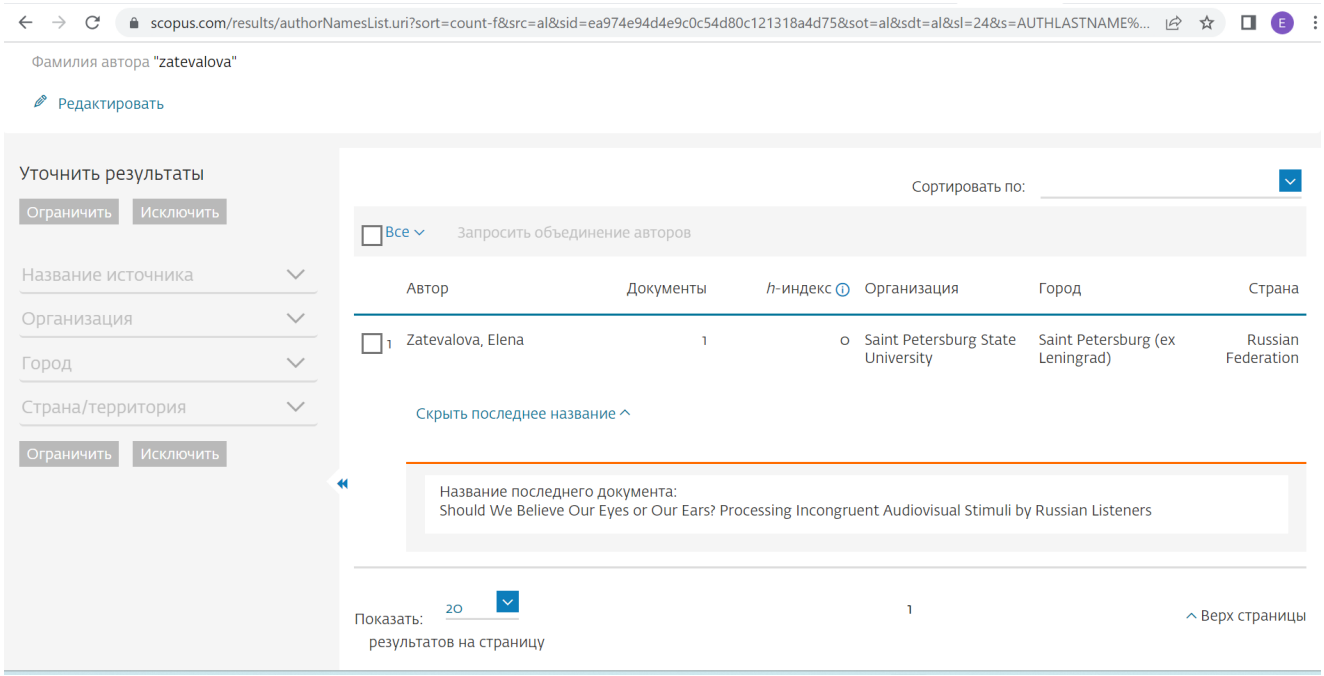
scopus.com/results/authorNamesList.uri?sort=count-f&src=al&sid=ea974e94d4e9c0c54d80c121318a4d75&sot=al&sdt=al&sl=24&s=AUTHLASTNAME%...
Фамилия автора "zatevalova"
Редактировать
Уточнить результаты
Ограничить
Исключить
Название источника
Организация
Город
Страна/территория
Ограничить
Исключить
Сортировать по:
Все
Запросить объединение авторов
Автор
Документы
h-индекс
Организация
Город
Страна
Zatevalova, Elena
1
0
Saint Petersburg State University
Saint Petersburg (ex Leningrad)
Russian Federation
Скрыть последнее название
Название последнего документа:
Should We Believe Our Eyes or Our Ears? Processing Incongruent Audiovisual Stimuli by Russian Listeners
Показать:
20
результатов на страницу
1
Верх страницы

S. R. Mahadeva Prasanna · Alexey Karpov ·
K. Samudravijaya · Shyam S. Agrawal (Eds.)

Speech and Computer

24th International Conference, SPECOM 2022
Gurugram, India, November 14–16, 2022
Proceedings

Springer

Editors
S. R. Mahadeva Prasanna
Indian Institute of Technology Dharwad
Dharwad, India

K. Samudravijaya
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India

Alexey Karpov
St. Petersburg Federal Research Center
of the Russian Academy of Sciences
St. Petersburg, Russia

Shyam S. Agrawal
KIIT Group of Colleges
Gurugram, India

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-031-20979-6 ISBN 978-3-031-20980-2 (eBook)
https://doi.org/10.1007/978-3-031-20980-2

LNCS Sublibrary: SL7 – Artificial Intelligence

© Springer Nature Switzerland AG 2022
This work is subject to copyright. All rights are reserved 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

Contents

Thematic Diversity of Everyday Russian Discourse: A Case Study Based on the ORD Corpus 1
Eleonora Akinshina and Tatiana Sherstinova

Neural Embedding Extractors for Text-Independent Speaker Verification 10
Jahangir Alam, Woohyun Kang, and Abderrahim Fathan

Deep Speaker Embeddings Based Online Diarization 24
Anastasia Avdeeva and Sergey Novoselov

Overlapped Speech Detection Using AM-FM Based Time-Frequency Representations 33
Shikha Baghel, S. R. M. Prasanna, and Prithwijit Guha

Significance of Dimensionality Reduction in CNN-Based Vowel Classification from Imagined Speech Using Electroencephalogram Signals 44
Oindrila Banerjee, D. Govind, Akhilesh Kumar Dubey, and Suryakanth V. Gangashetty

Study of Speech Recognition System Based on Transformer and Connectionist Temporal Classification Models for Low Resource Language 56
Shweta Bansal, Shambhu Sharan, and Shyam S. Agrawal

An Initial Study on Birdsong Re-synthesis Using Neural Vocoders 64
Rhythm Rajiv Bhatia and Tomi H. Kinnunen

Speech Music Overlap Detection Using Spectral Peak Evolutions 75
Mrinmoy Bhattacharjee, S. R. Mahadeva Prasanna, and Prithwijit Guha

Influence of Accented Speech in Automatic Speech Recognition: A Case Study on Assamese L1 Speakers Speaking Code Switched Hindi-English 87
Joyshree Chakraborty, Rohit Sinha, and Priyankoo Sarmah

ClusterVote: Automatic Summarization Dataset Construction with Document Clusters 99
Daniil Chernyshev and Boris Dobrov

Comparing Unsupervised Detection Algorithms for Audio Adversarial Examples 114
Shanatip Choosaksakunwiboon, Karla Pizzi, and Ching-Yu Kao

Celtic English Continuum in Pitch Patterns of Spontaneous Talk: Evidence of Long-Term Contacts 128
Maria Chubarova and Tatiana Shevchenko

Coherence Based Automatic Essay Scoring Using Sentence Embedding and Recurrent Neural Networks 139
Dadi Ramesh and Suresh Kumar Sanampudi

Analysis of Automatic Evaluation Metric on Low-Resourced Language: BERTScore vs BLEU Score 155
Goutam Datta, Nisheeth Joshi, and Kusum Gupta

DyCoDa: A Multi-modal Data Collection of Multi-user Remote Survival Game Recordings 163
Denis Dresvyanskiy, Yamini Sinha, Matthias Busch, Ingo Siegert, Alexey Karpov, and Wolfgang Minker

On the Use of Ensemble X-Vector Embeddings for Improved Sleepiness Detection 178
José Vicente Egas-López, Róbert Busa-Fekete, and Gábor Gosztolya

Multiresolution Decomposition Analysis via Wavelet Transforms for Audio Deepfake Detection 188
Abderrahim Fathan, Jahangir Alam, and Woohyun Kang

Automatic Rhythm and Speech Rate Analysis of Mising Spontaneous Speech 201
Parismita Gogoi, Priyankoo Sarmah, and S. R. M. Prasanna

An Electroglottographic Method for Assessing the Emotional State of the Speaker 214
Aleksey Grigorev, Anna Kurazhova, Egor Kleshnev, Aleksandr Nikolaev, Olga Frolova, and Elena Lyakso

Significance of Distance on Pop Noise for Voice Liveness Detection 226
Priyanka Gupta and Hemant A. Patil

CRIM's Speech Recognition System for OpenASR21 Evaluation with Conformer and Voice Activity Detector Embeddings 238
Vishwa Gupta and Gilles Boulianne

Joint Changes in First and Second Formants of /a/, /i/, /u/ Vowels in Babble Noise - a New Statistical Approach 252
Alisa P. Gvozdeva, Alexander M. Lunichkin, Larisa G. Zaytseva, Elena A. Ogorodnikova, and Irina G. Andreeva

Comparing NLP Solutions for the Disambiguation of French Heterophonic Homographs for End-to-End TTS Systems 265
Maria-Loulou Hajj, Martin Lenglet, Olivier Perrotin, and Gérard Bailly

Detection of Speech Related Disorders by Pre-trained Embedding Models Extracted Biomarkers 279
Attila Zoltán Jenei, Gábor Kiss, and Dávid Sztahó

Multi-label Dysfluency Classification 290
Melanie Jouaiti and Kerstin Dautenhahn

Harnessing Uncertainty - Multi-label Dysfluency Classification with Uncertain Labels 302
Melanie Jouaiti and Kerstin Dautenhahn

Continuous Wavelet Transform for Severity-Level Classification of Dysarthria 312
Aastha Kachhi, Anand Therattil, Priyanka Gupta, and Hemant A. Patil

Significance of Energy Features for Severity Classification of Dysarthria 325
Aastha Kachhi, Anand Therattil, Ankur T. Patil, Hardik B. Sailor, and Hemant A. Patil

An Analytic Study on Clustering-Based Pseudo-labels for Self-supervised Deep Speaker Verification 338
Woo Hyun Kang, Jahangir Alam, and Abderrahim Fathan

Investigation of Transfer Learning for End-to-End Russian Speech Recognition 349
Irina Kipyatkova

Prosodic Features of Verbal Irony in Russian and French: Universal vs. Language-Specific 358
Uliana Kochetkova, Pavel Skrelin, Rada German, and Daria Novoselova

Categorization of Threatening Speech Acts 372
Liliya Komalova and Lyubov Kalyuzhnaya

Assessment of Speech Quality During Speech Rehabilitation Based on the Solution of the Classification Problem 382
Evgeny Kostyuchenko, Ivan Rakhmanenko, and Lidiya Balatskaya

Multi-level Fusion of Fisher Vector Encoded BERT and Wav2vec 2.0 Embeddings for Native Language Identification 391
Dani Krebbers, Heysem Kaya, and Alexey Karpov

Fake Speech Detection Using OpenSMILE Features 404
Devesh Kumar, Pavan Kumar V. Patil, Ayush Agarwal, and S. R. Mahadeva Prasanna

Nonverbal Constituents of Argumentative Discourse: Gesture and Prosody Interaction 416
Anna Leonteva and Tatiana Sokoreva

Classifying Mahout and Social Interactions of Asian Elephants Based on Trumpet Calls 426
Seema Lokhandwala, Priyankoo Sarmah, and Rohit Sinha

Recognition of the Emotional State of Children with Down Syndrome by Video, Audio and Text Modalities: Human and Automatic 438
Elena Lyakso, Olga Frolova, Anton Matveev, Yuri Matveev, Aleksey Grigorev, Olesia Makhnytkina, and Nersisson Ruban

Fake Speech Detection Using Modulation Spectrogram 451
Raghav Magazine, Ayush Agarwal, Anand Hedge, and S. R. Mahadeva Prasanna

Self-Configuring Genetic Programming Feature Generation in Affect Recognition Tasks 464
Danila Mamontov, Wolfgang Minker, and Alexey Karpov

A Multi-modal Approach to Mining Intent from Code-Mixed Hindi-English Calls in the Hyperlocal-Delivery Domain 477
Jose Mathew, Pranjal Sahu, Bhavuk Singhal, Aniket Joshi, Krishna Reddy Medikonda, and Jairaj Sathyanarayana

Importance of Supra-Segmental Information and Self-Supervised Framework for Spoken Language Diarization Task 494
Jagabandhu Mishra and S. R. Mahadeva Prasanna

Low-Resource Emotional Speech Synthesis: Transfer Learning and Data Requirements 508
Anton Nesterenko, Ruslan Akhmerov, Yulia Matveeva, Anna Goremykina, Dmitry Astankov, Evgeniy Shuranov, and Alexandra Shirshova

Fuzzy Classifier for Speech Assessment in Speech Rehabilitation 522
Dariya Novokhrestova, Ilya Hodashinsky, Evgeny Kostyuchenko, Konstantin Sarin, and Marina Bardamova

Analysis-By-Synthesis Modeling of Bengali Intonation 533
Moumita Pakrashi and Shakuntala Mahanta

Neural Network Based Curve Fitting to Enhance the Intelligibility of Dysarthric Speech 545
K. S. Pavithra, H. M. Chandrashekar, and Veena Karjigi

Personalizing Retrieval-Based Dialogue Agents 554
Pavel Posokhov, Anastasia Matveeva, Olesia Makhnytkina, Anton Matveev, and Yuri Matveev

Forensic Identification of Foreign-Language Speakers by the Method of Structural-Melodic Analysis of Phonograms 567
Rodmonga Potapova, Vsevolod Potapov, and Irina Kuryanova

Logistics Translator. Concept Vision on Future Interlanguage Computer Assisted Translation 579
Rodmonga Potapova, Vsevolod Potapov, and Oleg Kuzmin

Analysis of Time-Averaged Feature Extraction Techniques on Infant Cry Classification 590
Aditya Pusuluri, Aastha Kachhi, and Hemant A. Patil

Should We Believe Our Eyes or Our Ears? Processing Incongruent Audiovisual Stimuli by Russian Listeners 604
Elena Riekhakaynen and Elena Zatevalova

Emotional Speech Recognition Based on Lip-Reading 616
Elena Ryumina and Denis Ivanko

Exploring the Use of Machine Learning for Resume Recommendations 626
Anna Shestakova and Andrea Corradini

The Role of Pause in Interaction: A Case of Polylogue 641
Tatiana Sokoreva and Tatiana Shevchenko

Dictionary with the Evaluation of Positivity/Negativity Degree of the Russian Words 651
Valery Solovyev, Musa Islamov, and Venera Bayrasheva

Effects of Depth of Field on Focus Using a Virtual Reality Escape Room 665
Nikolaos Tsiftsis, Konstantinos Moustakas, and Nikolaos Fakotakis

Dynamics of Frequency Characteristics of Visually Evoked Potentials of Electroencephalography During the Work with Brain-Computer Interfaces 676
Yaroslav Turovsky, Daniyar Wolf, Roman Meshcheryakov, and Anastasia Iskhakova

Device Robust Acoustic Scene Classification Using Adaptive Noise Reduction and Convolutional Recurrent Attention Neural Network 688
Spoorthy Venkatesh and Shashidhar G. Koolagudi

Comparison of Word Embeddings of Unaligned Audio and Text Data Using Persistent Homology 700
Zhandos Yessenbayev and Zhanibek Kozhirbayev

Low-Cost Training of Speech Recognition System for Hindi ASR Challenge 2022 712
Alexander Zatvornitskiy

Author Index 719

Should We Believe Our Eyes or Our Ears? Processing Incongruent Audiovisual Stimuli by Russian Listeners

Elena Riekhakaynen(✉) and Elena Zatevalova

Saint-Petersburg State University, Universitetskaya Emb. 7/9, 199034 St. Petersburg, Russia
e.riehakajnen@spbu.ru

Abstract. In this paper, we describe the pilot study aimed at finding out those combinations of auditory syllables and lip movements for which the misinterpretation of auditory information because of the incongruent visual one would be the strongest for Russian listeners. We conducted an experiment where 60 schoolchildren and 60 adults processed congruent and incongruent audiovisual stimuli (the syllables containing one of six Russian consonants /t/, /d/, /p/, /b/, /f/, /v/ and the vowel /a/ pronounced by one female speaker). Most often we observed the visual dominance in the pairs "labial stop consonant in the auditory channel – labiodental fricative in the visual channel", i.e., baVA and paFA. The labial stops were most often substituted in responses to other sounds. Audiovisual integration was more prominent in adults than in schoolchildren, although the average number of mistakes did not differ much. We did not observe the effect of the preferred perceptual modality on the recognition of auditory stimuli which supports the previous findings in the field. Further studies can include the experiments with the data from several speakers and with other Russian consonants. The results of the study contribute to better understanding of multichannel processing and can be presumably taken into account in automatic audiovisual recognition.

Keywords: Audiovisual integration · McGurk effect · Russian

1 Introduction

When perceiving speech in a situation of direct contact with a speaker, we not only use auditory information, but also process the gestures and facial expressions of the speaker and correlate them with what we hear. The interaction of different modalities is often discussed in studies that focus on learning [1, 2, etc.]. Many of these studies draw on the Cognitive Theory of Multimedia Learning (CTML) [3]. This theory, among other things, postulates the existence of two independent channels of information processing – auditory and visual. Both channels have limited bandwidth. Apparently, due to this limitation, in the process of natural communication, we try to combine the information coming through these two channels. This process is called a multimodal association – a synergistic use of information received from different modalities. The term can refer to any stage of the integration process where there is a combination of different sources

© Springer Nature Switzerland AG 2022
S. R. M. Prasanna et al. (Eds.): SPECOM 2022, LNAI 13721, pp. 604–615, 2022.
https://doi.org/10.1007/978-3-031-20980-2_51

of information. In recent studies carried out on the material of the Russian language, it was shown that a multimodal text that combines auditory text and its written summary is perceived better than only auditory or only written one [4, 5]. It is believed that cross-modal merging increases the reliability of the system (both a cognitive and an automatic one) in case of an error or failure [6].

Moreover, the interaction of auditory and visual modalities can occur not only at high levels of perception (such as text or word processing), but also at lower ones. In particular, it has been shown that visual information about the articulation of sounds affects their auditory perception. This influence is called audiovisual integration. The nature of this phenomenon is not fully understood (see [7] for the overview). For instance, it is not clear what factors can enhance or weaken audiovisual integration: whether it depends on specific sounds, as well as on the individual characteristics of the speaker and listener. Recent studies of the McGurk effect have shown that this effect is not automatic. Thus, the "two-stage model of audiovisual fusion" is developed which includes the binding stage that is followed by the fusion one [8]. The binding is believed to be highly contextual [9].

The problem of audiovisual integration is crucial not only for cognitive studies, but also for automatic speech synthesis [10–12] and recognition [13] as well as for such practical issue as dubbing. We believe that the experimental evidence on how auditory and visual information are interconnected at low levels (i.e., individual sounds and their articulation) can be used to improve automatic audiovisual speech recognition systems, and also be taken into account in audiovisual synthesis. For example, for audiovisual synthesis, information about whether all native speakers equally rely on both auditory and visual information while perceiving speech is useful. Speech recognition systems based on articulation can possibly benefit from the data about the most perceptually stable articulations, which determine the interpretation of what is heard, even if the auditory signal does not match the visual information.

2 Previous Experimental Studies of the Incongruent Audiovisual Stimuli Processing

Much of the experimental research that considers the interaction of auditory and visual information at low linguistic levels is based on the McGurk effect [14]. If this effect occurs, the listener cannot correctly determine what he/she hears if the movements of the speaker's lips do not correspond to the auditory signal. In the original experiment by McGurk and McDonald, participants interpreted the syllable /ba/ as /da/ if the articulation in the video they were shown along with the sound corresponded to the syllable /ga/. Later studies [15] have shown that participants in the experiment proposed one more interpretation – /gba/, i.e., a syllable that includes both the consonant that was pronounced and the one that was articulated.

Many studies were conducted in English [16], but more recently there have been experiments in Chinese [17, 18, etc.], Japanese [19, 20, etc.], Dutch [21, 22, etc.], Swedish [23], and other languages, some of which have been part of cross-linguistic studies of the McGurk effect. In most studies, the effect is tested on the material of consonants, but there is evidence that it can also appear on vowels (for example, in Swedish

[23] and Dutch [24]), as well as on the tones of the Chinese language [25]. When consonants are analyzed, they are most often presented in a syllable with the vowel /a/ (as was done in [14]). However, it was shown in [26] that the McGurk effect is most prominent in syllables with the /i/ vowel, and weakest in syllables with the /u/ vowel. As for the consonants used, they are usually bilabial, alveolar, and velar stops, i.e. consonants that coincide in the manner of articulation, but differ in the place of articulation.

This effect has been used in certain studies aimed at describing neural mechanisms of auditory and visual speech information processing [18, 27, 28, etc.]. The aim of quite numerous cross-linguistic studies was, first of all, to answer the question of whether the interaction of visual and auditory information in the processing of auditory speech is universal or language specific. In the future, the results of such studies should allow a better understanding of the cognitive mechanisms of information processing. A separate area of research, which is addressed in a number of papers, is the comparison of how stimuli based on the McGurk effect are processed by different groups of recipients. There is evidence that the effect of visual information on the perception of auditory information is the weakest in 4- to 6-year-old preschool children, while it is quite pronounced in older informants, as well as in infants under similar conditions (for review, see [29]). It can be also assumed that different people rely on information coming through different channels of perception not in one and the same way. There are those for whom auditory modality is more crucial and those who prefer visual information. This assumption has been thoroughly discussed within the theory of cognitive/learning styles. The current experimental studies, however, claim that there is not a great difference between so called verbal learners and visual learners [30] and that "none of the four learning styles (visual, auditory, read/write, or kinesthetic) predicted students' retention of the material" [31]. The study of the McGurk effect can provide new evidence on the role of cognitive styles and preferred perceptual modalities in audiovisual processing.

Despite a fairly large number of experimental studies conducted on the material of various languages, there is almost no experimental evidence for the McGurk effect in Russian listeners. Perhaps the only exception is [29], which describes the methodology of a cross-linguistic study using Russian-language material. However, the author does not present the results of the study, but only discusses what they could be. Thus, it seems promising to use the McGurk effect to study the processing of information coming simultaneously through the auditory and visual channels. Since there is no experimental data for the Russian language on how this effect manifests itself, in the pilot experiment that will be described in this paper, we tried to find out those combinations of audio and visual stimuli that will cause the most errors in the perception of auditory information, i.e. demonstrate the highest audiovisual integration.

3 Our Experiment

3.1 Goal

The main goal of our study is to find out those combinations of auditory syllables and lip movements for which the misinterpretation of auditory information because of the incongruent visual one would be the strongest for Russian listeners. At the same time, we tried to take into account the recipient factor and check whether the results would

differ depending on the individual characteristics of the participants: on the preferred perceptual modality and on age group.

3.2 Stimuli

Audiovisual stimuli were created specifically for this experiment. We asked a 20-year-old Russian native speaker to pronounce six syllables of the Russian language /ta/, /da/, /pa/, /ba/, /fa/ and /va/. The speaker is female. She is a linguist, but she is not a professional speaker. She does not have any pronunciation disorders. As our study was a pilot one, we used the data from only one speaker, although we understand that it is probably the most crucial limitation of the study.

In contrast to the classic experiments aimed at studying the McGurk effect, which use only stop consonants, we decided to choose consonants that are close in place of articulation and, at the same time, those whose articulation is easy to distinguish by the listener when he/she looks at the speaker. Therefore, in our experiment, we included bilabial stop consonants, labio-dental fricative consonants, and alveolar stop consonants. We used the syllables with the /a/ vowel as this vowel was used in the majority of previous studies of the McGurk effect.

The speaker repeated each syllable five times. The pronunciation of the syllables was recorded on video. Then, we compiled stimuli from the original videos, some of which were supposed to provoke the participants to experience an effect close to the McGurk effect, namely, to lead to misinterpretation of what they heard. To do this, the sound track of one syllable was combined with the video of another. The synchronization process was performed manually by one of the experimenters and then checked by the other. Records with voiceless consonants were combined only with voiceless consonants, and voiced ones – only with voiced ones. A total of 18 combinations were obtained: six initial ones, in which the auditory and visual information coincided (these were control stimuli) and 12 stimuli that were supposed to provoke audiovisual integration – six each for voiceless and voiced consonants.

3.3 Procedure

The experiment was conducted on the Google Forms platform. It consisted of two parts. Participants had to read the instructions for the experiment, provide consent to take part in it and some personal information (gender, age, year of study (for schoolchildren)). After that, participants had to choose one of four questionnaire options, which differed from each other only in the sequence of stimuli (thus, pseudo-randomization was achieved in order to reduce the influence of the order of stimuli presentation on the participants' answers).

In the first part of the experiment, participants had to carefully look and listen to each of the 18 stimuli and note what the speaker said, choosing one of the six proposed answers or writing down their own. It is the forced-choice paradigm that is used in most studies of the McGurk effect (see [29] for review), so we also chose it. But at the same time, we left the participants the opportunity to write their own answer if they believed that none of the proposed options was suitable. All 6 syllables that were used in the experiment appeared as suggested answers for all stimuli.

The second part of the Google form contained a questionnaire aimed at determining the preferred modality of perception. For this, a questionnaire by S. Efremtsev was used, consisting of 48 questions that must be answered "yes" or "no". This questionnaire is said to determine how strongly a person has expressed preferences for each of the three following modalities of perception: auditory, visual and kinesthetic. There are 16 questions for each modality in questionnaire. The more "yes" answers the participant gives to questions from the corresponding block, the greater the role for him/her in the process of perception plays precisely this channel of information transmission.

The experiment was conducted in accordance with the Declaration of Helsinki and the existing Russian and international regulations concerning ethics in research. It took participants around 10 min to pass the experiment. The experiment can be found at the link: https://forms.gle/riGhH1smjiPzchWdA At the end of the experiment, participants could leave their email address to get results of the Efremtsev's questionnaire.

3.4 Participants

Two groups of respondents took part in the experiment: schoolchildren from 14 to 17 years old and adults from 18 to 50 years old. Initially, there were 70 people in each group. We decided to analyze the results of schoolchildren (teenagers) separately, since the problem of learning and cognitive styles is often referred to school education.

3.5 The Principles of Data Analysis

We calculated the number of "correct" responses of each participant to each stimulus, which, within the framework of the study, were those answers that corresponded to the auditory token in the stimulus, regardless of the visual token. In further sections of the paper, we refer to the answers that did not corresponded to the auditory tokens as errors or the cases of audiovisual integration. The responses of those participants who made at least one error in their responses to control stimuli (where the audio and video corresponded to the same sound), were excluded from the analysis. The responses of those participants who, after completing the experiment, reported that they had problems with video playback, were also excluded. As a result, we analyzed the data from 60 schoolchildren (Me = 16.0; M = 15.9; SD = 0.9; 49 females) and 60 participants from 18 to 50 years old (Me = 20.5; M = 23.5; SD = 7.3; 46 females).

Using the chi-squared test, we compared the number of correct answers for different syllables. Using the Mann-Whitney test, the chi-squared test with continuity correction and the Spearman's rank correlation coefficient we compared the data obtained on schoolchildren and adults. Correlation analysis was also used to test the hypothesis about the dependence of the number of correct answers on the participant's preferences for one or another modality of perception. Only responses to target stimuli were taken into account, i.e., to those in which the sound of the stimulus did not coincide with the articulation in the video. Free software JASP (https://jasp-stats.org/) was used for statistical processing.

3.6 Results: Schoolchildren vs. Adults

We observed the influence of the group factor. Audiovisual integration was more prominent in adults than in schoolchildren: schoolchildren, on average, made significantly fewer mistakes than adults, although the average number of correct answers did not differ much (10.5 and 9.8 respectively, $W = 1414$, $p = 0.037$; see Fig. 1).

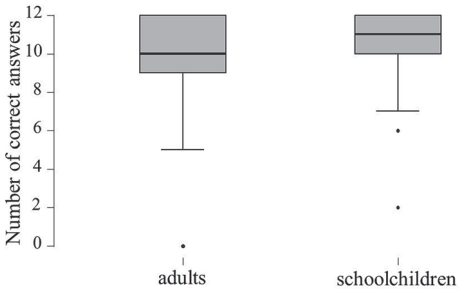

Fig. 1. The number of correct answers in two groups of participants.

The total number of incorrect answers among schoolchildren is significantly less than in the group of adults (91 (12.6%) and 135 (18.8%), respectively, for 720 answers in each of the groups; $X^2 = 9.704$; $p = 0.002$). The smallest number of correct answers given by one participant in the group of schoolchildren is 2 out of 12. In the adult group, two participants did not give a single correct response to the target stimuli (while they correctly identified all control stimuli). We decided not to exclude such participants from the sample because we were interested, among other things, in the influence of individual factors (in particular, the preferred perceptual modality) on the perception of target stimuli.

We also tested the hypothesis that the number of correct answers decreases with age. A significant weak negative correlation was found (Spearman's rho $= -0.222$; $p = 0.015$), but it should be noted that not the oldest participants in the experiment did not give a single correct answer, but two girls, whose age was 21 years old.

In the following sections, we will describe the results of recognition of all stimuli. The data will be given as a whole for all 120 participants, because all the trends that will be discussed below were the same for both groups of participants.

3.7 Results: Quantitative Analysis of Audiovisual Integration

A total of 1214 correct responses and 226 incorrect responses were received for 12 target stimuli. There is no significant difference in the number of correct answers between voiceless and voiced sounds ($X^2 = 0.635$; $p = 0.426$): participants made 119 errors in voiced consonant pairs and 107 errors in voiceless ones.

We found the influence on the number of errors both of the syllable that sounded ($X^2 = 91.240$; $df = 5$; $p =< 0.001$), the syllable that was articulated ($X^2 = 49.042$; $df = 5$; $p =< 0.001$), and the combination of spoken and articulated syllables ($X^2 = 196.987$; $df = 11$; $p =< 0.001$).

Most often, the audiovisual integration occurred when the participants heard the syllables /ba/ (76 errors) and /pa/ (60 errors). The integration was the greatest when the speaker articulated the syllables /va/ (65 errors) and /fa/ (51 errors) in the video. These results indicate that bilabial stop consonants are the most vulnerable to the impact of contradictory visual information (which was confirmed by statistical analysis: $X^2 = 87.052$; df = 2; p =< 0.001; see Table 1).

Table 1. The number of errors and correct answers to different types of auditory tokens.

	Type of the sound (auditory token)			
Answer	Alveolar	Labial	Labiodental	Total
AV integration	47	136	43	226
Correct	433	344	437	1214

And vice versa: the greatest number of errors in the perception of auditory tokens was provoked by the articulation of labiodental consonants in the video ($X_2 = 44.381$; df = 2; p =< 0.001; see Table 2).

Table 2. The number of errors and correct answers to different types of visual tokens.

	Type of the articulation (visual token)			
Answer	Alveolar	Labial	Labiodental	Total
AV integration	42	68	116	226
Correct	438	412	364	1214

As for the combinations of auditory and visual tokens, the most common errors were found when the syllable /ba/ sounded, while the speaker articulated /va/ in the video (57 errors (47%)), and when /pa/ sounded, while in the video the speaker pronounced /fa/ (48 errors (40%)) (see Table 3; the first two small letters in each stimulus correspond to what was heard (auditory token), whereas two big letters show what was articulated (visual token)). The least common errors occurred when /ta/ was pronounced and the video had /fa/ (3 errors) and vice versa (6 errors), as well as when /va/ sounded and the video had /da/ (5 errors; in the opposite combination, 8 errors were made).

Table 3. The number of different types of audiovisual integration for all the stimuli.

	Total	Visual dominance	Audiovisual serialization: two consonants in the answer	Audiovisual fusion: substitutions for another sound	Substitutions for the voiceless / voiced pair of the auditory token
baDA	19	4	4	11	0
baVA	57	50	7	0	0
daBA	14	6	8	0	0
daVA	8	3	4	1	0
faPA	16	13	2	0	1
faTA	6	1	4	1	0
paFA	48	43	5	0	0
paTA	12	2	3	7	0
taFA	3	3	0	0	0
taPA	22	6	15	0	1
vaBA	16	10	6	0	0
vaDA	5	1	0	3	1

3.8 Results: Qualitative Analysis of Audiovisual Integration

All the errors that were made by the participants in the experiment can be divided into four groups: 1) visual dominance: answers with the consonant, the articulation of which was in the video (or with its voiced/voiceless pair); 2) audiovisual serialization – answers in which two consonants occur: the one that sounded and the one that was in the video (in this case, different options are possible: *ba-da-ba-da-ba* or *bda-bda-bda-bda-bda* for the **baDA** stimulus), or a combination of one of the consonants from the stimulus with some other consonant (for example, *ba-va-ba-va-ba* for the **baDA** stimulus); 3) audiovisual fusion: answers containing only one consonant, which does not match either the one that sounded or the one that was in the video (for example, *va-va-va-va-va* for **baDA**); 4) responses containing a voiced/voiceless pair consonant to the consonant that sounded in the stimulus (for example, *va-va-va-va-va* for the **faPA** stimulus). The distribution of responses to each stimulus is presented in Table 3.

For the stimuli with the highest number of incorrect responses (**baVA** and **paFA**), we observed mostly visual dominance (substitutions for the sound shown in the video). Interestingly, the cases of audiovisual serialization (when there are two or more consonants in the response) were observed for almost all stimuli (except for **vaDA** and **taFA**, for which the smallest number of errors occurred). The largest number of responses containing a consonant that does not match either the one that sounded or the one that was in the video was obtained for the pairs **baDA** and **paTA**: in both cases, these were consonants, whose place of articulation is between the place of articulation of the auditory and visual tokens from the stimulus: /v/ and /f/ respectively.

3.9 Results: The Influence of the Preferred Perceptual Modality

Correlation analysis using Spearman's test did not reveal a relationship between the number of correct responses to target stimuli and a greater preference for any of the modalities of perception (according to Efremtsev's questionnaire), neither for all participants in general, nor separately for schoolchildren and adults (see Table 4).

Table 4. The results of the Spearman's rank correlation test for the number of correct answers and the number of answers "yes" to each of three perceptual modalities

Modality	All participants	Schoolchildren	Adults
Auditory	rho = −0.028 p = 0.759	rho = 0.012 p = 0.929	rho =−0.133 p = 0.311
Visual	rho = 0.012 p = 0.893	rho = 0.014 p = 0.916	rho = 0.077 p = 0.558
Kinesthetic	rho = −0.051 p = 0.583	rho = −0.019 p = 0.888	rho = −0.060 p = 0.649

4 Discussion and Conclusions

The experiment showed that the greatest audiovisual integration is in pairs "labial stop consonant in the auditory channel - labiodental fricative in the visual channel" (i.e., **baVA** and **paFA**) where we observed visual dominance. Such cases, strictly speaking, do not indicate the emergence of the McGurk effect, but they show that the labiodental articulation of Russian consonants is quite clear and even can lead to the misinterpretation of what was pronounced. The labial stops are the most vulnerable from the point of view of auditory perception: they were most often substituted in responses to other sounds.

Examples of the manifestation of the McGurk effect can be considered the answers with the combination of several consonants (primarily those that were presented in the stimulus, as reported in [15]), as well as those cases when the sound in the answer is the one articulated between the sounded consonant and the one that was in video (these are examples with the responses to **baDA** and **paTA** stimuli, see Sect. 3.8).

The fact that there is no difference in the processing of voiceless and voiced consonants in the experiment shows that in further similar studies with the participation of native speakers of the Russian language, both stimuli with voiced and voiceless consonants can be used.

The influence of the factor of the group of participants was revealed: schoolchildren gave more correct answers than those who had already graduated from school. At the same time, we understand that in our study the boundary between the two groups of participants is largely conditional: in the group of adult participants, there were 1st year University students, i.e., those who are 18 years old, and the average age of the adult group is 23.5 years. Therefore, the question of the influence of the age of participants

on the results of the study requires further study. Perhaps the greater number of correct answers in the group of schoolchildren is due to the fact that, on the whole, because of their habit of completing school assignments, they were more attentive and responsible in completing the experimental task.

The fact that we did not find the effect of the preferred modality of perception on the recognition of auditory stimuli can be explained by various reasons. For example, this may support the previous findings [30, 31] that people cannot be easily divided into auditory, visual and kinesthetic groups and cognitive styles are not that crucial for audiovisual interaction while processing speech. Or it can indicate the imperfection of the questionnaire that was chosen to identify these groups, although it is this questionnaire that is most often used to determine the preferred perceptual modality in the studies with Russian-speaking informants. In any case, this aspect of the study requires further development. Perhaps, in the future, it makes sense to test the hypothesis about the influence of the preferred modality of perception on multimodal processing only using the stimuli for which audiovisual integration is high in Russian speakers, but involving a larger number of speakers. An increase in the number of speakers is also necessary in order to make sure that the results obtained are not due to the individual articulatory characteristics of a particular speaker.

The experiment that we conducted included only six consonants, the articulation of which we considered the most obvious (noticeable) for the listener. In the future, we can expand the experiment to include other consonants (in particular, velar stops, as in most classic experiments on the McGurk effect). We believe that our further studies of the McGurk effect in Russian speakers will contribute to the discussion of the theoretical problem of audiovisual binding and audiovisual integration in general.

Acknowledgements. The study is supported by the research grant #21–18-00429 from the Russian Science Foundation.

References

1. Griban, O.N.: Application of educational presentations in the educational process: types, stages and structure of presentations. Historical Pedagogical Readings **20**(3), 23–32 (2016). (In Russian)
2. Svärdemo Åberg, E., Åkerfeldt, A.: Design and recognition of multimodal texts: selection of digital tools and modes on the basis of social and material premises? J. Computers Educ. **4**(3), 283–306 (2017). https://doi.org/10.1007/s40692-017-0088-3
3. Mayer, R.E.: Principles for managing essential processing in multimedia learning: Segmenting, pretraining, and modality principles. In: The Cambridge Handbook of Multimedia Learning. Cambridge University Press, Cambridge, pp. 169–182 (2005). http://dx.doi.org/https://doi.org/10.1017/cbo9780511816819.012
4. Petrova, T.E.: Text presentation and information processing in Russian In: ExLing 2021. 12th International Conference of Experimental Linguistics. In: International Society of Experimental Linguistics, pp, 164–167 (2021)
5. Riekhakaynen, E., Skorobagatko, L.: Written, not spoken or too much to read: How to present information more effectively? In: Neurobiology of Speech and Language. Proceedings of the 5th International Conference. Saint Petersburg, pp. 15–16 (2021)

6. Ivanko, D.V., Kipyatkova, I.S., Robzhin, A.L., Karpov, A.A.: Analysis of methods for multimodal information combination for audiovisual speech recognition // scientific and technical bulletin of information technologies. Mechanics and Optics **16**(3), 387–401 (2016). (In Russian)
7. Brown, V.A., Strand, J.F.: "Paying" attention to audiovisual speech: do incongruent stimuli incur greater costs? Atten. Percept. Psychophys. **81**(6), 1743–1756 (2019). https://doi.org/10.3758/s13414-019-01772-x
8. Berthommier, F.: A phonetically neutral model of the low-level audio-visual interaction. Speech Commun **44**(1–4), 31–41 (2004). https://doi.org/10.1016/j.specom.2004.10.003
9. Ganesh, A., Berthommier, F., Schwartz, J.-L.: Audiovisual binding for speech perception in noise and in aging. Lang. Learn. **68**(S1), 193–220 (2018). https://doi.org/10.1111/lang.12271
10. Lobanov, B.M., Tsyrulnik, L.I., Zhelezny, M., Krnoul, Z., Ronzhin, A., Karpov, A.: System of audiovisual synthesis of Russian speech. Informatics **4**(20), 67–78 (2008). (In Russian)
11. Thézé, R., Gadiri, M.A., Albert, L., Provost, A., Giraud, A.L., Mégevand, P.: Animated virtual characters to explore audio-visual speech in controlled and naturalistic environments. Sci. Rep. **10**(1), 1–12 (2020)
12. Almeida, N., Cunha, D., Silva, S., Teixeira, A.: Designing and deploying an interaction modality for articulatory-based audiovisual speech synthesis. In: Karpov, A., Potapova, R. (eds.) SPECOM 2021. LNCS (LNAI), vol. 12997, pp. 36–49. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87802-3_4
13. Ivanko, D., Ryumin, D., Axyonov, A., Kashevnik, A.: Speaker-dependent visual command recognition in vehicle cabin: methodology and evaluation. In: Karpov, A., Potapova, R. (eds.) SPECOM 2021. LNCS (LNAI), vol. 12997, pp. 291–302. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87802-3_27
14. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature **264**, 746–748 (1976). https://doi.org/10.1038/264746a0
15. Green, K.P., Gerdeman, A.: Cross-Modal discrepancies in coarticulation and the integration of speech information: the McGurk effect with mismatched vowels. J. Experiment Psychology: Human Perception Performance **21**(6), 1409–1426 (1995). https://doi.org/10.1037/0096-1523.21.6.1409
16. Summerfield, Q.: Some preliminaries to a comprehensive account of audiovisual speech perception. In: Dodd, B., Campbell, R. (eds.) Hearing by eye: Psychology of lipreading Hillsdale, pp. 3–51. Erlbaum, NJ (1987)
17. Sekiyama, K.: Cultural and linguistic factors in audiovisual speech processing: the McGurk effect in Chinese subjects. Percept. Psychophys. **59**(1), 73–80 (1997). https://doi.org/10.3758/BF03206849
18. Wu, J.: Speech perception and the McGurk effect: A cross cultural study using event-related potentials. Electronic Theses and Dissertations. Paper 1597 (2009). https://doi.org/10.18297/etd/1597
19. Sekiyama, K., Burnham, D.: Impact of language on development of auditoryvisual speech perception. Dev. Sci. **11**(2), 306–320 (2008). https://doi.org/10.1111/j.1467-7687.2008.00677.x
20. Sekiyama, K., Tohkura, Y.I.: Inter-language differences in the influence of visual cues in speech perception. J. Phon. **21**(4), 427–444 (1993). https://doi.org/10.1016/S0095-4470(19)30229-3de
21. de Gelder, B., Bertelson, P., Vroomen, J., Chen, H.C.: Inter-language differences in the McGurk effects for Dutch and Cantonese listeners. In: Eurospeech 1995: Proceedings of the Fourth European Conference on Speech Communication and Technology, Madrid, Spain, September 18–21, pp. 1699–1702 (1995)

22. Massaro, D.W., Cohen, M.M., Smeele, P.M.: Cross-linguistic comparisons in the integration of visual and auditory speech. Mem. Cognit. **23**(1), 113–131 (1995). https://doi.org/10.3758/BF03210561
23. Traunmüller, H., Öhrström, N.: Audiovisual perception of openness and lip rounding in front vowels. J. Phon. **35**(2), 244–258 (2007). https://doi.org/10.1016/j.wocn.2006.03.002
24. Valkenier, B., Duyne, J.Y., Andringa, T.C., Baskent, D.: Audiovisual perception of congruent and incongruent Dutch front vowels. J. Speech Lang. Hear. Res. **55**(6), 1788–1801 (2012). https://doi.org/10.1044/1092-4388(2012/11-0227)
25. Wang, R.: Audiovisual perception of Mandarin lexical tones. Doctoral dissertation, Bournemouth University (2018)
26. Shigeno, S.: Influence of vowel context on the audio-visual speech perception of voiced stop consonants. Jpn. Psychol. Res. **42**(3), 155–167 (2000). https://doi.org/10.1111/1468-5884.00141
27. Besle, J., Caclin, A., Mayet, R., Bauchet, F., Delpuech, C., Giard, M.H., et al.: Audiovisual events in sensory memory. J. Psychophysiol. **21**, 231–238 (2007). https://doi.org/10.1027/0269-8803.21.34.231
28. Kelly, S.D., Kravitz, C., Hopkins, M.: Neural correlates of bimodal speech and gesture comprehension. Brain Lang **89**(1), 253–260 (2004)
29. Yang, Z.: A cross-linguistic examination on the McGurk effect in different developmental states. Research Master's Thesis in Linguistics, Utrecht University (2021)
30. Massa, L.J., Mayer, R.E.: Testing the ATI hypothesis: should multimedia instruction accommodate verbalizer-visualizer cognitive style? Learn. Individ. Differ. **16**(4), 321–335 (2006). https://doi.org/10.1016/j.lindif.2006.10.001
31. Cuevas, J., Dawson, B.L.: A test of two alternative cognitive processing models: learning styles and dual coding. Theory Res. Educ. **16**(1), 40–64 (2018). https://doi.org/10.1177/1477878517731450