulation through the profound implementations of the early life development [28]. In order to test how depression develops in individuals, researchers use experiments of postnatal environmental manipulations in rodents and primates that can probably gain evidence about the early-life experience and behavior of emotion. Therefore, this model could be utilized to progress animal models for depression research. In paper, researchers review some evidence that neglects of environmental problems in rat pups and monkey infants leads to long-term depression [29]. So, the early-life environment may have the significant effect on the formation of depression.

2.4 Social Hierarchy Stress Model

Both animals as well as human beings exhibit a wide range of hierarchical and territorial behaviors [30]. One of the models to test the effect of social hierarchy on animal behavior is conventional mice and inhumane primate social stress models with an emphasis on social hierarchy models are utilized in laboratories [31]. A forced social status loss would induce depressive-like behaviors, to be added to the collection [1] As a result, we also detect how the level of ranking can influence physiology and health. Not surprisingly, the results showed that they differ depending on the social structure of various animals and groups. Besides, we also recognized how stressful characteristics affect the different fields of consequences, and how these findings applies to the human body [32].

Agonistic behavior at the start of the cycle between light and darkness was repeatedly seen in pairs of male Lister hooded mice in order to stablish permanent supremacy rankings. The aggressive mal Tryon Maze Dull (TMD) stain that defeated the dominant species resulted in the demise of dominance in the enclosure that persisted at least seven days. Additionally, consistent weekly loss to TMD species reduced both domestic home dominance actions and ingestion of a tasty solution containing sucrose [33]. And the research used the male tree shrew's model of persistent psychosocial strain, which has its foundation on territorial actions to show mechanisms of stress-related disorders and determine which one dominates or subordinates. Additionally, the subordinates show distinct stress when there are two males, and a series of biological and physiological changes occur with chronic social stress [34] such as raising plasma corticosterone level (CORT) and increasing corrected adrenal and spleen weights. The visible burrow systems (VBS), which is a free social interaction a environment that enable the establishment of a social ladder, a specialized framework for researching the behavioral, neurological,

and endocrine aspects of prolonged stress allows mice communities to establish natural, pressure-engendering relationships with one another [35].

3 Behavior and Physical Changes After Social Defeat

Long experience of daily social defeat with permanently living with aggressive males change the conducts that constitute submission of C57BL/6J male mice. The data are discussed the emergence of depression in obedient C57BL/6J male mice caused by chronic unavoidable social stress. [21] Examiners test the long-lasting behavioral consequences of a one stressful event in male rats, employing an aversion scenario. Using social defeat or electric shocks as stressors, the individual was positioned in the little space and given five minutes to look around the enclosed area. The experiment was conducted on Day1, Day5, and Day10. The inspection of a vacant enclosure was substantially fewer limited than the exploration of an enclosure containing the stimulus rat. Consequently, this paradigm could be utilized to investigate how various substances affect anxiety brought on by pressure [36]. The social aspect of the anxiety test was responsive to both anxiolytic and anxiogenic impacts, as well as to biological and environmental variables that modulate anxiety. This test already examined many biochemical changes. Thus, the application of this model was essential in revealing the neurological underpinnings of anxiety [37]. In the experiment, one pair of male mice were positioned in a test box for ten minutes. The substantial reduction in social contact that happened in rat who have not taken drugs as the light intensity or the unpredictability of the test box was raised was then partially avoided by persistent chlordiazepoxide (5 mg/kg) administration; in the latter scenario, it simply has calming effects. It is examined and contrasted to other studies to determine whether this test may be utilized as an laboratory model of anxiety. [38] More research examined what degree of autonomic and behavioral is stress animals can stand for, and to how much an animal that has suffered defeat could be shielded from stress? And people test the reactions with clonidine or metoprolol, two adrenergic agents with clinical anxiolytic implications. Only clonidine could prevent the tachycardia response to stress. Clonidine decreased some motor behavior (like rearing, and walking), but kept the protective responses at elevated levels and extended the ultrasonic vocalization (USV) [39]. This study examined the implications of stress-induced alterations in diurnal body temperature and physical activity rhythms, elucidating their significance in relation to the etiology of depressive disorders characterized by disturbances in biological rhythms [40]. Then the present study investigates the behavioral responses of C57BL/6J male mice, who have undergone repeated instances of either victories or defeats in daily agonistic interactions, through comprehensive assessments utilizing the plus-maze and partition tests (sociability test). The result showed that the parameters associated with the division and plus-maze evaluations in mice display a correlation with the alternative encounters of agonistic confrontations, whether positive or negative. The amalgamation of these two evaluations could be employed for the assessment of anxiety development in mice [41].

Besides, in the social defeat model, social avoidance and fear in mice could be assessed to measure anxiety and affective disorders [20]. After a social defeat exposure in the house-case of an aggressive conspecific, termed conditioned defeat, the male

hamsters were failing to guard their territory in the subsequent experiment even against the small, nonaggressive intruders. In experiment 1, the result shows that conditioned defeat occurred in all defeated male mice and lasted for at least 33 days. In the subsequent experiment, it shows whether conditioned defeat is related to sex differences. Then in Experiment 3, it suggests that the behavioral reactions of hamsters were not directly related to sex difference, but due to the release of plasma adrenocorticotropin [42].

In order to test social stress, we have four behavioral categories in animal models regarding social stress: Resident Dominants (RD), Resident Subordinates (RS), Intruder Dominants (InD), and Intruder Subordinates (InS). And researchers investigate some factors of rats, like their metabolic functions. To sum up, social hierarchy and territory ownership were elements that determined an individual's susceptibility when being exposed to pressure. Additionally, these paradigms might be useful for examining the biological causes of individual variations in how people or animal react to challenging situations [19].

Defeated mice (C57BL/6) displayed a depressive-like characteristics, as suggested by a rise in social withdrawal, and changes during social interaction, such as forced swimming, sucrose preference, and elevated plus-maze behavioral assays [43]. Specifically, the forced swimming test has been proposed to be a new behavioral test in rat that not only resemble depressive illness, but test the core component behavior in depression (Irwin Lucki 1997). But also give effective antidepressant drugs that acted through different pharmacological mechanisms since they are selectively sensitive. Besides, the antidepressant should produce different behavioral effects [44].

Research examines the effects of singular and dual social disputes on mice behaviors and the body mass index. Three experiments were conducted in order to evaluate the mode of functions of social defeat. 48 h after the most recent conflict, the preliminary experiment was carried out in an open are. In a subsequent experiment, rats were exposed to a strange, peaceful, experimental mice two days following a single or double dispute [40]. Social defeat triggers the amount of biological and physiological changes: In the 5 weeks of social stress exposed rats, reduced locomotor and exploratory activity, reducing preference for sucrose solution, lowering body weight gain, escalating adrenal weights, and lowered levels of testosterone in plasma have been reported. These effects could be rescued by fluoxetine [14] In another experiment carried out with 4-weeks defeated mice, hyperactivity of HPA, decreasing body mass, intake of sweet water, hippocampus volume and climbing inactivity time in the forced swimming test, together with a decline in the proliferation of cells in the dentate gyrus, increased cortical CCK release, have been reported. These effects could be rescued by CCK2 receptor antagonist CI-988 [10].

The research examined the behavioral, cardiovascular, and thermoregulatory responses of rats to temporary alternating agonistic encounters over the course of both the short and the long-term. In an experiment, two rats are separated for 10min, then they have a short encounter bringing to the defeat of the intruder and left in a ten minutes period. While an intruder is removed, they have 30-min-long conflicts are duplicated on five successive days. The outcome demonstrates that competitions cause attackers to experience quick, severe tachycardia and heat. Moreover, during the initial inspection of the resident's housing box, the defensive standing up is all but nonexistent [45].

Urine also contributes to the change of mice under social stress. Mouse could use urination as one type of behavioral response in the social context. The mouse could use urination to identify the conspecific's identity. Moreover, a previous study has suggested that male mice would urinate more in the presence of a female conspecific, compared with a male conspecific (Reynolds, 1971, *Nature*). Overall, urination could be used as a measurement of the animal's status. Urination has been used to assess the status of the depressed animals. It has been shown that urine scent marking would decrease in animals exposed to stress, either physical stress or social stress [46]. A novel animal model of rodents, urine scent marking (USM), is used to evaluate hedonic behaviors. The USM test is a highly sensitive, validated approach to measuring psychosocial stress and then examining the stress resiliency and vulnerability, also their neurobiological substrates [46]. The paper looked at how territorial urine marks and ultrasonic courting utterances in DBA/2 male mice were affected by chronic social anxiety. In the first experiment, male mice who experienced social defeat showed long-lasting suppression of tribal urine markings. In the following experiment, four weeks after being exposed to the social defeat, researchers noticed a reduction in urine markings [22].

Dominant Submissive Relationship (DSR) lead to two model, mania and depression, susceptible to mood-stabilizing medications, and the paper discusses the advantages and limitations of this paradigm compared to other animal paradigms. They used food-restricted animals and measure the order of food access [47]. The specific research tests the implications of persistent psychosocial stress on male mice in terms of heart rate, temperature, and physical activity. Male subordinates displayed tachycardia, mild hyperthermia, and low physical activity, in contrast to dominant males who displayed tachycardia, pronounced hyperthermia, and normal level of physical activity. The immediate and long-run autonomic reactions did not show any association with one another [48].

Using the resident-intruder paradigm, the research investigated some specific biological and behavioral changes. To be more specific, the experiment probes into the alterations of copulatory behavior in the male mice when they are experiencing the fighting, and also, determined the implications of naloxone on the integration between social conflict and copulatory distress. Compared to the control group without aggressive combat, mice who played a role as intruder that had been attacked one again and again demonstrated a numerous decline in copulatory behavior. Also, the endogenous opioid mechanism potentially do not taking account into this specific behavioral phenomenon [49].

The tail suspension test, a novel antidepressant testing approach, involves suspending a rat by its tail from a lever. There will be phases of agitation and immobility throughout the duration of the examination [50].

4 Long-Term Effect of Social Defeat Exposure

The research aimed to determine the enduring impacts of a really stressful event, such as social rejection followed by one's habitation. They found that the social stress-induced depressive-like behavior could last for up to 3 months after social defeat. Surprisingly, the implications that bring from social housing could mitigated the negative repercussions of the defeat [51].

Further research examined the effects that may last for a long time of experienced to single social defeat on open-field behavior, with a focus on the progression of alterations brought on by stress. In the experiment, tested rats were forced to experience social defeat by putting them with an aggressive male rat for 1h, after that they were putting back into their home cage with no cues. The result shows the stress-induced reduction in open field movement, food intake, and body weight, as well as social interaction with unfamiliar conspecifics[52].

5 Molecular or Cellular Changes Social Defeat Induced Depression

Much research examined the behavioral and biological effects of social hierarchy and mutual combat in the resident-intruder dyads of Long Evans male mice. The result is, compared to the rats from another experimental groups, rats who are in the subordinated status lost their body mass and had greater concentration of plasma corticosteroid. Furthermore, in compared to the superior rats, the subordinated members had smaller size of thymus glands. Thus some of the behavioral and physiological changes could be significantly influenced by altering the intensity of fierce confrontations [53].

California mice are capable to be a helpful and instructive animal model for researching sex differences in the transformations to stress from the society, especially in neurobiological mechanisms, like the brain activity, corticosterone, and a number of phosphorylated CREB cells, responses to social stimulus. And the responses last for 4 weeks [54]. In response to stress, the brain releases adrenal corticosteroid hormones. A bias for brain diseases linked to stress and pressure can be introduced by the discrepancy in the binary control mechanism [55]. In the hypothalamus nucleus, the rats in the subordination status that are also non-responsive displayed substantially fewer CRF mRNA granules on average per cell [56]. The members that failed in the confrontations may be revealed in the remarkably alterations persistent with biological index of the long exposure to the stress. The visible burrow systems (VBS), which is a free social interaction environment that enable the establishment of a social ladder, a specialized model allows the group of rats to establish natural, tension-endangering relationship [57]. We divided the subordinated rats into two categories: stress-responsive and non-responsive. Compared to the controls, binding to 5-HT1A receptors was reduced in stress-responsive subordinates, and binding to 5-HT2 receptors was raised in these two categories [35]. Both members in subordinate and dominant exhibit physiological alterations in comparison to non-group controls, manifesting as increased adrenal and spleen dimensions and diminished thymus proportions. Nonetheless, these alterations seem to be more pronounced in subordinate individuals. Notably, the CORT level exhibit s significant escalation, while the plasma testosterone experiences a substantial decline in subordinated rats. Additionally, the diminished presence of CORT binding globulin serves to further augment the concentration of free CORT, particularly in subordinate males [58]. Another research carries the experiments to investigate two themes: 1. the potential associations between the hormone levels and aggressive actions. 2. Hormonal and behavioral alterations after the exposure to social defeat or social stress. Using Tryon Maze Dull S-3 rats as the experimental rats, experimenters get blood samples to detect their hormonal levels [59]. One experimental result show that AVP is released within the PVN in response to an

emotional stimulator. The researchers assume an hypothesis that the intranuclear release which mentioned above provides an adverse tonus on ACTH secretion [60].

The social stress models could also bring the experiment with high-intensity stressors. And under huge social stress, the hippocampus and dentate gyrus, and HPA axis functioning change a lot [61]. The experiment shows that alterations in diet alone cannot adequately respond for the varied implications of social stress exposure on HPA activation [62]. In the 4 weeks of social defeat rats, hyperactivity of HPA and many fluctuating variables are reported. These effects could be rescued by CCK2 receptor antagonist CI-988 [10]. The potential role of transforming in glucocorticoid (GR) and mineralocorticoid receptor (MR) linked to the varied HPA activity after social defeat were tested, and been examined with standardized experiments [63]. Some of these parameters can be renormalized by antidepressants. According to the social stress model, the processes related to the transformation of neurons are what lead to behavioral changes that can be treated with antidepressants [64].

Further experiments pay more attention to the submissive reactions in the field of behavior, endocrine and neurochemistry in mice and rats that confront an agonistic opponent [65].

The dopaminergic response may vary in magnitude and location depending on the nature of the aversive stimulus. Experimental findings indicate that some biological changes, for example, alterations in accumbens, are not a result of motor activation. Instead, they likely reflect ascending attention towards the provocative stimulus or the intruder's endeavors. Both aversive and rewarding environmental stimuli can increase extracellular dopamine system [66]. Additionally, level of reactivity that rats performed to environmental stressors has been detected to build an relationship with the time it takes for self-administration of psychomotor stimulant drugs to begin. The experiment examining the behavioral and dopaminergic responses of rats to social defeat provides evidence supporting the hypotheses that we listed before [66].

The resident-intruder model of defeat was utilized as a social stressor in adult male Sprague Dawley rats, and we measure their biological and physiological changes over 7 days. Lastly, the result suggests that the prospective ramifications of social stress and susceptibility to depressive-like symptoms may be anticipated by inherent variations in stress reactivity [65]. One unique research focuses on the resident-intruder paradigm, and deeply investigates about that the operational capacity of the hippocampus in rats exposing to societal stress was assessed. The findings demonstrate that both a brief double social defeat and reduplicated social defeat stress manifest distinct impacts on dendritic remodeling in hippocampal CA3 neurons, ultimately influencing hippocampal a potentiation and depression for the extremely long time. Moreover, the reports illustrates that a substantial dynamic range in the brain's adaptive plasticity, enabling the animals to accommodate their behavior in response to previously experienced stressful situations [67].

In the days after a social defeat exposure in the cage of an aggressive conspecific, termed conditioned defeat, the male rats were failed to guard their own territory even when they faced small, mild intruder. In experiment 1, the result shows that conditioned defeat occurred in all defeated rats and lasted for at least 33 days. In the next experiment, it shows whether conditioned defeat related to sex differences. Then in Experiment 3,

it suggests that the behavioral reactions in hamsters are not directly related to sex difference, but due to the release of plasma adrenocorticotropin [42].

Acknowledging the intracellular chain of incidents underlying the shift from episodic to chronic social stress in adolescence and adulthood might offer knowledge about how to modify the fundamental reward mechanisms that are crucial for the development of addiction and depressive disorders [68]. The susceptible and unsusceptible subpopulations in mice subjected to social defeat have different behavior and physiological symptoms. The mesolimbic dopamine circuit's distinctive modifications that specifically related to vulnerability or insusceptibility. An experiment is already been constructed that not only evaluate the mechanisms of variation about stress resistance but also demonstrate the paramount importance of plasticity in the reward circuit [69].

Several other molecular components have also been suggested to participate in social stress-induced depression [70]. Meanwhile, BDNF contributes to social aversion and gene expression. The critical function of BDNF is in treating enduring neural and behavioral plasticity results from aversive social experiences [71]. Moreover, in the social defeat model, the brain CCKergic transmission increase, and the administration of CCK2 receptors by the specific antagonist CI-988 has been shown to have an anti-depressant effect [10].

The immune system has been studied as well. The investigators conducted an analysis on animal-based studies utilizing social interactive activities in rodents, revealing that the impact of short- and long-term social stress on immune system functions is not universally suppressing. Additionally, they propose that the extent and orientation of alterations in various components of the immune system exhibit a strong correlation-ship with the social status of the animal [72].

In the chronic social stress model in rats, it has been shown that fluoxetine can improve depression-related phenotypes [14].

6 The Mechanism for Initiating Fighting Between the Intruder and the Resident

This specific review paper investigates the intruder's social stress from another perspective that how the fight between the intruder and the resident was initiated. Several genes mainly influence animals' aggressive behavior, including Y chromosome. But the genetic effect on aggressive behavior primarily depends on the circumstances of background, house-cage which is the mutual environment for rats, and the aggressive or non-aggressive intruder [73]. Additionally, the trends of assault and defense show a variety of offensive and protective behavior comparable to that of wild rats. The data reveals that particular assault and protective responses have undergoes changes as well [74].

7 The Social Stress Test on Females

Female social stress models are always lacking in the field since male mice seem to have obvious and observable behavior since they are aggressive and territorial. But analyzing a female model of social defeat or social instability is also necessary. To be more specific,

there are three main concerns in the formation of mental diseases in the form of animal models: (1) health mood issues appear more frequently in female compared with male but present major researches focus on testing male rodents. (2) sex-differences in behavior ought to be included in the conceptual frame of sexual selection. (3) there is a significant difference between how often it is for different genders to experience social stress. We need more standardized female models. And in individually housed females, show a diminished tendency for inquiry and increased anxiety [75].

A female social defeat model has been developed by artificially activating the ventromedial hypothalamus using chemical genetic approaches to trigger male aggression towards females. The examiner builds a robust female social defeat model in order to identify and develop new medicinal substances to cure women's anxiety and depression: a female rat model of reduplicated social defeat stress (RSDS). The result is, females that are sensitive to RSDS exhibit social avoidance, anxiety-like action, weight loss, and raised levels of interleukin 6 in the blood. And this model could help us to examine the sex differences in neurobiological mechanisms of social defeat [76].

Additionally, using the female mice model, we investigate how male and female mice's anxiety and exploratory behavior are affected by social context in the housing habitat, also considering about estrous phase and social status. The result shows that individually housed females explore less and have a higher level of anxiety which is opposite to the phenomenon from male rats. Different housing procedures may result in social stress in both males and females variedly [75] In summary, different housing procedures may develop social stress in both male and female males in various level and aspects [75].

To investigate social stress in females, the social instability model is an excellent choice. To create social unrest, enrollment cycle was utilized in conjunction with alternating isolation and crowding society with both males and females. As a result of gaining weights, female beings only had an increase in their adrenal weight, thymus involution and other important variables [77] The research uses chronic social instability stress in paradigm in female rats to test affective disorders in females. The experiment lasts for four weeks and includes challenging social circumstances like periods of isolation and congestion which is totally opposite to each other. And physiological level of female rats changes a lot. And the result shows that female mice are allowed to be subjected to prolonged stress of weeks without habituation, and they finally induced a depressive-like phenotype [78]. Another carrying experiment uses the breakdown of the social interactions in a crowded environment in female rats to illustrate the impact of persistent social stress. In this specific paradigm, the mice's group membership is altered twice weekly for a total of seven weeks, spanning the teenage and early childhood years. The findings demonstrate that living in an unexpected social milieu caused acute chronic stress in female mice, and that after experiencing chronic social stress, there is an alternate manifestation of genes associated with the stress system in the female mice [28].

There are also researchers who used the same sex resident-intruder social defeat model, lactating dam was used in this study to administer social defeat towards female rats [79].

8 Resilience

In the face of social stress, animals may develop depression-related behaviors. However, they could also exhibit resilience. The reviewing paper summarized the literature in rodents focused on identifying widely varied fields of resilience. And the paper examines findings from both human and animals and also discuss treatments for mental disorders [80]. In order to test social stress, observers have four behavioral categories of social stress, especially regard with animal models. And researchers investigate some factors of rats, like their metabolic functions. To sum up, social class and territory sovereignty were variables that determined a person's or animal's susceptibility to stress exposure. And these paradigms may be reliable when looking into the biological and physiological causes of individual variations in how people react to stressful situations [19].

9 Cellular and Molecular Mechanisms to Explain Susceptibility and Insusceptibility

Rats experienced with social defeat could be clearly divided into two groups, one is susceptible, and the other is unsusceptible subpopulations. It has been shown that plasticity changes in the mesolimbic dopamine circuit could be underlying this vulnerability and insusceptibility difference [69].

10 Social Defeat and Addiction

Researchers used the same sex resident-intruder social defeat model, aggressive male, and lactating dam used in this study. Cross-sensitization to cocaine caused by periodic social defeat and is characterized by elevated physical movement and dopamine levels in the nucleus accumbens. And the experiment shows that, at every stage of the development that been addicted to cocaine, women are more susceptible than men [79]. The implications of social stress on the propensity to start self-administering cocaine was investigated in Sprague-Dawley rats that had been repeatedly subjected to agonistic assaults from a same-sex component. As a result, both sexes of rats' social class seems to have a significant impact on when they start using drugs [81].

And the results of the experiment which inspected the alterations in behavior and changes in dopamine level of mice to social defeat already proved these hypotheses [82].

11 Discussion

The collaborative effort may be needed to address the behavioral, epidemiology and neural substrates underlying social stress-related depression. A previous study has discussed nature versus nurture, opportunities and challenges in the collaboration between psychiatry, epidemiology, and neuroscience, and understanding of gene-environment interactions [83].

Through the concept of stress. It has been proposed that stress shall refer to the condition that exceeds the normal range of natural responses. The following work discussed the outcomes of a limited definition for stress research and interpret results in terms of the adaptive nature of the stress response, stress may be regarded as the feature of either the absence of an unpredictable reply or uncontrollable results [84].

References

1. Lang, U.E., Borgwardt, S.: Molecular mechanisms of depression: perspectives on new treatment strategies. *Cellular Physiol. Biochem.* (2013). <https://doi.org/10.1159/000350094>
2. Willner, P.: Validity, reliability and utility of the chronic mild stress model of depression: a 10-year review and evaluation. *Psychopharmacology* (1997). <https://doi.org/10.1007/s002130050456>
3. Willner, P.: The validity of animal models of depression. *Psychopharmacology* (1984). <https://doi.org/10.1007/bf00427414>
4. Nestler, E.J., Hyman, S.E.: Animal models of neuropsychiatric disorders. *Nat. Neurosci.* (2010). <https://doi.org/10.1038/nn.2647>
5. Nestler, E.J., Gould, E., Manji, H., Manji, H.K.: Preclinical models: status of basic research in depression. *Biol. Psychiat.* (2002). [https://doi.org/10.1016/s0006-3223\(02\)01405-1](https://doi.org/10.1016/s0006-3223(02)01405-1)
6. Kendler, K.S., Karkowski, L.M., Prescott, C.A., Prescott, C.A., Prescott, C.A., Prescott, C.A.: Causal relationship between stressful life events and the onset of major depression. *Am. J. Psychiatry* (1999). <https://doi.org/10.1176/ajp.156.6.837>
7. Kessler, R.C.: The effects of stressful life events on depression. *Ann. Rev. Psychol.* (1997). <https://doi.org/10.1146/annurev.psych.48.1.191>
8. Koolhaas, J.M., De Boer, S.F., De Rutter, A.J., Meerlo, P., Sgoifo, A.: Social stress in rats and mice. *Acta Physiologica Scandinavica* (1997)
9. Hollis, F., Kabbaj, M.: Social defeat as an animal model for depression. *IJAR Journal* (2014). <https://doi.org/10.1093/ilar/ilu002>
10. Becker, C., et al.: Repeated social defeat-induced depression-like behavioral and biological alterations in rats: involvement of cholecystokinin. *Mol. Psychiatry* (2008). <https://doi.org/10.1038/sj.mp.4002097>
11. Chaouloff, F., Chaouloff, F.: Social stress models in depression research: what do they tell us? *Cell Tissue Res.* (2013). <https://doi.org/10.1007/s00441-013-1606-x>
12. Martinez, M., Calvo-Torrent, A., Pico-Alfonso, M.A.: Social defeat and subordination as models of social stress in laboratory rodents: a review. *Aggressive Behav.* (1998). [https://doi.org/10.1002/\(sici\)1098-2337\(1998\)24:4<241::aid-ab1>3.0.co;2-m](https://doi.org/10.1002/(sici)1098-2337(1998)24:4<241::aid-ab1>3.0.co;2-m)
13. Björkqvist, K.: Social defeat as a stressor in humans. *Physiol. Behav.* (2001). [https://doi.org/10.1016/s0031-9384\(01\)00490-5](https://doi.org/10.1016/s0031-9384(01)00490-5)
14. Rygula, R., Abumaria, N., Domenici, E., Hiemke, C., Fuchs, E.: Effects of fluoxetine on behavioral deficits evoked by chronic social stress in rats. *Behav. Brain Res.* (2006). <https://doi.org/10.1016/j.bbr.2006.07.017>
15. Koolhaas, J.M., Coppens, C.M., de Boer, S.F., Buwalda, B., Meerlo, P., Timmermans, P.J.: The resident-intruder paradigm: a standardized test for aggression, violence and social stress. *J. Visualized Exp.* (2013). <https://doi.org/10.3791/4367>
16. Rygula, R., Abumaria, N., Flügge, G., Fuchs, E., Rütther, E., Havemann-Reinecke, U.: Anhedonia and motivational deficits in rats: Impact of chronic social stress. *Behav. Brain Res.* (2005). <https://doi.org/10.1016/j.bbr.2005.03.009>
17. Hollis, F., Wang, H., Dietz, D.M., Gunjan, A., Kabbaj, M.: The effects of repeated social defeat on long-term depressive-like behavior and short-term histone modifications in the hippocampus in male Sprague-Dawley rats. *Psychopharmacology* (2010). <https://doi.org/10.1007/s00213-010-1869-9>
18. Bartolomucci, A., et al.: Social factors and individual vulnerability to chronic stress exposure. *Neurosci. Biobehav. Rev.* (2005). <https://doi.org/10.1016/j.neubiorev.2004.06.009>
19. Bartolomucci, A., et al.: Behavioral and physiological characterization of male mice under chronic psychosocial stress. *Psychoneuroendocrinology* (2004). <https://doi.org/10.1016/j.psyneuen.2003.08.003>

20. Toth, I., Neumann, I.D.: Animal models of social avoidance and social fear. *Cell Tissue Res.* (2013). <https://doi.org/10.1007/s00441-013-1636-4>
21. Kudryavtseva, N.N., Bakshtanovskaya, I.V., Koryakina, L.A.: Social model of depression in mice of C57BL/6J strain. *Pharmacol. Biochem. Behav.* (1991). [https://doi.org/10.1016/0091-3057\(91\)90284-9](https://doi.org/10.1016/0091-3057(91)90284-9)
22. Lumley, L.A., Sipos, M.L., Charles, R.C., Charles, R.F., Meyerhoff, J.L.: Social stress effects on territorial marking and ultrasonic vocalizations in mice. *Physiol. Behav.* (1999). [https://doi.org/10.1016/s0031-9384\(99\)00131-6](https://doi.org/10.1016/s0031-9384(99)00131-6)
23. Huhman, L.H.: Social conflict models: can they inform us about human psycho-pathology? *Horm. Behav.* (2006). <https://doi.org/10.1016/j.yhbeh.2006.06.022>
24. Ruis, M.A.W., et al.: Housing familiar male wildtype rats together reduces the long-term adverse behavioural and physiological effects of social defeat. *Psychoneuroendocrinology* (1999). [https://doi.org/10.1016/s0306-4530\(98\)00050-x](https://doi.org/10.1016/s0306-4530(98)00050-x)
25. Kudryavtseva, N.N., Avgustinovich, D.F.: Behavioral and physiological markers of experimental depression induced by social conflicts (DISC). *Aggressive Behav.* (1998). [https://doi.org/10.1002/\(sici\)1098-2337\(1998\)24:4%3c271::aid-ab3%3e3.0.co;2-m](https://doi.org/10.1002/(sici)1098-2337(1998)24:4%3c271::aid-ab3%3e3.0.co;2-m)
26. Brown, K.J., Grunberg, N.E.: Effects of housing on male and female rats: crowding stresses males but calms females. *Physiol. Behav.* (1995). [https://doi.org/10.1016/0031-9384\(95\)02043-8](https://doi.org/10.1016/0031-9384(95)02043-8)
27. Valzelli, L.: The 'isolation syndrome' in mice. *Psychopharmacology* (1973). <https://doi.org/10.1007/bf00421275>
28. Schmidt, M.V., Wang, X.D., Meijer, O.C.: Early life stress paradigms in rodents: potential animal models of depression? *Psychopharmacology* (2011). <https://doi.org/10.1007/s00213-010-2096-0>
29. Pryce, C.R., et al.: Long-term effects of early-life environmental manipulations in rodents and primates: potential animal models in depression research. *Neurosci. Biobeh. Rev.* (2005). <https://doi.org/10.1016/j.neubiorev.2005.03.011>
30. Rohde, P.: The relevance of hierarchies, territories, defeat for depression in humans: hypotheses and clinical predictions. *J. Affect. Disord.* (2001). [https://doi.org/10.1016/s0165-0327\(00\)00219-6](https://doi.org/10.1016/s0165-0327(00)00219-6)
31. Tamashiro, K.L., Nguyen, M.M., Sakai, R.R.: Social stress: from rodents to primates. *Front. Neuroendocrinol.* (2005). <https://doi.org/10.1016/j.yfme.2005.03.001>
32. Sapolsky, R.M.: The influence of social hierarchy on primate health. *Science* (2005). <https://doi.org/10.1126/science.1106477>
33. Willner, P., et al.: Loss of social status: preliminary evaluation of a novel animal model of depression. *J. Psychopharmacol.* (1995). <https://doi.org/10.1177/026988119500900302>
34. van Kampen, M., Kramer, M., Hiemke, C., Flügge, G., Fuchs, E.: The chronic psychosocial stress paradigm in male tree shrews: evaluation of a novel animal model for depressive disorders. *Stress* (2002). <https://doi.org/10.1080/102538902900012396>
35. McKittrick, C.R., Blanchard, D.C., Blanchard, R.J., McEwen, B.S., Sakai, R.R.: Serotonin receptor binding in a colony model of chronic social stress. *Biol. Psychiatry* (1995). [https://doi.org/10.1016/0006-3223\(94\)00152-s](https://doi.org/10.1016/0006-3223(94)00152-s)
36. Haller, J., Bakos, N.: Stress-induced social avoidance: a new model of stress-induced anxiety? *Physiol. Behav.* (2002). [https://doi.org/10.1016/s0031-9384\(02\)00860-0](https://doi.org/10.1016/s0031-9384(02)00860-0)
37. File, S.E., Seth, P.: A review of 25 years of the social interaction test. *Euro. J. Pharmacol.* (2003). [https://doi.org/10.1016/s0014-2999\(03\)01273-1](https://doi.org/10.1016/s0014-2999(03)01273-1)
38. File, S.E., Hyde, J.R.: Can social interaction be used to measure anxiety. *Br. J. Pharmacol.* (1978). <https://doi.org/10.1111/j.1476-5381.1978.tb07001.x>
39. Tornatzky, W., Miczek, K.A.: Behavioral and autonomic responses to intermittent social stress: differential protection by clonidine and metoprolol. *Psychopharmacology* (1994). <https://doi.org/10.1007/bf02245339>

40. Meerlo, P., Overkamp, G.J.F., Daan, S., Van Den Hoofdakker, R.H., Koolhaas, J.M.: Changes in behaviour and body weight following a single or double social defeat in rats. *Stress* (1996). <https://doi.org/10.3109/10253899609001093>
41. Avgustinovich, D.F., Gorbach, O.V., Kudryavtseva, N.N.: Comparative analysis of anxiety-like behavior in partition and plus-maze tests after agonistic interactions in mice. *Physiol. Behav.* (1997). [https://doi.org/10.1016/s0031-9384\(96\)00303-4](https://doi.org/10.1016/s0031-9384(96)00303-4)
42. Huhman, K.L., et al.: Conditioned defeat in male and female Syrian hamsters. *Horm. Behav.* (2003). <https://doi.org/10.1016/j.yhbeh.2003.05.001>
43. Iñiguez, S.D., et al.: Social defeat stress induces a depression-like phenotype in adolescent male c57BL/6 mice. *Stress* (2014). <https://doi.org/10.3109/10253890.2014.910650>
44. Porsolt, R.D., Le Pichon, M., Jalfre, M.: Depression: a new animal model sensitive to antidepressant treatments. *Nature* (1977). <https://doi.org/10.1038/266730a0>
45. Tornatzky, W., Miczek, K.A.: Long-term impairment of autonomic circadian rhythms after brief intermittent social stress. *Physiol. Behav.* (1993). [https://doi.org/10.1016/0031-9384\(93\)90278-n](https://doi.org/10.1016/0031-9384(93)90278-n)
46. Lehmann, M.L., Geddes, C.E., Lee, J.L., Herkenham, M.: Urine scent marking (USM): a novel test for depressive-like behavior and a predictor of stress resiliency in mice. *PLoS ONE* (2013). <https://doi.org/10.1371/journal.pone.0069822>
47. Malatynska, E., Knapp, R.J.: Dominant-submissive behavior as models of mania and depression. *Neurosci. Biobehav. Rev.* (2005). <https://doi.org/10.1016/j.neubiorev.2005.03.014>
48. Bartolomucci, A., et al.: Chronic psychosocial stress persistently alters autonomic function and physical activity in mice. *Physiol. Behav.* (2003). [https://doi.org/10.1016/s0031-9384\(03\)00209-9](https://doi.org/10.1016/s0031-9384(03)00209-9)
49. Yoshimura, H., Kimura, N.: Ethopharmacology of copulatory disorder induced by chronic social conflict in male mice. *Neurosci. Biobeh. Rev.* (1991). [https://doi.org/10.1016/s0149-7634\(05\)80138-1](https://doi.org/10.1016/s0149-7634(05)80138-1)
50. Steru, L., Chermat, R., Thierry, B., Simon, P.: The tail suspension test: a new method for screening antidepressants in mice. *Psychopharmacology* (1985). <https://doi.org/10.1007/bf00428203>
51. Von Frijtag, J.C., Reijmers, L.G.J.E., Van der Harst, J.E., Leus, I.E., Van den Bos, R., Spruijt, B.M.: Defeat followed by individual housing results in long-term impaired reward- and cognition-related behaviours in rats. *Behav. Brain Res.* (2000). [https://doi.org/10.1016/s0166-4328\(00\)00300-4](https://doi.org/10.1016/s0166-4328(00)00300-4)
52. Meerlo, P., et al.: Changes in daily rhythms of body temperature and activity after a single social defeat in rats. *Physiol. Behav.* (1996). [https://doi.org/10.1016/0031-9384\(95\)02182-5](https://doi.org/10.1016/0031-9384(95)02182-5)
53. Raab, A., et al.: Behavioural, physiological and immunological consequences of social status and aggression in chronically coexisting resident-intruder dyads of male rats. *Physiol. Behav.* (1986). [https://doi.org/10.1016/0031-9384\(86\)90007-7](https://doi.org/10.1016/0031-9384(86)90007-7)
54. Trainor, B.C., et al.: Sex differences in social interaction behavior following social defeat stress in the monogamous California mouse (*Peromyscus californicus*). *PLOS ONE* (2011). <https://doi.org/10.1371/journal.pone.0017405>
55. De Kloet, E.R., Joëls, M., Holsboer, F.: Stress and the brain: from adaptation to disease. *Nat. Rev. Neurosci.* (2005). <https://doi.org/10.1038/nrn1683>
56. Albeck, D., et al.: Chronic social stress alters levels of corticotropin-releasing factor and arginine vasopressin mRNA in rat brain. *J. Neurosci.* (1997). <https://doi.org/10.1523/jneurosci.17-12-04895.1997>
57. Caroline Blanchard, D., et al.: Visible burrow system as a model of chronic social stress: behavioral and neuroendocrine correlates. *Psychoneuroendocrinology* (1995). [https://doi.org/10.1016/0306-4530\(94\)e0045-b](https://doi.org/10.1016/0306-4530(94)e0045-b)

58. Blanchard, D.C., Sakai, R.R., McEwen, B., Weiss, S.M., Blanchard, R.J.: Subordination stress: behavioral, brain, and neuroendocrine correlates. *Behav. Brain Res.* (1993). [https://doi.org/10.1016/0166-4328\(93\)90096-9](https://doi.org/10.1016/0166-4328(93)90096-9)
59. Schuurman, T.: Hormonal correlates of agonistic behavior in adult male rats. *Prog. Brain Res.* (1980). [https://doi.org/10.1016/s0079-6123\(08\)60079-5](https://doi.org/10.1016/s0079-6123(08)60079-5)
60. Wotjak, C.T., et al.: Release of vasopressin within the rat paraventricular nucleus in response to emotional stress: a novel mechanism of regulating adrenocorticotrophic hormone secretion? *J. Neurosci.* (1996). <https://doi.org/10.1523/jneurosci.16-23-07725.1996>
61. Blanchard, R.J., McKittrick, C.R., Blanchard, D.C.: Animal models of social stress: effects on behavior and brain neurochemical systems. *Physiol. Behav.* (2001). [https://doi.org/10.1016/s0031-9384\(01\)00449-8](https://doi.org/10.1016/s0031-9384(01)00449-8)
62. Bhatnagar, S., Vining, C., Iyer, V., Kinni, V.: Changes in hypothalamic-pituitary-adrenal function, body temperature, body weight and food intake with repeated social stress exposure in rats. *J. Neuroendocrinol.* (2006). <https://doi.org/10.1111/j.1365-2826.2005.01375.x>
63. Buwalda, B., et al.: Long-lasting deficient dexamethasone suppression of hypothalamic-pituitary-adrenocortical activation following peripheral CRF challenge in socially defeated rats. *J. Neuroendocrinol.* (1999). <https://doi.org/10.1046/j.1365-2826.1999.00350.x>
64. Fuchs, E., Flügge, G.: Social stress in tree shrews: Effects on physiology, brain function, and behavior of subordinate individuals, pharmacology. *Biochem. Behav.* (2002). [https://doi.org/10.1016/s0091-3057\(02\)00795-5](https://doi.org/10.1016/s0091-3057(02)00795-5)
65. Wood, S.K., Walker, H.E., Valentino, R.J., Bhatnagar, S.: Individual differences in reactivity to social stress predict susceptibility and resilience to a depressive phenotype: role of corticotropin-releasing factor. *Endocrinology* (2010). <https://doi.org/10.1210/en.2009-1026>
66. Tidey, J.W., Miczek, K.A.: Social defeat stress selectively alters meso-corticolimbic dopamine release: an in vivo microdialysis study. *Brain Res.* (1996). [https://doi.org/10.1016/0006-8993\(96\)00159-x](https://doi.org/10.1016/0006-8993(96)00159-x)
67. Buwalda, B., et al.: Long-term effects of social stress on brain and behavior: a focus on hippocampal functioning. *Neurosci. Biobehav. Rev.* (2005). <https://doi.org/10.1016/j.neubio.rev.2004.05.005>
68. Miczek, K.A., Yap, J.J., Covington, H.E.: Social stress, therapeutics and drug abuse: preclinical models of escalated and depressed intake. *Pharmacol. Ther.* (2008). <https://doi.org/10.1016/j.pharmthera.2008.07.006>
69. Krishnan, V., et al.: Molecular adaptations underlying susceptibility and resistance to social defeat in brain reward regions. *Cell* (2007). <https://doi.org/10.1016/j.cell.2007.09.018>
70. Miczek, K.A., Thompson, M.L., Shuster, L.: Opioid-like analgesia in defeated mice. *Science* (1982). <https://doi.org/10.1126/science.7199758>
71. Berton, O., et al.: Essential role of BDNF in the mesolimbic dopamine pathway in social defeat stress. *Science* (2006). <https://doi.org/10.1126/science.1120972>
72. Bohus, B., Koolhaas, J.M., De Ruiter, A.J., Heijnen, C.J.: Stress and differential alterations in immune system functions: conclusions from social stress studies in animals. *Neth. J. Med.* (1991)
73. Miczek, K.A., Maxson, S.C., Fish, E.W., Faccidomo, S.: Aggressive behavioral phenotypes in mice. *Behav. Brain Res.* (2001). [https://doi.org/10.1016/s0166-4328\(01\)00298-4](https://doi.org/10.1016/s0166-4328(01)00298-4)
74. Blanchard, R.J., Blanchard, D.C.: Aggressive behavior in the rat. *Behav. Biol.* (1977). [https://doi.org/10.1016/s0091-6773\(77\)90308-x](https://doi.org/10.1016/s0091-6773(77)90308-x)
75. Palanza, P., Gioiosa, L., and Parmigiani, S.: Social stress in mice: gender differences and effects of estrous cycle and social dominance. *Physiol. Behav.* (2001). [https://doi.org/10.1016/s0031-9384\(01\)00494-2](https://doi.org/10.1016/s0031-9384(01)00494-2)
76. Takahashi, A., et al.: Establishment of a repeated social defeat stress model in female mice. *Sci. Rep.* (2017). <https://doi.org/10.1038/s41598-017-12811-8>

77. Haller, J., Fuchs, E., Halasz, J., Makara, G.B.: Defeat is a major stressor in males while social instability is stressful mainly in females: towards the development of a social stress model in female rats. *Brain Res. Bull.* (1999). [https://doi.org/10.1016/s0361-9230\(99\)00087-8](https://doi.org/10.1016/s0361-9230(99)00087-8)
78. Herzog et al., C.J.: Chronic social instability stress in female rats: a potential animal model for female depression. *Neuroscience* (2009). <https://doi.org/10.1016/j.neuroscience.2009.01.059>
79. Holly, E.N., Shimamoto, A., DeBold, J.F., Miczek, K.A.: Sex differences in behavioral and neural cross-sensitization and escalated cocaine taking as a result of episodic social defeat stress in rats. *Psychopharmacology* (2012). <https://doi.org/10.1007/s00213-012-2846-2>
80. Russo, S.J., Murrough, J.W., Han, M.-H., Charney, D.S., Nestler, E.J.: Neurobiology of resilience. *Nat. Neurosci.* (2012). <https://doi.org/10.1038/nn.3234>
81. Haney, M., Haney, M., Maccari, S., Le Moal, M., Simon, H.: And Pier Vincenzo Piazza, social stress increases the acquisition of cocaine self-administration in male and female rats. *Brain Res.* (1995). [https://doi.org/10.1016/0006-8993\(95\)00788-r](https://doi.org/10.1016/0006-8993(95)00788-r)
82. Tidey, J.W., Miczek, K.A.: Acquisition of cocaine self-administration after social stress: role of accumbens dopamine. *Psychopharmacology* (1997). <https://doi.org/10.1007/s002130050230>
83. Caspi, A., Moffitt, T.E.: Gene–environment interactions in psychiatry: joining forces with neuroscience. *Nat. Rev. Neurosci.* (2006). <https://doi.org/10.1038/nrn1925>
84. Koolhaas, J.M., et al.: Stress revisited: a critical evaluation of the stress concept. *Neurosci. Biobehav. Rev.* (2011). <https://doi.org/10.1016/j.neubiorev.2011.02.003>



VLSI Floorplanning Algorithm Based on Reinforcement Learning with Obstacles

Shenglu Yu^{1,2} and Shimin Du^{1,2}(✉)

¹ Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, Zhejiang, China

dushimin@nbu.edu.cn

² College of Science and Technology, Ningbo University, Ningbo 315300, Zhejiang, China

Abstract. In typical Very Large Scale Integration Circuit (VLSI) designs, some modules have predetermined placements, while the placement of other modules cannot overlap with these pre-placed modules. The presence of fixed modules can complicate floorplanning and make it more challenging. To address this obstructed floorplanning problem, a Reinforcement Learning (RL)-based algorithm for obstructed VLSI floorplanning is proposed. This algorithm uses Sequence Pair (SP) encoding to represent the floorplan structure and leverages RL's ability to learn autonomously and generalize to perform floorplanning. The proposed algorithm is tested on MCNC and GSRC benchmarks, and the experimental results demonstrate that it produces better floorplan solutions.

Keywords: VLSI · Floorplanning · Reinforcement learning · Sequence pair

1 Introduction

According to Moore's Law [1], the number of transistors on a chip doubles every 18 months, this increases the complexity of circuits and greatly reduces the characteristic size of integrated circuits, making floorplanning much more difficult. This problem is known as an NP-hard problem [2, 3]. Floorplanning is a crucial step in VLSI physical design, which involves placing macro modules in appropriate locations and optimizing various design metrics such as area, wire length, and temperature. In Very Large Scale Integration (VLSI) [4] floorplan design, certain macro modules (such as RAM and CPU cores) often need to be placed in specific locations, while other macro modules are placed in the remaining areas of the chip. This will increase the difficulty of the floorplan and lead to the addition of Dead Space (DS) [5], which refers to the gaps between macro modules in the floorplan optimization process.

Heuristic algorithms are an important method for solving VLSI floorplanning problems, and traditional heuristic algorithms such as Simulated Annealing (SA) [6], Genetic Algorithm (GA) [7], and Particle Swarm Optimization (PSO) [8] have been widely used in VLSI floorplanning. However, these methods face challenges as circuit size and complexity increase, particularly in obtaining high-quality solutions for large-scale chip floorplanning problems. Therefore, we need to explore new methods to improve

the efficiency and accuracy of VLSI floorplanning. Reinforcement Learning (RL)-based methods [9] provide a new direction, traditional algorithms operate based on rules, logic, and mathematical models that can solve repeatable problems, but they often struggle to adapt to uncertain and dynamic environments. In contrast, RL is a machine learning technique that can learn from interactive trial-and-error processes and is suitable for handling uncertain and changing environments. As an emerging intelligent algorithm, RL has been widely applied in recent years and has also shown great potential in the field of VLSI floorplanning. Through training, RL agents can place macro modules on chips [10], and Ref. [11] uses graph convolutional networks and RL for floorplanning. He et al. [12] use Q-learning to train agents to learn how to select optimal floorplan neighborhood solutions.

According to the representation of geometric relationships between modules, the floorplan structure can be divided into two types. One is a slicing structure [13], which is widely represented by a binary tree. The other is the more commonly used non-slicing structure, whose floorplan structure representation includes O-tree [14], B*-tree [15], Sequence Pair (SP) [16], etc.

This paper uses the SP method to represent the floorplan structure and proposes a VLSI floorplanning algorithm based on obstructed reinforcement learning. Under certain constraints where some modules have already been placed, this algorithm defines the RL algorithm's state, action, and reward function. The experimental results show that compared with traditional SA algorithms, the proposed algorithm can effectively reduce the floorplan's total area and total wire length, resulting in better floorplan results.

2 Preliminaries

In this section, the background of reinforcement learning is introduced, and the problem of floorplanning is described in detail.

2.1 Reinforcement Learning

Reinforcement learning is the learning of a mapping from states to actions to maximize numerical reward. Essentially all reinforcement learning satisfies the Markov Decision Process (MDP), which includes four key elements:

- 1) State S : A finite set of environmental states.
- 2) Action A : A finite set of actions the reinforcement learning agent takes.
- 3) State transition model $P(s, a, s')$: Indicates the probability of transitioning from state $s \in S$ to the next state $s' \in S$, given action $a \in A$.
- 4) Reward function $R(s, a)$: Represents the numerical reward for taking action $a \in A$ in a state $s \in S$, this reward can be positive, negative, or zero.

MDP relies on the Markov assumption, which states that the future state distribution depends only on the current state. A typical Markov decision process is shown in Fig. 1, where at each time step t , the agent begins at state s_t , takes action a_t , reaches new state s_{t+1} , and receives reward r_t from the environment.

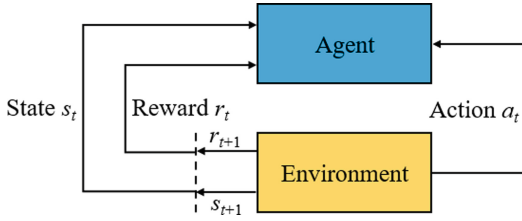


Fig. 1. A typical framework of an MDP.

The objective of an MDP is to find a policy π that maximizes the cumulative total reward. The expression for the total cumulative reward is as follows:

$$R_t = \sum_t^{\infty} \gamma^t r_t \quad (1)$$

where γ is the reward discount factor, t represents the time step, and r represents the reward value at time step t . The state value function $V_{\pi}(s)$ in an MDP is defined as the expected discounted reward for state s under policy π , as defined in Eq. (2):

$$\begin{aligned} V_{\pi}(s) &= E_{\pi}[R_t | s_t = s] \\ &= E_{\pi}\left[\sum_t^{\infty} \gamma^t r_t | s_t = s\right] \end{aligned} \quad (2)$$

where E_{π} represents the expected value of the reward function under policy π . Similarly, the state-action value function $Q_{\pi}(s, a)$ is defined as the expected reward value when taking action a in state s under policy π , the definition is as follows:

$$\begin{aligned} Q_{\pi}(s, a) &= E_{\pi}[R_t | s_t = s, a_t = a] \\ &= E_{\pi}\left[\sum_t^{\infty} \gamma^t r_t | s_t = s, a_t = a\right] \end{aligned} \quad (3)$$

2.2 Description of Floorplanning Problem

Let $B = \{b_i | 1 \leq i \leq n\}$ be a set of rectangular blocks, where each block b_i has a specified width w_i and height h_i , and $N = \{n_i | 1 \leq i \leq m\}$ represents a netlist that describes the connections between the blocks. Floorplanning is to assign a set of coordinates to each block b_i while satisfying the following conditions:

- 1) There is no overlapping between any two blocks.
- 2) Minimize the total chip area and reduce the total wire length.

The optimization objectives based on the minimum rectangle area A and the total wire length W are defined as follows:

$$\min_F A(F) + \alpha W(F) \quad (4)$$

The estimation method for bus length commonly used is the Half-Perimeter Wire Length (HPWL) model [17], where F is a feasible layout solution, α is weight in coefficients between 0 and 1, and min represents minimizing the weighted sum of area A and bus length W as much as possible in the feasible floorplan solution.

Definition 1 (Sequence Pair). A sequence pair (S_+, S_-) is a pair of sequences containing n blocks, with the following constraint relationship applied to each pair of blocks:

$$\begin{aligned} (\langle \dots b_i \dots b_j \dots \rangle, \langle \dots b_i \dots b_j \dots \rangle) &\rightarrow b_i \text{ is left to } b_j, \\ (\langle \dots b_i \dots b_j \dots \rangle, \langle \dots b_j \dots b_i \dots \rangle) &\rightarrow b_i \text{ is below } b_j. \end{aligned}$$

Figure 2 shows a floorplan corresponding to a sequence pair $(\langle 3, 0, 2, 5, 4, 1 \rangle, \langle 5, 0, 1, 2, 3, 4 \rangle)$ consisting of six blocks. The size of each block is as follows: 0(4×3), 1(4×2), 2(3×3), 3(7×3), 4(3×7), and 5(6×3). It can be seen from the figure that all blocks satisfy the above constraint relationship.

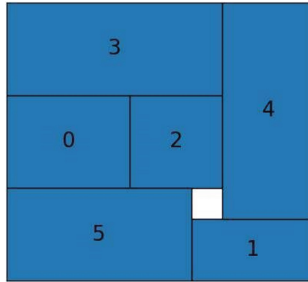


Fig. 2. A floorplan representation by SP.

3 Our Method

The reinforcement learning algorithm used in this paper is the Proximal Policy Optimization (PPO) algorithm, which is a policy gradient (PG) algorithm based on the Actor-critic (AC) architecture. Like Q-learning [18], the Actor-network has both old and new neural networks and periodically updates the old neural network to maximize cumulative rewards. To explore better floorplan solutions, the following MDP is defined.

- 1) State space S : For the floorplanning problem, $s \in S$ is a floorplan solution that includes a complete sequence of pairs (S_+, S_-) and the direction of each block.
- 2) Action space A : A neighboring solution of a floorplan is generated by predefined perturbations in the action space. Here, the following four types of perturbations are defined:
 - a) Swap a random pair of blacks in S_+ .
 - b) Swap a random pair of blacks in S_- .
 - c) Swap a random pair of blacks in both S_+ and S_- .
 - d) Rotate a randomly selected block by 90° .
- 3) State transition P : Given a state, any of the above perturbations will cause the agent to transition to another state, simplifying the probability setting in MDP.

- 4) Reward r : Assigning rewards for an action taken in a state is key to reinforcement learning. In the floorplanning problem in this paper, the objective is to minimize the area and wire length, so we use the reduction of the target cost as the reward. A positive reward is assigned whenever the agent explores a better solution; otherwise, no reward is given. The reward function is defined as follows:

$$r = \begin{cases} F(s) - F(s'), & F(s) \geq F(s') \\ 0, & F(s) < F(s') \end{cases} \quad (5)$$

where r is the reward value, representing the reward value of the current floorplan from state s to state s' through perturbation. F is the optimization objective function defined in Eq. (4).

Our reinforcement learning agent here is an Actor-network that predicts the index of candidate blocks. The action value taken by the agent is calculated by the Critic network, where the state value function is used to calculate the advantage of the action taken using generalized advantage estimation [19], the Actor and Critic networks together form the policy network.

Algorithm 1 Actor-Critic for Floorplan.

Input: Two parameterized functions: policy network $\pi_\theta(a | s)$ and state value network $V_\Phi(s)$

Output: An Optimal policy π^*

- 1: Initialize parameter vectors θ and Φ
 - 2: Initialize learning rates α^θ and α^Φ
 - 3: for $n \in \{1, \dots, N\}$
 - 4: Initialize s_1 (the first state of n)
 - 5: for $t \in \{1, \dots, T\}$
 - 6: Sample an action $a_t \sim \pi_\theta(a_t | s_t)$ from the policy network
 - 7: Receive reward r_t and next state s_{t+1}
 - 8: Calculate the temporal difference error for this step
 - 9: Update the value network
 - 10: Update the policy network
 - 11: end for
 - 12: end for
-

The pseudocode for the Actor-critic algorithm is shown in Algorithm 1. First, the policy parameters θ and state weights Φ are initialized, where the learning rates α^θ and α^Φ are set to 2×10^{-4} and 10^{-3} , respectively. Then, at the beginning of each episode n a random initial state s_1 is generated. At each step t , given the current state s_t , the Actor samples an action a_t from the probability distribution $\pi_\theta(a_t | s_t)$, receives a scalar reward r_t as feedback, and transitions to the next state s_{t+1} according to the reward function and state transition model. The Critic further estimates the temporal difference error based on the reward r_t and its state value estimation update. Finally, the Actor updates the policy parameters θ based on the temporal difference error received from the Critic, while the Critic also uses the same error to update the state weights Φ .

4 Experimental Results and Discussions

4.1 Simulation Settings

The experiments are conducted on a computer with a 3 GHz CPU and 16.00 GB of memory, the proposed algorithm is implemented using the PyTorch library in Python [20]. Adam [21] is the optimizer to train the neural network model, with a discount factor γ set to be 0.992. To save time and avoid overfitting during training, a stopping mechanism was employed where the training would stop if no better solution is found in the last 40 steps.

4.2 Benchmark Test

The proposed algorithm is tested on two standard benchmark circuit sets, MCNC and GSRC, where three blocks are selected from each test circuit to serve as RAM, ROM, and CPU and placed in predetermined locations as obstacles. The results of the placement are compared with SA. In the experiment, all blocks in the test circuits are hard blocks of fixed size, with their basic information shown in Tables 1 and 2.

Table 1. Basic information of the MCNC test circuit.

MCNC	Blocks	I/O pad	Nets	Area ($\times 10^6$)
apte	9	73	97	46.56
xerox	10	2	203	19.35
hp	11	309	83	8.83
ami33	33	42	123	1.16
ami49	49	22	408	35.45

Table 2. Basic information of the GSRC test circuit.

GSRC	Blocks	I/O pad	Nets	Area
n100	100	334	885	179501
n200	200	564	1585	175696
n300	300	569	1893	273170

Optimization of area and wire length is conducted on the MCNC benchmark circuits, and the experimental results are shown in Table 3. DS represents the proportion of gaps between blocks within the floorplan border to the floorplan area. As can be seen from the table, the proposed algorithm has a smaller floorplan area, DS, and wire length than the SA algorithm for all five test circuits. The average improvement of the proposed algorithm over the SA algorithm is 9.2% and 3.4% for wire length and DS, respectively.

Therefore, the proposed algorithm has certain advantages over the SA algorithm in optimizing floorplan area and wire length for the MCNC benchmark circuits.

Table 4 presents a comparison of experimental results for three GSRC test circuits, from which it can be seen that the algorithm proposed in this paper outperforms SA in terms of the floorplan area, wire length, and DS. In the comparison with SA, the algorithm presented in this paper has improved by an average of 11.2% and 8.5% on wire length and DS, respectively. As can be seen from Tables 3 and 4, as the size of the test circuit increases, the wire length and DS for both methods also increase, making floorplan even more difficult. However, the performance of the algorithm proposed in this paper further improves in large-scale circuits compared to SA, with more significant advantages in optimizing floorplan area and wire length.

Table 3. Comparison of experimental results of MCNC test circuit.

MCNC	Area ($\times 10^6$)		Wirelength ($\times 10^5$)		DS (%)	
	Ours	SA	Ours	SA	Ours	SA
apte	47.82	48.73	0.52	0.61	2.63	4.45
xerox	20.75	21.59	0.57	0.68	6.75	10.38
hp	9.26	9.52	0.41	0.55	4.66	7.25
ami33	1.27	1.32	0.68	0.81	8.66	12.12
ami49	38.95	41.52	6.47	6.79	8.99	14.62
Average	23.46	24.51	1.73	1.89	6.34	9.76
Normalization	1.000	1.045	1.000	1.092	1.000	1.034

Table 4. Comparison of experimental results of GSRC test circuit.

GSRC	Area ($\times 10^5$)		Wirelength ($\times 10^5$)		DS (%)	
	Ours	SA	Ours	SA	Ours	SA
n100	1.98	2.23	1.97	2.21	9.09	19.28
n200	2.15	2.37	3.46	3.87	18.14	25.74
n300	3.42	3.78	5.03	5.66	20.18	27.78
Average	2.52	2.79	3.49	3.91	15.80	24.27
Normalization	1.000	1.107	1.000	1.112	1.000	1.085

4.3 Floorplan Visualization

On the MCNC benchmark, the floorplan of the ami49 test circuit generated by our algorithm (a) and SA algorithm (b) is shown in Fig. 3. The DS of the floorplan generated by our algorithm is only 9.0%, while that of the SA algorithm is 14.6%. On the GSRC

benchmark, the floorplan of the n100 test circuit generated by our algorithm (c) and SA algorithm (d) is shown in Fig. 4. The DS of the floorplan generated by our algorithm is only 9.1%, while that of the SA algorithm is 19.3%, thus verifying the effectiveness of our algorithm.

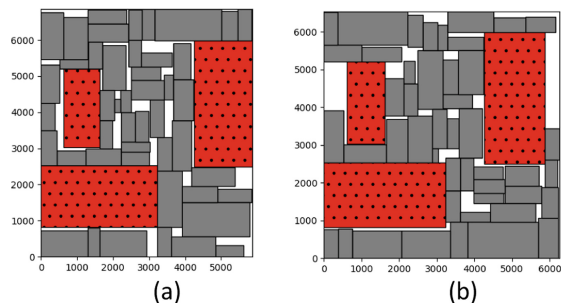


Fig. 3. Comparison of ami49 circuit floorplan results.

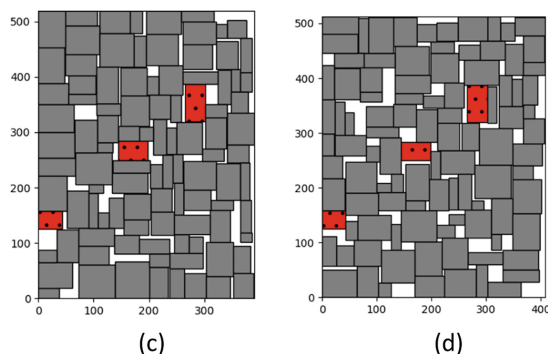


Fig. 4. Comparison of n100 circuit floorplan results.

5 Conclusion

This paper researches the floorplanning problem in the integrated circuit design process and proposes a VLSI floorplanning algorithm based on reinforcement learning with obstacles. The constructed reinforcement learning framework and designed reward function can be effectively implemented to generate better-performing floorplan designs. Through experiments with a test circuit set, the results show that the proposed algorithm is superior to the traditional SA algorithm, and can effectively reduce the total chip area and wire length. In recent years, machine learning-based methods have been increasingly introduced into the EDA field, in the next step, we will explore the use of graph neural network methods to solve the floorplanning problem.

Acknowledgement. This research is funded by the Graduate Education Practice Project of Ningbo University (YJD202305), the Science and Technology Innovation 2025 Major Project of Ningbo City (2022Z203).

References

1. Moore, G.E.: Cramming more components onto integrated circuits. *Proc. IEEE* **86**(1), 82–85 (1998)
2. Sherwani, N.A.: *Algorithms for VLSI Physical Design Automation*. Springer Science & Business Media (2012)
3. Wong, D.F., Liu, C.L.: Floorplan design of VLSI circuits. *Algorithmica* **4**(1–4), 263–291 (1989)
4. Mead, C., Conway, L.: *Introduction to VLSI systems* (1980)
5. Singh, R.B., Baghel, A.S.: Dead space reduction of floorplan using simulated annealing algorithm. In: *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pp. 2108–2112. IEEE (2017)
6. Murata, H., Fujiyoshi, K., Nakatake, S., et al.: VLSI module placement based on rectangle-packing by the sequence-pair. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **15**(12), 1518–1524 (1996)
7. Hung, W.L., Xie, Y., Vijaykrishnan, N., et al.: Thermal-aware floorplanning using genetic algorithms. In: *Sixth International Symposium on Quality Electronic Design (ISQED'05)*, pp. 634–639. IEEE (2005)
8. Prakash, A., Lal, R.K.: Floorplanning for area optimization using parallel particle swarm optimization and sequence pair. *Wireless Pers. Commun.* **118**(1), 323–342 (2021)
9. Sutton, R.S., Barto, A.G.: *Reinforcement Learning: An Introduction*. MIT Press (2018)
10. Mirhoseini, A., Goldie, A., Yazgan, M., et al.: A graph placement methodology for fast chip design. *Nature* **594**(7862), 207–212 (2021)
11. Xu, Q., Geng, H., Chen, S., et al.: GoodFloorplan: graph convolutional network and reinforcement learning-based floorplanning. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **41**(10), 3492–3502 (2021)
12. He, Z., Ma, Y., Zhang, L., et al.: Learn to floorplan through acquisition of effective local search heuristics. In: *2020 IEEE 38th International Conference on Computer Design (ICCD)*, pp. 324–331. IEEE (2020)
13. Otten, R.H.J.M.: Automatic floorplan design. In: *19th Design Automation Conference*, pp. 261–267. IEEE (1982)
14. Guo, P.N., Cheng, C.K., Yoshimura, T.: An O-tree representation of non-slicing floorplan and its applications. In: *Proceedings of the 36th Annual ACM/IEEE Design Automation Conference*, pp. 268–273 (1999)
15. Shanthi, J., Rani, D.G.N., Rajaram, S.: An enhanced memetic algorithm using SKB tree representation for fixed-outline and temperature driven non-slicing floorplanning. *Integration* **86**, 84–97 (2022)
16. Tang, X., Wong, D.F.: FAST-SP: a fast algorithm for block placement based on sequence pair. In: *Proceedings of the 2001 Asia and South Pacific Design Automation Conference*, pp. 521–526 (2001)
17. Shahookar, K., Mazumder, P.: VLSI cell placement techniques. *ACM Comput. Surv. (CSUR)* **23**(2), 143–220 (1991)
18. Watkins, C.J.C.H., Dayan, P.: Q-learning. *Mach. Learn.* **8**, 279–292 (1992)
19. Schulman, J., Moritz, P., Levine, S., et al.: High-dimensional continuous control using generalized advantage estimation. *arXiv preprint [arXiv:1506.02438](https://arxiv.org/abs/1506.02438)* (2015)

20. Paszke, A., Gross, S., Chintala, S., et al.: Automatic differentiation in PyTorch (2017)
21. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014). Author, F.: Article title. Journal **2**(5), 99–110 (2016)

Author Index

A

Abdivakhidov, Kamaliddin I-453
Adib, Abdellah I-159
Aida, Saori I-21, I-27
Aleksandrov, A. A. I-595
Aleksandrov, Aleksander A. I-14
Alicea, Bradly I-33
Almeida, David Alberto Vique I-912
Altshuler, Julia I-41
Alva, Ruth A. Bastidas I-83
Alyushin, Alexander M. I-50
Alyushin, Mikhail V. I-56
Alyushin, Victor M. I-50
Anchekov, Murat I-382
Anchokov, Murat I-151, I-545
Andreyuk, Denis S. I-577
Andreyuk, Denis I-832
Andronenko, Andrey I-62
Antonov, E. V. I-884
Apshev, Artur I-706
Arakelova, Irina I-643, I-651
Arora, Sankalp I-995
Artamonov, A. A. I-884
Atalikov, Boris I-706
Atreides, Kyrтин I-70
Avshalumov, Mikhail I-62

B

Balykov, Maksim I-767
Bazhenkov, Nikolay I-169
Belozerov, Aleksandr I-96
Bespalova, Natalia I-104
Bitney, Vladislav D. I-56
Blagosklonov, Nikolay A. I-113
Boboshko, Maria Yu. I-14
Boiko, Sofia A. I-512
Boltuc, Peter I-3
Bondarev, Sergey I-651
Budzko, Vladimir I-121, I-135
Burlakov, Evgenii I-904
Bylevsky, Pavel I-453

Bystrov, Oleg V. I-143
Bzhikhatlov, Kantemir I-151, I-382, I-706

C

Cai, Xiaojie I-177
Chai, Yanjie I-188, I-197, I-231
Chaivanov, Dmitry B. I-216
Chakhtouna, Adil I-159
Chaplinskaia, Nadezhda I-169
Chartier, Sylvain I-730
Chekulaeva, Ekaterina I. I-577
Chen, Song I-177
Chen, Yang I-390
Chen, Yi I-177
Chen, Yongqi I-390
Cheng, Jiajun I-188, I-197
Chepin, Evgenii I-208
Chistikov, Matvei I-832
Chudina, Yuliya A. I-216
Coletta, Luiz Fernando I-750
Colombini, Esther Luna I-750
Conteron, Luis Armando Chicaiza I-912
Costa, Paula Dornhofer Paro I-750

D

da Silva Simões, Alexandre I-750
da Silva, Anderson Anjos I-750
Dai, Xuyao I-231
Davydov, Yury I-741
Del Signore, Ken I-242
Demarev, Andrey I-277
Demareva, Valeriia I-41, I-277
Demin, Vyacheslav I-62
Diamant, Emanuel I-285
Didkovsky, Nikolay A. I-577
Dionne, Karley I-294
Dolenko, Sergei A. I-348
Dolenko, Sergey I-406
Dolgikh, Anatoly A. I-312
Dorokhov, Vladimir B. I-894
Du, Shimin I-1034

Dvoryankin, Sergey V. I-50
 Dzhivelikian, Evgenii I-330

E

Enes, Ahmed I-382
 Erlou, Margarita I-208
 Ershov, Alexey I-104

F

Fadeicheva, Galina I-643
 Fahlman, Scott E. I-3
 Fedorov, Klim A. I-627
 Filippova, Ekaterina A. I-577
 Fu, Jiaqi I-340

G

Gadzhiev, Ismail M. I-348
 Gallardo, Edison Gonzalo Espinosa I-912
 Gamez, David I-422
 Gapanyuk, Yuriy E. I-356
 Garbaruk, Ekaterina S. I-14
 Gavrilkina, Anastasia I-367
 Ghatnekar, Vedant I-845
 Golitsina, Olga I-367
 Gorbov, Evgeniy A. I-695
 Gridnev, Alexander I-208
 Grig, Aliya I-373
 Gudwin, Ricardo Ribeiro I-750
 Guo, Cheng Jun I-962
 Gurtovoy, Konstantin I-904, I-956
 Gurtueva, Irina I-382, I-545

H

Hendrikse, Sophie I-3
 Huang, Wanqing I-390
 Huyck, Chris I-422

I

Ignatev, Andrey I-932
 Ionkina, K. V. I-884
 Isaev, Igor I-406
 Ismailova, Larisa I-973
 Ivanov, Mikhail I-104, I-453
 Ivlev, Vitaly Y. I-356
 Ivlev, Vladimir I. I-414

J

Ji, Yuehu I-422
 Jiang, Zhidi I-980

K

Kaganov, Yuriy T. I-356
 Kalmykova, M. V. I-659
 Kaminchenko, Dmitry I. I-695
 Kankulov, Sultan I-706
 Karabulatova, Irina S. I-356
 Karachurin, Raul I-504
 Kartashov, S. I. I-462, I-659
 Kartashov, Sergey I. I-666
 Kartashov, Sergey I-477, I-612
 Katti, Sudaman I-845
 Kazakov, M. A. I-429
 Kazakova, E. M. I-436
 Keyer, Petr I-135
 Khabarov, Dmitry I-444
 Khazova, Maria L. I-894
 Khidirova, Mohiniso I-453
 Klimov, Valentin I-96, I-553
 Knyazeva, V. M. I-595
 Knyazeva, Veronika M. I-14
 Kobrinskii, Boris A. I-113
 Kolobashkina, Lyubov V. I-50, I-56
 Kolonin, Anton I-14
 Korosteleva, A. N. I-462
 Korosteleva, Anastasia N. I-566
 Kosikov, Sergey I-973
 Kostyuk, Georgy P. I-577
 Kostyuk, Georgy I-485, I-612
 Kotov, A. A. I-462
 Kotov, Artemy I-932
 Kozlov, Maksim I-946
 Kozlov, Stanislav I-469, I-477
 Krupina, Ekaterina I-485
 Krynskiy, Sergey A. I-577
 Kuderov, Petr I-330
 Kudriashov, Anton V. I-495
 Kudryashov, N. A. I-759
 Kudryashov, Nikolay I-504
 Kulik, Sergey D. I-143
 Kulik, Sergey I-815, I-853
 Kumar, Akash I-995
 Kunitsyn, Dmitry I-767
 Kupriyanov, Dmitry I-104, I-453

L

Ladygin, Stanislav I-504
 Laird, John I-3
 Leonov, Pavel Y. I-512, I-520, I-860
 Leonovich, Daria I-528
 Li, Yuejiao I-671

Lieto, Antonio I-3
 Liu, Tingting I-3, I-188, I-197, I-231, I-787
 Liu, Zhen I-188, I-197, I-231, I-787
 Lyutikova, L. A. I-536

M

Makarov, Alexander S. I-348
 Makoeva, Dana I-545
 Maksimov, Nikolay I-367, I-553
 Malakhov, Denis G. I-894
 Malanchuk, I. G. I-595
 Malanchuk, Irina G. I-216, I-566
 Malashenkova, Irina K. I-577
 Malynov, Andrey I-588
 Mamedova, Galina I-485
 Manzhurtsev, Andrei I-485
 Mao, Xiaoyan I-390
 Martynova, Olga I-904
 Medennikov, Victor I-121, I-135
 Medina, José Luis Carrillo I-912
 Medvedeva, Olga I-651
 Memetova, K. S. I-595
 Mishra, Amit Kumar I-604, I-688
 Misyurin, Sergej Yu. I-414
 Moloshnikov, I. A. I-759
 Morozov, Nikolay I-635, I-643, I-651
 Mosina, Larisa I-485, I-612

N

Nagoeva, Olga I-151, I-382, I-545
 Nazarko, Mikhail Yu. I-627
 Nazarov, Nicolay I-277
 Nechaev, Sergey I-104
 Nifontov, D. R. I-759
 Nikolaev, Andrey A. I-216
 Norkina, Anna I-635, I-643, I-651
 Nosova, Svetlana I-635, I-643, I-651

O

Obornev, Eugeny I-406
 Obornev, Ivan I-406
 Ogawa, Yuui I-27
 Ogurtsov, Daniil P. I-577
 Okhlopov, Andrey V. I-56
 Orlov, V. A. I-659
 Orlov, Vyacheslav A. I-577, I-666, I-894
 Orlov, Vyacheslav I-469, I-477, I-612
 Oseledchik, Mikhail B. I-356
 Osetrova, Maria I-485

Osipov, Alexey I-453

P

Pan, Tiejun I-340, I-671
 Panov, Aleksandr I. I-330
 Pavlenko, Ivan A. I-679
 Pederson, Thomas I-688
 Peña, Frank W. Zarate I-83
 Petrunin, Yuri I-832
 Petukhov, Alexandr Y. I-695
 Pleshakova, Ekaterina I-453
 Polevaya, Sofia A. I-695
 Poma, Angie L. Herrera I-83
 Poyda, A. A. I-659
 Poyda, Alexey A. I-666
 Poyda, Alexey I-469, I-477
 Prokhorov, Igor I-588
 Pshenokova, Inna I-706

R

Radygin, Victor I-104, I-453
 Raghavachary, Saty I-716
 Red'ko, Vladimir G. I-723
 Rizzo, Anastasia I-373
 Robertson, Paul I-3
 Rodionov, Eugeny I-406
 Rolon-Mérette, Damiem I-730
 Rolon-Mérette, Thaddé I-730
 Romanovsky, Valentin A. I-520
 Ryabov, Pavel I-504
 Rybka, R. B. I-759
 Rybka, Roman I-741

S

Sakabe, Eduardo Yuji I-750
 Samsonovich, Alexei V. I-3, I-14, I-312,
 I-348, I-444, I-627, I-679
 Savin, Konstantin I-867
 Sboev, A. G. I-759
 Sboev, Alexander I-741, I-767
 Schneider, Howard I-775
 Sekkate, Sara I-159
 Semenov, Dmitry V. I-356
 Serenko, Alexey I-741, I-767
 Shan, Yuduo I-787
 Shevaldova, Olga I-800
 Shibzukhov, Z. M. I-807
 Shilnikov, Kirill I-504
 Shimelevich, Mikhail I-406

Shtanko, Alexander I-815
 Shumsky, Sergey I-823
 Shuranova, Anna I-832
 Singh, Shweta I-845
 Sitnikov, Sergey I-104
 Sofronov, Ivan I-853
 Stankevich, L. N. I-595
 Stepanenkova, Margarita A. I-512
 Sun, Ron I-3
 Sushkov, Viktor M. I-512, I-520, I-860
 Suslina, Anastasiia I-867
 Suslina, Irina I-867

T

Takeno, Junichi I-3
 Talanov, Max I-875
 Tao, Chenchen I-177
 Terekhov, Valery I. I-356
 Tikhomirova, Anna I-924
 Tikhomirova, Daria V. I-348
 Treur, Jan I-3, I-294
 Trushchelev, Sergey A. I-577
 Trushchelev, Sergey I-485, I-612
 Tukumbetova, R. R. I-884

U

Ublinskiy, Maxim I-485
 Ulizko, M. S. I-884
 Untilova, Arina I-946
 Ushakov, Vadim L. I-577, I-659, I-666,
 I-894
 Ushakov, Vadim I-469, I-485, I-612, I-832

V

Vanina, Margarita I-104
 Vartanov, Alexander I-528, I-800
 Vasilyeva, Marina J. I-14
 Verkhlyutov, Vitaly I-904, I-956
 Verma, Kaveri I-995
 Vermeer, Maya I-294

Villa, Valeryia E. Perez I-83
 Vishnevsky, Stanislav V. I-520
 Vitko, Kirill I-924
 Vlasov, Danila I-741
 Volkova, Liliya I-932, I-946
 Vvedensky, Victor I-904, I-956

W

Wang, Chao I-962
 Wang, Chong I-177
 Wolfengagen, Viacheslav I-973

X

X.Duan, Simon I-323
 Xue, Yuhui I-980
 Xue, Zichu I-340

Y

Yadav, Sourav I-995
 Yakovlev, Andrei N. I-679
 Yan, Chenyang I-1007
 Yang, Shuya I-1016
 Yarnykh, Vasily I-485
 Yu, Feiyu I-390
 Yu, Jinjie I-671
 Yu, Mudan I-980
 Yu, Shenglu I-1034

Z

Zakharova, Natalia V. I-577
 Zakharova, Natalia I-485, I-612
 Zakirov, Arthur D. I-679
 Zavertyaev, S. V. I-759
 Zayceva, Irina I-277
 Zhang, Tao I-390
 Zhemchuzhnikov, Artur I-477
 Zheng, Leina I-340, I-671
 Zhigulina, Polina E. I-894
 Zhong, Yi I-604