24th INTERNATIONAL CONGRESS ON ACOUSTICS

ICA 2022

October 24(Mon) - 28(Fri), 2022 | Gyeongju, Korea

PROCEEDINGS

Hosted by The Acoustical Society of Korea | ICA

ICA 2022

24th INTERNATIONAL CONGRESS ON ACOUSTICS

October 24(Mon) - 28(Fri), 2022 | Gyeongju, Korea

A15: SPEECH

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0338

Automatic Classification of the Emotional State of Atypically Developing Children

Yuri MATVEEV[1]; Elena LYAKSO[1]; Anton MATVEEV[1]; Olga FROLOVA[1]; Aleksey GRIGOREV[1]; Aleksandr NIKOLAEV[1]

[1] The Child Speech Research Group, St. Petersburg State University, St. Petersburg, Russia

ABSTRACT

To study of the emotional state reflection in the voice and speech of 6-12 years old children with autism spectrum disorders (ASD), Down syndrome (DS) and typical development (TD), the automatic classification of children's emotional speech on the states "comfort – neutral – discomfort" were conducted. Child speech was recorded in model situations a dialogue with the experimenter and playing with a standard set of toys and annotated by three experts. Automatic classification of children's speech on three states was performed using automatically extracted sets of acoustic features GeMAPS and extended eGeMAPS. As classifiers, we used classifiers based on Gaussian Mixture Models (GMM) and Support Vector Machine (SVM). The state of discomfort is classified better for children with ASD (0.523; 0.305 – precision and recall) and DS (0.504; 0.564); the state of comfort – for TD children (0.546; 0.241). The minimal GeMAPS feature set gives better results (accuracy - 0.687, 0.725, 0.641 – for ASD, DS, TD children) than the extended eGeMAPS feature set (0.671, 0.717, 0.631), that indicates the importance of low-level features. A comparison with the data of an auditory perceptual experiment (100 listeners) was made. Listeners better recognized discomfort in children with ASD and DS (78%) and comfort state in TD children (58%).

Keywords: Emotions, Speech, Children with Atypical development, Perceptual Experiment, Automatic Classification

1. INTRODUCTION

In the past ten to fifteen years, affective computing or emotion artificial intelligence became a focus of intense research. Since 2009, the Interspeech Computational Paralinguistics ChallengE (ComParE) is held annually, and the IEEE Transactions on Affective Computing journal is issued since 2015. Reviewing the results coming from the ComParE challenge, we can note some of the more common approaches to solving emotion-related tasks (1):

The machine learning techniques based on the extraction of common handicraft features empirically figured out in the recent 20 to 30 years and the application of the established stochastic models such as Gaussian mixture models (GMM) or classifiers such as Support-vector machine (SVM) are employed to detect emotional states. These methods are still highly relevant, which is proven by the performances of some of the participants in the ComParE challenge.

The machine learning techniques based on the deep neural networks (DNN) which, in recent years, became massively popular for detection and recognition of objects of various natures, including emotional states, often outperform the traditional approaches but require significantly larger amounts of training data to learn the deep features truly representing the essence of the process, which is not always available for the emotion-related problems. Additionally, comparing to other areas such as object detection or automatic speech recognition, the potential of deep learning in emotion recognition is not yet fully developed. Partially, it emerges from the higher level of abstraction of the concept and vagueness of expression of emotions in speech and voice: though many modern databases now have annotations from multiple experts, the labels are often conflicting or unreliable. Also, since some emotions are more socially acceptable and often expressed, when collected in the wild, the datasets are

1 yunmatveev@gmail.com
[1] lyakso@gmail.com

often unbalanced. Nevertheless, the main obstacle is the volume of data in the annotated datasets. Research shows that the lack of data is the bottleneck for efficiency for most of the deep learning methods. Eventually, with the growth of the volume of the available for training data, the deep learning models would become accessible for feature extraction similar to the expert feature extractors, however, at this moment, even with augmentation and transfer learning, the issue of the lack of data is not yet overcome.

The aim of the study is exploring the possibility of automatic recognition of "comfort – neutral – discomfort" states in speech of 6-12 years old children with autism spectrum disorders (ASD), Down syndrome (DS) and typical development (TD).

2. METHODS

2.1 Auditory perceptual evaluation

For the auditory perceptual evaluation by adult listeners, 3 test sequences containing a speech material of TD children (test-1), children with DS (test-2), and children with ASD (test-3) were created. Each test sequence included 30 samples from 10 children annotated with comfort, neutral, or discomfort labels. Tests were presented individually to each listener.

The listeners were 100 adults (age 17-32 years, 19.3±2.3 years; 81 women, 19 men): pediatric students (n=50) and biology students (n=50). The task for listeners was to determine the state of the children as comfort, neutral, or discomfort when listening to the test material.

2.2 Automatic Classification

For automatic classification of "comfort-neutral–discomfort" states in children's speech, two models are used: Gaussian Mixture Models (GMM) and Support Vector Machine (SVM) models.

2.2.1 Dataset

The speech material was obtained by database "AD_CHILD.RU" (2). The "AD_CHILD.RU" database (Atypical Development_Child.Ru) contains the speech material of 392 children (265 children with different diagnoses): 96 children with autism spectrum disorders (ASD, F84 - according to ICD-10), 49 children with Down syndrome (DS, Q90), 52 children with mixed specific developmental disorders (MSDD, F83), 49 children with intellectual disabilities – mental retardation (ID, F70, 71), 7 children with cerebral palsy (CP, G80), 12 children with mild neurological disorders (F80, 90), 127 typically developing (TD) children (control). The database size is 1.5 Tb. The speech material is represented by long original files; recordings in model situations (dialogue with parents, dialogue with an experimenter, retelling of a fairy tale or cartoon, story based on a picture), emotional speech. Speech files are presented in the WAV format, video files are in AVI. The files are annotated (phonemic description, and phonetic description for a part of the material). The speech material is accompanied by information about the psychophysiological and psychoneurological state of the children, and includes the results of tests and questionnaires.

We selected 6594 audio files for 6 to 12 years old children (n=73) with ASD, DS, and TD. All speech files were stored in .wav format, 48,100 Hz, 16 bits per sample.

Annotation of the child's emotional speech was made on three categories (based on video recording and protocol of the recording situation) "comfort - neutral – discomfort" by two experts. The speech samples were assigned to the corresponding emotional category when there was an agreement between the two annotators.

2.2.2 Feature sets

The research in the field of speech technologies in the recent decades produced a bevy of different expert acoustic feature sets, which are used arbitrary and often defined and extracted inconsistently. Since there are many independent research groups working on different problems in the field, to have a shared standard is to guarantee the ability to objectively compare the research results and to enable an integration between various feature extraction and recognition systems.

The most popular is the ComParE set, which consists of more than 6,000 features from the openSMILE library (3) extracted on three levels (4):

- Low-level descriptors (LDD);
- Low-level descriptors with deltas (LLD deltas);
- Functionals of LDD.

The success of ComParE features can be attributed to several factors: they are easy to calculate,

widely known in the community, and well designed for specific applications of voice-based affect recognition.

The authors of (5) introduced a baseline acoustic feature set Geneva Minimalistic Acoustic Parameter Set (GeMAPS) applicable to miscellaneous speech analysis problems, including a paralinguistic speech analysis. Unlike for the amalgamated ComParE acoustic feature set, it is suggested to employ a minimalistic voice feature set.

The features are selected based on: a) their capability to reflect the affective physiological processes in speech formation, b) their proven in previous works efficiency and the accessibility of methods of their automatic extraction, and c) their theoretical value.

The authors of (5) demonstrate the comparison of efficiencies between GeMAPS and the large baseline feature set ComParE_2016 for paralinguistic tasks. The results show the advantages of the GeMAPS feature set in recognition accuracy relative to the sizes of the feature sets. There is a tool to extract GeMAPS features which helps to ensure the features are obtained according to the standard: openSmile library v. 2.0. There are three standard feature sets available: ComParE_2016 with more than 6 thousand parameters, and GeMAPS and eGeMAPS (extended GeMAPS). The number of features for each level is listed in Table 1.

The GeMAPS feature set with 62 parameters is often used in different paralinguistic challenges due to the variety of acoustic and prosodic features contained in this set. In our experiments we used version 2 of the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) with 26 extra parameters, and in total, eGeMAPSv02 contains 88 parameters.

Table 1 – Number of features in datasets on each of three levels

Feature sets	Number of features for each of three levels
ComParE_2016	65 / 65 / 6373
GeMAPSv01b	62 / 13 / 62
eGeMAPSv01b	10 / 13 / 88

2.2.3 Classification based on Gaussian Mixture Models

Since we have a dataset of small size and solve a close-set identification task with a small number of classes, we build a recognition system based on a simple state-of-the-art scheme: extracting Mel-Frequency Cepstral Coefficients (MFCC) from audio recordings, training class-depended Gaussian Mixture Models (GMM) on these features and performing the classification using the Maximum Likelihood Estimation (MLE) method. To extract MFCC from audio recordings we used a standard schema with a frame of 25 ms length, 10 ms overlap between frames, 1024-point Fourier transform, 13-component MFCC vector. All MFCC vectors extracted from the recordings of the same speaker are concatenated.

To perform the classification, we calculate the log-likelihood scores for each frame of each speaker's recording belonging for each emotional state. The likelihood of the frame originating from a speech sample of the speaker belonging to a specific class is calculated using the class-depended GMMs. This procedure is performed for each of the Gaussian components of the GMMs, and the weighted sum of these likelihoods from the components is calculated. The logarithm of the sum obtained gives the logarithmic probability value for the frame. This is repeated for all frames of the recording, and the probabilities of all frames are summed. We predict the speaker to belong to the class with the highest total likelihood.

Training and testing. For each generated collection split into a training and a testing parts, we perform the following process:

1. Extract 13 component MFCC vectors from each audio recording.
2. Train 8 component GMMs based on extracted MFCC vectors for each emotion state.
3. For each recording and for each MFCC vector on each frame calculate log-likelihood for each GMM (each class), sum them up, and predict the class with the highest total likelihood. Repeat 20 times and concatenate the results.

2.2.4 Classification based on Support Vector Machine

We conducted experiments for automatic recognition of three emotional states: comfort, neutral, and discomfort based on automatically extracted GeMAPS and eGeMAPS acoustic feature sets.

For training, we used 6594 audio records with speech material of 73 children six to twelve years old with TD, ASD, and DS.

We split the data into the collections for training (80%) and testing (20%) and employed C-Support Vector Classification from sklearn library (6) for classification.

2.3 Classification performance metrics

To evaluate the performance of the automatic classification of emotional states and ages, we utilize the performance metrics commonly accepted in international challenges for the evaluation of the automatic paralinguistic and biometric recognition systems:

1) Overall accuracy: the ratio of the correctly identified samples.
2) Per class precision: the ratio of the correctly identified samples to the total number of positive predictions.
3) Per class recall: the ratio of the correctly identified samples to the total number of predictions of that class.

3. EXPERIMENTAL RESULTS

3.1 Results of the auditory perceptual evaluation

The results of the auditory perceptual evaluation showed that in the analyzed groups of children, emotional states are worse determined in the speech of TD children. The state of comfort is recognized better by speech of children with DS (70% of correct answers), worse – by the speech of TD children (58% of correct answers). Discomfort state is recognized better by speech of children with DS and ASD (78%) than by speech of TD children (56%), neutral state - by the speech of children with ASD than by speech of TD children (54%) and children with DS (52%), Tables 2, 3, 4.

Table 2 – Confusion matrixes for emotion classification by the speech of children with ASD

	comfort	neutral	discomfort
comfort	**67**	17	16
neutral	17	**67**	16
discomfort	15	**7**	**78**
Accuracy	0.78	0.81	0.82
Recall	0.67	0.67	0.78
Precision	0.68	0.74	0.71

Table 3 – Confusion matrixes for emotion classification by the speech of children with DS

	comfort	neutral	discomfort
comfort	**70**	22	8
neutral	27	**52**	21
discomfort	9	13	**78**
Accuracy	0.78	0.72	0.83
Recall	0.70	0.52	0.78
Precision	0.66	0.60	0.73

Table 4 – Confusion matrixes for emotion classification by the speech of TD children

	comfort	neutral	discomfort
comfort	**58**	26	16
neutral	30	**54**	16
discomfort	14	30	**56**
Accuracy	0.71	0.66	0.75
Recall	0.58	0.54	0.56
Precision	0.57	0.49	0.64

In children with ASD and DS, listeners recognize the state of discomfort (78% of answers) better than the comfort state (70% and 67% - for children with DS and ASD, correspondingly).

The average accuracy of automatic classification of children's emotional speech for ASD Children is 0.795.

The average accuracy of automatic classification of children's emotional speech for DS children is 0.750.

The average accuracy of automatic classification of children's emotional speech for TD children is 0.685.

The average accuracy of automatic classification of children's emotional speech for all diagnoses is 0.743.

3.2 Results of the automatic classification based on GMM model

Table 5 demonstrates the results of the experiments on automatic classification of children's emotional speech into three states: comfort, neutral, and discomfort based on Gaussian Mixture Models.

Table 5 – Precision and recall of the automatic classification of children's emotional speech into three states: comfort, neutral, and discomfort

Diagnosis	Ground truth emotional state	precision	recall
ASD	Comfort	0.476	0.529
	Discomfort	**0.523**	0.305
	Neutral	**0.539**	0.652
DS	Comfort	0.325	0.348
	Discomfort	**0.504**	0.564
	Neutral	0.459	0.394
TD	Comfort	**0.546**	0.241
	Discomfort	0.361	0.718
	Neutral	0.472	0.423

For ASD, the accuracy is 0.675, the total precision is 0.512, and the total recall is 0.495.

For DS, the accuracy is 0.617, the total precision is 0.429, and the total recall is 0.435.

For TD, the accuracy is 0.617, the total precision is 0.459, and the total recall is 0.461.

The average accuracy of automatic classification of children's emotional speech for all diagnoses is 0.636.

Overall, the precision and the recall are higher for ASD children, and the difference between DS and TD children is insignificant.

3.3 Results of the automatic classification based on SVM model

The results of the experiments on automatic classification of comfort, neutral, and discomfort states in children's emotional speech based on GeMAPS and eGeMAPS feature sets. We calculated only the overall accuracy across all classes to discover comparative efficiency of the feature sets.

1. The average accuracy of automatic classification of comfort, neutral, and discomfort states in children's emotional speech for ASD Children:

GeMAPSv01b accuracy = 0.687;
eGeMAPSv01b accuracy = 0.671;

2. The average accuracy of automatic classification of comfort, neutral, and discomfort states in children's emotional speech for DS children:

GeMAPSv01b accuracy = 0.725;
eGeMAPSv01b accuracy = 0.717.

3. The average accuracy of automatic classification of comfort, neutral, and discomfort states in children's emotional speech for TD children:

GeMAPSv01b accuracy = 0.641;
eGeMAPSv01b accuracy = 0.631.

4. The average accuracy of automatic classification of comfort, neutral, and discomfort states in children's emotional speech for all diagnoses:

GeMAPSv01b accuracy = 0.694;
eGeMAPSv01b accuracy = 0.686.

The results show that the minimalistic feature set GeMAPS produces slightly better accuracy than the extended feature set eGeMAPS. One reason for that might be that low-level features are highly representative for our task. It also highlights the importance of not simply collecting all available features but approaching the feature selection process with an expert understanding of the underlying meaning of those features and choosing the ones that are representative for the specific domain.

4. DISCUSSION AND CONCLUSIONS

Our first experiments on the recognition of "comfort-neutral–discomfort" states in children's speech on the EmoChildRu dataset (7) showed that the performance of automatic classification based on the SVM classifier is very similar to the human perception, and both are higher than chance level, i.e. 50% for the three-class problem (8). During these experiments it was observed better classification performance with IS 2010 features compared to IS 2013 features, both are subsets of the ComParE_2016 feature set. The features were extracted with the same openSMILE toolkit, where the IS 2010 feature set (1582 acoustic features) (9) is smaller than the IS 2013 feature set (6373 features) (10).

Based on this finding, in our experiments based on the SVM classifier we used even smaller feature sets GeMAPS (62 features) и eGeMAPS (88 features) (5). Our experimental results confirmed the trend, for the smaller feature set GeMAPS we observed slightly better classification performance compared to the eGeMAPS feature set. Moreover, the classification performance for both feature sets is much higher than chance level.

In addition, we observed that the classification performance for different diagnosis (ASD, DS, and TD) separately remains approximately at the same high level. However, we have found that the state of comfort is recognized better than for TD children, the state of discomfort is recognized better for ASD children, and the neutral state is recognized better for ASD children.

Our experiments with another feature set (MFCC) and classifier (GMM) showed worse results compared to GeMAPS feature set and SVM classifier for all emotion states.

The algorithms for the automatic detection of the age, gender, and emotional state of a child in combination with the algorithms for the generation of emotional speech can be utilized for building edutainment systems (11), and systems for assisted or alternative communication for children with development disorders.

ACKNOWLEDGEMENTS

The study was financially supported by the Russian Science Foundation (projects № 18-18-00063 – database "AD_CHILD.RU" development and preliminary results of experiments, № 22-45-02007 - detailed data processing).

REFERENCES

1. Lyakso EE, Ruban N, Frolova OV, Gorodnyi VA, Matveev YN. Approbation of a method for studying the reflection of emotional state in children's speech and pilot psychophysiological experimental data. International Journal of Advanced Trends in Computer Science and Engineering. 2020;9(1):649-656.
2. Lyakso E, Frolova O, Kaliyev A, Gorodnyi V, Grigorev A, Matveev Y. AD-Child.Ru: Speech corpus for Russian children with atypical development. Lecture Notes in Computer Science. 2019;1658:299-308.
3. openSMILE Python [cited 2022 July 14]. Available from: https://github.com/audeering/opensmile-python
4. openSMILE Python [cited 2022 July 14]. Available from: https://audeering.github.io/opensmile-python/
5. Eyben F, Scherer KR, Schuller BW, Sundberg J, Andre E, Busso C, Devillers, LY, Epps J, Laukka P, Narayanan SS, Truong K. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. 2016;7:190–202.
6. sklearn.svm.SVC [cited 2022 July 14]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
7. Lyakso E, Frolova O, Dmitrieva E, Grigorev A, Kaya H, Salah AA, Karpov A. EmoChildRu: Emotional child Russian speech corpus. Lecture Notes in Computer Science. 2015;9319:144–152.
8. Kaya H, Ali Salah A, Karpov A, Frolova O, Grigorev A, Lyakso E. Emotion, age, and gender classification in children's speech by humans and machines. Computer Speech & Language. 2017;46:268-283.
9. Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, Müller C, Narayanan SS. The INTERSPEECH 2010 paralinguistic challenge. Proc INTERSPEECH 2010; 26-30 September 2010; Makuhari, Chiba, Japan 2010. p. 2794-2797.
10. Schuller B, Steidl S, Batliner A, Vinciarelli A, Scherer K, Ringeval F, Chetouani M, Weninger F., Eyben F, Marchi E, Mortillaro M, Salamin H, Polychroniou A, Valente F, Kim S. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. Proc INTERSPEECH 2013; 25-29 August 2013; Lyon, France, 2013. p. 148-152.
11. Guran AM, Cojocar GS, Diosan LS. The Next Generation of edutainment applications for young children—A Proposal. Mathematics. 2022;10(4):645.

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0435

CNN based Emotion Detection in Cross Linguistic Children speech

Elena LYAKSO[1]; Olga FROLOVA[1]; Nersisson RUBAN[2]; A. Mary MEKALA[3]; Alex Noel JOSEPH RAJ[4]; Anton MATVEEV[1]; Yuri MATVEEV[1]

[1] The Child Speech Research Group, St. Petersburg State University, St. Petersburg, Russia

[2] School of Electrical Engineering, VIT, Vellore, India

[3] School of Information Technology & Engineering, VIT, Vellore, India

[4] Key Laboratory of Digital Signal and Image Processing of Guangdong Province, Department of Electronics Engineering, College of Engineering, Shantou University, China

ABSTRACT

The goal of this cross-linguistic study is to analyze the speech of children in the age group of 8-12 years and automatic classification of children's emotions from various speech data in Russian and Tamil languages. A standardized approach for the collection of speech (spontaneous and acting speech - emotional words, phrases, and meaningless texts) was used. Two emotional child speech corpora include the emotional speech of 95 Russian children and 40 Indian children were created. Annotation of the emotional speech was carried out in four categories “joy - neutral - sadness – anger” by two speech specialists of the same nationality of the child. The set of 2505 labeled audio files of Russian children and 418 labeled audio files of Tamil children were used. The SVM classifier shows slightly better results in Russian language, and the MLP classifier in Tamil. Inter-Cultural approach (mixed dataset of Russian and Indian speech) revealed that the accuracy of recognition of all emotions remains higher. In Cross-Cultural approach, on samples of Tamil speech, anger state is recognized better, and in samples of Russian speech – sad state. Using 5 Layered CNN Model, the accuracy was revealed for Russian speech is higher than for Indian children.

Keywords: Child Speech, Emotion Detection, Cross-Linguistic Study

1. INTRODUCTION

In recent decades, the problem of automatic emotions recognition in speech has been widely studied, which is associated with a practical focus on the fields of social sphere, medicine, and education. Most of the work on automatic recognition of emotional speech has been carried out on the speech of adults (1-4). The reason for lower number of works on automatic recognition of children’s emotional speech is primarily due to (5): difficulties in recording children’s acting speech and the fact that systems trained to recognize emotions from adult speech show rather low efficiency when used to recognize emotional children’s speech (6). There are fewer child speech databases available to the scientific community, for example, (7-14). Works on the recognition of emotional children’s speech for Russian children are rare (7), as well as for Indian children (15). No cross-cultural data on automatic speech recognition of Russian and Tamil children are available.

The aim of the study is automatic recognition of the emotional state of children aged 8-12 years by

[1] lyakso@gmail.com
[1] olchel@yandex.ru
[1] aymatveev@gmail.com
[1] yunmatveev@gmail.com
[2] nruban@vit.ac.in
[3] amarymekala@vit.ac.in
[4] jalexnoel@stu.edu.cn

speech in Russian and Tamil.

2. METHODS

2.1 Dataset

To study the cross-linguistic recognition of the emotional state of Indian and Russian children of 8-12 years of age via their speech by humans and machine, two language-specific corpora of emotional speech were created. Each corpus contains records of spontaneous and acting speech of 8-12 years old children. Russian corpus includes emotional speech of 95 children; Indian corpus contains speech data of 40 children (17). Speech recording was performed according to a standardized protocol. Each recording session for every child included a dialogue with the experimenter with a standard set of questions (16) and the acting speech – words and phrases reflecting the emotional state, and meaningless texts.

Acting speech: Russian and Tamil children pronounced words and phrases in their native language, reflecting different emotional states. The emotional state was reflected in the lexical meaning of words and phrases (17). The children pronounced the speech, manifesting the emotional state. Russian children pronounced a meaningless text – the first quatrain of “Jabberwocky”, the poem by Lewis Carroll (18), and the meaningless (sentence) by L.V. Shcherba “glokaya kuzdra”, 1930 (19); Indian children spoke the meaningless text about Grandpa (20) and Tamil meaningless phrases.

The recording time varied from 30 min to 60 min and included a training session for pronouncing meaningless texts. The place of recording of child’s speech and behavior was a laboratory condition without special soundproofing. The recordings of speech of children were made by the “Marantz PMD660” recorder with an external microphone “SENNHEIZER e835S” with the following settings: the sampling rate was set to 16,000 Hz and the mono audio channel was used in all the recording sessions. Parallel with the recording of the speech, the child's behavior and facial expression were recorded using a video camera “SONY HDR-CX560E”. The distance from the child’s face to the microphone was 30 – 50 cm. All speech files were stored in .wav format, 44.100 Hz, 16 bits per sample.

The child’s emotional speech was annotated into four categories (based on video recording and protocol of the recording situation) “joy - neutral - sad – anger” by two speech experts for each language. The speech samples were assigned to the corresponding emotional category when there was an agreement between the two annotators.

2.2 Experiment 1: Automatic Classification

To train and test our automatic emotion recognition system we used the set of 2505 labeled audio files with acting emotional speech of Russian speaking children and 418 labeled audio files with acting emotional speech of Tamil speaking children.

2.2.1 Features and Classifiers

In experiments on automatic recognition, we used the DisVoice python framework (21) designed to compute features from speech files and the openSMILE (22). Dis Voice computes glottal, phonation, articulation, prosody and phonological features, and contains 103 features, computed based on pitch, energy and duration. Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (23) designed for speech emotion recognition was used. The eGeMAPSv02 feature set contains 88 parameters.

To implement SVM and MPLP classifiers we have used the scikit-learn machine learning library (24, 25) SVM is a deterministic and MLP is a non-deterministic supervised classification learning algorithm that are efficient in emotion recognition tasks, see e.g. (26, 27).

Total 2505 Russian data were used. We separated the data into 80% (2004 samples) for training and 20% (501 samples) for testing our classification models. Total 417 Indian data were used. We separated the data into 80% (323 samples) for training and 20% (84 samples) for testing our classification models.

2.2.2 Validation procedure

The K-fold cross-validation method was used. Experiments with K=6-fold cross-validation (5:1 ratio of training dataset to test dataset size), as well as cross-validation on individual objects (Leave-One-Out, LOO or Leave-One-Subject-Out, LOSO) were conducted. For K=6-fold cross-validation we have used Stratified K-Folds cross-validator from scikit-learn library (28), which provides a good balance of classes, especially for datasets of small size. For cross-validation by

individual objects Leave-One-Out cross-validator from scikit-learn library was used (29).

2.2.3 Evaluation setup

In our experiments, we used evaluation metrics per class Accuracy, Precision, Recall, F1-score, and Unweighted Average Precision (UAP).

2.2.4 Design

1. Classifiers are trained and tested on data of the same language/culture – Intra-lingual /Intra-Cultural (30). The goal is to evaluate the quality of training on each of the datasets used in the experiments and adjust the parameters of the selected classifiers in cross-validation procedures, etc.

2. Evaluation of the effectiveness of emotion recognition when training classifiers on a combined dataset of Russian and Tamil speech. A framework in which classifiers are trained and tested on data from different languages/cultures is called Inter - Lingual /Inter - Cultural (30).

3. Evaluation of the effectiveness of emotion recognition in a framework in which classifiers are trained on one set of language/culture data and tested on another set of language/culture data, the so-called Cross-Lingual/Cross-Cultural (31).

2.3 Experiment 2: Automatic Recognition Using Deep Learning Algorithms

For Automatic recognition, we have used popular deep learning algorithms i.e. CNN (Convolutional Neural Network) 1D layer Architecture to build the network model. Convolutional Neural Network (CNN) is a Deep Learning based technology that has the capabilities to achieve high precision in recognition. CNN has multiple layers where each layer performs a specific transformation function. Convolutional is the first layer to extract features from the input audio. The convolutional will then preserves the relationship between pitch and stress by learning audio feature using small squares of input data. Convolution of an audio with different filters can perform operations such as edge detection, blur and sharpen by applying filters. The purpose of an activation layer is to introduce non-linearity in ConvNet. The real data would want our ConvNet to learn would be non-negative linear values. Next, the pooling layer functions to reduce the number of parameters when the audio is too large. Spatial pooling also called subsampling or down sampling which reduces the dimensionally of each map but retains important information. Spatial pooling can be of different types which are max pooling, average pooling or sum pooling. Full connected layer is flattened the matrix into vector and feed it into a fully connected layer like a neural network. Figure 1 shows CNN architecture.

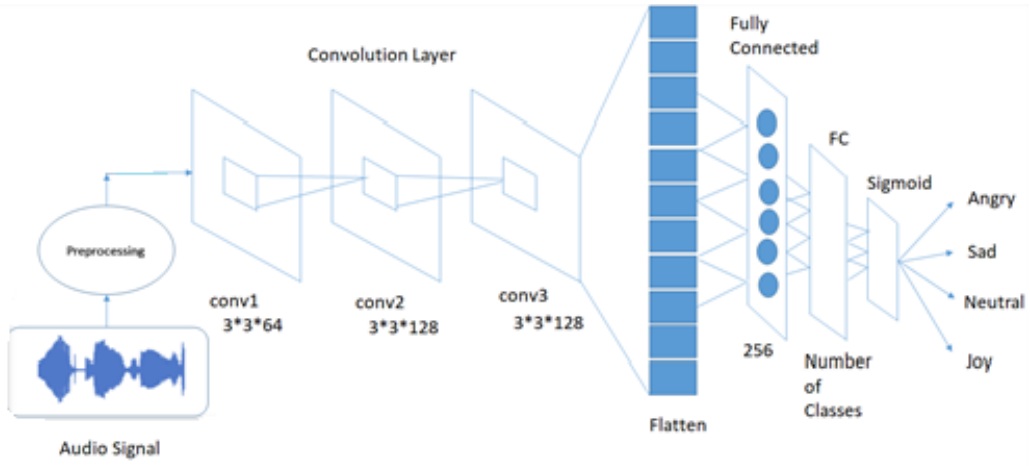


Figure 1– CNN Architecture

2.3.1 Pre-processing

Pre-processing is an important part of preparing data to achieve model accuracy and efficiency. In this phase, we clean the audio signals to remove the background noises, silent portion and other irrelevant information from speech signal using the adaptive threshold-based pre-processing. In this method, we find the relationship of energy with amplitude in speech signal using direct relation policy. The energy amplitude relationship is that the amount of energy passed by a wave is correlated to the amplitude of the wave. A high energy wave is considered by high amplitude; a low energy wave is considered by low amplitude. The amplitude of a wave mentions the extreme amount of displacement

of an element in the middle from its rest location. The logic underlying the energy-amplitude relationship is as follows to remove the silent and unnecessary particle from speech signals. Three steps are included in this process; first, read the audio file step by step with 1600 sampling rates. In the last step, we reconstruct a new audio file with the same sample rates without any noise and silent signals.

2.3.2. Implementation of CNN model

In this model, we have used Adam optimizer, to change the attributes of the neural network i.e. to reduce the loss value. And we used the Sigmoid Activation function, it returns value of Sigmoid Function is mostly in the range of values between 0 and 1 or -1 and 1. 2505 speech samples of Russian children were used. 70% of the data (1753 samples) were used for training and 30% (752 samples) for testing classification models: 80% a (334 samples) and 20 % (84 samples) - of Indian data for training and testing. Machine learning used XG Boost Classifier. XG Boost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.

3. RESULTS

3.1 Experiment 1

3.1.1 Intra-Cultural approach

Separate experiments were conducted for the SVM classifier on datasets of Russian and Tamil speech. Similar results were obtained for the MLP classifier. The results of experiments for the SVM classifier trained on the set of prosodic features of the DisVoice library, which expands the set of features used in expert testing (17), showed that all emotions are recognized in Russian with a probability of more than 50%, with better recognition of the state of anger, in Tamil - the state of anger (52%) only. Confusion matrices for emotion classification separately for Russian and Indian children, SVM and MLP classifiers, eGeMAPS feature set, cross-validation method Leave-One-Out, LOO or Leave-One-Subject-Out (LOSO) presented in Table 1.

Table 1 – Confusion matrices for emotion classification separately for Russian and Tamil children, SVM and MLP classifiers, eGeMAPS feature set, LOO cross-validation method: % automatic system responses

Classifier	Language	Actual emotion state	Predicted emotion state			
			Anger	Joy	Neutral	Sadness
SVM	Russian	Anger	**77.62**	11.58	4.85	5.95
		Joy	15.03	**71.28**	6.08	7.60
		Neutral	6.37	5.57	**64.65**	23.41
		Sadness	5.88	6.35	16.25	**71.52**
	Tamil	Anger	**52.63**	22.81	16.67	7.89
		Joy	36.84	**28.42**	14.74	20.00
		Neutral	20.19	25.96	**31.73**	22.12
		Sadness	9.52	21.90	24.76	**43.81**
MLP	Russian	Anger	**74.96**	12.52	6.73	5.79
		Joy	14.02	**70.27**	7.77	7.94
		Neutral	6.53	6.69	**66.08**	20.70
		Sadness	4.64	6.35	19.97	**69.04**
	Tamil	Anger	**52.63**	21.93	16.67	8.77
		Joy	24.21	**28.42**	25.26	22.11
		Neutral	11.54	30.77	**33.65**	24.04
		Sadness	10.48	21.90	20.00	**47.62**

In Tamil, both the SVM and MLP classifiers have prediction accuracy above 50% for the emotional state of anger only (52.63% - SVM & MLP).

Table 2 – Per-class performance of multi-class classification, eGeMAPS feature set

Classifier	SVM							
Language	Russian				Tamil			
Class	Anger	Joy	Neutral	Sad	Anger	Joy	Neutral	Sad
Accuracy	0.876	0.869	0.844	0.836	0.715	0.644	0.689	0.734
Recall	0.776	0.713	0.646	0.715	0.526	0.284	0.317	0.438
Precision	0.740	0.752	0.704	0.659	0.442	0.287	0.361	0.467
Classifier	MLP							
Accuracy	0.874	0.862	0.829	0.837	0.766	0.635	0.679	0.732
Recall	0.750	0.703	0.661	0.690	0.526	0.284	0.337	0.476
Precision	0.748	0.733	0.657	0.667	0.532	0.276	0.352	0.464

Average performance of multiclass classification, eGeMAPS feature set: Overall Accuracy (SVM – 0.856 - for Russian, 0.696 – for Tamil; MLP - 0.850 & 0.703 – correspondently), UAR (SVM – 0.713 - for Russian, 0.391 – for Tamil; MLP - 0.701 & 0.406 – correspondently).

The worst performance of emotion recognition in Tamil speech on all feature sets and on all classifiers can be explained by the fact that there are almost six times fewer data available for training automatic classifiers in Tamil than in Russian - 418 and 2505, respectively. The SVM classifier shows slightly better results in Russian, and the MLP classifier in Tamil.

3.1.2 Inter-Cultural approach

Automatic recognition of emotions by the SVM and MLP classifiers is approximately the same in terms of accuracy. Therefore, further experimental results are presented only for the SVM classifier. The recognition performance of the mixed dataset of Russian and Indian speech decreases in proportion to the proportion of Indian speech samples in the mixed dataset, but the accuracy of recognition of all emotions remains above chance (Table 3).

Average performance of multi-class classification, SVM classifier, eGeMAPS feature set, bilingual dataset: Overall Accuracy - 0.820 (0.842 –for Russian, 0.691 – for Tamil); UAR - 0.641 (0.683 – for Russian, 0.383 – for Tamil); UAP - 0.642 (0.686, 0.385 – for Russian &Tamil). The accuracy for recognizing emotions does not differ significantly between intracultural and intercultural approaches, which is consistent with the results of studies by other authors (32).

Table 3 – Per-class performance of multi-class classification, SVM classifier, eGeMAPS feature set, bilingual dataset

Classifier	SVM			
Language	Russian + Tamil			
Class	Anger	Joy	Neutral	Sad
Accuracy	0.832	0.838	0.808	0.804
Recall	0.717	0.632	0.578	0.636
Precision	0.649	0.693	0.625	0.601

3.1.3 Cross-Cultural approach

The task of cross-lingual/cross-cultural emotion recognition is one of the most difficult tasks when creating SER systems due to differences in emotion expression in different languages and cultures. When training SER on samples of Russian speech and recognizing emotions on Tamil speech samples, only one emotion - Anger (0.69-SVM; 0.57-MLP) is recognized with a probability above 50%, but it is

often confused with the emotion that reflects the state of joy (0.6-SVM). When training SER on Tamil speech samples and emotion recognition on Russian speech samples, only the state of sadness (0.778-SVM; 0.797 – MLP) is recognized with a probability higher than 50%, but it is also mixed with the neutral state (0.82-MLP) (Table 4).

Table 4 – Per-class performance of multi-class classification, eGeMAPS feature set, bilingual dataset

Classifier	SVM							
Language	Russian -Tamil				Tamil-Russian			
Class	Anger	Joy	Neutral	Sad	Anger	Joy	Neutral	Sad
Accuracy	0.515	0.683	0.652	0.743	0.701	0.743	0.698	0.701
Recall	0.693	0.232	0.221	0.038	0.279	0.162	0.045	0.279
Precision	0.298	0.317	0.264	0.361	0.370	0.461	0.149	0.370
Classifier	MLP							
Accuracy	0.612	0.672	0.647	0.745	0.714	0.749	0.699	0.571
Recall	0.570	0.253	0.375	0.152	0.399	0.186	0.083	0.797
Precision	0.337	0.309	0.323	0.466	0.423	0.495	0.224	0.345

Experiments with the cross-cultural approach showed low recognition accuracy results, although in experiments with training and testing separately by language the recognition accuracy of all emotions was above 50%. This indicates that there is a difference in the expression of emotions in the speech of Russian and Tamil speaking children, correlated with cultural characteristics.

3.2 Experiment 2: CNN based Emotion Detection

Automatic recognition of the Russian children's emotion includes four classes: anger, joy, neutral, sad. Total 2505 Russian data into 70% (1753 samples) for training and 30% (752 samples) for testing the classification models were separated. For automatic recognition of emotions by speech of Indian children by Tamil speech was separated the data into 80% (334 samples) for training and 20% (84 samples) for testing the classification models (Table 5).

Table 5 - Testing: Confusion matrix for 5 Layered CNN Model, Russian language

Language	Russian				Tamil			
	Anger	Joy	Neutral	Sad	Anger	Joy	Neutral	Sad
Anger	**126**	28	24	14	**10**	2	5	6
Joy	25	**113**	22	17	4	**8**	4	3
Neutral	5	8	**136**	41	0	4	**9**	8
Sad	8	6	31	**148**	1	4	2	**14**

Table 6 - Classification Report for CNN model

	Russian				Tamil			
	Anger	Joy	Neutral	Sad	Anger	Joy	Neutral	Sad
Recall	0.66	0.64	0.72	0.77	0.43	0.42	0.43	0.67
Precision	0.77	0.73	0.64	0.67	0.67	0.44	0.45	0.45
F1-Score	0.71	0.68	0.67	0.72	0.53	0.43	0.44	0.54
Accuracy	0.70				0.49			

In separation of the Russian data into 70:30 for training and testing have approximately the accuracy of 70% for testing and 91% of training. In separation of the Indian data into 80:20 for training and testing have approximately the accuracy of 49% for testing and 68% of training.

4. CONCLUSIONS

The results of automatic classification of emotional acting speech of Russian and Tamil children (Intra-Cultural approach) were obtained using modern algorithms/classifiers of machine learning methods on the collected data sets. The best accuracy of emotion recognition for the Russian speech vs Tamil speech on all sets and classifiers is obtained. The SVM classifier shows slightly better results in Russian language, and the MLP classifier in Tamil. The Inter-Cultural approach revealed that the recognition performance of the mixed dataset of Russian and Indian speech falls corresponding to the proportion of Indian speech samples in the mixed dataset, but the accuracy of recognition of all emotions remains above chance. In the Cross-Cultural approach, on samples of Tamil speech, anger state is recognized above chance, on samples of Russian speech – the sad state. It indicates that there is a difference in the expression of emotions in the speech of Russian- and Tamil-speaking children, associated with cultural characteristics. The accuracy of emotion recognition using CNN for the speech of Russian children is higher than for Indian children.

ACKNOWLEDGEMENTS

The research was financially supported by the Russian Science Foundation – RSF-DST (project 22-45-02007) – for Russian researchers, and Department of Science and Technology (DST) INT/RUS/RFBR/382) - for Indian researchers.

REFERENCES

1. Akçay MB, Oguz K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun. 2020;166:56–76.
2. Schuller DM, Schuller BW. A Review on five recent and near-future developments in computational processing of emotion in the human voice. Emotion Review. 2021;13(1):44-50.
3. Rouast Ph, Marc A, Raymond Ch. Deep learning for human affect recognition: Insights and new developments. IEEE Transactions on Affective Computing. 2021;12:524-543.
4. Swain M, Routray A, Kabisatpathy P. Databases, features and classifiers for speech emotion recognition: A review. Int. J. Speech Technol. 2018;21:93–120.
5. Onwujekwe D. Using Deep Leaning-Based Framework for Child Speech Emotion Recognition. PhD Thesis, Virginia Commonwealth University, Richmond, VA, USA, 2021.
6. Palo HK, Mohanty MN, Chandra M. Emotion analysis from speech of different age groups. Proc Second International Conference on Research in Intelligent and Computing in Engineering; Vol. 10; 24-26 March 2017; Gopeshwar, Uttrakhand, India 2017. p. 283–287.
7. Kaya H, Ali Salah A, Karpov A, Frolova O, Grigorev A, Lyakso E. Emotion, age, and gender classification in children's speech by humans and machines. Computer Speech & Language. 2017;46:268-283.
8. Kennedy J, Lemaignan S, Montassier C, Lavalade P, Irfan B, Papadopoulos F, Senft E, Belpaeme T. Child speech recognition in human-robot interaction: Evaluations and recommendations. Proc 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI'17); 06-09 March 2017; Vienna, Austria 2017. p. 82-90.
9. Pérez-Espinosa H, Reyes-García C, Villaseñor-Pineda L. EmoWisconsin: An Emotional Children Speech Database in Mexican Spanish. Proc 4th International Conference on Affective Computing and Intelligent Interaction (ACII); 9-12 October 2011; Memphis, TN, USA 2011. p. 62–71.
10. Steidl S. Automatic Classification of Emotion Related User States in Spontaneous Children's Speech. Berlin, Germany: Logos Verlag; 2009.
11. Batliner A, Blomberg M, D'Arcy S, Elenius D, Giuliani D, Gerosa M, Hacker C, Russell MJ, Steidl S, Wong M. The PF_STAR children's speech corpus. Proc INTERSPEECH 2005; 4-8 September 2005; Lisbon, Portugal 2005. p. 2761–2764.
12. Batliner A, Steidl S, Noth E. Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus. Proc LREC-2008 Workshop of on Corpora for Research on Emotion and Affect; 26 May 2008; Marrakech, Morocco 2008. p. 28–31.
13. Gerosa M, Giuliani D, Brugnara F. Acoustic variability and automatic recognition of children's speech.

Speech Commun. 2007;49(10-11):847–860.
14. Bell L, Boye J, Gustafson J, Heldner M, Lindstrom A, Wiren M. The Swedish NICE Corpus-spoken dialogues between children and embodied characters in a computer game scenario. Proc INTERSPEECH 2005; 4-8 September 2005; Lisbon, Portugal 2005. p. 2765–2768.
15. Mohanty MN, Palo HK. Child emotion recognition using probabilistic neural network with effective features. Measurement. 2020;152(3):107369.
16. Lyakso E, Ruban N, Frolova O, Gorodnyi V, Matveev Yu. Approbation of a method for studying the reflection of emotional state in children's speech and pilot psychophysiological experimental data. International Journal of Advanced Trends in Computer Science and Engineering. 2020;9(1):649-656.
17. Lyakso E, Frolova O, Ruban N, Mekala AM. The Child's emotional speech classification by human across two languages: Russian & Tamil. Lecture Notes in Computer Science. 2021;12997:384-396.
18. Carrol L. Through the Looking-Glass and What Alice Found There. London, UK: Macmillan and Co; 1872.
19. GLOKAYA KUZDRA [cited 2022 July 09]. Available from: http://languagehat.com/glokaya-kuzdra
20. Heyman M, Satpathy S, Ravishankar A. The Tenth Rasa: An Anthology of Indian Nonsense. New Delhi, India: Penguin Books; 2007.
21. Disvoice's documentation [cited 2022 July 09]. Available from: https://disvoice.readthedocs.io/en/latest/index.html
22. openSMILE Python [cited 2022 July 09]. Available from: https://github.com/audeering/opensmile-python
23. Eyben F, Scherer KR, Schuller BW, Sundberg J, Andre E, Busso C, Devillers, LY, Epps J, Laukka P, Narayanan SS, Truong K. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. 2016;7:190–202.
24. Support Vector Machines [cited 2022 July 09]. Available from: https://scikit-learn.org/stable/modules/svm.html#svm
25. Multi-layer Perceptron classifier [cited 2022 July 09]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html
26. Javaheri B. Speech & song emotion recognition using multilayer perceptron and standard vector machine. ArXiv. Preprints 2021; 2021050441. https://doi.org/10.20944/preprints202105.0441.v1
27. Farooq M, Hussain F, Baloch NK, Raja FR, Yu H, Zikria YB. Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. Sensors. 2020;20(21):6008.
28. Stratified K-Folds cross-validator [cited 2022 July 09]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html
29. Leave-One-Out cross-validator [cited 2022 July 09]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.LeaveOneOut.html
30. Neiberg D, Laukka P, Elfenbein HA. Intra-, inter-, and cross-cultural classification of vocal affect. Proc INTERSPEECH 2011; 27-31 August 2011; Florence, Italy 2011. p. 1581-1584.
31. Sun J, Ahn H, Park ChY, Tsvetkov Y, Mortensen DR. Cross-cultural similarity features for cross-lingual transfer learning of pragmatically motivated tasks. Proc 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume; 19-23 April 2021; online 2021. p. 2403–2414.
32. Neumann M, Vu NT. Cross-lingual and multilingual speech emotion recognition on English and French. Proc IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 15-20 April 2018; Calgary, AB, Canada 2018. p. 5769-5773.

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022, in Gyeongju, Korea

Extended abstract

ABS-0720

Phonetic variation in final consonants in Thai: The case of Thai preschool children

Weena WUTTHICHAMNONG[1]

[1] Department of Thai, Faculty of Arts, Silpakorn University, Thailand

ABSTRACT

Children's and adults' pronunciations are somewhat different. Previous studies indicated that children's pronunciation is unclear at an early age. This article aims at examining phonetic variation in final consonants pronounced by Thai preschool children. The data elicited are from sixteen video-recorded episodes of a Thai variety show named SUPER 10 Season 1-5 which participants are under-three-year-old preschool children. The findings reveal that there are three patterns of phonetic variations in final consonant pronunciation, namely 1) correct pronunciation (94.61%), 2) deletion (3.06%), which usually occur with closed syllables, especially syllables with /n/ in the final position, and 3) substitution (2.33%), which usually occur with final phonemes that have the same manner of articulation, but a different place of articulation. The results indicate that manner of articulation might affect the phonetic variations in the final consonant pronunciation of Thai preschool children.

Keywords: Phonetic variation, Final consonant, Thai

1. INTRODUCTION

Children's and adults' pronunciations are somewhat different. Previous studies indicated that children's pronunciation is still unclear at an early age, in prekindergarten and kindergarten. Preschool children are in the process of developing their speech production organs, such as the tongue, teeth, and palate, hence poor pronunciation is common in this age group. (3).

According to a study of phonological development of young Thai children, the tone system, the vowel sound system, and the consonant sound system would develop respectively (5).

In terms of the Thai consonant sound system, Thai syllables can end with a long vowel sound (open syllable), or one of the nine final consonants sounds as shown in Table 1.

Table 1 - Thai final consonant sound system

	Labial	Alveolar	Palatal	Velar	Glottal
Stop	p	t		k	ʔ
Nasal	m	n		ŋ	
Semivowel	w		j		

Table 1 shows the Thai final consonant sound system. According to Table 1, there are nine consonantal phonemes in the final position in Thai, namely /p/, /t/, /k/, /ʔ/, /m/, /n/, /ŋ/, /w/, /j/.

SUPER 10 is a Thai variety show broadcasting on Workpoint Entertainment Television Channel and WorkpointOfficial YouTube Channel. The main content is letting children under fifteen years old show their talents such as sports, music, and special interests. The children would receive some missions they have to complete. There will be two or three commentators who judge whether the children can pass the mission or not. If the children can do the mission, they will receive what they desired. There are various talented children of various ages who participate in this program. According to these characteristics of the program, the children have many opportunities to interact with the moderator and the commentators. Consequently, SUPER 10 can be a good database for collecting data

[1] weena.wtcn@gmail.com

about the pronunciation of Thai children.

Although there are some studies on Thai children's pronunciation (1, 2, 4), studies on phonetic variation may be quite not much. Therefore, this article aims to examine phonetic variation in final consonants pronounced by Thai preschool children.

2. METHODS

The samples of this study consist of under-three-year-old Thai preschool children who participate in SUPER 10 Season 1-5 as shown in Table 2.

Table 2 - Samples of the study

Age (years)	Sex		Number of samples
	Male	Female	
2	7	2	9
3	6	1	7
	Total		16

According to the qualification of the samples, sixteen video-recorded episodes of SUPER 10 were conducted. The samples' utterances were then transcribed into Thai according to the Thai phonetic system. Error analysis was adopted to examine differences between standard pronunciation and the pronunciation that the samples pronounced.

3. RESULTS

According to the data elicited, there are 3,431 syllables in total that all of the samples pronounced. The findings reveal three patterns of phonetic variations in final consonant pronunciation, namely correct pronunciation, deletion, and substitution as shown in Table 3.

Table 3 - Phonetic variations in final consonant pronunciation

Final consonantal phonemes	Patterns of phonetic variations in final consonant pronunciation										Number of syllables
	Correct pronunciation	Deletion	Substitution								
			/p/	/t/	/k/	/m/	/n/	/ŋ/	/w/	/j/	
Open-syllable and /ʔ/	1,246	-	-	-	-	-	-	-	-	-	1,246
/p/	363	12	-	2	-	-	-	-	-	1	378
/t/	209	8	1	-	12	-	-	-	-	-	230
/k/	170	8	-	8	-	1	-	1	-	-	188
/m/	135	9	1	1	-	-	2	5	2	-	155
/n/	348	31	-	-	-	7	-	25	-	-	411
/ŋ/	241	26	-	-	-	2	9	-	-	-	278
/w/	166	6	-	-	-	-	-	-	-	-	172
/j/	368	5	-	-	-	-	-	-	-	-	373
Total	3,246	105	80								3,431

3.1 Correct Pronunciation

Correct pronunciation is pronouncing final consonantal phonemes accurately, similar to Thai standard pronunciation. According to the data, 3,246 syllables (94.61%) were pronounced final consonantal phonemes correctly.

Correct pronunciation	Meaning	Child's pronunciation
/sìː/	'four'	/sìː/
/cèp/	'to hurt'	/cèp/
/t^hɔ̀ːt/	'to take off'	/t^hɔ̀ːt/
/jâːk/	'difficult'	/jâːk/
/k^hàʔ/	'Thai final particle for female'	/k^hàʔ/
/sôm/	'orange'	/sôm/
/nɔːn/	'to sleep'	/nɔːn/
/liŋ/	'monkey'	/liŋ/
/mɛːw/	'cat'	/mɛːw/
/jaːj/	'grandmother'	/jaːj/

3.2 Deletion

Deletion is omitting consonantal phonemes at the end of a syllable. It usually occurs with closed syllables especially syllables with /n/ in the final position. According to the data, 105 syllables (3.06%) were omitted in final consonantal phonemes.

Correct pronunciation	Meaning	Child's pronunciation
/c^hôːp/	'to like'	/c^hôː/
/bɔ̀ːt/	'board'	/bɔ̀ː/
/jàːk/	'to want'	/jàː/
/cam/	'to remember'	/ca/
/jɯːn/	'to stand'	/jɯː/
/t^hɔːŋ/	'gold'	/t^hɔː/
/léːw/	'already'	/léː/
/jaŋ/	'not yet'	/ja/

3.3 Substitution

Substitution is replacing a correct final consonantal phoneme with another sound. The correct one usually is replaced by the final consonantal phoneme which has the same manner of articulation, but a different place of articulation. According to the data, 80 syllables (2.33%) were replaced by other final consonantal phonemes.

Correct pronunciation	Meaning	Child's pronunciation	Substitution
/k^hráp/	'Thai final particle for male'	/ʔát/	p (stop) -----> t (stop)
/nùat/	'mustache'	/nùak/	t (stop) -----> k (stop)
/bɔ̀ːk/	'to tell'	/bɔ̀ːt/	k (stop) -----> t (stop)

Correct pronunciation	Meaning	Child's pronunciation	Substitution
/pʰrɔ́ːm/	'ready'	/pʰɔ́ːŋ/	m (nasal) -----> ŋ (nasal)
/kʰǒn/	'to carry'	/kʰǒŋ/	n (nasal) -----> ŋ (nasal)
/ŋûaŋ/	'sleepy'	/ŋûan/	ŋ (nasal) -----> n (nasal)

4. CONCLUSIONS

This article aims to analyze the pronunciation of Thai final consonant sounds pronounced by sixteen Thai preschool children. According to the data, three phonetic variations in final consonant pronunciation were found; correct pronunciation, deletion, and substitution.

The results indicate that manner of articulation might affect the phonetic variations in the final consonant pronunciation of Thai preschool children as obviously shown in the pattern of substitution.

The findings appear to conform to the previous studies. According to Rattanatamrong, Kongmeesub, Dittaporn, Siwahansaphan, and Chatarupa (3), the first syllable of a two-syllable word in Thai is usually unstressed, so children often pronounce the first syllable without the final consonant phoneme.

In addition, the data elicited in this study demonstrated that Thai preschool children can pronounce initial consonants more accurately than the final ones. It also supports the concept of the pronunciation development process that children can pronounce the final consonant after they learn to articulate the initial one, as a consequence, kids acquire open syllables before closed syllables (3).

REFERENCES

1. Juajun S. A study of Thai consonants variations of hard impaired children. [theses]. Nakhon Pathom: Silpakorn University; 2011.
2. Liamprawat S. The Thai pronunciation problems of junior primary school students of Karen and Mon ethnic groups in Kanchanaburi. Vannavidas. 2015; 15:319-344.
3. Rattanatamrong P, Kongmeesub O, Dittaporn T, Siwahansaphan N, Chatarupa S. Thai preschooler speech recognition for voice enabled interactive counting exercises. Proc 19th International Joint Conference on Computer Science and Software Engineering (JCSSE) [Internet]. 2022 [cited 2022 August 9]; 1-6. Available from: DOI: 10.1109/JCSSE54890.2022.9836310.
4. Saksiriphol D, Pangpong R, Petmunee R. A survey of pronunciation in grade 1 students in demonstration school. Journal of Research and Curriculum Development. 2019; 9(1):156-165.
5. Tuaycharoen P. A linguistic analysis of articulatory disorders among Thai children. Thai Khadi Research Institute, Bangkok: Thammasat University; 1989.

PROCEEDINGS of the
24th International Congress on Acoustics
October 24 to 28, 2022 in Gyeongju, Korea

ABS-1011

An analysis of automatic techniques for recognizing human's affective states by speech and multimodal data

Anastasia DVOYNIKOVA; Maxim MARKITANTOV; Elena RYUMINA; Mikhail UZDIAEV;
Alena VELICHKO; Ildar KAGIROV; Irina KIPYATKOVA; Elena LYAKSO; Alexey KARPOV
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia

ABSTRACT

In this paper, we present an analytical survey of state of the art models, methods and techniques for automatic recognition of affective states by analyzing human's natural speech and multimodal data such as acoustic and visual signals, as well as textually transcribed conversational speech. We consider computer based processing means for such human's affective states as natural emotions, sentiment, aggression, depression (and some other human's characteristics that may be indicators of a possible mental disease or communication disorder) suitable for analyzing both adults and young people. We also review existing speech and multimodal electronic resources, challenges and datasets available for creation and machine learning of computational models of various individual affective states. Additionally, we propose a novel methodological approach for a complex simultaneous analysis and multimodal recognition of multiple human's affective states in a parallel manner.

Keywords: Affective Computing, Speech Analysis, Multi-modal Recognition, Multi-task Recognition

1. INTRODUCTION

Over the last decade, much progress has been made in the area of affective computing as a part of artificial intelligence studies. Affective computing is a field of artificial intelligence that elaborates methods, algorithms, systems and devices in order to investigate human's affects in interaction with another person or machine (robot) (1). The notion of affect, as found within the field of affective computing, slightly differs from that coined within psychology and criminology. In fact, it is impossible to find a sound definition of affect in the affective computing works, and this calls for an overt explanation of what is understood by the term "affect" in affective computing. One can conclude, that "affect" is understood as a manifestation of psychological reactions to a stimulus, which can be either a short-term or long-term state, and have a different intensity. Based on this definition, one can identify several affect classes, as can be seen in Figure 1.

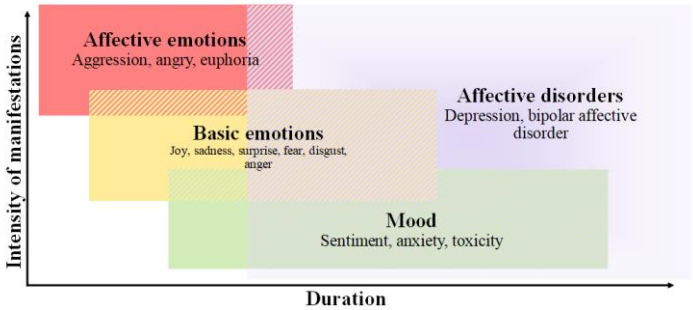


Figure 1 – Affect types systematics within the framework of affective computing

Affects in the area of data mining are usually divided into 4 classes: affective emotions, basic emotions, mood, and affective disorders. Affective emotions are characterized by a high intensity of manifestations, they are immediate and uncontrolled reactions to events, and tend to last over short periods of time. Basic emotions, in contrast to affective ones, have a lower intensity and a longer duration. Mood is an emotional state that can last for a long period. Affective disorders are psychological disorders that are characterized by an unconscious mood change, shifting to the

negative polarity, which can find their manifestation in threats, obscene language, insults, etc. Affective disorders can last from weeks up to several years. Various affective states can be closely related to each other.

Affective states are always manifested through acoustic and linguistic features of a person's speech, mimics, postures, gestures, as well as through physiological phenomena, such as heart rate, blood pressure, etc. When dealing with the automatic recognition of affective states, it is important to take into account both verbal and non-verbal features, which always have been of high importance in everyday communication acts. It is common to analyze physiological signals with the help of sensors, such as heart rate monitors, electrodes, etc. However, one should always bear in mind, that even non-invasive methods can lead to a change in the affective state of the patient. Therefore, it is video, audio, and textual modalities, that must be considered as relevant for building an automatic system for affective states analysis. The analysis of several communicative channels (multi-modal approach) has strong advantages over unimodal analysis (2), among which are high accuracy and resistance to missing data (recording errors or noisy environment).

2. EXISTING INFORMATION RESOURCES

To date, there is a large amount of electronic information resources for the analysis of affective states. In this paper, only multimodal (more than two modalities) data corpora are considered, their description is provided below. The corpora are classified in accordance with affective states, also a comparative table of the main characteristics of the corpora is proposed.

2.1 Emotions and Sentiment

The majority of current emotional corpora were collected in the course of dyadic interaction between several participants, among them are Interactive Emotional Dyadic Motion Capture (IEMOCAP) (3) (https://sail.usc.edu/iemocap/iemocap_release.htm), REmote COLlaborative and Affective interactions (RECOLA) (4) (https://diuf.unifr.ch/main/diva/recola/download.html), Russian Acted Multimodal Affective Set (RAMAS) (5). Some corpora are sets of videos hosted on YouTube (in the wild - uncontrolled conditions): CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) (6) (https://github.com/A2Zadeh/CMU-MultimodalSDK), or "Friends" TV sitcom: Multimodal EmotionLines Dataset (MELD) (7) (https://affective-meld.github.io/). Subjects who took part in recording sessions for SEWA corpus (8) first watched commercials and then discussed them between each other.

There are few multimodal corpora with sentiments: Multimodal Opinion Utterances Dataset (MOUD) (9) (http://multicomp.cs.cmu.edu/resources/moud-dataset), Multimodal Opinion-Level Sentiment Intensity (MOSI) (10) (http://multicomp.cs.cmu.edu/resources/cmu-mosi-dataset), YOUTUBE Dataset (11) (http://multicomp.cs.cmu.edu/resources/youtube-dataset-2), CH-SIMS Dataset (12) (https://github.com/thuiar/MMSA) are collections of YouTube videos or TV series/shows.

Since emotions and sentiment are related concepts, the comparative characteristics of corpora for emotion and sentiment recognition are presented together in Table 1. Usually, a set of 5-7 basic emotions is analyzed, among them anger, disgust, fear, sadness, joy, surprise and a neutral state. 3 sentiment levels mentioned in Table, are basic sentiment classes (negative, positive, and neutral), while 7 sentiment levels are an expansion of them, involving strong/weak positive and negative classes.

As one can see from Table 1, there exists a large number of multimodal corpora for emotion recognition. There are significantly fewer data sets for the recognition of sentiment through different modalities. The latter fact can be explained by the general trend in our society to express attitude via text messages. Most corpora were also recorded in the wild, containing spontaneous speech samples, and this provides an extra advantage for any automatic system for affective states analysis.

2.2 Aggression

The issues of collecting and evaluating data sets containing manifestations of aggressive behavior are addressed in a number of papers by researchers. In the works by the Delft University of Technology research group, multimodal data sets are presented, containing samples of aggressive behavior in passenger cars and at railway stations in the Netherlands (The Dataset of Aggression in Trains - TR) (13), as well as aggressive behavior in different locations of railway stations in potentially conflict situations (The Stress at Service Desk Dataset – SD) (14), and interactions between two persons in stressful situations (Negative Affect and Aggression – NAA) (15). Table 2 presents the comparative characteristics of multimodal corpora for aggression recognition.

Table 1 – Characteristics of multimodal corpora collected for emotions and sentiment analysis (V – video, A – audio, T – text, P – physiological signals)

Corpus name	Modalities	Volume, h	Labels	Language
Emotions				
IEMOCAP (3)	V, A, T	11.5	5 emotions, valence, activation, dominance	En
RECOLA (4)	V, A, P	4	Valence, intensity, dominance, consent, involvement, diligence, mutual understanding	Fr
RAMAS (5)	V, A, P	7	7 emotions, domination, submission	Ru
SEWA (8)	V, A, T	44	Valence, intensity	Zh, En, De, El, Hu, Sr
Emotions and Sentiment				
CMU-MOSEI (6)	V, A, T	66	7 emotions, 7 levels of sentiments	En
MELD (7)	V, A, T	-	7 emotions, 3 levels of sentiments	En
Sentiment				
MOUD (9)	V, A, T	0.6	3 levels of sentiments	Es
MOSI (10)	V, A, T	-	7 levels of sentiments	En
YOUTUBE (11)	V, A, T	0.3	3 levels of sentiments	En
CH-SIMS (12)	V, A, T	-	5 levels of sentiments	Zh

Table 2 – Characteristics of multimodal corpora collected for aggression analysis (V – video, A – audio, T – text, P – physiological signals)

Corpus name	Modalities	Volume, h	Labels	Language
TR (13)	V, A, T	0.6	3 levels of aggression	Nl
SD (14)	V, A, T	0.5	5 levels of stress, 3 levels of aggression	En, Nl
NAA (15)	V, A, T	-	5 levels of aggression, fear, intensity, 9 levels of valence	Nl

A comparative analysis of approaches to corpora recording confirms a weak elaboration of annotation criteria: common stereotypes of aggression are often used to find and annotate aggression, aggressive behavior is not always dissociated from anger, demonstrative behavior, etc. Moreover, the total number of informants that took part in the recordings of the considered multimodal corpora, is quite modest.

2.3 Depression

The current multimodal corpora for automatic recognition of depression were collected on the basis of clinical interviews - Pitt (16) and Distress Analysis Interview Corpus (DAIC) (17). Table 3 presents a comparative description of corpora for depression recognition.

As regards depression datasets, researchers often face the problem of data scarcity because there exist few speech corpora containing manifestations of depression; this situation is natural and caused by a range of issues. The process of collecting such specific data is labor-intensive and time-consuming, and it is not always possible to record data in the wild. These issues drastically decrease the number of corpora and their content in terms of amount of data.

Table 3 – Characteristics of multimodal corpora collected for depression analysis (V – video, A – audio, T – text, P – physiological signals)

Corpus name	Modalities	Volume, h	Labels	Language
Pitt (16)	V, A, T	5.9	5 levels of depression	En
DAIC (17)	V, A, T	73.2	5 questionnaires	En

3. CLASSIFICATION OF METHODS FOR AFFECTIVE STATE AUTOMATIC RECOGNITION

Video, audio and textual modalities are the most relevant for the analysis of affects. Figure 2 shows the main methods for preprocessing, feature extraction and classification used for each modality.

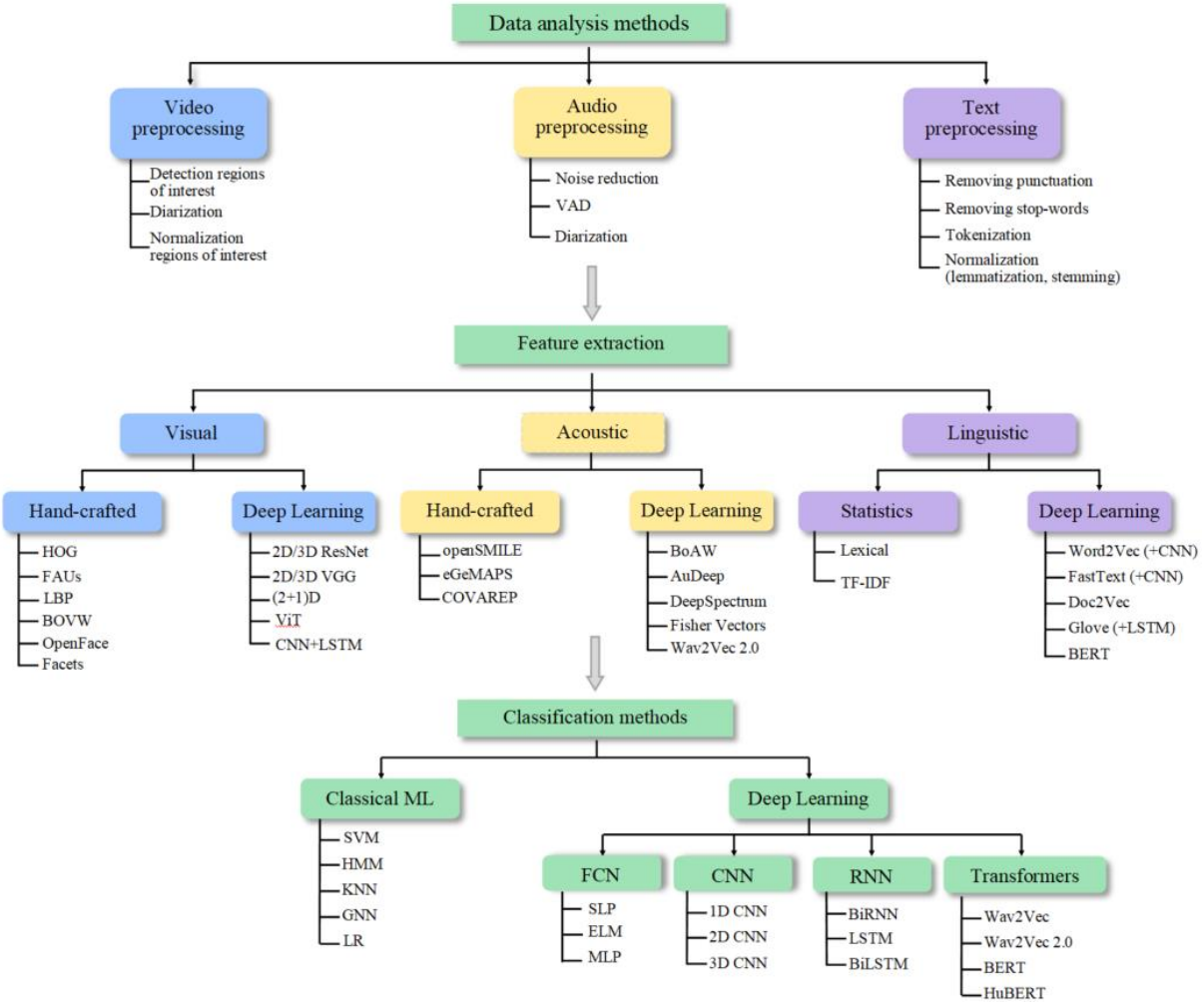


Figure 2 – Taxonomy of methods for preprocessing, feature extraction, and classification of video, audio and textual modalities

In Figure 2, information related to video, audio, and textual modalities is highlighted in blue, yellow, and purple respectively, while green highlights mark information related to all the modalities. In this paper, only methods for analyzing audio modality will be considered in detail.

3.1 Audio Modality. Methods for Acoustic Features Extraction

Acoustic features can be divided into two types: expert- and neural network-related. The classification is shown in Figure 2. Further a detailed description of the acoustic features used to analyze affective states is provided.

Expert / handcrafted feature sets are based on the knowledge about the acoustic properties of speech signals. These features are usually defined at two levels: segments of the audio signal (Low Level Descriptor - LLD) and the entire utterance. LLDs are extracted from short audio segments and provide instantaneous information about the audio signal. Utterance-level features are obtained by applying statistical functionals to the resulting descriptors.

Low-level descriptors can be roughly classified into the following clusters: energetic, prosodic, vocalized and spectral. Each group of indicators corresponds to certain features of the human's voice and finds its application in the task of affective state recognition. Prosodic features are the features of speech associated with the melodic, temporal and timbre characteristics of the voice, as well as the rhythm of utterances. Vocalized features reflect such characteristics as signal deviations of frequency (jitter) and amplitude (shimmer), as well as the ratio of pitch harmonics to noise. These features qualitatively affect the perception of the voice and are associated with prosodic characteristics. Spectral features characterize the speech signal in physical and mathematical sense, based on periodic (tonal) and non-periodic (noise) spectral components. They reflect the specific features of the spectrum of vocal impulses and the dynamics of the articulatory organs. Energy cues convey information concerning the level of energy carried by an audio signal and are relevant for recognition of the intensity of an expressed emotion. They can be represented as the sum of the spectrum, the null-crossing frequency, and so on.

Neural network features are automatically extracted by machine learning algorithms and often cannot be interpreted by users. The generated features can be defined at two different levels: the level of short audio segments and the level of utterances.

The openSMILE feature sets (18) are knowledge-based, and actually represent the standard used as a basis in various tasks of computer paralinguistics (19). Depending on the configuration, the set of features may include from 30 to 65 low-level descriptors, for example, the sum of acoustic spectra (loudness); Mel-frequency cepstral coefficients (MFCC) harmonic / noise ratio, jitter, shimmer, etc.

The Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) is a popular set of features manually tuned by experts for speech emotion recognition tasks (20). It consists of two functional descriptors, the arithmetic mean and the coefficient of variation of a set of 42 descriptors.

An alternative to the features above is the use of machine feature encoding algorithms, such as Bag-of-Audio-Words, (BoAW) (21) and Fisher Vectors (FV) (22). The BoAW approach performs clustering over 12 MFCCs, as well as the logarithmic signal energy. FV provides supra-segment coding of low-level descriptors, such as MFCC and RASTA-like coefficients of Perceptual Linear Prediction (PLP), by their deviation from the distribution, which can be modeled using Gaussian Mixture Model (GMM).

The emergence of neural networks capable of early feature analysis made it possible to use automatic feature extraction. Among examples are Deep Spectrum neural network feature sets extracted with the use of a deep neural network, and AuDeep. These feature sets are extracted using a deep recurrent neural network architecture (20), and TRILL (23). The features obtained using Wav2Vec 2.0 have also gained popularity recently (24).

The repository of speech processing algorithms COVAREP (Cooperative Voice Analysis Repository) (25), is used to store original implementations of published algorithms. For example, the creators of the CMU-MOSEI corpus (6) extracted 12 MFCC, the ratio of pitch harmonics to noise, segmentation functions for (un)voiced sounds, peak slope parameters and peak dispersion coefficients. All selected features are associated with emotions and speech tone.

3.2 Methods for Affective State Classification

Early systems for automatic recognition of affective states were based on such methods as K-Nearest Neighbor (KNN), Gaussian Mixture Model (GMM), Hidden Markov models (HMM), Support Vector Machine (SVM), and Logistic Regression (LR) (Figure 2).

With the emergence of large amounts of data, training such systems has become a time-consuming process. Therefore, they were replaced by Fully Connected Networks (FCN), in particular, a single-layer perceptron (SLP), an Extreme Learning Machine (ELM) and a multilayer perceptron (MLP). Then, deep neural networks (DNN) emerged, which have become more efficient every year, this applies in particular to Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). A variation of RNN, called the Long Short-Term Memory (LSTM) model, gained popularity for its ability to model long-term sequences. As a result, LSTM turned out to be more efficient than simple RNN models for emotion recognition (26). Other varieties of recurrent networks, such as bidirectional

RNNs and LSTMs, are able to predict both the previous and the future contexts (27). RNNs are also often combined with other types of neural networks such as CNNs. Attention technique is a concept that can be used in RNNs to improve the memory mechanism (28). It enables focusing on the most important features and discard less important ones. Self-attention is a type of attention, aimed at identifying patterns only between input data. This technique proved to be so efficient in the area of machine translation, that it allowed to abandon RNNs in favor of CNNs combined with self-attention in the Transformer architecture (29). This made it possible to speed up the algorithm, since now each segment can be processed in parallel, in contrast to sequential processing in RNN. This mechanism is also successfully used in applications for affective states recognition (30). Well-known transformer models are Wav2Vec 2.0 (31) and HuBert (32). In addition, there are End-2-End (E2E) approaches that allow analyzing the raw signal (wave form) using 1D CNN and LSTM (2). However, these approaches, like transformers, require a large amount of training data, therefore, the transfer learning is often used in combination with transformers (33). The basic idea is to train the neural network to perform one task when a lot of data are available, and then to tune the last layers in the pre-trained network for the next target task. This method is effective because the early layers of neural networks typically analyze representations of low-level features that are basic for understanding speech signals in general.

In current works in the field of affective computing, both classical, expert methods and neural network methods are used for feature extraction and classification. The research trend is that neural network-based approaches are gradually replacing expert-based ones, due to greater accuracy in affect recognition tasks and their ability to quickly process a large amount of data. The multimodal approach for automatic analysis of affective states makes it possible to increase the accuracy of affect recognition compared to unimodal ones.

4. COMPARISON OF METHODS FOR AFFECTIVE STATE RECOGNITION

In this section, automatic systems for emotion, sentiment, and aggression recognition are discussed. Discussion of approaches to depression recognition can be found in (34).

4.1 Emotions and Sentiment

For the following analysis, experimental studies conducted only on the English multimodal CMU-MOSEI corpus (6) data are considered. Among the benefits of this corpus are: annotation both for emotions and sentiment; natural recording conditions and spontaneous speech; a large number of informants (1000); predominantly one participant in the frame; annotation by full utterances. In different works, sentiment is classified into different number of classes: 2 (negative, positive), 3 (negative, neutral, positive), 5 (strongly negative, negative, neutral, positive, strongly positive) and 7 classes (from -3 to 3), according to the annotation principles proposed by the CMU-MOSEI authors. Table 4 presents a comparative analysis of automatic emotion and sentiment recognition systems for the CMU-MOSEI corpus.

In many papers, the COVAREP features, Facets and Glove, respectively, were used for the analysis of acoustic, visual and linguistic information. As for CMU-MOSEI, all the features are publicly available (https://github.com/A2Zadeh/CMU-MultimodalSDK).

Classification methods can be conditionally divided into groups depending on the neural network architectures used: graph neural networks - Graph Memory Fusion Network (Graph-MFN) (6) and Adversarial Representation Graph Fusion (ARGF) (42); recurrent neural networks – Multi-Modal Multi-Utterance - Bi-Modal Attention (MMMU-BA) (40), Multi-task Multi-modal Emotion and Sentiment (MTMM-ES) (39), Interaction Canonical Correlation Network (ICCN) (38), Hierarchical Feature Fusion Network (HFFN) (41) and Inter-modal Interactive Module for Multi-modal Sentiment and Emotion Recognition (IIM-MMSE) (35); transformers - Transformer-Based Joint-Encoding (TBJE) (36) and Multimodal Transformer (MulT) (37). At the same time, the best accuracy in terms of mWAcc for recognizing 6 categories of emotions was achieved using TBJE and MulT transformer models-based classification methods. However, not only the use of transformers affected the efficiency of these classification methods, but also the use of CNN for visual features, Mel-spectrograms and 40-D Log-Filter bank energy for acoustic features. In addition, TBJE proved to be the most efficient for the 2-class sentiment recognition task. Classification methods based on recurrent neural networks IIM-MMSE (35) and ICCN (38), respectively, have proven their efficiency for the recognition of 5 and 7 classes of sentiment in terms of WAcc.

Table 4 – Comparison of systems for automatic emotion and sentiment recognition (WAcc –weighted accuracy, WF – weighted F-score, m – WAcc/WF mean for 6 classes)

Work	Features	Classification method	Emotions		Sentiment		
			mWAcc, %	mWF, %	Number of classes	WAcc, %	WF, %
(6)	Facets, OpenFace, CNN COVAREP Glove	Graph-MFN	62.3	76.3	2 5 7	76.9 45.1 45.0	77.0
(35)	-	IIM-MMSE	63.0	**79.0**	2 **5** 7	80.4 **49.2** 50.1	78.2
(36)	CNN Mel-spectrograms Glove	TBJE	**80.7**	76.7	**2** 7	**81.5** 44.4	
(37)	CNN 40-D Log-Filter bank energy Glove	MulT	67.4	78.6			
(38)	Facets COVAREP BERT	ICCN			**7**	**51.6**	
(39)	Facets COVAREP Glove	MTMM-ES	62.8	78.6	2	80.5	78.8
(40)		MMMU-BA			2	79.8	
(41)		HFFN			3	60.4	59.1
(42)		ARGF			**3**	**60.9**	**59.5**

One can conclude, that the best emotion and sentiment recognition performance can be achieved with the use of RNN and transformer-based machine classification.

4.2 Aggression

Nowadays, the area of automatic aggression recognition is very little developed. No experimental studies have been conducted on the base of aforementioned corpora, thus a comparative analysis of methods for aggression recognition is difficult, if not impossible. However, one can provide a classification of methods for aggression recognition, which is shown in Figure 3.

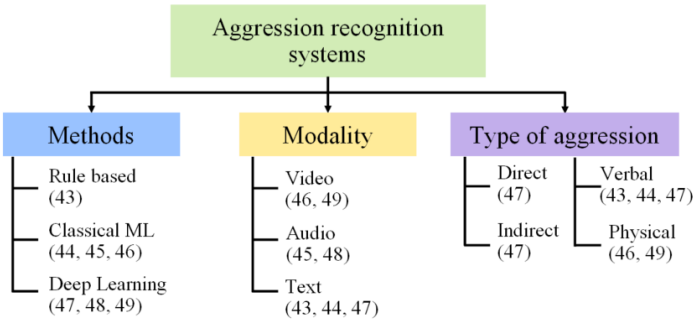


Figure 3 – Classification of approaches to aggression recognition

Based on the analysis of the abovementioned papers, one can distinguish the main clusters of methods and models used for aggression recognition. The following features can be considered as classification criteria: (1) modality used; (2) the type of aggression at issue; (3) recognition methods.

All the considered corpora containing multimodal aggression data are restricted of access. These corpora, however, contain small amounts of data and not representative in terms of the number of informants. In this regard, there is an acute need for collection and annotation of a new multimodal

corpus containing aggressive behavior samples. In the future, the authors of this paper are going to collect a corpus of aggressive behavior manifested through different modalities: motor and mimic activity, non-verbal behavior in the flow of audio signal, verbal speech behavior in textual modality. Among the other features of this corpus should be a large number of informants, participation of Russian native speakers, natural behavior, ensured by recording only spontaneous, live broadcasts on the Internet.

One can conclude that the area of emotion and sentiment analysis has gained much attention, and there exists a large number of emotional and sentiment corpora, both unimodal and multimodal, and automatic systems of sentiment and emotion recognition achieve an accuracy of about 80%. However, the problem of depression and aggression recognition is still far from solution, this applies in particular to quality and amount of available data and processing techniques. Today there exist no multi-task systems that deal with all the four basic affective states (emotions, sentiment, aggression and depression) simultaneously.

5. MULTI-TASK AND MULTI-MODAL APPROACH FOR AFFECTIVE STATE RECOGNITION

Multi-task systems that recognize the four basic affective states: emotions, sentiment, aggression and depression have not been found to date. Therefore, there is a need to develop a new multi-task and multi-modal approach for the recognition of affective states. Figures 4 and 5 show a pipeline of the proposed approach, backed with the analysis of visual, acoustic and linguistic information.

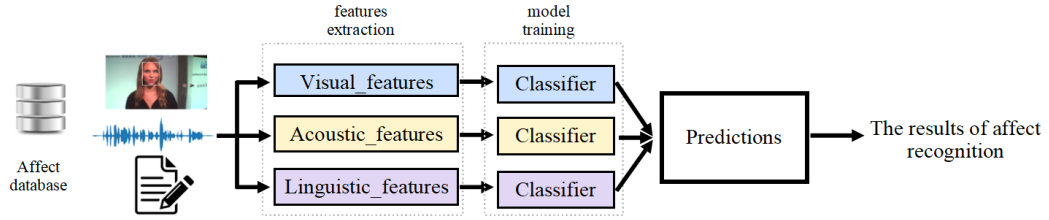


Figure 4 – Multi-modal approach for affective state recognition task

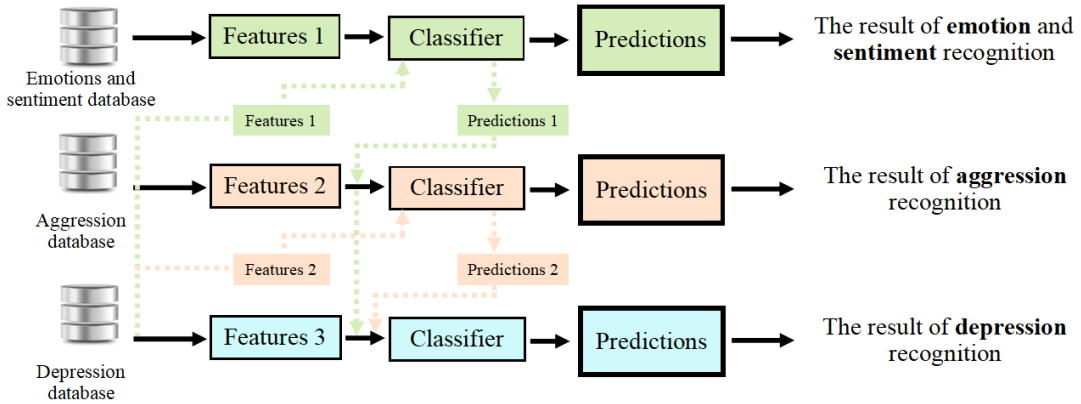


Figure 5 – Multi-task approach for affective state recognition task

Figure 4 sketches a multimodal approach for affective states recognition. Visual, acoustic and linguistic features are extracted from the database, then they are fed into the input of the classifier, while the output is probabilistic predictions of affective state recognition. In order to analyze video data by affective states, it is planned to use feature extractors based on 2D CNNs with LSTM, similar to (2), and 3D CNNs. For the analysis of acoustic modality, Wav2vec methods, neural networks with a transformer architecture, such as Hubert, will be applied. Clove and Bert methods are to be used for linguistic information extraction and text data classification.

It is planned to use a hierarchical approach for the main affective states recognition (Figure 5). The first step is to train models for emotions and sentiment recognition, after that visual, acoustic and linguistic features are extracted from the data contained in the aggression corpus (Features 1), on the basis of which probabilistic predictions for recognizing emotions and sentiment are further obtained.

These predictions are then fed along with feature vectors (Features 2) to the input of classifiers for learning aggression recognition, so that finally the result of aggression recognition is obtained. This approach lets recognize aggression based on the knowledge gained from trained models. Moreover, this approach allows solving the problem of multi-task analysis of affective states and achieving high performance (50). It is planned to use the following hierarchy of affective states recognition: first, emotions and sentiment are recognized, then, based on these data, aggression is recognized, and finally depression. It is worth mentioning that the recognition of depression can be carried out both on the basis of the knowledge gained from the recognition of aggression, and emotions with sentiment, together or separately.

It is planned to use such corpora as CMU-MOSEI, RAMAS, and DAIC to recognize emotions, sentiment, and depression. For aggression recognition, it is planned to collect a new data corpus. The advantages of CMU-MOSEI corpus are that it annotated for both emotions and sentiment, and the data were collected in natural conditions. This corpus will be suitable for elaboration of a multi-task system for recognizing affective states. The RAMAS multimodal emotional corpus is currently the only Russian language corpus, which will allow the system to work with different languages. The DAIC corpus is the largest in terms of data amount, compared to other multimodal (video, audio, and text) corpora for depression analysis.

6. CONCLUSIONS

This paper addresses the current multimodal corpora collected for the analysis of such affective states as emotions, sentiment, aggression and depression. There is a large amount of data for the analysis of emotions and sentiment, however, depression and aggression corpora are significantly inferior in terms of data amount and quality. To date, there exist yet no multimodal corpora that would provide data for the recognition of a wide range of affective states. It is only MELD and CMU-MOSEI corpora that are exception to some extent, being annotated both for emotions and sentiment. These issues are obstacles for the elaboration of automatic affect recognition systems. Besides that, most of corpora contain data in English, while French, German, Chinese and Russian data are covered to a much lesser extent. Only SEWA emotional corpus provides data in six different languages. This is the reason of the current situation in affective computing, when the overwhelming majority of researches is based on data in the English language. Such corpora as CMU-MOSEI, RAMAS, and DAIC are most relevant for the elaboration of the systems for automatic affective state analysis.

A classification of state-of-the-art methods for multimodal (video, audio, text) data analysis is provided within this paper as well. A detailed description of feature extraction and classification methods is discussed only for the audio modality. This study has shown that much of papers in the field of affective computing are based on researches carried out within the framework of COMPARE paralinguistics challenge. The authors of these works tend to use OpenSMILE acoustic features, as well as neural network-based features (AuDeep, BOAW, DeepSpectrum) together with CNNs or RNNs of various topologies, for elaboration of their own approaches to affect recognition. More recently, however, other approaches have been gaining popularity, namely, the approaches implying the use of transformers with attention, combined with neural network features and transfer learning.

This work also provides an overview of current research trends in the area of emotion, sentiment and aggression recognition. The research area of emotion and sentiment analysis shows most advanced development and provides a large amount of works if compared to aggression and depression research area. The reason is that the number and quality of emotional corpora greatly exceeds these of other affective states corpora. Based on the present survey, one can state, that RNNs and transformers help to achieve the best results when applied to emotion and sentiment recognition tasks (up to 80% accuracy).

Besides, a new multi-task and multi-modal approach is proposed to recognize different affective states. Audio, video and textual modalities have proven to be the most representative for affective state analysis. This leads to the conclusion, that these modalities must be used within the multimodal approach, implying a fusion of modalities on the level of predictions. The use of the hierarchical approach was also proposed in this paper for affective state analysis. At first emotions and sentiment are recognized, and the results obtained are further used for aggression and depression recognition, consequently. The proposed approach is a theoretical model. In the future, this model will be applied in practice, tested on selected databases, after which conclusions will be drawn about the effectiveness of the proposed approach.

ACKNOWLEDGEMENTS

This study was supported by the Russian Science Foundation, project No. 22-11-00321.

REFERENCES

1. Picard R. Affective Computing for HCI. HCI (1); 1999. p. 829-833.
2. Dresvyanskiy D, Ryumina E, Kaya H, Markitantov M, Karpov A, Minker W. End-to-end Modelling and Transfer Learning for Audiovisual Emotion Recognition in the Wild. Multimodal Technologies and Interaction: MDPI. 2022;6(2):1-23.
3. Busso C, Bulut M, Lee C, Kazemzadeh A, Mower E, Kim S, et al. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation. 2008;42(4):335-359.
4. Ringeval F, Sonderegger A, Sauer J, Lalanne D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. Proceedings of 10th IEEE international conference and workshops on automatic face and gesture recognition (FG): IEEE; 2013. p. 1-8.
5. Perepelkina O, Kazimirova E, Konstantinova M. RAMAS: Russian multimodal corpus of dyadic interaction for affective computing. Proceedings of International Conference on Speech and Computer: Springer, Cham; 2018. p. 501-510.
6. Zadeh A, Liang P, Poria S, Cambria E, Morency LP. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; 2018. p. 2236-2246.
7. Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019. p. 527-536.
8. Kossaifi J, Walecki R, Panagakis Y, Shen J, Schmitt M, Ringeval F, et al. SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. IEEE Transactions on Pattern Analysis & Machine Intelligence. 2021;43(03):1022-1040.
9. Pérez-Rosas V, Mihalcea R, Morency LP. Utterance-level multimodal sentiment analysis. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics; 2013. p. 973-982.
10. Zadeh A, Zellers R, Pincus E, Morency LP. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. IEEE Intelligent Systems. 2016;31(6):82-88. doi: 10.1109/MIS.2016.94.
11. Morency LP, Mihalcea R, Doshi P. Towards multimodal sentiment analysis: Harvesting opinions from the web. Proceedings of the 13th international conference on multimodal interfaces; 2011. p. 169-176.
12. Yu W, Xu H, Meng F, Zhu Y, Ma Y, Wu J, et al. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020. p. 3718-3727.
13. Lefter I, Rothkrantz L, Burghouts G, Yang Zh, Wiggers P. Addressing multimodality in overt aggression detection. International Conference on Text, Speech and Dialogue: Springer; Berlin, Heidelberg 2011. p. 25-32.
14. Lefter I, Rothkrantz L. Multimodal cross-context recognition of negative interactions. 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW): IEEE; 2017. p. 56-61.
15. Lefter I, Jonker C, Tuente S, Veling W, Bogaerts S. NAA: A multimodal database of negative affect and aggression. 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII): IEEE; 2017. p. 21-27.
16. Yang Y, Fairbairn C, Cohn J. Detecting depression severity from vocal prosody. IEEE Transactions on Affective computing. 2013;4(2):142–150.
17. Gratch J, Artstein R, Lucas G, Stratou G, Scherer S, Nazarian A, et al. The Distress Analysis Interview Corpus of Human and Computer Interviews. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); Reykjavik, Iceland 2014. p. 3123-3128.
18. Eyben F, Weninger F, Gross F, Schuller B. Recent developments in opensmile, the munich open-source multimedia feature extractor. Proceedings of ACM International Conference on Multimedia; 2013. p. 835–838.
19. Schuller B, Batliner A, Bergler C, Mascolo C, Han J, Lefter I, et al. The INTERSPEECH 2021

Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. Proceedings of INTERSPEECH; 2021. p. 431–435.
20. Eyben F, Scherer K, Schuller B, Sundberg J, Andr´e E, Busso C, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing. 2015;7(2):190–202.
21. Schmitt M, Ringeval F, Schuller B. At the border of acoustics and linguistics: Bag-of-Audio-Words for the recognition of emotions in speech. Proceedings of INTERSPEECH; 2016. p. 495–499.
22. Kaya H, Karpov A, Salah A. Fisher vectors with cascaded normalization for paralinguistic analysis. Proceedings of INTERSPEECH; 2015. p. 909–913.
23. Shor J, Jansen A, Maor R, Lang O, Tuval O, de Chaumont Quitry F, et al. Towards Learning a Universal Non-Semantic Representation of Speech. Proceedings of INTERSPEECH; 2020. p. 140–144.
24. Wagner J., Triantafyllopoulos A., Wierstorf H., Schmitt M., Eyben F., Schuller B. Dawn of the transformer era in speech emotion recognition: closing the valence gap. arXiv preprint arXiv:2203.07378; 2022. p. 1-25.
25. Degottex G, Kane J, Drugman T, Raitio T, Scherer S. COVAREP – A collaborative voice analysis repository for speech technologies. Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2014. p. 960-964.
26. Wang J, Xue M, Culhane R, Diao E, Ding J, Tarokh V. Speech emotion recognition with dual-sequence lstm architecture. Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020. p. 6474–6478.
27. Chen Q, Huang G. A novel dual attention-based blstm with hybrid features in speech emotion recognition. Engineering Applications of Artificial Intelligence. 2021;102:104277.
28. Xie Y, Liang R, Liang Z, Huang C, Zou C, Schuller B. Speech emotion classification using attention-based lstm. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2019;27(11):1675–1685.
29. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017. p. 1-11.
30. Ho NH, Yang HJ, Kim SH, Lee G. Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. IEEE Access. 2020;8:61672-61686.
31. Baevski A, Zhou Y, Mohamed A, Auli M. Wav2vec 2.0: A framework for self-supervised learning of speech representations. Proceedings of Advances in Neural Information Processing Systems (NeurIPS); 2020. p. 12449-12460.
32. Hsu WN, Bolte B, Tsai YH, Lakhotia K, Salakhutdinov R, Mohamed A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021;29:3451–3460.
33. Siriwardhana S, Reis A, Weerasekera R, Nanayakkara S. Jointly fine-tuning "bert-like" self supervised models to improve multimodal speech emotion recognition. arXiv preprint arXiv:2008.06682; 2020. p. 1-5.
34. Velichko A, Karpov A. Analytical Review of Automatic Systems for Depression Detection by Speech. Informatics and Automation. 2021;20:497-529. in Rus.
35. Chauhan DS, Akhtar MS, Ekbal A, Bhattacharyya P. Context-aware interactive attention for multi-modal sentiment and emotion analysis. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019. p. 5647-5657.
36. Delbrouck JB, Tits N, Brousmiche M, Dupont S. A transformer-based joint-encoding for emotion recognition and sentiment analysis. In Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML); 2020. p. 1-7.
37. Khare A, Parthasarathy S, Sundaram S. Self-Supervised learning with cross-modal transformers for emotion recognition. 2021 IEEE Spoken Language Technology Workshop (SLT): IEEE; 2021. p. 381-388.
38. Sun Z, Sarma P, Sethares W, Liang Y. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. Proceedings of the AAAI Conference on Artificial Intelligence. 2020;34(05):8992-8999.
39. Akhtar MS, Chauhan DS, Ghosal D, Poria S, Ekbal A, Bhattacharyya P. Multi-task learning for

multi-modal emotion recognition and sentiment analysis. arXiv preprint arXiv:1905.05812; 2019. p. 1-10.
40. Ghosal D, Akhtar MS, Chauhan D, Poria S, Ekbal A, Bhattacharyya P. Contextual inter-modal attention for multi-modal sentiment analysis. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018. p. 3454-3466.
41. Mai S, Hu H, Xing S. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019. p. 481-492.
42. Mai S, Hu H, Xing S. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. Proceedings of the AAAI Conference on Artificial Intelligence. 2020;34(1):164-172.
43. Zaib S, Asif M, Arooj M. Development of Aggression Detection Technique in Social Media. International Journal of Information Technology and Computer Science. 2019;5(8):40-46.
44. Al-Garadi M, Varathan K, Ravana S. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Computers in Human Behavior. 2016; 63:433-443.
45. Potharaju Y, Kamsali M, Kesavari C. Classification of Ontological Violence Content Detection through Audio Features and Supervised Learning. International Journal of Intelligent Engineering and Systems. 2019;12(3):20-230.
46. Hassner T, Itcher Y, Kliper-Gross O. Violent flows: Real-time detection of violent crowd behavior. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops: IEEE; 2012. p. 1-6.
47. Tommasel A, Rodriguez J, Godoy D. Textual Aggression Detection through Deep Learning. TRAC@ COLING 2018; 2018. p. 177-187.
48. Santos F, Durães D, Marcondes F, Hammerschmidt N, Lange S, Machado J, et al. In-car violence detection based on the audio signal. International Conference on Intelligent Data Engineering and Automated Learning: Springer, Cham; 2021. p. 437-445.
49. Liang Q, Li Y, Chen B, Yang K. Violence behavior recognition of two-cascade temporal shift module with attention mechanism. Journal of Electronic Imaging. 2021;30(4):043009.
50. Velichko A, Markitantov M, Kaya H, Karpov A. Complex Paralinguistic Analysis of Speech: Predicting Gender, Emotions and Deception in a Hierarchical Framework. Proceedings of INTERSPEECH-2022; Incheon, Korea 2022.

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022, in Gyeongju, Korea

Extended abstract

ABS-0705

Using Majority Vote Decision Fusion From Multi-Performance CNN Classifier To Enhance Vocal Cancer Detection Accuracy From Patient's Voice

Cheolwoo JO[1]; Soo-Geun WANG[2]; Ickhwan KWON[3]

[1] Changwon National University, Korea

[2] Pusan National University, Korea

[3] Pusan National University, Korea

ABSTRACT

This paper is a study on how to improve the classification rate of laryngeal disordered voice by CNN and ensemble learning methods. In general, laryngeal disordered voice data are small in size, so even if identifiers are configured by statistical methods, degradation in identification rates can occur due to overfitting depending on training methods. In this work, we attempt and confirm the results by combining results derived from CNN models trained to have varying degree of accuracy in multiple voting ways to achieve improved classification efficiency compared to the original trained model. To train and validate the algorithm, we used the PUSAN National University Hospital (PNUH) dataset. The dataset contains data on normal, seven benign tumors and malignant tumors. In the experiment, an attempt was made to distinguish between non-cancer group (normal and benign tumors) and malignant tumors. As a result of the experiment, as an evaluation index of the fused results by the Hard Voting and Soft Voting methods, values of accuracy of 94% and precision of 97%, specificity of 98% and sensitivity of 87% were obtained. This result showed overall improved classification results compared to the case of a single trained model.

Keywords: Classifier, Voice, CNN, Fusion, Ensemble

1. INTRODUCTION

In this paper, we report a case of improving the performance of the laryngeal disability speech identifier by ensemble methods. The purpose of the laryngeal disorder voice identifier is to analyze the voice when a voice mutation occurs due to a disease in the larynx, identify the presence or absence of a disease, and diagnose it in early stage. Various studies have been conducted on the distinction and identification of voice with laryngeal disorder by means of various methods of machine learning. In recent years, as various tools of artificial neural networks are disseminated and common data acquisition is facilitated, identification cases with convolutional neural networks (CNNs) have been frequently reported [1][2][3][4]. These studies are mainly conducted as a process of constructing appropriate CNN networks and verifying the accuracy of the classification within the given data to classify the obtainable disordered voice data. In this study, as a way to increase the identification rate of the identifier configuration using the existing CNN, it is intended to derive a way to obtain improved results by integrating CNN models with various accuracy obtained during the CNN training. In addition, as a focus of this study, it was performed by focusing on the distinction between non-cancer(normal and benign diseases) and malignant tumors rather than the classification of various diseases. This is based on the judgment that early screening of malignant diseases is much more necessary than distinguishing between benign diseases when considering practical application cases in the medical field. Therefore, the composition of the identifier in this study was performed with an emphasis on distinguishing between non-cancer group and malignant tumors.

[1] cwjo@changwon.ac.kr
[2] entwangsg@daum.net
[3] kanikwon@pusan.ac.kr

2. DATA AND METHODS

The voice data used in the experiment is a PNUH dataset collected by the ENT department of Pusan National University Hospital, and includes /a/ voice data pronounced 4 seconds. It includes 221 cases of normal and benign tumors and 146 cases of malignant tumors, and includes cases of normal and six types of benign tumors and malignant tumors. Table 1 shows the configuration of the voice dataset used.

Table 1. Structure of Voice Dataset

Label: Group		Number
Cancer		146
Non-Cancer	normal	49
	cyst	16
	palsy	45
	edema	46
	polyp	42
	nodule	20
	others	3
	sub-total	221
Total		367

Voice data were collected from 1998 to 2020 for those who visited the ENT department at Pusan National University Hospital and were diagnosed with disease names by a specialist. Cancer data were collected from men over the age of 40, and normal and non-cancer data were collected from men over the age of 40. The reason for using male data is that laryngeal cancer is frequent in middle-aged men and rare in women. The equipment for voice recording and analysis was the Kay Computer Speech Lab (CSL) (Model 4300B and 4150B; KayPENTAX, Montvale, NJ, USA). The voice signals were sampled at 16 bits and 44.1 KHz, and were collected at a distance of 10-15 cm from the speaker using the Shure-Prolog microphone. When vocalizing, the speaker was asked to vocalize the /a/ voice for more than 4 seconds in a comfortable state. The collected voices were validated after being reviewed by an otolaryngologist. For the analysis, 40th MFCC (Mel-frequency Cepstral Coefficient) analysis was performed at 20ms intervals to obtain the MFCC coefficient. For analysis, Tensorflow [5] and Python library Librosa [6] packages for voice analysis were used.

3. EXPERIMENT AND RESULTS

In this work, CNN was used as the primary identifier and the architecture of the CNN model used is summarized in Figure 1. In the literature [3], various methods such as SVM, HMM, GMM, and CNN have been attempted as identifiers for various cases, including laryngeal diseases, but it is difficult to determine the superiority of identifier performance with absolute identification rates. Recently, a similar study that identified laryngeal disease as negative reported the result of Kim [4] trying various methods using its own dataset and constructing a classifier with 85% sensitivity using 1D-CNN. In this study, identification by the CNN model, which has recently been in the spotlight, was attempted. The 2D-CNN used in the experiment consists of a four-layer structure. The model includes three convolutional layers, one maximum pooling layer, two fully connected layers, and one output layer. Input layer accepts 40x40 MFCC coefficients calculated from voice signal. The first

convolution layer has 16 filters of size 3×3, and then a batch normalization layer. The second and third convolution layers have 32 and 64 filters of size 3×3, respectively. The last layer provides two layers that are fully connected by 128 neurons and 64 neurons through planarization. The final output layer is a single neuron for binary classification with sigmoid activation function.

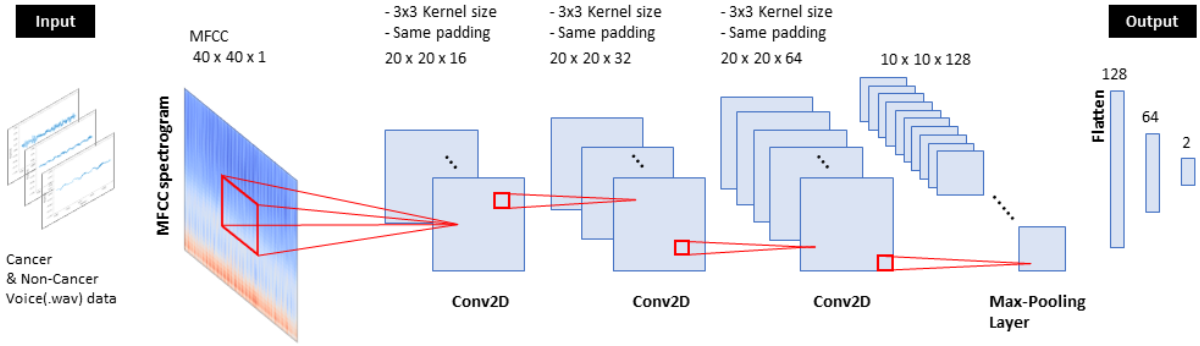


Figure 1. Structure of Baseline CNN Classifier

In CNN's training, 100 Epochs were set up by experiments, and the results of sufficient convergence for the given data were obtained. In this paper, we use CNN classifier models with varying accuracy instead of decision trees as the base classifier and use methods to obtain an ensemble classifier from the results obtained from each model.

The ensemble of identifiers in this study was constructed through the process shown in Figure 2. Hard and soft decision algorithm was applied [7]. First, five classifier models with varying classification rate are constructed using the MFCC analyzed dataset. We classify the input datasets through five classifier models (CNN1 through CNN5) and use the results as inputs for the ensemble model.

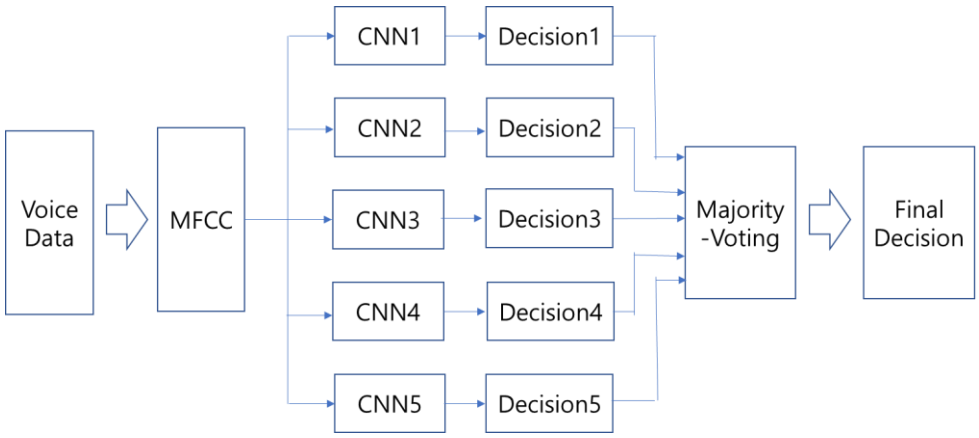


Figure2. Structure of Majority Voting

Table 2. Parameter Value Changes for Various Classifier

	Accuracy	Precision	Specificity	Sensitivity
CNN5 (Best)	0.9346	0.9621	0.9774	0.8699
Max5	0.8256	0.8015	0.8778	0.7466
Max4	0.8992	0.9360	0.9638	0.8014
Max3	**0.9373**	**0.9695**	**0.9819**	**0.8699**
Soft Vote	**0.9373**	**0.9695**	**0.9819**	**0.8699**

Summarizing the experimental results, it was confirmed that the ensemble results obtained by

Max3 of Hard Voting and Soft Voting showed 0.0027 improvements in accuracy compared to the accuracy of Model CNN5, which showed the best performance among the five pre-trained models, as shown in Table 1. It was confirmed that there was an improvement of up to 0.0354 from CNN1 with the lowest accuracy. In addition, improvements of 0.0074 and 0.0045 were observed in precision and specificity, respectively, and the same values were measured for sensitivity. Therefore, it was confirmed that the performance of the ensemble model was improved when constructing an identifier for laryngeal malignancy using CNN models trained with various accuracy, and in this experiment, the ensemble method applied improved from minimum value of 0.0074 to maximum value of 0.0354 in accuracy.

4. CONCLUSIONS

In this paper, we apply the ensemble method by voting to experiments that classify voice into non-cancer(normal and benign) and malignant diseases and examine the results by ensemble of different accuracy CNN models. It has been confirmed that the ensemble method is effective as a way to improve the performance of the identifier while resolving overfitting by a small amount of data when the number of data is inevitably small, such as voice with laryngeal disease. The ensemble model showed improved values not only in accuracy but also in precision, sensitivity, and specificity. Although the improved numerical values were small, it was an experiment that confirmed that various parameters could be improved by the ensemble method.

Subsequent research in this study should be conducted in the following direction. In order to improve with identifiers that are meaningful for actual diagnosis in the future, research should focus on how identification rates can be improved for external data rather than the data used for training. Furthermore, research is needed on identification model adaptation methods that can effectively improve existing models by utilizing additionally collected data. In addition, standardization of procedures for the process of collecting data from patients by clinical professionals and purification criteria is paramount. To this end, it is also necessary to solve various non-research problems, such as cooperation between medical institutions interested in similar research.

ACKNOWLEDGEMENTS

This paper was supported by the Changwon National University Research Fund in 2021 and 2022.

REFERENCES

1. Wu, H., Soraghan, J., Lowit, A., Di Catrina, G. Convolutional Neural Networks for Pathological Voice Detection, *Annu Int. Conf. IEEE Eng. Med. Biol.* Soc, 20018; p. 1-4.
2. Fang., S.H., Tsao, Y., Hsiao, M.J., Chen, J.Y., Lai, Y.H., Lin, F.C., Wang, C.T. Detection of Pathological Voice Using Cepstrum Vectors: A Deep Learning Approach, *Journal of Voice*, 2018; 33(5), 634-641.
3. Hegde, S., Shetty, S., Rai, S., Dodderi, T. A. Survey of Machine Learning Approaches for Automatic Detection of Voice Disorders, *Journal of Voice*, 2019; 33(6): 947.e11-e13.
4. Kim, H.B., Jeon, J., Han, Y.J., Joo, Y.H., Lee, J.H., Lee, S., Im, S. Convolutional Neural Network Classifies Pathological Voice Change in Laryngeal Cancer with High Accuracy, *Journal of Clinical Medicine, 2020;* 9, 3415.
5. Tensorflow. Available online: http://www.tensorflow.org/ (accessed on 1 October 2021), 2021.
6. Librosa . Available online: http://librosa.org/ (accessed on 1 October 2021), 2021.
7. Ruta, D., Gabrys, B. An Overview of Classifier Fusion Methods, *Computing and Informations Systems,* 2000; 7, 1-10.

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0033

Multi-channel end-to-end neural network for speech enhancement, source localization, and voice activity detection

Yuan Chen[1]; Yicheng Hsu[1]; Mingsian R. Bai[1,2]

[1] Department of Power Mechanical Engineering, National Tsing Hua University, Taiwan

[2] Electrical Engineering, National Tsing Hua University, Taiwan

ABSTRACT

Speech enhancement and source localization has been active research for several decades with a wide range of real-world applications. Recently, the Deep Complex Convolution Recurrent network (DCCRN) has yielded impressive enhancement performance for single-channel systems. In this study, a neural beamformer consisting of a beamformer and a novel multi-channel DCCRN is proposed for speech enhancement and source localization. Complex-valued filters estimated by the multi-channel DCCRN serve as the weights of beamformer. In addition, a one-stage learning-based procedure is employed for speech enhancement and source localization. The proposed network composed of the multi-channel DCCRN and the auxiliary network models the sound field, while minimizing the distortionless response loss function. Simulation results show that the proposed neural beamformer is effective in enhancing speech signals, with speech quality well preserved. The proposed neural beamformer also provides source localization and voice activity detection (VAD) functions.

Keywords: multi-channel speech enhancement, source localization, deep learning

1. INTRODUCTION

The goal of speech enhancement is to extract the target speech from the noisy signal. Since interfering noise and reverberation are pervasive in real-world, speech enhancement is essential in many applications to generate the enhanced speech waveform. Recently, deep neural network (DNN) has achieved impressive results in speech enhancement problems. Monaural processing with DNN is well-known approach to speech enhancement, which exploits the information in time domain [1-3] or time-frequency domain [4-6]. Time-frequency masking-based approaches in [4,5] employing supervised learning are proved to be effective in speech enhancement problems. Based on the success in time-frequency masking approach, DCCRN that exploits complex operations performs competitively over other previous networks [6].

However, in far-field applications, such as hands-free teleconferencing and smart speaker, the enhancement performance often degrades due to interference, reverberation, noise, etc. Motivated by fabulous results in DNN-based monaural speech enhancement, several multi-channel DNN-based speech enhancement approaches have been studied. Multi-channel methods benefit from spatial features given by inter-channel information. Employing multi-channel signals, a typical strategy is to combine DNN with conventional beamforming methods. In [7], DNN estimates time-frequency (T-F) masks to specify the dominant T-F bins for the signal of interest and noise. T-F masks are used to calculate the spatial covariance matrix for beamformer's weights such as minimum variance distortionless response (MVDR) [8] and generalized eigenvalue (GEV) [9]. [10] proposes an all deep learning MVDR (ADL-MVDR) beamformer, where the inverse operation in the aforementioned DNN method is unnecessary. Without reference to conventional beamformers, filter-based approachs applying filter-and-sum beamformer are proposed either in time domain [11] and time-frequency domain [12-14]. Previous studies have shown the importance of using time-frequency domain beamforming techniques in sensor array systems [15]. A state-of-the-art multiple-in-multiple-out (MIMO) U-net neural beamformer in time-frequency domain is proposed in [14] to produce a complex beamformer. The U-net neural beamformer utilizes spatial features implicitly by analyzing spectral features. This system ranked first in the ConferencingSpeech 2021 Challenge.

[1] cya10230223@gmail.com; shane.ychsu@gmail.com; msbai@pme.nthu.edu.tw

In this study, inspired by the concept in the U-net beamformer [14] and DCCRN [6], the MIMO-DCCRN neural beamformer is proposed. In addition, a one-stage learning based procedure based on MIMO-DCCRN is formulated for both speech enhancement and localization. The neural beamformer incorporates the traditional beamforming structure with a DNN model by adding a filter-and-sum operation at the end of the beamformer, similar to the U-net structure in [14]. The complex encoder in MIMO-DCCRN extracts the feature from both spectral and spatial perspectives. The framework is then equipped with complex long short-term memory (LSTM) in the bottleneck to model temporal context. The end of the network is complex decoder to estimate complex spatial filters. Comparing to the existing neural beamformers, the proposed neural beamformer deliberates the complex operation additionally. Further, depending on complex filters estimation, two localization solutions are derived from distortionless constraint. The first solution is to maximize the magnitude of steered beamformer response with free-field steering vectors [8] across the selected zones. Second, combine MIMO-DCCRN with the auxiliary network which models the sound field in the proposed network. Accordingly, the loss function is formulated to correlate with the enhancement quality and localization accuracy given distortionless constraint.

Note that in contrast to conventional beamformers such as MVDR [8], that the second-order spatial statistics and accurate direction of arrival (DOA) require to be explicitly calculated, our approache s directly estimate beamformer's weights and further locate the position of the target speaker. VAD can also be conducted from localization results. Consequently, the proposed network is able to handle enhancement, localization and VAD simultaneously. The simulation results show that this novel neural beamformer outperforms the existing DCCRN and MIMO U-net in enhancement. It is even shown that multi-tasking training, including speech enhancement and localization, improves enhancement in perceptual quality and intelligibility.

2. PROBLEM FORMULATION

Given a M-microphone time-frequency domain signal $\mathbf{y}(t,f)=\left[Y_0(t,f) \ \ldots \ Y_{M-1}(t,f)\right]^T \in \mathbb{C}^{M\times 1}$ recorded in a reverberant and noisy environment, where t = 1,…,T denotes the time frame, f =1,…,F denotes the frequency bin, the signal model in the short-time Fourier transform (STFT) domain is formulated as:

$$\mathbf{y}(t,f)=\mathbf{s}(t,f)+\mathbf{h}(t,f)+\mathbf{n}(t,f), \tag{1}$$

where $\mathbf{s}(t,f)=\left[S_1(t,f) \ \ldots \ S_M(t,f)\right]^T \in \mathbb{C}^{M\times 1}$ denotes the direct sound and early reflection components of a single target speech received by microphones, and $\mathbf{h}(t,f)$, $\mathbf{n}(t,f) \in \mathbb{C}^{M\times 1}$ denotes late reverberation components and reverberant noises, respectively. (1) can be expressed as

$$\mathbf{y}(t,f)=\mathbf{s}(t,f)+\mathbf{v}(t,f), \tag{2}$$

where $\mathbf{v}(t,f)=\mathbf{h}(t,f)+\mathbf{n}(t,f)$ is the undesired signal. The proposed approach estimates the signal of interest $\hat{S}(t,f)$ by filter-and-sum beamforming:

$$\hat{S}(t,f)=\sum_{m=0}^{M-1}\left\{w_m^*(t,f)\cdot Y_m(t,f)\right\}=\mathbf{w}^H(t,f)\mathbf{y}(t,f). \tag{3}$$

Thus, the neural beamformer system aims to estimate complex filter weights $\mathbf{w}(t,f)=\left[w_0(t,f) \ \ldots \ w_{M-1}(t,f)\right] \in \mathbb{C}^{M\times 1}$ for M microphones.

3. THE PROPOSED SYSTEM

Our proposed framework is illustrated in Fig. 1(a). The system is composed of an enhancement and a localization module. The signal from each channel is transformed into time-frequency domain representation as the input of our system. These features first passed through MIMO-DCCRN (Sec 3.1), which is the enhancement module, to generate filter coefficients. The estimated filter weights are then fed to the localization module (Sec 3.2) to analyze information from each channels and output localization results.

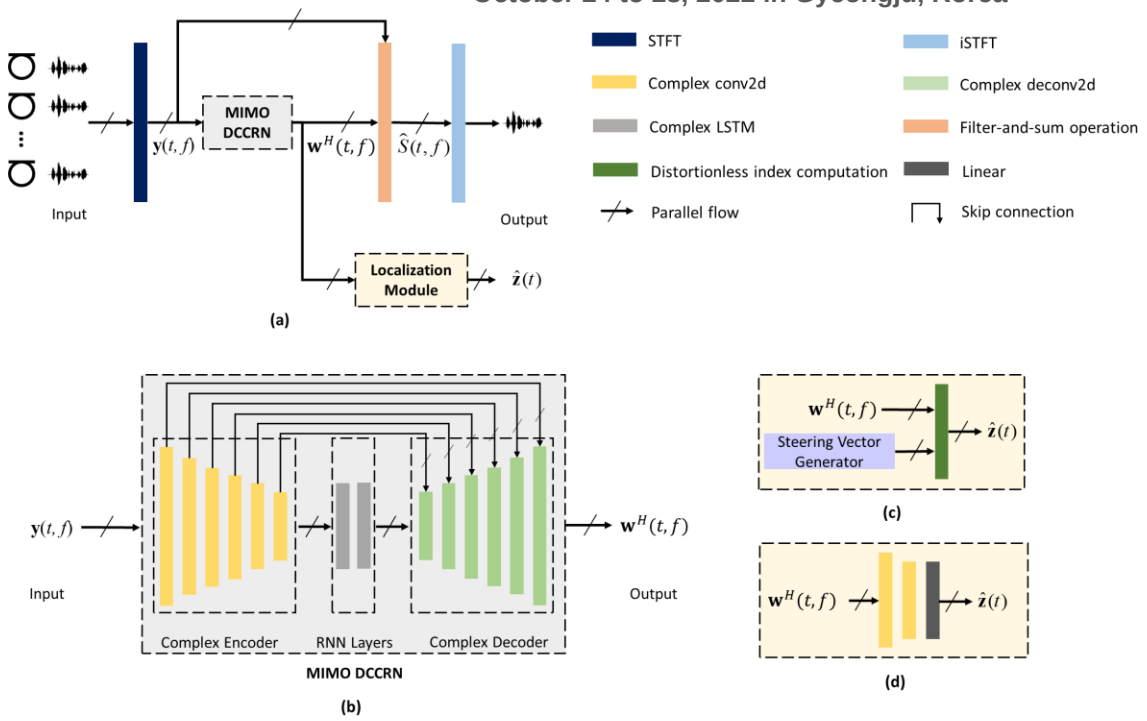


Figure 1 – Block diagram of (a) the proposed system, (b) MIMO-DCCRN, (c) the signal processing localization module (SPLM), and (d) the neural localization module (NLM).

3.1 MIMO DCCRN

DCCRN in [6] is the complex network modified by CRN in [5]. In this study, we extend DCCRN to MIMO-DCCRN. Signals captured by different channels in time-frequency domain through STFT are taken as the MIMO-DCCRN input. The real and imaginary parts of complex are stacked together to obtain a 2-channel tensor of size $2 \times T \times F$, where T represents the time steps and F represents the number of frequency bins, which are used as input features. Considering all M channels, it leads to a $2M \times T \times F$ dimensional input $\mathbf{y}$ to the framework. The complex encoder consists of complex convolution layers followed by a complex batch normalization (BN) and PReLU activation function. In the multi-input complex encoder, the input across different channels is interpreted both in spectral and spatial aspects. Features extracted by the complex encoder are utilized in complex LSTM to be further interpreted in temporal information. The complex decoder is designed to be the structure symmetrical to the complex encoder to estimate $\mathbf{w}^H$ based on the above feature extraction. Skip-connection is conductive to flowing the gradient by concentrating the complex encoder and decoder. Fig. 1(b) illustrates the proposed architecture for multi-channel speech enhancement.

3.2 Localization and VAD

Aiming to locate the target speaker position, complex filters estimated in the aforementioned system are passed through the localization module. The speech enhancement and localization problems are jointly tackled via multi-task learning. As shown in Fig. 1(a), a localization module is cascaded with the output of the MIMO-DCCRN. The main goal of an array beamformer is to restore the source signal from a target direction as a spatial filter, while suppressing noise and interference from other directions. To achieve this goal, the distortionless constraint is imposed to maintain a constant response for the target source direction. This motivates the proposed localization module to utilize the following distortionless constraint as a localization criterion:

$$\mathbf{w}^H \mathbf{a} = 1, \tag{4}$$

where $\mathbf{a}$ is a steering vector pertaining to the target speaker direction.

3.2.1 The Signal Processing Localization Module (SPLM)

The signal processing localization module (SPLM) is illustrated in Fig. 1 (c). By assuming N candidate zones in azimuthal angles, the free-field plane wave steering vector

$\mathbf{a}_n(f) \in \mathbb{C}^{M\times 1}$, $n = 1,\ldots,N$ is predesignated according to the pre-specified angular grid, which is the center of zone-*n*,

$$\mathbf{a}_n(f) = \left[e^{-j\mathbf{k}_n\cdot\mathbf{r}_1} \quad e^{-j\mathbf{k}_n\cdot\mathbf{r}_2} \quad \cdots \quad e^{-j\mathbf{k}_n\cdot\mathbf{r}_M} \right]^T, \tag{5}$$

where $\mathbf{r}_m$ is the position vector of the *m*th microphone, $\mathbf{k}_n = -k\boldsymbol{\kappa}_n = -(\omega/c)\boldsymbol{\kappa}_n = -(2\pi f/c)\boldsymbol{\kappa}_n$ is the wave vector, $\boldsymbol{\kappa}_n$ denotes the direction of arrival (DOA) vector that is a unit vector pointing to the look direction, *k* denotes the wave number, and *c* is the speed of sound. Here we define the estimated localization mapping be $\hat{\mathbf{z}}(t) = \left[\hat{z}_1(t) \quad \ldots \quad \hat{z}_N(t)\right] \in \mathbb{R}^N$, which can be viewed as probabilities, composed of distortionless index. The distortionless index $\hat{z}_n(t)$ of the *n*-direction and time frame *t* is formulated based on the estimated filter coefficients $\mathbf{w}^H$ in MIMO-DCCRN and predefined steering vectors,

$$\hat{z}_n(t) = \frac{1}{F}\sum_{f=1}^{F} \left| \mathbf{w}^H(t,f)\mathbf{a}_n(f) \right|. \tag{6}$$

The active zone can be determined by peak finding of $\hat{z}_n(t)$

$$\hat{n}(t) = \arg\max_n \hat{z}_n(t). \tag{7}$$

In addition, the peak of $\hat{z}_n(t)$ give the information of voice activity

$$\mathrm{VAD}(t) = \max_n \hat{z}_n(t). \tag{8}$$

Hereafter, we denote MIMO-DCCRN-SPLM-*N* as the signal processing localization module with predefined *N* zones cascaded with MIMO-DCCRN.

3.2.2 The Neural Localization Module (NLM)

In Eq. (5), steering vectors are derived based on the free-field plane wave assumption, which does not well accommodate in the real-world application. Thus, the auxiliary network is further proposed to model the realistic sound field. The neural network consisting of the complex convolutional layer with complex BN, PReLU activation and the linear layer with sigmoid activation function replaces coefficients of steering vector. As shown in Fig. 1(d), filter coefficients are processed through the localization network to estimate $\hat{\mathbf{z}}(t)$. The direction and VAD are computed according to Eq. (7) and Eq. (8), respectively. Hereafter, the localization module employing the learning-based sound field with predefined *N* zones for localization in training and cascaded with the MIMO-DCCRN is denoted as MIMO-DCCRN-NLM-*N*.

3.3 Loss Functions

In this study, different loss functions are introduced for enhancement and localization problems. For speech enhancement, the objective is to minimize the negative scale-invariant source-to-noise ratio (SI-SNR), which has commonly used as an evaluation metric for enhancement replacing the mean square error (MSE) and signal-to-distortion ratio (SDR) [16]. SI-SNR is defined as

$$\begin{aligned} \mathbf{s}_{\text{target}} &= \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s}}{\|\mathbf{s}\|_2^2} \\ \text{SI-SNR} &= 20\log_{10} \frac{\|\mathbf{s}_{\text{target}}\|_2^2}{\|\hat{\mathbf{s}} - \mathbf{s}_{\text{target}}\|_2^2}, \end{aligned} \tag{9}$$

where $\hat{\mathbf{s}}$ and $\mathbf{s}$ are the estimated and original clean time-domain waveform, respectively. $\langle \cdot,\cdot \rangle$ denotes the dot product between two vectors and $\|\cdot\|_2$ is Euclidean norm. $loss_{\text{SI-SNR}}$ is denoted as the average of negative SI-SNR in time index.

For DOA estimation, if the target source is activated at azimuth angle θ and time frame *t*, the ground truth of localization mapping is defined as $\mathbf{z}(t) = \left[z_1(t) \quad \ldots \quad z_N(t)\right] \in \mathbb{R}^N$, where

$$z_n(t)=\begin{cases}1, & -\frac{360}{2N}+(n-1)\cdot\frac{360}{N}<\theta\le-\frac{360}{2N}+n\cdot\frac{360}{N}\\ 0, & \text{otherwise}\end{cases}. \tag{10}$$

While the target source is not activated, $z_n(t)=0$, $n=1,\ldots,N$. The localization problem is formulated as a classification task with $\mathbf{z}(t)$ and $\hat{\mathbf{z}}(t)$. As mentioned above, with $z_n(t)\in\{0,1\}$ and $\hat{z}_n(t)\in[0,1]$, the binary cross-entropy is minimized in training:

$$loss_{\text{BCE}}=-\frac{1}{TN}\sum_{t=1}\sum_{n=1}z_n(t)\log\hat{z}_n(t)+\left(1-z_n(t)\right)\log\left(1-\hat{z}_n(t)\right), \tag{11}$$

Therefore, the total loss of the multi-task network is

$$loss=loss_{\text{BCE}}+\gamma loss_{\text{SI-SNR}}, \tag{12}$$

where γ is the weighting constant and is set to 1 in this paper.

4. SIMULATION STUDY

4.1 Dataset Preparation

To validate the proposed system, we conduct a simulation data based on four open datasets. The clean utterances for training and testing are selected from the train-clean-360 and dev-clean subsets of the LibriSpeech corpus [17], which contains utterances from 921 and 40 speakers, respectively. The noisy training and testing data are synthesized with noise corpus using the MS-SNSD dataset [18], FSDnoisy18k dataset [19] and the Free Music Archive (FMA) [20]. From MS-SNSD dataset, noise data excluded babble noise is chosen for directional interference noises. In FSDnoisy18k dataset, noises which are lack of directionality, such as the environment noise, are not considered for data preprocessing in this work. In this paper, we process the audio both contained stationary noises, such as music and engine, and transient noises, such as glass breaking and typing noise.

The duration of mixed noisy audio signals prepared for training and testing is 6 -s. Each mixture is composed of 4-s clean speech audio clips, which were randomly inserted into 6-s duration, Gaussian white noise, which is used as sensor noises, and interference noises. A 6-element uniform circular array (UCA) with a radius of 5 cm is employed. In the training phase, the signal-to-noise ratio (SNR) of sensor noise ranged from 10 to 30 dB. The signal-to-interference ratio (SIR) of inference noise is selected from -5 to 15 dB. The room size, the distance between sound source and microphone array, and T60 are sampled from $\{4\times4\times3,\ 5\times5\times3,\ 6\times6\times3\}$ m^3, {[1, 1.5], [1, 2], [1, 2.5]} m, and {[0.16, 0.32], [0.32, 0.48], [0.48, 0.64]} second, respectively. In the test dataset, the room size and T60 are selected to be $5\times5\times3$ m^3 and 0.32-s. The distance between the target speaker and the microphone array is fixed to 1 m, and the distance between the interference noise and the microphone array is fixed to 2 m. In the simulation, the microphone array is placed at the center of the room. The target speaker is placed sequentially on a semicircle centered at the microphone, from 0° to 180° with 1° increments, whereas the interference is located sequentially on a semicircle centered at the microphone, from 180° to 360° at increments of 1°. The noisy signals received by the microphone array are generated by convolving the anechoic clean signals with the respective room impulse responses (RIRs) simulated using the image-source method [21]. We generate 37000, 460 and 600 mixtures for the training, validation, and test dataset, respectively.

4.2 Training Setup and Baselines

All the waveforms are sampled at 16 kHz. The STFT settings are a Hanning window with 25-ms length, a 6.25-ms hop size, and a 512-point fast Fourier transform. In the training stage, Adam optimizer is adopted with the learning rate as 0.001. All models are trained for 100 epochs.

In our proposed system, the number of channels of the complex convolution layers in the encoder of MIMO-DCCRN are 16, 32, 64, 128, 256, 256, respectively. The number of channels of the decoder are the same of the encoder in reversed order. All complex (de-)convolution of the encoder (decoder) use 5×2 kernel size, with stride 2×1. For RNN layers, 256 hidden units are applied in complex

LSTM. In NLM, the complex convolution layers are 2*N*, 2*N*, and the joined linear layers are 32, 1. In the following, we compare the improvements of MIMO-DCCRN to simply enhance speech, MIMO-DCCRN-SPLM-12 and MIMO-DCCRN-NLM-12 which are combining localization on 12 zones, and MIMO-DCCRN-SPLM-36 and MIMO-DCCRN-NLM-36 to further localize on 36 zones for higher accuracy.

Two baselines are adopted in the following, including MIMO U-net [14] and DCCRN [6]. MIMO U-net is the multi-channel speech enhancement system that achieved impressive performance in neural beamforming. DCCRN outperforms other speech enhancement methods in the monaural scenario.

4.3 Results

4.3.1 Speech Enhancement Performance

To quantify the signal enhancement performance, perceptual evaluation of speech quality (PESQ) [22] and short-time objective intelligibility (STOI) [23] are employed. Table 1 summarizes the PESQ and STOI scores for the test dataset. First, we examine the effects of the novel MIMO-DCCRN on the level of performance improvement. Table 1 shows that our proposed methods yield better performance than baseline methods. Comparing MIMO-DCCRN and DCCRN, multi-channel speech enhancement attains significantly improvements than monaural enhancement, especially the STOI score at low SIR. For multi-channel enhancement, the MIMO-DCCRN system consisting of complex operations achieves great improvement than MIMO U-net. The results proved that MIMO-DCCRN is more robust to scenarios at different SIR.

Second, we examine the effects of combining localization to speech enhancement. According to Table 1, MIMO-DCCRN-NLM-36 performs the best among all methods. On the contrary, PESQ and STOI decrease slightly in the MIMO-DCCRN-SPLM system without learning-based sound field. NLM thus is proved to be more reliable than SPLM. Furthermore, MIMO-DCCRN-SPLM-12 and MIMO- DCCRN-NLM-12, which using small numbers of zones in localization, degrade the enhanced signal quality, since insufficient number of zones designed in localization leads to misalignment in neural beamforming. The result suggests that appropriate localization module aids beamforming in maintaining the desired signal at specific direction. Thus, MIMO-DCCRN-NLM-36 can further improve the enhanced signal quality.

Table 1 – Speech enhancement performance

Performance Metric	PESQ					STOI (in %)				
SIR (dB)	-10	-5	0	10	Avg.	-10	-5	0	10	Avg.
Noisy	1.58	1.68	1.85	2.03	1.79	70.81	75.85	82.26	86.29	78.80
DCCRN	1.69	1.86	2.12	2.38	2.01	72.75	78.94	85.94	89.42	81.76
MIMO U-net	1.90	2.08	2.38	2.63	2.25	77.93	82.48	87.91	90.66	84.75
MIMO-DCCRN	**2.43**	**2.71**	**3.07**	**3.31**	**2.88**	86.64	90.63	94.56	96.07	91.98
MIMO-DCCRN-SPLM-12	2.33	2.61	2.98	3.25	2.79	86.26	90.37	94.43	95.96	91.76
MIMO-DCCRN-NLM-12	2.30	2.59	2.95	3.21	2.76	86.40	90.58	94.58	96.05	91.90
MIMO-DCCRN-SPLM-36	2.33	2.59	2.93	3.19	2.76	86.31	90.37	94.32	95.91	91.73
MIMO-DCCRN-NLM-36	**2.43**	**2.71**	**3.07**	**3.31**	**2.88**	**86.94**	**90.87**	**94.66**	**96.11**	**92.15**

4.3.2 Localization and VAD Performance

To evaluate the performance of the proposed DOA estimation, we define three evaluation metrics: total accuracy (ACC), adjacent error rate (AER), and other error rate (OER) in Eq. (12).

$$ACC = \frac{L_{\text{true}}}{L_{\text{all}}},\ AER = \frac{L_{\text{adjacent}}}{L_{\text{all}}},\ OER = \frac{L_{\text{false}}}{L_{\text{all}}}, \tag{12}$$

where L_{true} is the number of true estimations, L_{adjacent} is false estimations on adjacent zones, L_{false} is false estimations except on adjacent zones, and L_{all} is the total number of localization estimations. In this evaluation, we only estimate localization when the target source was activated. The localization

results are illustrated in Table 2, where different localization module and different number of zones are compared. As the number of zones increase, the performance of SPLM decrease considerably. This result reveals that NLM is more robust in application than SPLM.

Moreover, Fig. 2 shows how the VAD output has learned from localization module. The system enables the model to learn only the interfering voice activity while ignoring the interfering voice activity. Further evaluation on the quality of the VAD output is beyond the scope of this work.

Table 2 – Localization performance

Performance Metric	ACC (%)					AER (%)					OER (%)				
SIR (dB)	-10	-5	0	10	Avg.	-10	-5	0	10	Avg.	-10	-5	0	10	Avg.
MIMO-DCCRN-SPLM-12	80.1	83.3	85.7	86.2	83.8	15.8	14.1	12.5	12.5	13.7	4.0	2.5	1.7	1.3	2.4
MIMO-DCCRN-NLM-12	**90.9**	**94.1**	**95.6**	**96.1**	**94.2**	**6.8**	**4.7**	**3.8**	**3.4**	**4.7**	**2.2**	**1.1**	**0.7**	**0.5**	**1.1**
MIMO-DCCRN-SPLM-36	55.0	57.4	59.4	60.4	58.1	36.8	37.1	37.3	37.1	37.1	7.7	5.1	3.1	2.4	4.6
MIMO-DCCRN-NLM-36	**80.8**	**85.1**	**88.8**	**90.0**	**86.2**	**15.5**	**12.6**	**9.9**	**9.0**	**11.8**	**3.8**	**2.3**	**1.4**	**1.0**	**2.1**

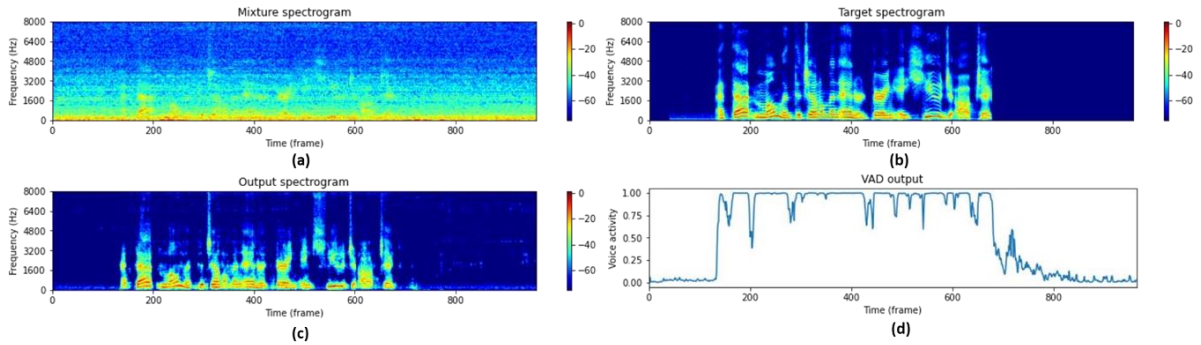


Figure 2 – Spectrogram of (a) noisy mixture signal, (b) clean speech (c) enhanced signal by MIMO-DCCRN-NLM-12, (d) VAD output.

5. CONCLUSIONS

In this paper, a DCCRN-based neural beamformer, is proposed for multi-channel speech enhancement and localization. MIMO-DCCRN exhibits superior performance in speech enhancement. By employing distortionless constraint in loss function, the neural beamformer also provides effective localization and VAD functions. Combining neural beamformer and the learning-based sound field for localization with a sufficiently fine grid proves useful in enhancing speech quality. In brief, the combined MIMO-DCCRN-NLM-36 system has achieved significantly improved performance in speech enhancement, as compared to baselines.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Science and Technology (MOST), Taiwan, under the project number 110-2221-E-007-027-MY3.

REFERENCES

1. C. Macartney and T. Weyde, “Improved speech enhancement with the wave-u-net”, 2018.
2. R. Giri, U. Isik and A. Krishnaswamy, “Attention wave-U-Net for speech enhancement”, *Proc. WASPAA*, pp. 4049-4053, 2019.
3. Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation”, *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 27, no. 8, pp. 1256-1266, 2019.
4. A. Narayanan and D. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech

recognition", *Proc. IEEE ICASSP*, pp. 7092-7096, 2013.
5. K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement", *Proc. Interspeech*, pp. 3229-3233, 2018.
6. Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, et al., "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement", *Proc. Interspeech*, pp. 2472-2476, 2020.
7. J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming", *Proc. IEEE ICASSP*, pp. 196-200, 2016.
8. B. D. van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering", *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4-24, Apr. 1988.
9. E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition", *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 5, pp. 1529-1539, 2007.
10. Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All deep learning MVDR beamformer for target speech separation", *Proc. IEEE ICASSP*, pp. 6089-6093, 2021.
11. Y. Luo, E. Ceolini, Cong Han, Shih-Chii Liu, and N. Mesgarani, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing", *IEEE Automatic Speech Recognition and Understanding Workshop*, 2019.
12. Z. Q. Wang and D. L. Wang, "All-neural multi-channel speech enhancement", *Proc. Interspeech*, pp. 3234-3238, 2018.
13. A. Li, W. Liu, C. Zheng, and X. Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement", *arXiv preprint arXiv:2109.00265*, 2021.
14. X. Ren, X. Zhang, L. Chen, X. Zheng, et al., "A causal U-Net based neural beamforming network for real-time multichannel speech enhancement", *Proc. Interspeech 2021*, Sep. 2021.
15. U. Hamid, R. A. Qamar, and K. Waqas, "Performance comparison of time-domain and frequency-domain beamforming techniques for sensor array processing", *Int. Bhurban Conference on Appl. Sci. & Technology*, pp. 379-385, Jan. 2014.
16. E. Vincent, R. Gribonval, and C. F´evotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
17. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books", in *Proc. IEEE ICASSP*, South Brisbane, Australia, pp. 5206–5210, 2015.
18. C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework", *Proc. Interspeech*, pp. 1816-1820, 2019.
19. E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels", *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 21-25, 2019.
20. M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *Proc. Int. Society for Music Information Retrieval Conf.*, Suzhou, China, pp. 316–323, 2017.
21. J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *JASA*, vol. 65, pp. 943–950, 1979.
22. A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE ICASSP*, Salt Lake City, Utah, USA, pp. 749-752, 2001.
23. C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE ICASSP*, Dallas, pp. 4214–4217, 2010.

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0214

Evaluating the dereverberation-separation capability of multi-frame full-rank spatial covariance analysis

Hiroshi SAWADA[(1)], Rintaro IKESHITA[(1)], Keisuke KINOSHITA[(1)], Tomohiro NAKATANI[(1)]

[(1)]NTT Communication Science Laboratories, NTT Corporation, Japan

ABSTRACT

Full-rank spatial covariance analysis (FCA) is a method for blind source separation (BSS) that can be applied to underdetermined situations where the number of sources outnumbers the number of microphones. Recently, the authors proposed multi-frame FCA (mfFCA) as an extension of FCA to improve the performance of BSS when the room reverberations are so long that multiple frames are needed to cover the dominant part of the reverberations. The authors have already demonstrated that mfFCA performed better for BSS than the existing FCA-related methods. However, the dereverberation capability of mfFCA in BSS tasks has not been evaluated so far. This paper newly evaluates the dereverberation-separation capability of mfFCA. We perform experimental comparisons with FCA and also weighted prediction error (WPE), a well-established dereverberation method for over-determined situations where microphones outnumber sources.

Keywords: Blind source separation (BSS), Blind dereverberation (BD), Full-rank spatial covariance analysis (FCA), Weighted prediction error (WPE)

1 INTRODUCTION

Aiming to capture sound sources distant from microphones as clear as possible, blind source separation (BSS) and blind dereverberation (BD) have been studied for several decades [1–4]. Among such methods, independent component analysis (ICA) [5–7] and weighted prediction error (WPE) [8] are well-established ones for BSS and BD, respectively. One of the fundamental limitations of ICA is that it can only be applied to over-determined cases where the number N of sources does not exceed the number M of microphones. Full-rank spatial covariance analysis (FCA) [9–11], on the other hand, can be applied to underdetermined cases where $N > M$. Since ICA and WPE are based on linear transformations, WPE has the same limitation as ICA in principle. However, in practice, WPE can be applied to underdetermined cases for the purpose of dereverberation (not separation) with tolerating some estimation errors.

The original FCA is basically a BSS method as ICA, and does not perform BD. There have been proposed FCA extensions to perform BD [12–14]. However, their proposed methods are combined with WPE and have been experimentally confirmed only for over-determined cases.

Recently, we have proposed multi-frame FCA (mfFCA) and experimentally demonstrated that mfFCA solves the BSS tasks better than FCA especially in reverberant situations [15]. According to its multi-frame nature, mfFCA can naturally perform BD like the FCA extensions above. However, the BD capability of mfFCA has not yet been evaluated. Thus, in this paper, we newly evaluate the BD capability of mfFCA in BSS tasks. To the best of our knowledge, this is the first experimental evaluation of joint BD and BSS tasks in underdetermined situations ($N > M$).

The next section describes the model and the algorithm of mfFCA, together with the way of source separation and dereverberation. Section 3 reports the experimental procedures and results for blind dereverberation-separation tasks. Section 4 concludes the paper.

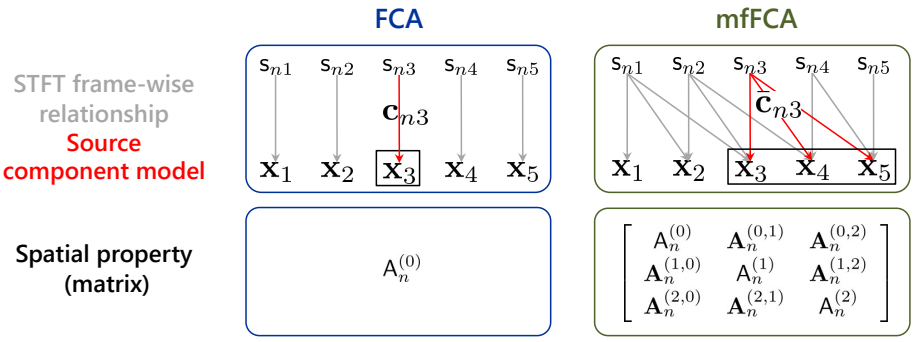


Figure 1. Illustrations of FCA and mfFCA models. The set of time lags for mfFCA is $\mathscr{L} = \{1,2\}$.

2 MULTI-FRAME FCA (mfFCA)

2.1 Model

Let us begin with the original FCA model. Suppose that $n = 1,\ldots,N$ sources are mixed and observed at $m = 1,\ldots,M$ microphones. Let the observations at a time frame $t \in \{1,\ldots,T\}$ be denoted by an M-dimensional complex vector $\mathbf{x}_t \in \mathbb{C}^M$ with $\mathbf{x}_t = [x_{1t},\ldots,x_{Mt}]^\mathsf{T}$. Throughout this paper, unless otherwise noted, we consider complex-valued time-frequency signal representations obtained by applying a short-time Fourier transform (STFT) to time-domain sound signals. In the original FCA model, $\mathbf{x}_t$ is assumed to be the sum $\mathbf{x}_t = \sum_{n=1}^N \mathbf{c}_{nt}$ of N source components $\mathbf{c}_{nt} \in \mathbb{C}^M$, each of which follows a zero-mean multivariate complex Gaussian distribution $p(\mathbf{c}_{nt}) = \mathcal{N}(\mathbf{c}_{nt} \mid \mathbf{0}, \mathsf{C}_{nt})$ with covariance matrices $\mathsf{C}_{nt} = \mathsf{s}_{nt}\mathsf{A}_n^{(0)}$. By these Sans-serif fonts, the lower and upper cases represent that the parameters are positive scalars and Hermitian positive definite matrices, respectively. $\mathsf{A}_n^{(0)}$ encodes the time-invariant spatial property from source n to all M microphones. s_{nt} represents the time t dependent power of source n. The FCA parameters are summarized as $\theta = \{\{\{\mathsf{s}_{nt}\}_{t=1}^T, \mathsf{A}_n^{(0)}\}_{n=1}^N\}$.

mfFCA is an extension of FCA (see Fig. 1), in which we consider a set $\mathscr{L} = \{l_1,\ldots,l_L\}$ of time lags and employ multi-frame vectors for observations as

$$\bar{\mathbf{x}}_t = [\mathbf{x}_t^\mathsf{T}, \mathbf{x}_{t+l_1}^\mathsf{T}, \ldots, \mathbf{x}_{t+l_L}^\mathsf{T}]^\mathsf{T} \in \mathbb{C}^{M(L+1)}, \tag{1}$$

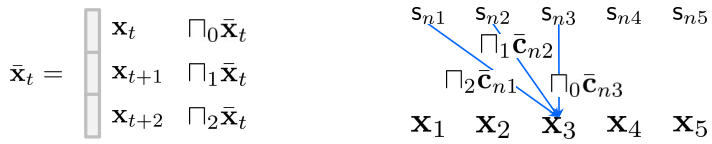
together with multi-frame source components $\bar{\mathbf{c}}_{nt} \in \mathbb{C}^{M(L+1)}$ of the same dimensionality as shown in Fig. 1. We assume that the source components $\bar{\mathbf{c}}_{nt}$ occur at time t and are observed at time $t, t+l_1, \ldots, t+l_L$ due to reverberations. We model $\bar{\mathbf{c}}_{nt}$ by zero-mean Gaussian distributions $p(\bar{\mathbf{c}}_{nt}) = \mathcal{N}(\bar{\mathbf{c}}_{nt} \mid \mathbf{0}, \bar{\mathsf{C}}_{nt})$ with covariance matrix $\bar{\mathsf{C}}_{nt} = \mathsf{s}_{nt}\bar{\mathsf{A}}_n$, where

$$\bar{\mathsf{A}}_n = \begin{bmatrix} \mathsf{A}_n^{(0)} & \cdots & \mathbf{A}_n^{(0,l_L)} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_n^{(l_L,0)} & \cdots & \mathsf{A}_n^{(l_L)} \end{bmatrix} \tag{2}$$

is of size $M(L+1) \times M(L+1)$ and encodes the time-invariant spatial property from source n to all M microphones with all the considered time lags. The set of mfFCA parameters is given as $\theta = \{\{\{\mathsf{s}_{nt}\}_{t=1}^T, \bar{\mathsf{A}}_n\}_{n=1}^N\}$, which can be reduced to the FCA parameter set by letting $\mathscr{L} = \emptyset$.

A multi-frame observation $\bar{\mathbf{x}}_t$ consists mostly of the summation $\sum_{n=1}^N \bar{\mathbf{c}}_{nt}$ of multi-frame source components. However, $\bar{\mathbf{x}}_t$ has additional components as Fig. 1 shows oblique grey lines coming into the $\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5$'s box. For more specific definitions, let us introduce an operator $\sqcap_i$ that extracts the $(i+1)$-th subvector of a multi-frame vector. Then, as Fig. 2 illustrates, we model each subvector of a multi-frame observation as

$$\sqcap_j \bar{\mathbf{x}}_t = \sum_{n=1}^N \left\{ \sqcap_0 \bar{\mathbf{c}}_{n(t+l_j)} + \sum_{i=1}^L \sqcap_i \bar{\mathbf{c}}_{n(t+l_j-l_i)} \right\}. \tag{3}$$



$$\mathbf{x}_3 = \sqcap_0\bar{\mathbf{x}}_3 = \sqcap_1\bar{\mathbf{x}}_2 = \sqcap_2\bar{\mathbf{x}}_1 = \textstyle\sum_{n=1}^{N}\left(\sqcap_0\bar{\mathbf{c}}_{n3} + \sqcap_1\bar{\mathbf{c}}_{n2} + \sqcap_2\bar{\mathbf{c}}_{n1}\right)$$

Figure 2. Upper left: $\sqcap_i$ operator that extracts the $(i+1)$-th subvector of a multi-frame vector. Upper right: mixture model for $\mathbf{x}_3$. Bottom: various expressions of $\mathbf{x}_3$.

We have derived the probabilistic model of mfFCA with some additional mild assumptions [15]. The multi-frame observations can be modeled as zero-mean Gaussian distributions $p(\bar{\mathbf{x}}_t \mid \theta) = \mathcal{N}(\bar{\mathbf{x}}_t \mid \mathbf{0}, \bar{\mathsf{X}}_t)$ whose covariance matrices are expressed as

$$\bar{\mathsf{X}}_t = \textstyle\sum_{n=1}^{N}\bar{\mathsf{C}}_{nt} + \begin{bmatrix} \check{\mathsf{X}}_t & & \\ & \ddots & \\ & & \check{\mathsf{X}}_{t+l_L} \end{bmatrix}. \tag{4}$$

The first term depends on s_{nt} because $\bar{\mathsf{C}}_{nt} = \mathsf{s}_{nt}\bar{\mathsf{A}}_n$. The second term depends on $\mathsf{s}_{n(t-l)}$ and $\mathsf{s}_{n(t+l)}$ with the time lags l in $\mathscr{L}$ as the block diagonal components are defined by

$$\check{\mathsf{X}}_t = \textstyle\sum_{n=1}^{N}\left(\sum_{l\in\mathscr{L}}\mathsf{C}_{nt}^{(l)}\right),\ \ldots,\ \check{\mathsf{X}}_{t+l_L} = \sum_{n=1}^{N}\left(\mathsf{C}_{n(t+l_L)}^{(0)} + \sum_{l\in\mathscr{L}-\{l_L\}}\mathsf{C}_{n(t+l_L)}^{(l)}\right),\ \ \mathsf{C}_{nt}^{(l)} = \mathsf{s}_{n(t-l)}\mathsf{A}_n^{(l)}. \tag{5}$$

2.2 EM algorithm

By assuming the conditional independence of multi-frame observation vectors $\bar{\mathbf{x}}_t$ under fixed parameters θ, the objective function of mfFCA is given as $\sum_{t=1}^{T-l_L}\log p(\bar{\mathbf{x}}_t \mid \theta)$. The parameters θ can be optimized from some initial values by locally maximizing the objective function by the EM algorithm [16]. In the **E-step**, we calculate the conditional distributions of the multi-frame source components $\bar{\mathbf{c}}_{nt}$ as

$$p(\bar{\mathbf{c}}_{nt} \mid \bar{\mathbf{x}}_t, \theta) = \mathcal{N}(\bar{\mathbf{c}}_{nt} \mid \bar{\mu}_{nt}, \bar{\Sigma}_{nt}), \tag{6}$$

$$\bar{\mu}_{nt} = \bar{\mathsf{C}}_{nt}\bar{\mathsf{X}}_t^{-1}\bar{\mathbf{x}}_t,\ \ \bar{\Sigma}_{nt} = \bar{\mathsf{C}}_{nt} - \bar{\mathsf{C}}_{nt}\bar{\mathsf{X}}_t^{-1}\bar{\mathsf{C}}_{nt}. \tag{7}$$

The part $\bar{\mathsf{C}}_{nt}\bar{\mathsf{X}}_t^{-1}$ for the mean $\bar{\mu}_{nt}$ calculation is called multi-frame multichannel Wiener filter in [17]. In the **M-step**, we optimize the parameters in θ, with some approximation detailed in [15], as

$$\bar{\mathsf{A}}_n \leftarrow \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\mathsf{s}_{nt}}\widetilde{\mathsf{C}}_{nt},\ \ \mathsf{s}_{nt} \leftarrow \frac{1}{M(L+1)}\mathrm{tr}\left(\bar{\mathsf{A}}_n^{-1}\widetilde{\mathsf{C}}_{nt}\right),\ \ \widetilde{\mathsf{C}}_{nt} = \bar{\mu}_{nt}\bar{\mu}_{nt}^{\mathsf{H}} + \bar{\Sigma}_{nt}. \tag{8}$$

2.3 Source separation and dereverberation

Once the parameters θ are optimized by the EM algorithm, we can separate a multi-frame observation $\bar{\mathbf{x}}_t$ into $\bar{\mathbf{c}}_{1t},\ldots,\bar{\mathbf{c}}_{Nt}$ by applying the multi-frame multichannel Wiener filters $\bar{\mathsf{C}}_{nt}\bar{\mathsf{X}}_t^{-1}$, $n=1,\ldots,N$ to $\bar{\mathbf{x}}_t$ as in (7) to obtain $\bar{\mu}_{1t},\ldots,\bar{\mu}_{Nt}$. Then, we calculate the separated signal for source n at time t as

$$\mathbf{y}_{nt} = \sqcap_0\bar{\mu}_{nt} + \textstyle\sum_{i=1}^{L}\sqcap_i\bar{\mu}_{n(t-l_i)} \tag{9}$$

by accumulating all the source n components observed at time t. Because the expression (9) consists of the direct component (the first term) and the delayed components (the second term), the dereverberated signal can simply be obtained as

$$\mathbf{d}_{nt} = \sqcap_0\bar{\mu}_{nt}. \tag{10}$$

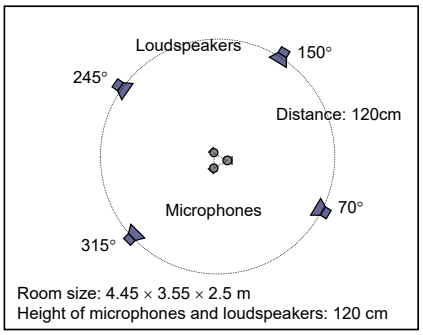


Figure 3. Experimental setup

3 EXPERIMENTS

3.1 Conditions and task

We performed experiments to separate and dereverberate $N = 4$ speech sources with $M = 3$ microphones. We measured the impulse responses from the sources (loudspeakers) to the microphones under the room conditions shown in Fig. 3. The room reverberation time varied from 200 ms to 450 ms. For each reverberation time, time-domain 8 mixtures at the microphones were constructed by convolving the impulse responses and 6-second English speech sources. In order to evaluate separation and dereverberation performances, we additionally made time-domain source n images $\mathrm{img}_{mn}^{(cut)}$ at each microphone m by convolving impulse responses that were cut to 64 ms. The performances were measured in terms of signal-to-distortion ratios (SDRs) [18] by setting $\mathrm{img}_{mn}^{(cut)}$ as reference signals. In other words, the task of joint BSS and BD was to recover $\mathrm{img}_{mn}^{(cut)}$ from the reverberant mixtures as clearly as possible.

3.2 Full-band signal construction

In the descriptions of Sec. 2, we intentionally omit frequency dependency f for notational simplicity. We actually have microphone observations x_{mtf} for frequency bins $f = 1, \ldots, F$. In the experiments, the sampling frequency was 8 kHz, and the STFT window width and shift were 128 ms and 32 ms. Consequently, the numbers of time frames and frequency bins were $T = 208$ and $F = 513$, respectively.

The solutions of mfFCA have permutation ambiguities among frequency bins. We need to align the permutations for blindly separating sources in a full-band manner. We aligned permutations first by the post-processing approach [19] at EM iterations 5, 50, 100, 150, 200. Then, from 201 to 500 iterations, we shared the source power parameters s_{ntf} among adjacent four frequency bins. Specifically, we averaged s_{ntf} among the corresponding frequency bins after the update (8). The sharing is effective for permutation alignment [20, 21] and also for mitigating the rank deficient problem of FCA/mfFCA [15, 22].

3.3 Methods

We examined and compared eight methods for the BSS and BD task. The first one was FCA, which did not perform dereverberation. The second to fourth ones were mfFCA with sets of time lags $\mathscr{L} = \{2\}$, $\mathscr{L} = \{2,4\}$ and $\mathscr{L} = \{2,4,6\}$, respectively. These sets of time lags were considered to be effective for the quarter shift of STFT windows. The fifth one was WPE with a set of time lags $\{2,3,4,5,6\}$ followed by FCA. We expected that WPE removed reverberations to a certain degree despite that the number M of microphones was insufficient for the number N of sources. The sixth to eighth ones were WPE followed by mfFCA with the same sets of time lags above. The labels for these eight methods are listed on the right edge of Fig. 4. All the methods were coded with Python using CuPy [23] for computational acceleration.

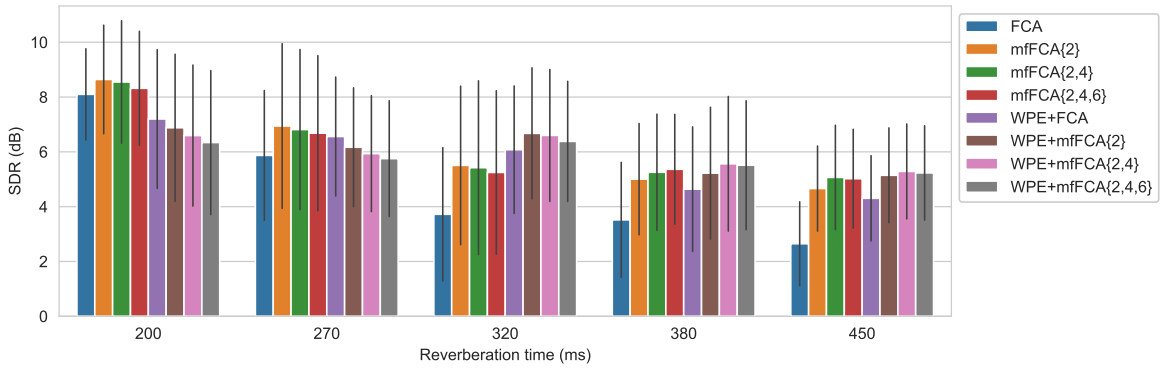


Figure 4. Mean SDR values with the examined eight methods under various reverberation times. Each vertical line shows the standard deviation of 32 SDR values (8 mixture combinations of $N = 4$ sources).

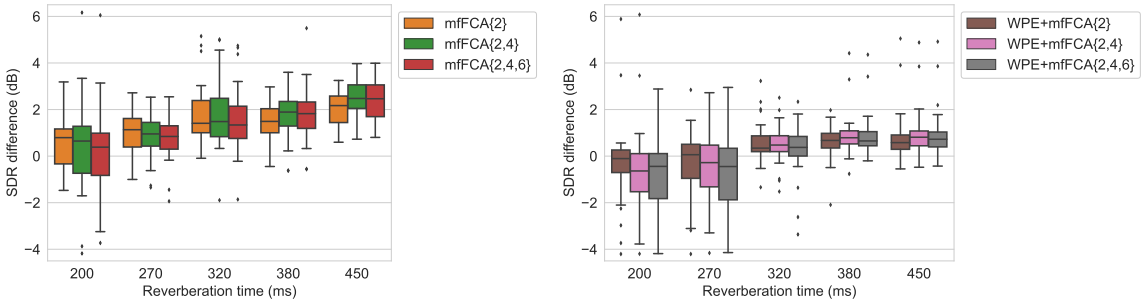


Figure 5. SDR differences to the first baseline FCA (left) and the second baseline WPE+FCA (right). Each box plot shows the distribution of 32 SDR differences (8 mixture combinations of $N = 4$ sources).

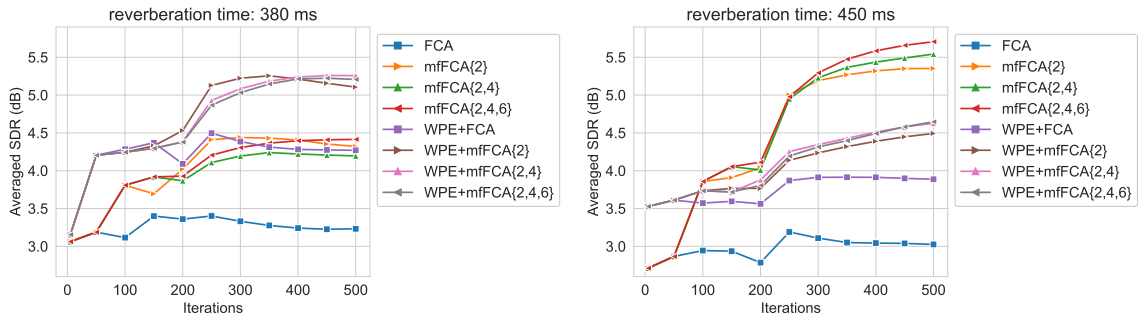


Figure 6. Typical convergence behaviors.

3.4 Results

Figure 4 summarizes the experimental results. We observe that employing mfFCA instead of FCA generally improved the results. To examine how much the results improved, we calculated SDR differences to two baselines, FCA and WPE+FCA. Figure 5 shows the distributions of SDR differences. From the left plot, we observe that the improvements by replacing FCA with mfFCA increased as the reverberation times became longer, and achieved more than 2 dB improvements with the 450 ms reverberation time. The right plot shows the SDR differences when WPE+FCA was the baseline. We observe that in the high reverberant cases (320, 380, 450 ms) there were improvements despite the tendency to be weaker than the left plot. We understand these high reverberant situations as follows. The dereverberation results by WPE were not perfect due to the insufficient number of microphones, and there were residual reverberations in the WPE results. Then, mfFCA removed the residual reverberations to get better results.

Figure 6 shows typical convergence behaviors as examples. The horizontal axes show the iteration numbers of the EM algorithms for FCA/mfFCA. The plots begin with the 5th iteration. Before this point, WPE was applied if specified, the FCA parameters were initialized by the procedure shown in [24], the FCA parameters were updated by 5 EM iterations, and permutation ambiguities were aligned. Then, mfFCA inherited the FCA parameters and iterated the EM updates by augmenting the time lag set from an empty set to the specified set $\mathscr{L}$. The vertical axes show the average of $N = 4$ SDRs for a source combination. Using a large set of time lags, .e.g, $\mathscr{L} = \{2,4,6\}$, mfFCA performed the best for these specific cases with long reverberations, although it took many iterations (500 in these cases). As contrasting between the left and right plots, it was hard to state that applying WPE before mfFCA was always better than not applying WPE.

4 CONCLUSIONS

We have evaluated the dereverberation capability of mfFCA in underdetermined BSS tasks. From the experimental results, mfFCA with an appropriate set of time lags for the reverberation time performed well for the joint BSS and BD tasks. Even in underdetermined situations, WPE can remove reverberations to a certain degree. Therefore, it is nice to combine WPE and FCA/mfFCA as already approached [12–14]. Future work includes the development of such methods where WPE and mfFCA are well coordinated.

References

[1] C. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal processing*, 24(1):1–10, 1991.

[2] S. Haykin, editor. *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*. John Wiley & Sons, 2000.

[3] S. Makino, T.-W. Lee, and H. Sawada, editors. *Blind Speech Separation*. Springer, 2007.

[4] E. Vincent, T. Virtanen, and S. Gannot. *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.

[5] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[6] A. Bell and T. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[7] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[8] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang. Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1717–1731, 2010.

[9] N.Q.K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Trans. Audio, Speech, and Language Processing*, 18(7):1830–1840, September 2010.

[10] S. Arberet, A. Ozerov, N.Q.K. Duong, E. Vincent, R. Gribonval, F Bimbot, and P. Vandergheynst. Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation. In *Proc. ISSPA 2010*, pages 1–4, May 2010.

[11] N. Ito, R. Ikeshita, H. Sawada, and T. Nakatani. A joint diagonalization based efficient approach to underdetermined blind audio source separation using the multichannel wiener filter. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1950–1965, 2021.

[12] M. Togami. Multi-channel speech source separation and dereverberation with sequential integration of determined and underdetermined models. In *Proc. ICASSP*, pages 231–235, 2020.

[13] K. Sekiguchi, Y. Bando, A.A. Nugraha, M. Fontaine, and K. Yoshii. Autoregressive fast multichannel nonnegative matrix factorization for joint blind source separation and dereverberation. In *Proc. ICASSP*, pages 511–515, 2021.

[14] K. Sekiguchi, Y. Bando, A.A. Nugraha, M. Fontaine, K. Yoshii, and T. Kawahara. Autoregressive moving average jointly-diagonalizable spatial covariance analysis for joint source separation and dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2368–2382, 2022.

[15] H. Sawada, R. Ikeshita, K. Kinoshita, and T. Nakatani. Multi-frame full-rank spatial covariance analysis for underdetermined bss in reverberant environments. In *Proc. ICASSP*, pages 496–500, 2022.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–22, 1977.

[17] Z.-Q. Wang, H. Erdogan, S. Wisdom, K. Wilson, D. Raj, S. Watanabe, Z. Chen, and J.R. Hershey. Sequential multi-frame neural beamforming for speech separation and enhancement. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 905–911. IEEE, 2021.

[18] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N.Q.K. Duong. The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges. *Signal Processing*, 92(8):1928–1936, August 2012.

[19] H. Sawada, S. Araki, and S. Makino. Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE Trans. Audio, Speech, and Language Processing*, 19(3):516–527, March 2011.

[20] A. Hiroe. Solution of permutation problem in frequency domain ICA using multivariate probability density functions. In *Proc. ICA 2006 (LNCS 3889)*, pages 601–608. Springer, March 2006.

[21] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee. Blind source separation exploiting higher-order frequency dependencies. *IEEE Trans. Audio, Speech and Language Processing*, 15(1):70–79, January 2007.

[22] H. Sawada, R. Ikeshita, and T. Nakatani. Experimental analysis of EM and MU algorithms for optimizing full-rank spatial covariance model. In *Proc. EUSIPCO 2020*, pages 885–889, January 2021.

[23] R. Okuta, Y. Unno, D. Nishino, S. Hido, and C. Loomis. CuPy: A NumPy-compatible library for NVIDIA GPU calculations. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)*, 2017.

[24] H. Sawada, R. Ikeshita, N. Ito, and T. Nakatani. Computational acceleration and smart initialization of full-rank spatial covariance analysis. In *Proc. EUSIPCO*, pages 1–5, 2019.

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0675

Switching independent vector extraction and its joint optimization with weighted prediction error dereverberation

Tomohiro NAKATANI, Rintaro IKESHITA, Keisuke KINOSHITA, Hiroshi SAWADA, Naoyuki KAMO, and Shoko ARAKI

NTT Corporation, Japan, tnak@ieee.org

ABSTRACT

A switching filter is a versatile technique to improve the accuracy of linear filtering that enhances audio source signals from a noisy reverberant sound mixture with time-varying characteristics. It clusters time frames of an observed signal into several groups, each of which has relatively time-invariant characteristics. It achieves accurate estimation by applying different time-invariant filters separately to individual groups. Recently, the switching filter was successfully introduced into a blind source separation algorithm, Independent Vector Analysis (IVA). The algorithm is referred to as switching IVA (swIVA). This paper newly introduces the switching filter into a blind source extraction algorithm, Independent Vector Extraction (IVE). The new algorithm is referred to as switching IVE (swIVE). swIVE can more accurately estimate individual sources included in their noisy mixtures than conventional IVE. In comparison with swIVA, swIVE can perform the source extraction in a more computationally efficient way without degrading the estimation accuracy. This paper further develops an algorithm for jointly optimizing swIVE with Weighted Prediction Error dereverberation (WPE). It can simultaneously perform blind source extraction and dereverberation from noisy reverberant sound mixtures. Experiments confirm the effectiveness of the proposed algorithms.

Keywords: Blind source separation, source extraction, dereverberation, microphone array, switching filter

1 INTRODUCTION

Blind Source Separation (BSS) separates mixed audio signals captured by multiple microphones into a given number of source signals with no prior knowledge of the acoustic transfer functions from the sources to the microphones [1, 2]. BSS improves the quality of captured audio signals and thus can be used as preprocessing of speech applications, such as hands-free teleconferencing and automatic speech recognition.

Researchers have actively studied Independent Vector Analysis (IVA) [3, 4, 5] as an effective way of achieving BSS under determined conditions, i.e., when the number of sources N equals the number of microphones M ($N = M$). It performs BSS assuming that the source signals in the short-time Fourier transformation (STFT) domain can be modeled by mutually independent vectors. Independent Vector Extraction (IVE) is an extension of IVA [6, 7, 8]. From noisy mixtures, IVE can extract N source signals that are fewer than M microphones (i.e., $N < M$) in a computationally much more efficient way than IVA. For IVA to perform the same processing, it needs to separate M signals first, requiring much more computing cost, especially when $M \gg N$, and then to select N source signals from the separated M signals. In reverberant environments, applying Weighted Prediction Error-based dereverberation (WPE) [9] as preprocessing of IVA and jointly optimizing WPE and IVA have also been shown to be effective in improving the estimation accuracy of BSS [10]. We refer to this joint optimization as a blind convolutional beamforming algorithm with IVA (CIVA). WPE can also be jointly optimized with IVE. We refer to this joint optimization algorithm as CIVE in this paper [11, 12]. (It is also referred to as IVE-conv [12]).

A severe limitation of the above BSS algorithms is that they use time-invariant filters to estimate the source signals. In general, the statistical characteristics of the captured signals can be different over different time frames. Thus the use of time-invariant filters hinders the possibility of more effective estimation that could be achieved by optimally adapting the filter coefficients according to such time-varying characteristics of the signals. For example, when the sources of interest are time-varying signals like speech, the number of active sources at each time-frequency (TF) point is dependent on the activities of individual sources. The noise characteristics can also vary over time frames when the noise is not entirely stationary.

A switching filter has recently been proposed to overcome the above limitation. With a switching filter, we cluster time frames of the captured signals at each frequency so that each cluster contains relatively stationary signals composed of a relatively small number of sources and estimate and apply time-invariant filters separately to respective clusters. It can accomplish more accurate estimation than simply applying a time-invariant filter to a whole captured signal. The clustering can be performed based on the same criterion as that for filter estimation, and thus we can accomplish overall optimal estimation. The switching filter was first introduced into Minimum-Variance Distortionless Response (MVDR) beamformer to enable it to handle underdetermined mixtures (i.e. $M > N$) [13]. It was then successfully introduced into WPE [14], IVA, and CIVA [15] to improve their estimation accuracy. WPE, IVA, and CIVA with switching filters are referred to as swWPE, swIVA, and swCIVA hereafter.

Considering the success of switching filters, this paper newly develops IVE and CIVE with switching filters, referred to as swIVE and swCIVE. We can derive swIVE and swCIVE mostly following the derivation of IVE, CIVE, swIVA, and swCIVA, but we need to take special care of spatial covariance matrices (SCMs) of separated noise signals. Unlike the conventional IVE and CIVE, we need to calculate them explicitly for updating WPE and clustering time frames. Experiments will show that swIVE and swCIVE can improve the estimation accuracy of IVE and CIVE and that swIVE and swCIVE can greatly reduce the computational cost of swIVA and swCIVA when $M \gg N$ without degrading the estimation accuracy.

In the remainder of this paper, we develop swIVE and swCIVE in Section 2 and evaluate their performance in Section 3. Concluding remarks are presented in Section 4.

2 PROPOSED METHOD

In this section, we mainly derive the optimization algorithm, swCIVE. The optimization algorithm, swIVE, is a particular case of swCIVE, and can be obtained in a way similar to obtaining swIVA from swCIVA [15].

2.1 Model of observed signals

Suppose that N speech signals are captured by M distant microphones with reverberation and diffuse background noise. We assume $M \geq N$ in this paper. Let $x_{m,t,f}$ be the captured signal at the mth microphone and a TF point (t,f) in the STFT domain for $1 \leq t \leq T$ and $1 \leq f \leq F$, where T and F are the numbers of time frames and frequency bins, and let $(\cdot)^\top$ denote a non-conjugate transpose. Then the captured signal at all the microphones, $\mathbf{x}_{t,f} = [x_{1,t,f}, \ldots, x_{M,t,f}]^\top \in \mathbb{C}^M$, is modeled by

$$\mathbf{x}_{t,f} = \sum_{n=1}^{N} \mathbf{d}_{n,t,f} + \sum_{n=1}^{N} \mathbf{l}_{n,t,f} + \mathbf{v}_{t,f}, \quad \text{and} \quad \mathbf{d}_{n,t,f} = \mathbf{h}_{n,f} s_{n,t,f} \quad \text{for all } n, \tag{1}$$

where $\mathbf{d}_{n,t,f} = [d_{n,1,t,f}, \ldots, d_{n,M,t,f}]^\top \in \mathbb{C}^M$ is the direct signal plus the early reflections of the nth source, $\mathbf{l}_{n,t,f}$ is the source's late reverberation, and $\mathbf{v}_{t,f}$ is the diffuse noise. This paper deals with $\mathbf{d}_{n,t,f}$ for each n as a signal to be estimated, called a desired signal, and models it by a product of a time-invariant acoustic transfer function $\mathbf{h}_{n,f} \in \mathbb{C}^M$ and the nth clean source signal $s_{n,t,f} \in \mathbb{C}$.

2.2 Definition of switching convolutional beamformer (swCBF)

This subsection presents a structure of a convolutional beamformer (CBF) with a switching mechanism, hereafter referred to as a swCBF. This paper uses the same structure for swCBF as that used for swCIVA.

First, let us explain swCBF in its expanded form using Fig. 1. A thorough introduction of the form can be found in our previous paper [15]. It comprises a set of time-invariant Multichannel Linear Prediction (MCLP) filters, a set of time-invariant separation matrices, and a switch. It first dereverberates the captured signal $\mathbf{x}_{t,f}$ using each MCLP filter, and then performs source separation/extraction using each separation matrix. Individual MCLP filters and separation matrices perform different filtering, and the switch selects one output from a pair of an MCLP filter and a separation matrix as swCBF's output $\mathbf{y}_{t,f}$ at each TF point. Time-varying filtering is realized by letting the switch select different inputs at each TF point.

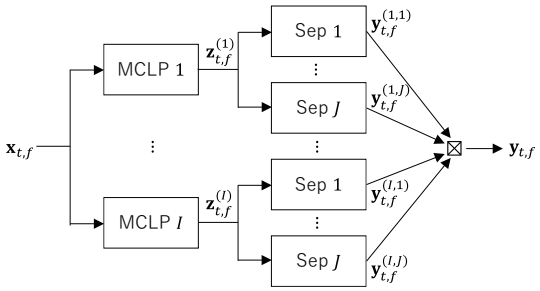


Figure 1. swCBF in its expanded form, receiving a captured signal $\mathbf{x}_{t,f} \in \mathbb{C}^M$ and yielding a separated signal $\mathbf{y}_{t,f} \in \mathbb{C}^M$. It comprises a set of MCLP filters, separation matrices (Sep), and a switch, indicated by a small square with a cross inside. The switch selectively outputs one of its inputs at each TF point. Separation matrices labeled with identical number share the same filter coefficients.

Mathematically, swCBF in the expanded form is defined:

$$\mathbf{z}_{t,f}^{(i)} = \mathbf{x}_{t,f} - (\mathbf{G}_f^{(i)})^{\mathsf{H}} \bar{\mathbf{x}}_{t,f} \quad \text{for} \quad 1 \le i \le I, \tag{2}$$

$$\mathbf{y}_{t,f}^{(i,j)} = (\mathbf{W}_f^{(j)})^{\mathsf{H}} \mathbf{z}_{t,f}^{(i)} \quad \text{for} \quad 1 \le i \le I \quad \text{and} \quad 1 \le j \le J, \tag{3}$$

$$\mathbf{y}_{t,f} = \sum_{i=1}^{I} \sum_{j=1}^{J} \beta_{t,f}^{(i,j)} \mathbf{y}_{t,f}^{(i,j)} \quad \text{where} \quad \beta_{t,f}^{(i,j)} \in \{0,1\} \quad \text{and} \quad \sum_{i=1}^{I} \sum_{j=1}^{J} \beta_{t,f}^{(i,j)} = 1. \tag{4}$$

Eq. (2) is the ith MCLP filter that yields a dereverberated signal $\mathbf{z}_{t,f}^{(i)}$ from the current captured signal $\mathbf{x}_{t,f}$ and the past captured signal sequence $\bar{\mathbf{x}}_{t,f} = [\mathbf{x}_{t-D,f}^{\top}, \ldots, \mathbf{x}_{t-L+1,f}^{\top}]^{\top} \in \mathbb{C}^{M(L-D)}$ using a prediction matrix $\mathbf{G}_{t,f}^{(i)} \in \mathbb{C}^{M(L-D)\times M}$. L and D (≥ 1) are prediction order and delay of MCLP [9]. Eq. (3) applies the jth separation matrix $\mathbf{W}_f^{(j)} \in \mathbb{C}^{M\times M}$ to $\mathbf{z}_{t,f}^{(i)}$ to yield the (i,j)th internal output of swCBF, $\mathbf{y}_{t,f}^{(i,j)}$, containing a set of separated signals. I ($\ll T$) and J ($\ll T$) are the numbers of switching states for MCLP filters and separation matrices. $\beta_{t,f}^{(i,j)}$ is a switching weight that takes a binary value and is sum to one. By definition, $\beta_{t,f}^{(i,j)}$ has the value 1 only for a state, e.g., (i',j'), among all combinations of the switching states, (i,j), at each TF point. Accordingly, Eq. (4) selects $\mathbf{y}_{t,f}^{(i',j')}$ as the swCBF's output $\mathbf{y}_{t,f}$. Note that letting the switching weight have the value 1 at a state (i',j') corresponds to selecting the i'th MCLP filter and the j'th separation matrix in swCBF.

2.3 Probabilistic model for swCIVE

Let $\mathbf{y}_{t,f} = [y_{1,t,f}, \ldots, y_{N,t,f}, \mathbf{u}_{t,f}^{\top}]^{\top} \in \mathbb{C}^M$ be an output of a desired swCBF at a TF point (t,f), where $y_{n,t,f}$ is an estimate of the nth source and $\mathbf{u}_{t,f} \in \mathbb{C}^{M-N}$ is an estimated noise vector. Then, to derive the Maximum Likelihood (ML) objective, we introduce the probabilistic models of $\mathbf{y}_{t,f}$ as

1. $y_{n,t,f}$ for $1 \le n \le N$ and $\mathbf{u}_{\mathbf{v},t,f}$ are mutually independent over all TF points:

$$p(\{\mathbf{y}_{t,f}\}_{t,f}) = \prod_{t=1}^{T} \prod_{f=1}^{F} \left(p(\mathbf{u}_{t,f}) \prod_{n=1}^{N} p(y_{n,t,f}) \right). \tag{5}$$

2. Each $y_{n,t,f}$ is modeled by a time-varying Gaussian with a mean zero and a time-varying variance $\lambda_{n,t,f}$:

$$p(y_{n,t,f}; \lambda_{n,t,f}) = \frac{1}{\pi \lambda_{n,t,f}} \exp\left(-\frac{|y_{n,t,f}|^2}{\lambda_{n,t,f}} \right) \quad \text{for} \quad 1 \le n \le N. \tag{6}$$

3. $\mathbf{u}_{t,f}$ is modeled by a multivariate Gaussian with a mean zero and a covariance matrix $\Omega_{t,f}$, where $\Omega_{t,f}$ is determined based on a state-dependent covariance matrix $\Omega_f^{(j)}$ and the switch $\beta_{t,f}^{(i,j)}$:

$$p(\mathbf{u}_{t,f}; \Omega_{t,f}) = \frac{\pi^{N-M}}{\det \Omega_{t,f}} \exp\left(-\mathbf{u}_{t,f}^{\mathsf{H}} \left(\Omega_{t,f} \right)^{-1} \mathbf{u}_{t,f} \right), \quad \text{where} \quad \Omega_{t,f} = \sum_{i=1}^{I} \sum_{j=1}^{J} \beta_{t,f}^{(i,j)} \Omega_f^{(j)}. \tag{7}$$

Note here that we assumed the mutual independence of $y_{n,t,f}$ between different frequencies in Eq. (5) because it turned out crucial for swIVA and swCIVA [15]. With this setting, however, BSS is known to suffer from the frequency permutation problem. As a practical solution, swIVA and swCIVA used a frequency-independent source model, which is usually adopted by conventional IVA, only for updating the separation matrices [15]. This paper also takes this technique and details it in Section 2.4.2.

Obtaining $\mathbf{y}_{t,f}^{(i,j)} = [y_{1,t,f}^{(i,j)}, \ldots, y_{N,t,f}^{(i,j)}, (\mathbf{u}_{t,f}^{(i,j)})^\top]^\top$ by Eqs. (2) and (3) dependent on $\mathbf{G}_f^{(i)}$ and $\mathbf{W}_f^{(j)}$ and considering that $\beta_{t,f}^{(i,j)}$ takes the value 1 only for a state at each TF point, the log likelihood function of swCIVE is derived:

$$\mathscr{L}(\mathscr{G}, \mathscr{W}, \Lambda, \mathscr{B}) = \sum_{t,f,i,j} \beta_{t,f}^{(i,j)} \mathscr{L}_{t,f}^{(i,j)} \left(\mathbf{G}_f^{(i)}, \mathbf{W}_f^{(j)}, \Lambda_{t,f}, \Omega_f^{(j)} \right), \tag{8}$$

$$\mathscr{L}_{t,f}^{(i,j)} \left(\mathbf{G}_f^{(i)}, \mathbf{W}_f^{(j)}, \Lambda_{t,f}, \Omega_f^{(j)} \right) = 2\log|\det \mathbf{W}_f^{(j)}| - \sum_{n=1}^{N} \left(\frac{|y_{n,t,f}^{(i,j)}|^2}{\lambda_{n,t,f}} + \log \lambda_{n,t,f} \right) - \left(\mathbf{u}_{t,f}^{(i,j)} \right)^{\mathsf{H}} \left(\Omega_f^{(j)} \right)^{-1} \mathbf{u}_{t,f}^{(i,j)} - \log\det \Omega_f^{(j)}, \tag{9}$$

where $\mathscr{G} = \{\mathbf{G}_f^{(i)}\}_{i,f}$, $\mathscr{W} = \{\mathbf{W}_f^{(j)}\}_{j,f}$, $\Lambda = \{\{\Lambda_{t,f}\}_{t,f}, \{\Omega_f^{(j)}\}_{j,f}\}$, $\Lambda_{t,f} = \{\lambda_{n,t,f}\}_{1 \le n \le N}$, and $\mathscr{B} = \{\beta_{t,f}^{(i,j)}\}_{i,j,t,f}$.

2.4 Optimization algorithm swCIVE

We now derive the algorithm, swCIVE, which optimizes a swCBF by the ML objective in Eqs. (8) and (9). Because no closed form solution has been obtained, we use iterative estimation based on a coordinate ascent method [16]. It alternately updates one of $\mathscr{G}$, $\mathscr{W}$, Λ, and $\mathscr{B}$ by fixing the other parameters, and iterates the update until a convergence is obtained. The following describes each update step in the iteration.

2.4.1 $\mathscr{G}$ update

First, we extract the terms related with $\mathbf{G}_f^{(i)}$ from Eqs. (8) and (9) and obtain

$$\mathscr{L}_{\mathbf{G}_f^{(i)}} = -\sum_{j,t} \beta_{t,f}^{(i,j)} \left\| \left(\mathbf{W}_f^{(j)} \right)^{\mathsf{H}} \left(\mathbf{x}_{t,f} - (\mathbf{G}_f^{(i)})^{\mathsf{H}} \bar{\mathbf{x}}_{t,f} \right) \right\|_{\Xi_{t,f}^{(j)}}^2 \quad \text{where} \quad \Xi_{t,f}^{(j)} = \left[\begin{array}{ccc|c} \lambda_{1,t,f}^{-1} & & O & \\ & \ddots & & O \\ O & & \lambda_{N,t,f}^{-1} & \\ \hline & O & & \left(\Omega_f^{(j)} \right)^{-1} \end{array} \right] \tag{10}$$

and $\|\mathbf{x}\|_\Xi^2 = \mathbf{x}^{\mathsf{H}} \Xi \mathbf{x}$. Since the above equation is a simple quadratic form in terms of $\mathbf{G}_f^{(i)}$, we can update it by a closed form when fixing the other parameters. Let $\mathbf{g}_f^{(i)} = \mathrm{vec}(\mathbf{G}_f^{(i)}) \in \mathbb{C}^{M^2(L-D)}$, where $\mathbf{a} = \mathrm{vec}(\mathbf{A})$ is an operation to reshape matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_M]$ to vector $\mathbf{a} = [\mathbf{a}_1^\top, \ldots, \mathbf{a}_M^\top]^\top$. Then $\mathbf{g}_f^{(i)}$ is updated by

$$\mathbf{g}_f^{(i)} \leftarrow (\Psi_f^{(i)})^{-1} \mathrm{vec}(\Phi_f^{(i)}), \tag{11}$$

where $\Psi_f^{(i)} \in \mathbb{C}^{M^2(L-D) \times M^2(L-D)}$ and $\Phi_f^{(i)} \in \mathbb{C}^{M(L-D) \times M}$ are calculated:

$$\Psi_f^{(i)} = \sum_{j=1}^{J} \sum_{n=1}^{N} \left(\mathbf{w}_{n,f}^{(j)} (\mathbf{w}_{n,f}^{(j)})^{\mathsf{H}} \right)^* \otimes \mathbf{R}_{n,f}^{(i,j)} + \sum_{j=1}^{J} \left(\mathbf{W}_{\mathbf{u},f}^{(j)} (\Omega_f^{(j)})^{-1} (\mathbf{W}_{\mathbf{u},f}^{(j)})^{\mathsf{H}} \right)^* \otimes \mathbf{R}_{\mathbf{x},f}^{(i,j)}, \tag{12}$$

$$\Phi_f^{(i)} = \sum_{j=1}^{J} \sum_{n=1}^{N} \mathbf{P}_{n,f}^{(i,j)} \left(\mathbf{w}_{n,f}^{(j)} (\mathbf{w}_{n,f}^{(j)})^{\mathsf{H}} \right) + \sum_{j=1}^{J} \mathbf{P}_{\mathbf{x},f}^{(i,j)} \left(\mathbf{W}_{\mathbf{u},f}^{(j)} (\Omega_f^{(j)})^{-1} (\mathbf{W}_{\mathbf{u},f}^{(j)})^{\mathsf{H}} \right). \tag{13}$$

$(\cdot)^*$ is a complex conjugate, $\otimes$ is a Kronecker product, $\mathbf{w}_{n,f}^{(j)}$ is the nth column of $\mathbf{W}_f^{(j)}$, and $\mathbf{W}_{\mathbf{u},f}^{(j)}$ is the last $M-N$ columns of $\mathbf{W}_f^{(j)}$. $\mathbf{R}_{n,f}^{(i,j)}$ and $\mathbf{R}_{\mathbf{x},f}^{(i,j)} \in \mathbb{C}^{M(L-D) \times M(L-D)}$ and $\mathbf{P}_{n,f}^{(i,j)}$ and $\mathbf{P}_{\mathbf{x},f}^{(i,j)} \in \mathbb{C}^{M(L-D) \times M}$ are spatio-

temporal covariance matrices of the nth source and the captured signal, obtained by

$$\mathbf{R}_{n,f}^{(i,j)} = \sum_{t=1}^{T} \frac{\beta_{t,f}^{(i,j)}}{\lambda_{n,t,f}} \bar{\mathbf{x}}_{t,f} \bar{\mathbf{x}}_{t,f}^{\mathsf{H}}, \quad \mathbf{R}_{\mathbf{x},f}^{(i,j)} = \sum_{t=1}^{T} \beta_{t,f}^{(i,j)} \bar{\mathbf{x}}_{t,f} \bar{\mathbf{x}}_{t,f}^{\mathsf{H}}, \quad \mathbf{P}_{n,f}^{(i,j)} = \sum_{t=1}^{T} \frac{\beta_{t,f}^{(i,j)}}{\lambda_{n,t,f}} \bar{\mathbf{x}}_{t,f} \mathbf{x}_{t,f}^{\mathsf{H}}, \text{ and } \quad \mathbf{P}_{\mathbf{x},f}^{(i,j)} = \sum_{t=1}^{T} \beta_{t,f}^{(i,j)} \bar{\mathbf{x}}_{t,f} \mathbf{x}_{t,f}^{\mathsf{H}}. \tag{14}$$

2.4.2 $\mathscr{W}$ update

As mentioned in Section 2.3, as a practical technique to solve the frequency permutation problem, we adopt a frequency-independent source model only for the update of the separation matrices $\mathscr{W}$. This can be done by replacing the frequency-dependent variance $\lambda_{n,t,f}$ in Eq. (6) to a frequency-independent variance $\lambda_{n,t}$, which can be updated by $\lambda_{n,t} = \frac{1}{F}\sum_{f=1}^{F} \lambda_{n,t,f}$. Then, extracting terms related with $\mathbf{W}_f^{(j)}$ from Eqs. (8) and (9) yields

$$\mathscr{L}_{\mathbf{W}_f^{(j)}} = 2T_f^{(j)} \log|\det \mathbf{W}_f^{(j)}| - \sum_{n=1}^{N} (\mathbf{w}_{n,f}^{(j)})^{\mathsf{H}} \Sigma_{n,f}^{(j)} \mathbf{w}_{n,f}^{(j)} - \mathrm{tr}((\mathbf{W}_{\mathbf{u},f}^{(j)})^{\mathsf{H}} \Sigma_{\mathbf{z},f}^{(j)} \mathbf{W}_{\mathbf{u},f}^{(j)} (\Omega_f^{(j)})^{-1}) - T_f^{(j)} \log\det \Omega_f^{(j)}, \tag{15}$$

$$\Sigma_{n,f}^{(j)} = \sum_{i=1}^{I}\sum_{t=1}^{T} \frac{\beta_{t,f}^{(i,j)}}{\lambda_{n,f}} \mathbf{z}_{t,f}^{(i)} (\mathbf{z}_{t,f}^{(i)})^{\mathsf{H}}, \quad \Sigma_{\mathbf{z},f}^{(j)} = \sum_{i=1}^{I}\sum_{t=1}^{T} \beta_{t,f}^{(i,j)} \mathbf{z}_{t,f}^{(i)} (\mathbf{z}_{t,f}^{(i)})^{\mathsf{H}}, \quad \text{and} \quad T_f^{(j)} = \sum_{i=1}^{I}\sum_{t=1}^{T} \beta_{t,f}^{(i,j)}, \tag{16}$$

where $\mathbf{z}_{t,f}^{(i)}$ is the output of the ith MCLP filter. Because the above objective is in the same form as that of IVE [8], we can apply iterative optimization techniques proposed for it. This paper employs Iterative Projection (IP) [5] and updates each beamformer for separating the nth source for $1 \le n \le N$ (or the nth column of $\mathbf{W}_f^{(j)}$):

$$\mathbf{w}_{n,f}^{(j)} \leftarrow \left((\mathbf{W}_f^{(j)})^{\mathsf{H}} \Sigma_{n,f}^{(j)}\right)^{-1} \mathbf{e}_n, \quad \text{and} \quad \mathbf{w}_{n,f}^{(j)} \leftarrow \mathbf{w}_{n,f}^{(j)} / \left((\mathbf{w}_{n,f}^{(j)})^{\mathsf{H}} \Sigma_{n,f}^{(j)} \mathbf{w}_{n,f}^{(j)}\right)^{1/2}, \tag{17}$$

where $\mathbf{e}_n$ is the nth column of identity matrix $\mathbf{I}_M \in \mathbb{R}^{M\times M}$. Then, $\mathbf{W}_{\mathbf{u},f}^{(j)}$ is updated by obtaining a stationary point of Eq. (15) [7, 8]:

$$\mathbf{W}_{\mathbf{u},f}^{(j)} \leftarrow \begin{bmatrix} -((\mathbf{W}_{\mathbf{s},f}^{(j)})^{\mathsf{H}} \Sigma_{\mathbf{z},f}^{(j)} \mathbf{E}_{\mathbf{s}})^{-1} (\mathbf{W}_{\mathbf{s},f}^{(j)})^{\mathsf{H}} \Sigma_{\mathbf{z},f}^{(j)} \mathbf{E}_{\mathbf{u}} \\ \mathbf{I}_{M-N} \end{bmatrix}, \tag{18}$$

where $\mathbf{W}_{\mathbf{s},f}^{(j)}$ is the first N columns of $\mathbf{W}_f^{(j)}$, and $\mathbf{E}_{\mathbf{s}}$ and $\mathbf{E}_{\mathbf{u}}$ are the first N and the remaining columns of $\mathbf{I}_M$.

2.4.3 Λ and $\mathscr{B}$ updates

After updating $y_{n,t,f}$ and $\mathbf{u}_{t,f}^{(i,j)}$, the variance $\lambda_{n,t,f}$ and the covariance matrix $\Omega_f^{(j)}$ are updated based on the likelihood function in Eq. (8):

$$\lambda_{n,t,f} \leftarrow |y_{n,t,f}|^2 + \varepsilon_0 \quad \text{and} \quad \Omega_f^{(j)} \leftarrow \frac{1}{T_f^{(j)}} \sum_{i,t} \beta_{t,f}^{(i,j)} \mathbf{u}_{t,f}^{(i,j)} \left(\mathbf{u}_{t,f}^{(i,j)}\right)^{\mathsf{H}}, \tag{19}$$

where ε_0 is a small positive scalar to avoid zero division during the optimization.

Then, according to Eq. (8), the switching weight can be updated:

$$\beta_{t,f}^{(i,j)} \leftarrow \begin{cases} 1 & \text{if } \{i,j\} = \underset{\{i',j'\}}{\arg\max}\, \mathscr{L}_{t,f}^{(i,j)}(\mathbf{G}_f^{(i')}, \mathbf{W}_f^{(j')}, \Lambda_{t,f}, \Omega_f^{(j')}) \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

2.5 Major differences between swCIVE and swCIVA that affect computational cost

As shown in the above derived algorithm, swCIVE (and swIVE) can separate source signals from the noise vector without further separating the noise into $M-N$ noise signals, unlike swCIVA (and swIVA). This change has significant impact on the computational cost reduction of swCIVA (and swIVA) as detailed in the following.

1. At each update of the spatio-temporal covariance matrices in Eq. (14), swCIVA calculates $\mathbf{R}_{n,f}^{(i,j)}$ and $\mathbf{P}_{n,f}^{(i,j)}$ $M-N$ times for noise, i.e., for $N+1 \leq n \leq M$, while swCIVE calculates $\mathbf{R}_{\mathbf{x},f}^{(i,j)}$ and $\mathbf{P}_{\mathbf{x},f}^{(i,j)}$ only one time. This has substantial impact on the computational cost reduction for $\mathscr{G}$ update when $M > N+1$.

2. At each update of separation matrices, swCIVA calculates Eq. (17) $M-N$ times also for noise, i.e., for $N+1 \leq n \leq M$, while swCIVE calculates Eq. (18) only one time. Because Eq. (18) is much less computationally demanding than Eq. (17), this has substantial impact on the computational cost reduction for $\mathscr{W}$ update when $M > N$.

3. Unlike swCIVA, swCIVE needs to update SCMs of separated noise signals, $\{\Omega_f^{(j)}\}_{j,f}$, in Eq. (19), and their inversions and determinants in Eq. (9). This has a certain negative impact on the computational cost reduction when $M > N+1$, but the impact is relatively minor in comparison with the first two items.

In summary, the above first and second items have the largest impacts, respectively, on swCIVE and swIVE. Thus, we consider that swCIVE and swIVE have the advantage on the computational cost reduction over swCIVA and swIVA, respectively, when $M > N+1$ and when $M > N$.

3 EXPERIMENTS

We experimentally evaluated the proposed methods, swIVE and swCIVE. Our goal here is to show that swIVE and swCIVE outperform conventional IVE and CIVE in terms of estimation accuracy, and that they can largely reduce the computational cost of swIVA and swCIVA without degrading the estimation accuracy.

3.1 Dataset, analysis conditions, and evaluation metric

In this experiment, we used a dataset, TIMIT-ConvMix [15], composed of simulated noisy reverberant mixtures. Each mixture in the dataset is composed of 3 speakers and diffuse noise recorded by 8 microphones. The dataset contains 40 mixtures with the average length being 12.6 s. To generate each mixture, we mixed 3 clean speech utterance sequences randomly extracted from the TIMIT corpus [17] and five different additive noise signals extracted from the CHiME-3 dataset [18] after individually reverberating them using room impulse responses (RIRs) extracted from JR1 in the RWCP dataset [19]. Its RT60 was 0.6 s. We set the power ratio of each reverberant speech signal to the sum of the additive noise signals to 10 dB.

We set the frame length and the shift to 32 and 8 ms and used a Hann window for analysis. The sampling frequency was 16 kHz. For WPE, the prediction delay and order were set at $D = 2$ and $L = 10$. In the iterative optimization, we adopted different iteration numbers for updating IVA/IVE and for updating WPE to compensate their different convergence property [15]. We updated IVA and IVE 50 times in total, and updated WPE once in five updates of IVA/IVE. To solve the scale ambiguity of BSS, we applied projection back [20] for all the methods. To solve the inter-state permutation problem [15], which occurs in BSS with switching filter, we adopted the blind single-state initialization technique [15]. We used the first and remaining 25 iterations, respectively, for the single-state initialization and for optimization with switching filter.

In this evaluation, we adopted Signal-to-Distortion Ratio (SDR) as the evaluation metric of the estimation accuracy [21], which is widely used in source separation research. We used the MUSEVAL V4 toolkit [22] with its bss_eval_images configuration. As reference signals, we used clean utterance sequences that were convolved with the initial 32 ms part of the RIRs used for generating the corresponding mixtures.

3.2 Comparison of IVE, swIVE, CIVE, and swCIVE in terms of SDR improvement

Figure 2 shows the SDR improvement obtained by applying IVE, swIVE, CIVE, and swCIVE to TIMIT-ConvMix using various number of microphones (#Mics) and switching states (#Switching_states), i.e., $J = 2$ and 3 for swIVE, and $(I,J) = (2,2)$ and $(3,3)$ for swCIVE.

First, comparing IVE and swIVE in Fig. 2 (a), the estimation accuracy was consistently improved by introducing the switching filter and increasing the number of switching states under all #Mics conditions.

Next, comparing CIVE and swCIVE in Fig. 2 (b), the estimation accuracy was improved by introducing the switching filter and increasing the number of switching states for #Mics< 6. In contrast, the improvement was marginal for #Mics≥ 6, and became even negative with $(I,J) = (3,3)$ for #Mics$= 8$. The accuracy degradation might be caused by overfitting of the switching filters because the number of filter coefficients in swCIVE becomes very high as #Mics and #Switching_states are increased.

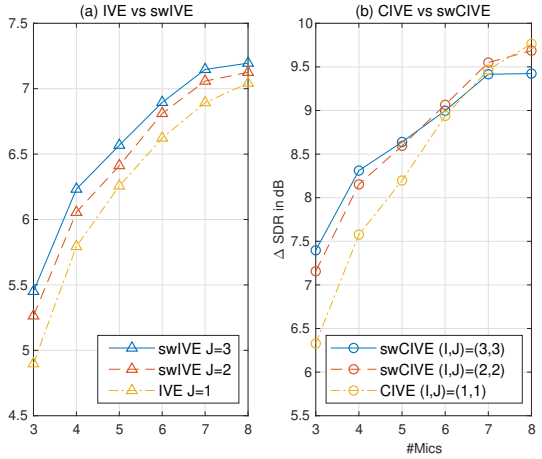


Figure 2. Comparison of IVE, swIVE, CIVE, and swCIVE in terms of SDR improvements using various number of microphones (#Mics) and various number of switching states (#Switching_states).

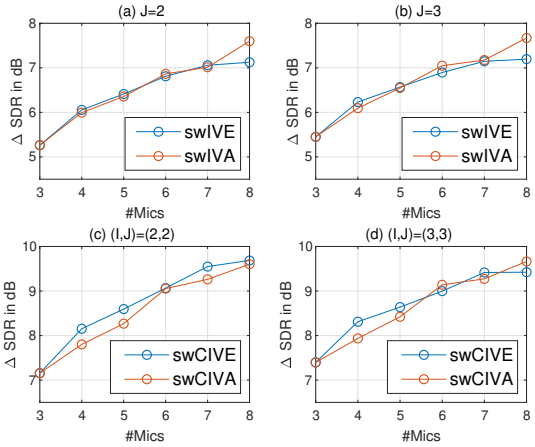


Figure 3. Comparison of swIVA, swIVE, swCIVA, and swCIVE in terms of SDR improvement using various #Mics and #Switching_states. Source selection post-processing was also applied to swIVA and swCIVA.

Table 1. Computing times in s required for processing a mixture with length of 13.1 s. Boldface letters indicate the shortest times required under each #Switching_states and #Mics condition.

	IVA	IVE	swIVA	swIVE	swIVA	swIVE		CIVA	CIVE	swCIVA	swCIVE	swCIVA	swCIVE
J	1		2		3		(I,J)	(1,1)		(2,2)		(3,3)	
#Mics							#Mics						
3	4.6	-	6.6	-	8.4	-	3	10.2	-	24.2	-	42.3	-
4	7.2	**6.8**	10.5	**10.3**	13.5	**13.4**	4	17.8	**17.0**	**41.5**	41.9	**73.7**	75.2
5	11.5	**10.0**	16.6	**14.6**	21.9	**18.9**	5	30.7	**28.9**	81.0	**70.1**	154.2	**110.1**
6	15.9	**13.8**	22.9	**22.3**	30.2	**26.7**	6	48.1	**44.4**	116.4	**109.1**	208.2	**157.4**
7	21.7	**14.5**	30.5	**22.0**	41.1	**28.7**	7	79.2	**53.1**	213.0	**131.5**	429.8	**231.5**
8	30.0	**18.6**	44.0	**27.7**	63.6	**36.5**	8	163.7	**101.1**	415.7	**252.5**	797.2	**468.1**

In summary, swIVE and swCIVE successfully improved SDRs of the separated signals in comparison with IVE and CIVE except for cases using large #Mics and #Switching_states with swCIVE.

3.3 Comparison of swIVA, swCIVA, swIVE, and swCIVE

3.3.1 In terms of SDR improvement

Figure 3 shows the SDR improvement obtained by applying swIVA, swIVE, swCIVA, and swCIVE when using various #Mics and #Switching_states, i.e., (a) $J=2$ and (b) $J=3$ for swIVA and swIVE, and (c) $(I,J)=(2,2)$ and (d) $(I,J)=(3,3)$ for swCIVA and swCIVE.

For fair comparison we applied source selection post-processing after swIVA and swCIVA separated as many signals as microphones at each #Mics conditions. We selected the signals that maximize the likelihood function defined for the optimization of IVE (or CIVE) as the target sources.

As shown in the figure, swIVE achieved the almost same performance as swIVA for all #Mics conditions except for #Mics$=8$. Also, swCIVE was comparable to or even slightly better than swCIVA.

3.3.2 In terms of computational cost

Table 1 shows computing times required for all the methods to process a mixture with length of 13.1 s. The processing code was implemented by Python version 3.7 and ran on a linux computer with a single thread.

First, swIVE and swCIVE almost always reduced the computing times in comparison with swIVA and swCIVA under each #Mics and #Switching_states condition. The reduction becomes more significant when

#Mics and #Switching_states increased. When using swCIVE with #Mics= 4, the computing times were increased from using swCIVA, but the increase was rather minor.

The results shown in Fig. 3 and Table 1 indicate that swIVE and swCIVE reduced the computational cost of swIVA and swCIVA without degrading the estimation accuracy. In particular, the proposed methods were effective to mitigate the rapid increase in the computational cost by swIVA and swCIVA occurring as the increase of #Mics and #Switching_states.

4 CONCLUDING REMARKS

This paper newly developed swIVE and swCIVE by introducing the switching filter to two BSS algorithms, IVE and CIVE. The switching filter is a recently proposed versatile technique for improving the accuracy of linear filtering that estimates audio sources from a captured signals with time-varying characteristics. The technique has been shown to work effectively with conventional BSS techniques, IVA and CIVA. Experiments showed that the proposed methods successfully improved the accuracy of source extraction in comparison with IVE and CIVE in terms of SDR improvement. Also, the proposed methods greatly reduced the computational cost required for swIVA and swCIVA without degrading the accuracy of the estimation. These results clearly demonstrate the superiority of the proposed methods over the conventional methods.

REFERENCES

[1] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: John Wiley & Sons, 2001.

[3] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Speech, and Audio Processing*, vol. 15, no. 1, pp. 70–79, 2006.

[4] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in *Independent Component Analysis and Blind Signal Separation*, 2006, pp. 601–608.

[5] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," in *LVA/ICA*. Springer, 2010, pp. 165–172.

[6] Z. Koldovský and P. Tichavský, "Gradient algorithms for complex non-gaussian independent component/vector extraction, question of convergence," *IEEE Transactions on Signal Processing*, vol. 67, no. 4, pp. 1050–1064, 2018.

[7] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *Proc. IEEE WASPAA*, 2019.

[8] R. Ikeshita, T. Nakatani, and S. Araki, "Block coordinate descent algorithms for auxiliary-function-based independent vector extraction," *IEEE Trans. Signal Processing*, vol. 69, pp. 3252–3267, 2021.

[9] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[10] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Computationally efficient and versatile framework for joint optimization of blind speech separation and dereverberation," in *Proc. Interspeech*, 2020, pp. 91–95.

[11] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Blind and neural network-guided convolutional beamformer for joint denoising, dereverberation, and source separation," in *IEEE ICASSP*, 2021, pp. 6129–6133.

[12] R. Ikeshita and T. Nakatani, "Independent vector extraction for fast joint blind source separation and dereverberation," *IEEE Signal Processing Letters*, vol. 28, pp. 972–976, 2021.

[13] K. Yamaoka, N. Ono, S. Makino, and T. Yamada, "Time-frequency-bin-wise switching of minimum variance distortionless response beamformer for underdetermined situations," in *Proc. IEEE ICASSP*, 2019, pp. 7908–7912.

[14] R. Ikeshita, N. Kamo, and T. Nakatani, "Blind signal dereverberation based on mixture of weighted prediction error models," *IEEE Signal Processing Letters*, vol. 28, pp. 399–403, 2021.

[15] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, N. Kamo, and S. Araki, "Switching independent vector analysis and its extension to blind and spatially guided convolutional beamforming algorithms," *IEEE/ACM Trans. ASLP*, vol. 30, pp. 1032–1047, March 2022.

[16] S. J. Wright, "Coordinate descent algorithms," *Mathematical Programming*, vol. 151, no. 1, pp. 3–34, 2015.

[17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *TIMIT acoustic-phonetic continuous speech corpus*. Philadelphia: Linguistic Data Consortium, 1993.

[18] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE ASRU-2015*, 2015, pp. 504–511.

[19] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. 2nd International Conference on Language Resources and Evaluation*, 2000.

[20] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1–4, p. 1–24, Oct. 2001.

[21] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[22] "Museval," https://github.com/sigsep/sigsep-mus-eval.

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0776

Adaptive Convolutional Beamforming for Joint Dereverberation, Interferer and Noise Reduction

Henri GODE, Simon DOCLO

Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany,
{henri.gode, simon.doclo}@uni-oldenburg.de

ABSTRACT

Background noise, interfering sources and reverberation may degrade speech quality and speech intelligibility in hearing aid applications. In this contribution, we consider binaural hearing aids and compare several multi-microphone algorithms for joint dereverberation, interferer and noise reduction and preservation of binaural cues based on convolutional beamforming. Convolutional beamforming is a multi-frame filtering approach which can be interpreted as a combination of multi-channel linear prediction (focusing on dereverberation) and linearly constrained minimum power beamforming (focusing on interferer and noise reduction). We propose to adaptively optimize the convolutional beamformer coefficients by incorporating an exponential window into a sparsity-promoting ℓ_p-norm cost function, enabling to track a moving target speaker. In this contribution, we compare different model-based methods to estimate the required parameters for the convolutional beamformer, in particular the relative transfer functions and covariance matrices of the coherent sources. For a static and a moving target speaker the performance of the considered estimation methods is evaluated based on objective measures for dereverberation, interferer and noise reduction.

Keywords: Dereverberation, Noise & Interferer Reduction, Adaptive Convolutional Beamforming, Non-Stationary Scenario

1 INTRODUCTION

In many hands-free communication systems such as hearing aids, mobile phones and smart speakers, interfering sounds, ambient noise and reverberation may degrade the speech quality and intelligibility of the recorded microphone signals [3]. To enhance speech quality and intelligibility, many multi-microphone speech enhancement methods aiming at noise and interferer reduction and dereverberation have been proposed in the last decades [5, 21]. For many of these methods, both non-adaptive versions with time-invariant parameters as well as adaptive versions with time-varying parameters exist. In this contribution we specifically consider binaural hearing aids, where besides enhancing the desired speech signal it is often also desired to preserve the binaural cues, which provide spatial awareness of the acoustic scene for the listener [13].

A commonly used multi-microphone noise reduction method is the minimum power distortionless response (MPDR) beamformer [20], which aims at minimizing the output power while leaving the desired speech component undistorted. The linearly constrained minimum power (LCMP) beamformer generalizes the MPDR beamformer by providing the possibility to impose multiple linear constraints, e.g., to perform controlled reduction of the interfering sources [8]. Often the constraints are formulated in terms of the relative transfer function (RTF) vectors of the target speaker and the interfering sources [14].

To achieve dereverberation, the weighted prediction error (WPE) method [17] and its generalization using sparse priors [11] are commonly employed. WPE uses a convolutional filter, applied to a number of past frames in the short-time Fourier transform (STFT) domain, to estimate and subtract the late reverberation component from the reference microphone signal. Since the WPE cost function does not have an analytic solution, it has been proposed to use iterative alternating optimization schemes. In [10, 22] adaptive versions of the WPE algo-

rithm have been proposed, e.g., by incorporating an exponential window into the cost function and incorporating an additional constraint to prevent overestimation of the late reverberation component [10].

Aiming at joint dereverberation and noise reduction, it has been proposed to perform multiple-input multiple-output (MIMO)-WPE before MPDR beamforming in a cascade system [4]. By unifying the optimization of the convolutional WPE filter and the MPDR beamformer, the so-called weighted power minimization distortionless response (WPD) beamformer [16] and its generalization using sparse priors [7] were shown to outperform cascade systems. The unified WPD beamformer can be optimized similarly to the WPE filter with an additional distortionless constraint using the relative transfer functions (RTFs) of the target speaker. In [15] two adaptive versions of the WPD beamformer have been proposed.

Aiming at joint dereverberation, reduction of interfering sources and noise and preservation of the binaural cues of all sources, the weighted binaural linearly constrained minimum power (wBLCMP) beamformer in [1] generalizes the WPD beamformer by unifying the optimization of the convolutional WPE filter and the LCMP beamformer. To enable tracking of spatial changes in the acoustic scene, e.g., a moving target speaker, an adaptive version was derived in [6] by incorporating an exponential window into the cost function. A generalization of the wBLCMP beamformer was proposed in [6, 7]. This generalization incorporates sparse priors to explicit control the sparsity of the STFT coefficients using an ℓ_p-norm cost function. In this contribution, we evaluate different versions of the wBLCMP beamformer in terms of objective speech enhancement performance measures. More in particular, we compare different initializations (single-channel & multi-channel), different sparsity levels and a non-adaptive version with an adaptive version using different adaptation speeds.

2 SIGNAL MODEL

We consider J acoustic sources captured by a binaural hearing aid setup with $M/2$ microphones on each hearing aid in a noisy and reverberant acoustic environment (with $J < M$). Without loss of generality, the first source ($j = 1$) is considered to be the target speaker and the remaining $J-1$ sources are considered to be interfering sources. The STFT coefficients of the microphone signals at time frame t are denoted as

$$\mathbf{y}_t = \begin{bmatrix} y_{1,t} & \cdots & y_{M,t} \end{bmatrix}^{\mathrm{T}} \in \mathbb{C}^{M\times 1}, \tag{1}$$

with $(\cdot)^{\mathrm{T}}$ denoting the transpose operator. In (1) the frequency index has been omitted since it is assumed that each frequency subband is independent and hence can be processed individually. Similarly as in [1, 7, 11, 15, 16], the multi-channel signal $\mathbf{y}_t$ in (1) is modeled as the sum of each source signal $s_{j,t}$ convolved with its (possibly time-varying) multi-channel convolutive transfer function (CTF) matrix $\mathbf{A}_{j,t} = \begin{bmatrix} \mathbf{a}_{j,t,0} & \cdots & \mathbf{a}_{j,t,L_a-1} \end{bmatrix} \in \mathbb{C}^{M\times L_a}$ plus background noise $\mathbf{n}_t \in \mathbb{C}^{M\times 1}$, where L_a denotes the number of taps of the CTFs. By splitting the CTFs into early reflections and late reverberation using the integer parameter τ, the reverberant signal for the j-th source can be decomposed into its direct component $\mathbf{d}_{j,t} \in \mathbb{C}^{M\times 1}$ (including early reflections) and its late reverberation component $\mathbf{r}_{j,t} \in \mathbb{C}^{M\times 1}$, i.e.

$$\mathbf{y}_t = \sum_{j=1}^{J}\sum_{l=0}^{L_a-1} \mathbf{a}_{j,t,l}s_{j,t-l} \;+\; \mathbf{n}_t = \underbrace{\sum_{j=1}^{J}\sum_{l=0}^{\tau-1} \mathbf{a}_{j,t,l}s_{j,t-l}}_{:=\mathbf{d}_{j,t}} \;+\; \underbrace{\sum_{j=1}^{J}\sum_{l=\tau}^{L_a-1} \mathbf{a}_{j,t,l}s_{j,t-l}}_{:=\mathbf{r}_{j,t}} \;+\; \mathbf{n}_t. \tag{2}$$

We assume that the direct component for the j-th source $\mathbf{d}_{j,t}$ can be approximated using the multiplicative transfer function (MTF) vector $\mathbf{v}_{j,t} \in \mathbb{C}^{M\times 1}$ as [2]

$$\mathbf{d}_{j,t} \approx \mathbf{v}_{j,t}s_{j,t} = \tilde{\mathbf{v}}_{j,m,t}d_{j,m,t} \quad \text{with} \quad \tilde{\mathbf{v}}_{j,m,t} = \mathbf{v}_{j,t}/v_{j,m,t} \in \mathbb{C}^{M\times 1} \quad \text{and} \quad m \in \{1,\ldots,M\}, \tag{3}$$

where $d_{j,m,t}$ denotes the direct component of the j-th source in the reference microphone m at time frame t and the vector $\tilde{\mathbf{v}}_{j,m,t}$ denotes the (possibly time-varying) RTF vector for the j-th source, where $v_{j,m,t}$ is the m-th entry of $\mathbf{v}_{j,t}$.

3 SPARSE WBLCMP BEAMFORMER

To obtain an estimate of the direct component of the target speaker $d_{1,\nu,t}$ in the left and right reference microphone, denoted by $\nu \in \{L,R\}$, it has been proposed in [1, 6, 7, 15, 16] to apply a convolutional filter $\bar{\mathbf{h}}_{\nu,t} \in \mathbb{C}^{M(L_h-\tau+1)\times 1}$ to the stacked noisy STFT vector $\bar{\mathbf{y}}_t$, i.e.

$$\hat{d}_{1,\nu,t} = \bar{\mathbf{h}}_{\nu,t}^{\mathrm{H}} \bar{\mathbf{y}}_t \quad \text{with} \quad \bar{\mathbf{y}}_t = \begin{bmatrix} \mathbf{y}_t^{\mathrm{T}} & | & \mathbf{y}_{t-\tau}^{\mathrm{T}} & \cdots & \mathbf{y}_{t-L_h+1}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \in \mathbb{C}^{M(L_h-\tau+1)\times 1}, \tag{4}$$

where $(\cdot)^{\mathrm{H}}$ denotes the conjugate transpose operator and L_h denotes the filter length. It should be noted that the stacked noisy STFT vector $\bar{\mathbf{y}}_t$ only includes a subset of the L_h most recent frames, i.e. it includes the current frame but excludes the preceding $\tau-1$ frames, aiming at preserving the early reflections.

3.1 Conventional non-adaptive wBLCMP

Assuming that all CTFs and MTFs and the convolutional filter $\bar{\mathbf{h}}_{\nu,t}$ do not change over time, i.e. $\bar{\mathbf{h}}_{\nu,t} = \bar{\mathbf{h}}_{\nu}$ for all time frames $t \in \{1,\dots,T\}$, a non-adaptive version of the wBLCMP beamformer aiming at joint dereverberation, noise and interferer reduction has been derived in [1]. Assuming that the direct component of the target speaker follows a zero mean complex circular Gaussian distribution with a time-varying variance $\lambda_n = |d_{1,\nu,n}|^2$, the convolutional filter in (4) is computed by minimizing the negative log-likelihood function subject to a linear constraint for each source using their RTFs defined in (3), i.e.

$$\underset{\bar{\mathbf{h}}_{\nu}}{\operatorname{argmin}} \sum_{n=1}^{T} \ln \lambda_n + \frac{|\hat{d}_{1,\nu,n}|^2}{\lambda_n} \quad \text{s.t.} \quad \bar{\mathbf{h}}_{\nu}^{\mathrm{H}} \bar{\mathbf{v}}_{j,\nu} = \beta_j \quad \forall j \in \{1,\dots,J\} \quad \text{with} \quad \bar{\mathbf{v}}_{j,\nu} = \begin{bmatrix} \tilde{\mathbf{v}}_{j,\nu}^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}, \tag{5}$$

where $\mathbf{0}$ denotes a vector containing $M(L_h-\tau)$ zeros and β_j denotes a scaling factor for the direct component of the j-th source. The scaling factor β_1 is usually set to 1, corresponding to a distortionless constraint for the target speaker, whereas all other scaling factors are usually chosen to be close to 0, aiming at suppressing the interfering sources.

3.2 Generalized sparse wBLCMP

A generalization of the cost function in (5) was proposed in [7], aiming at explicitly taking into account that the STFT coefficients of the direct component of the target speaker are sparser than the STFT coefficients of the noisy reverberant mixture recorded by the microphones. In [7], the convolutional filter in (4) is optimized using an ℓ_p-norm cost function instead of (5), i.e.

$$\underset{\bar{\mathbf{h}}_{\nu}}{\operatorname{argmin}} \sum_{n=1}^{T} |\hat{d}_{1,\nu,n}|^p = \sum_{n=1}^{T} |\bar{\mathbf{h}}_{\nu}^{\mathrm{H}} \bar{\mathbf{y}}_n|^p \tag{6}$$

where $p \in (0,2]$ denotes the so-called shape parameter. This parameter determines the sparsity of the cost function, where small values of p promote sparsity. It should be noted that for $0 < p < 1$ this cost function is non-convex.

3.3 Adaptive sparse wBLCMP

To deal with time-varying acoustic scenarios, e.g., moving sources, an adaptive version of the wBLCMP beamformer was derived in [6] by incorporating an exponential window into the cost function in (6). The resulting minimization problem for each time frame t is given by

$$\boxed{\underset{\bar{\mathbf{h}}_{\nu,t}}{\operatorname{argmin}} \sum_{n=1}^{t} \gamma^{t-n} |\hat{d}_{1,\nu,n}|^p = \sum_{n=1}^{t} \gamma^{t-n} |\bar{\mathbf{h}}_{\nu,t}^{\mathrm{H}} \bar{\mathbf{y}}_n|^p \quad \text{s.t.} \quad \bar{\mathbf{h}}_{\nu,t}^{\mathrm{H}} \bar{\mathbf{v}}_{j,\nu,t} = \beta_j \quad \forall j \in \{1,\dots,J\},} \tag{7}$$

where the smoothing parameter $\gamma \in (0,1]$ allows adaptation to time-varying CTFs and MTFs. Note that the cost function in (7) reduces to the cost function in (6) for $\gamma = 1$ and $t = T$. Therefore, the following derivations for the adaptive version in (7) also hold for the non-adaptive version in (6).

3.4 Filter Optimization

In [6, 7, 10] an iteratively reweighted least squares (IRLS) procedure has been presented to solve the minimization problem in (7). The basic idea is to replace the non-convex ℓ_p-norm minimization problem with a series of convex ℓ_2-norm minimization subproblems, which have an analytic solution.

3.4.1 Constrained ℓ_2-Norm Subproblem Minimization

In each frame, the non-convex cost function in (7) is replaced with a convex weighted ℓ_2-norm cost function, i.e.

$$\underset{\bar{\mathbf{h}}_{v,t}}{\operatorname{argmin}} \sum_{n=1}^{t} \gamma^{t-n} w_n \left|\hat{d}_{1,v,n}\right|^2 = \sum_{n=1}^{T} \gamma^{t-n} w_n \left|\bar{\mathbf{h}}_{v,t}^{\mathrm{H}} \bar{\mathbf{y}}_n\right|^2, \tag{8}$$

where the weights w_n are real-valued and positive. The filter minimizing (8) subject to the linear constraints in (7) is equal to

$$\boxed{\bar{\mathbf{h}}_{v,t} = \bar{\mathbf{R}}_{y,t}^{-1} \bar{\mathbf{C}}_t \left(\bar{\mathbf{C}}_t^{\mathrm{H}} \bar{\mathbf{R}}_{y,t}^{-1} \bar{\mathbf{C}}_t\right)^{-1} \mathbf{B} \bar{\mathbf{C}}_t^{\mathrm{H}} \mathbf{e}_v} \quad \text{with} \quad \bar{\mathbf{R}}_{y,t} = \sum_{n=1}^{t} \gamma^{t-n} w_n \bar{\mathbf{y}}_n \bar{\mathbf{y}}_n^{\mathrm{H}} \quad \text{and} \quad \bar{\mathbf{C}}_t = \begin{bmatrix} \bar{\mathbf{v}}_{1,v,t} & \cdots & \bar{\mathbf{v}}_{J,v,t} \end{bmatrix}, \tag{9}$$

where $\bar{\mathbf{R}}_{y,t}$ denotes the weighted noisy spatio-temporal covariance matrix (STCM) of the stacked microphone signals, $\bar{\mathbf{C}}_t$ denotes the constraint matrix containing the RTF vectors for all sources, $\mathbf{B} = \operatorname{diag}\left(\begin{bmatrix} \beta_1 & \cdots & \beta_J \end{bmatrix}^{\mathrm{T}}\right)$ contains the scaling factors for all sources, and $\mathbf{e}_v$ is a selection vector with its entry corresponding to the left or right reference microphone equal to 1 and all other entries equal to 0. Assuming that the weights w_n of past frames $n \in \{1, \ldots, t-1\}$ are well estimated, the weighted noisy STCM $\bar{\mathbf{R}}_{y,t}$ in (9) in frame t can be effectively computed by a recursive update, i.e. $\bar{\mathbf{R}}_{y,t} = \gamma \bar{\mathbf{R}}_{y,t-1} + w_t \bar{\mathbf{y}}_t \bar{\mathbf{y}}_t^{\mathrm{H}}$. Since only the inverse of the weighted noisy STCM is required in (9), it is more efficient to use an update formula for $\bar{\mathbf{R}}_{y,t}^{-1}$ based on the Woodbury matrix identity, i.e.

$$\bar{\mathbf{R}}_{y,t}^{-1} = \frac{1}{\gamma} \left(\bar{\mathbf{R}}_{y,t-1}^{-1} - \frac{w_t \bar{\mathbf{R}}_{y,t-1}^{-1} \bar{\mathbf{y}}_t \bar{\mathbf{y}}_t^{\mathrm{H}} \bar{\mathbf{R}}_{y,t-1}^{-1}}{\gamma + w_t \bar{\mathbf{y}}_t^{\mathrm{H}} \bar{\mathbf{R}}_{y,t-1}^{-1} \bar{\mathbf{y}}_t} \right) \tag{10}$$

3.4.2 Weight Estimation / Update

Similarly as in [7, 11], in each frame the weight w_t in (10) is estimated as

$$w_t = \left(\sum_v \left|\hat{d}_{1,v,t}\right|^2 \right)^{\frac{p}{2}-1} = \left(\sum_v \left|\bar{\mathbf{h}}_{v,t}^{\mathrm{H}} \bar{\mathbf{y}}_t\right|^2 \right)^{\frac{p}{2}-1}, \tag{11}$$

such that (8) is a first-order approximation of (7). It should be noted that the shape parameter p only affects the weight update in (11). Setting $p = 0$ corresponds to the conventional wBLCMP beamformer described in Section 3.1 using the time-varying Gaussian model in (5).

Weight Initialization In each iteration of the IRLS procedure, first the convolutional filter in (9) is estimated, based on which the weights in (11) are updated. These weight updates modify the estimation of the convolutional filter in the next iteration. However, the update equation (11) depends on the estimate of the direct component of the target speaker, which is obviously not available in the first iteration. A single-channel (SC) initialization and a multi-channel (MC) initialization using the noisy reverberant microphone signals are discussed in [7] i.e.

$$w_{t,1}^{(\mathrm{SC})} = \frac{1}{|y_{v,t}|^{2-p}} \quad \text{and} \quad w_{t,1}^{(\mathrm{MC})} = \frac{M}{\|\mathbf{y}_t\|_2^{2-p}}. \tag{12}$$

3.5 RTF Estimation

The wBLCMP beamformer in (9) requires estimates of the RTFs for each source, which can be obtained using the covariance whitening method [14]. It has been shown in [15] that performing RTF estimation on multi-channel dereverberated signals $\mathbf{z}_t$, obtained by a MIMO-WPE preprocessing stage, is beneficial, since the MTF-based model in (3) assumes short transfer functions for the direct component. The RTF vector of the j-th source can then be estimated based on the generalized eigenvalue decomposition of the dereverberated noisy covariance matrix $\mathbf{R}_{j,t}$ of that source and the dereverberated covariance matrix $\mathbf{R}_{v,j,t}$ of all other sources and the background noise.

4 EXPERIMENTAL RESULTS

In this section we evaluate the performance of different versions of the wBLCMP beamformer. In Section 4.1 we consider a stationary acoustic scenario with a static target speaker, whereas in Section 4.2 we consider a non-stationary scenario with a switching target speaker.

4.1 Static Target Speaker

In this section, we compare the performance of the non-adaptive version of the conventional wBLCMP beamformer with the sparse wBLCMP beamformer for an acoustic scene featuring a static target speaker and diffuse noise. More in particular, we evaluate the influence of the shape parameter p and different weight initializations.

4.1.1 Dataset, Evaluation Metrics and Analysis Conditions

We used the simulated data of the development set of the REVERB challenge [12] with sampling frequency $f_s = 16\,\text{kHz}$. The dataset simulates a circular microphone array with 8 channels in six different reverberation conditions resulting from two speaker-to-microphone distances (50 cm and 200 cm) and three different reverberation times ($T_{60} \in \{0.3\,\text{s}, 0.6\,\text{s}, 0.7\,\text{s}\}$). After convolving clean utterances with one of the six room impulse responses, stationary diffuse background noise was added at a signal-to-noise ratio of 20 dB. As objective performance measures we computed perceptual evaluation of speech quality (PESQ) and frequency-weighted segmental signal-to-noise ratio (FWSSNR) scores [19, 9], where we used the clean speech signal s_t as the reference signal. The algorithm used an STFT-framework with 32 ms sqrt-Hann windows and 25% overlap. The prediction delay and filter length were set to $\tau = 4$ frames (corresponding to 32 ms) and $L_h = 12$ frames (corresponding to 96 ms), respectively. The RTF vector $\tilde{\mathbf{v}}_m$ was estimated blindly using the covariance whitening (CW) method [14], assuming that noise-only frames are present in the first 225 ms and the last 75 ms were used to estimate the noise covariance matrix $\mathbf{R}_n$.

4.1.2 Results

Fig. 1 shows the average PESQ and FWSSNR improvement vs. the number of iterations of the ℓ_p-norm WPD beamformer for different shape parameters p and weight initializations (see Section 3.4.2). First, the results show that for all considered parameter choices the speech quality is improved in terms of PESQ and FWSSNR compared to the noisy reference microphone signal. Second, the results after $I = 10$ iterations show that for both initializations a shape parameter of $p = 0.5$ outperforms the conventional method with $p = 0$, which stronger promotes sparsity, and $p = 1$, which promotes sparsity less, in terms of PESQ and FWSSNR improvement, except for the FWSSNR improvement of the conventional method for the multi-channel initialization. Third, it can be observed that the multi-channel initialization consistently outperforms the single-channel initialization in terms of convergence speed and for the conventional method ($p = 0$) also in terms of performance after $I = 10$ iterations.

4.2 Switching Target Speaker

In this section, we compare the performance of the adaptive version of the wBLCMP beamformer (Sec. 3.3) with the non-adaptive version (Sec. 3.1) using different shape parameters p for a spatially non-stationary acoustic scenario where the target speaker suddenly switches position.

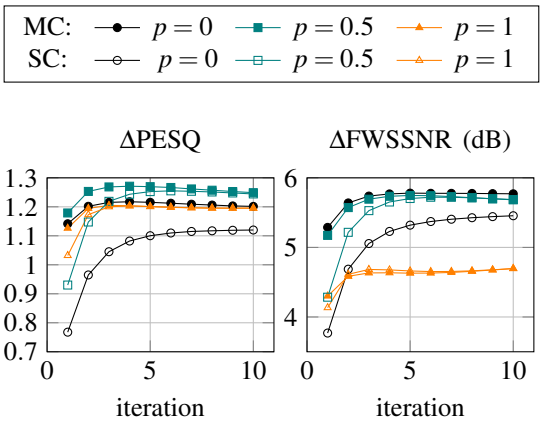


Figure 1. Average PESQ and FWSSNR improvement vs. number of iterations for different shape parameters p. Filled markers correspond to multi-channel (MC) weight initialization, while empty markers correspond to single-channel (SC) weight initialization.

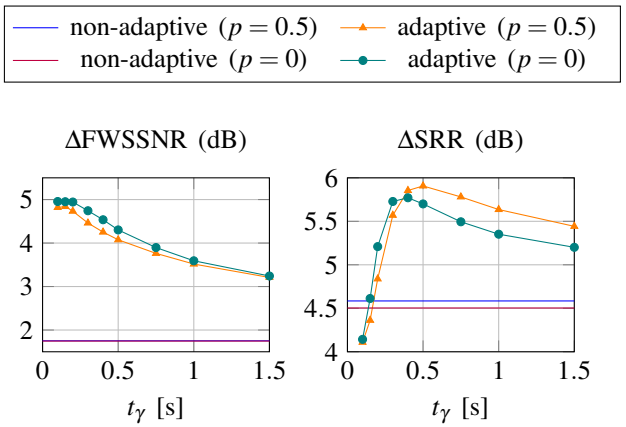


Figure 2. Average FWSSNR and SRR improvement vs. time constant t_γ for different values of the shape parameter p. Note that the non-adaptive method obviously does not have a time constant.

4.2.1 Acoustic scenario

We considered 2 behind-the-ear hearing aids with 2 microphones each, mounted on a dummy head located approximately in the center of an acoustic laboratory ($7\,\mathrm{m} \times 6\,\mathrm{m} \times 2.7\,\mathrm{m}$) with a reverberation time $T_{60} \approx 510\,\mathrm{ms}$. The acoustic scenario consists of one target speaker (which suddenly switches position), one static interfering speaker and background noise. The target and interfering speech components at the microphones were generated by convolving clean speech signals with room impulse responses measured from loudspeakers at about 2 m from the dummy head at a sampling frequency of $f_s = 16\,\mathrm{kHz}$. The target speaker at position 1 (0°, front of dummy head) is a male speaker which is active in the interval $[2\,\mathrm{s}, 20.4\,\mathrm{s}]$, whereas the target speaker at position 2 (90°, right of dummy head) is a female speaker which is active in the interval $[20.4\,\mathrm{s}, 39\,\mathrm{s}]$. The interfering speaker is a male speaker which is located at $-120°$ and is active in the interval $[1\,\mathrm{s}, 39\,\mathrm{s}]$. Quasi-diffuse babble noise, which is constantly active, was generated by playing back cafeteria noise using 4 loudspeakers facing the corners of the laboratory. The noisy mixture is constructed at a broadband signal-to-noise ratio (SNR) of 0 dB and a broadband signal-to-interferer ratio (SIR) of 0 dB for both target positions. Note that there is a noise-only period in the interval $[0\,\mathrm{s}, 1\,\mathrm{s}]$ and a noise-plus-interferer period in the interval $[1\,\mathrm{s}, 2\,\mathrm{s}]$.

4.2.2 Algorithm Settings

The wBLCMP beamformer used an STFT-framework with 32 ms sqrt-Hann windows and 50% overlap. We compared the performance of two shape parameters ($p = 0$ and $p = 0.5$). The filter length L_h and the prediction delay τ in (4) were set to 16 frames (corresponding to 256 ms) and 3 frames (corresponding to 48 ms), respectively. The scaling factors of the target speaker and the interfering speaker in (5) were set to $\beta_1 = 0\,\mathrm{dB}$ and $\beta_2 = -20\,\mathrm{dB}$, respectively. For the adaptive versions, different time constants were evaluated between $t_\gamma = [100\,\mathrm{ms}, 1500\,\mathrm{ms}]$, where the corresponding smoothing parameter was computed as $\gamma = e^{-t_s/t_\gamma}$. The noise-plus-interferer covariance matrix $\mathbf{R}_{v,2}$ and the RTF vector of the interfering source $\tilde{\mathbf{v}}_{2,v}$ were fixed after the first 2 s, whereas the covariance matrix and the RTF vector of the target speaker were adaptively tracked.

4.2.3 Objective Speech Enhancement Measures

As objective performance measures we used the FWSSNR [9] and the signal-to-reverberation ratio (SRR) [18], averaged across the left and right output signal. As reference signal for FWSSNR and SRR we used the direct target speech component including early reflections (first 50 ms of the room impulse responses) at the left and right reference microphones.

4.2.4 Results

Fig. 2 compares the FWSSNR and SRR improvements (difference between scores for input and output signals) for different time constants t_γ of the adaptive version and the non-adaptive version of the wBLCMP beamformer using two different shape parameters ($p = 0$ and $p = 0.5$). It can be clearly observed that for the considered switching-target scenario the adaptive version of the wBLCMP beamformer outperforms the non-adaptive version in terms of both performance measures for almost all time constants. The best SRR improvement is obtained using a time constant of roughly $t_\gamma = 450\,\text{ms}$, whereas the FWSSNR improvement is generally higher for shorter time constants. The shape parameter $p = 0.5$ yields better SRR improvements especially for larger time constants, whereas the shape parameter $p = 0$, corresponding to the conventional cost function in (5), yields slightly better FWSSNR improvements.

5 CONCLUSIONS

In this contribution we evaluated different versions of the wBLCMP beamformer in terms of weight initialization, sparsity promoting shape parameter, and adaptation speed. It was shown that multi-channel weight initialization yields faster and better convergence than single-channel weight initialization. Furthermore, using a shape parameter $p = 0.5$ slightly improves performance in terms of PESQ and SRR compared to the conventional method with $p = 0$. Finally, for a non-stationary scenario with a switching target speaker the results clearly show that the adaptive version of the wBLCMP beamformer outperforms the non-adaptive version.

ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 390895286 – EXC 2177/1.

REFERENCES

[1] A. Aroudi, M. Delcroix, T. Nakatani, K. Kinoshita, S. Araki, and S. Doclo. Cognitive-Driven Convolutional Beamforming Using EEG-Based Auditory Attention Decoding. In *Proc. IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, Espoo, Finland, Sept. 2020.

[2] Y. Avargel and I. Cohen. On Multiplicative Transfer Function Approximation in the Short-Time Fourier Transform Domain. *IEEE Signal Processing Letters*, 14(5):337–340, May 2007.

[3] R. Beutelmann and T. Brand. Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 120(1):331–342, July 2006.

[4] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, S. Araki, T. Hori, and T. Nakatani. Strategies for distant speech recognition in reverberant environments. *EURASIP Journal on Advances in Signal Processing*, 2015(1):1–15, July 2015.

[5] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm. Multichannel Signal Enhancement Algorithms for Assisted Listening Devices: Exploiting spatial diversity using multiple microphones. *IEEE Signal Processing Magazine*, 32(2):18–30, Mar. 2015.

[6] H. Gode and S. Doclo. Adaptive dereverberation, noise and interferer reduction using sparse weighted linearly constrained minimum power beamforming. In *Proc. European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, Aug. 2022.

[7] H. Gode, M. Tammen, and S. Doclo. Joint multi-channel dereverberation and noise reduction using a unified convolutional beamformer with sparse priors. In *Proc. ITG Conference on Speech Communication*, pages 144–148, Kiel, Germany, Sep. 2021.

[8] N. Gößling, D. Marquardt, I. Merks, T. Zhang, and S. Doclo. Optimal binaural LCMV beamforming in complex acoustic scenarios: Theoretical and practical insights. In *Proc. International Workshop on Acoustic Signal Enhancement*, pages 381–385, Tokyo, Japan, 2018.

[9] Y. Hu and P. C. Loizou. Evaluation of Objective Quality Measures for Speech Enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229–238, Jan. 2008.

[10] A. Jukić, T. van Waterschoot, and S. Doclo. Adaptive Speech Dereverberation Using Constrained Sparse Multichannel Linear Prediction. *IEEE Signal Processing Letters*, 24(1):101–105, Jan. 2017.

[11] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo. Group sparsity for MIMO speech dereverberation. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–5, New Paltz NY, USA, Oct. 2015.

[12] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj. The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4, New Paltz NY, USA, Oct. 2013.

[13] M. Lavandier and J. F. Culling. Speech segregation in rooms: Monaural, binaural, and interacting effects of reverberation on target and interferer. *The Journal of the Acoustical Society of America*, 123(4):2237–2248, Apr. 2008.

[14] S. Markovich, S. Gannot, and I. Cohen. Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment With Multiple Interfering Speech Signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1071–1086, Aug. 2009.

[15] T. Nakatani and K. Kinoshita. Simultaneous Denoising and Dereverberation for Low-Latency Applications Using Frame-by-Frame Online Unified Convolutional Beamformer. In *Proc. Interspeech*, pages 111–115, Graz, Austria, Sept. 2019.

[16] T. Nakatani and K. Kinoshita. A Unified Convolutional Beamformer for Simultaneous Denoising and Dereverberation. *IEEE Signal Processing Letters*, 26(6):903–907, June 2019.

[17] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang. Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction. *IEEE Trans. on Audio, Speech, and Language Processing*, 18(7):1717–1731, Sept. 2010.

[18] P. A. Naylor, N. D. Gaubitch, and E. A. P. Habets. Signal-Based Performance Evaluation of Dereverberation Algorithms. *Journal of Electrical and Computer Engineering*, 2010:e127513, Jan. 2010.

[19] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 749–752, Salt Lake City, UT, USA, May 2001.

[20] B. D. Van Veen and K. M. Buckley. Beamforming: a versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24, Apr. 1988.

[21] E. Vincent, T. Virtanen, and S. Gannot. *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, Oct. 2018.

[22] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi. Adaptive dereverberation of speech signals with speaker-position change detection. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3733–3736, Taipei, Taiwan, Apr. 2009.

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0831

Independent low-rank matrix analysis based on the Sinkhorn divergence source model for blind source separation

Jianyu Wang[(1)], Shanzheng Guan[(1)], Jingdong Chen[(1)], and Jacob Benesty[(2)]

[(1)]Center of Intelligent Acoustics and Immersive Communication, Northwestern Polytechnical University, Xi'an, 710072, China
alexwang96@mail.nwpu.edu.cn, gshanzheng@mail.nwpu.edu.cn, jingdongchen@ieee.org

[(2)]INRS-EMT, University of Quebec, 800 de la Gauchetiere Ouest, Suite 6900, Montreal, QC H5A 1K6, Canada
jacob.benesty@inrs.ca

ABSTRACT

The so-called independent low-rank matrix analysis (ILRMA) has demonstrated a great potential for dealing with the problem of determined blind source separation (BSS) for audio and speech signals. This method assumes that the spectra from different frequency bands are independent and the spectral coefficients in any frequency band are Gaussian distributed. The Itakura-Saito divergence is then employed to estimate the source model related parameters. In reality, however, the spectral coefficients from different frequency bands may be dependent, which is not considered in the existing ILRMA algorithm. This paper presents an improved version of ILRMA, which considers the dependency between the spectral coefficients from different frequency bands. The Sinkhorn divergence is then exploited to optimize the source model parameters. As a result of using the cross-band information, the BSS performance is improved. But the number of parameters to be estimated also increases significantly, and so is the computational complexity. To reduce the algorithm complexity, we apply the Kronecker product to decompose the modeling matrix into the product of a number of matrices of much smaller dimensionality. An efficient algorithm is then developed to implement the Sinkhorn divergence based BSS algorithm and the complexity is reduced by an order of magnitude.

Keywords: Independent low-rank matrix analysis (ILRMA), Blind source separation (BSS), Sinkhorn distance, Kronecker product.

1 INTRODUCTION

Multichannel blind source separation (BSS) refers to the problem of estimating source signals from their mixtures observed by an array of sensors without using any prior information about the mixing system [1]. For audio and speech applications [2], the problem can be divided into two cases: underdetermined and determined. The former refers to the case where the number of sensors in the array is less than the number of sources. In this case, the problem cannot be solved without additional information or constraints [3, 4]. The latter refers to the scenario where the number of sensors is greater than or equal to the number of sources. In this case, separation can be achieved by identifying the demixing system from only the observation signals. This work focus on the latter case, i.e., the determined BSS for audio and speech signals.

In audio and speech applications, the signal observed at every sensor is a mixture of all the source signals convolved with the corresponding acoustic channel impulse responses. As the acoustic channel impulse responses are usually very long (it is not uncommon to have a few thousands of points), this convolutive mixing process make it challenging and difficult to achieve source separation directly in the time domain from the perspectives of accuracy, robustness, and complexity. A widely adopted approach to circumventing this issue is to transform the time-domain signals into the time-frequency domain using the short-time fourier transform (STFT), thereby converting the convolutive mixing problem into one of instantaneous mixing. Consequently, majority of efforts in audio and speech BSS have been focused in the STFT domain. Many methods and algorithms have been developed in this domain over the last few decades and the representative ones include the so-called independent

component analysis (ICA) [8] and independent vector analysis (IVA) [9, 10]. In comparison, IVA based methods are more appropriate than ICA for dealing with audio BSS in the STFT domain as it dramatically mitigates the permutation problem. While they have demonstrated reasonably good performance, the classical IVA algorithms do not take advantage of the structural information in the source spectra, which are useful to improve BSS performance. To exploit such information, Daichi *et al.* proposed an independent-low-rank-matrix-analysis (ILRMA) method [5], which utilizes nonnegative matrix factorization (NMF) to decompose the given spectrogram as the product between basis and temporal activation matrices. By assuming that the spectral components from different frequency bands are independent and the spectral coefficients in any frequency band are Gaussian distributed, this method employs the Kullback-Leibler (KL) or Itakura-Saito (IS) divergence as the cost function to estimate the parameters of the NMF-based source model.

However, the spectral components of the same source from different frequency bands may be correlated as demonstrated in the literature of noise reduction [6, 7], which is not considered in the ILRMA algorithm. This paper presents an improved version of ILRMA, which takes advantage of the cross-band dependency of spectra to improve BSS performance. We adopt the Sinkhorn divergence [11], [13], [14] as the cost function to optimize the parameters of the NMF-based source model, resulting in a Sinkhorn divergence based ILRMA (SDILRMA) algorithm. Since the cross-band information is used, SDILRMA is able to improve the BSS performance. But the number of parameters to be estimated also increases significantly, and so is the computational complexity. To reduce the number of parameters and the algorithm complexity, we subsequently apply the Kronecker product tool [15, 16] to decompose the modeling matrix into the product of a number of matrices of much smaller dimensionality, leading to a simplified SDILRMA, which is computationally more efficient than its original counterpart and is able produce better performance than ILRMA.

2 SIGNAL MODEL AND PROBLEM FORMULATION

Suppose that there are N sources in the sound field and we use a microphone array consisting of M sensors to pick up the signals. The observation signal at the mth microphone and time index j is then

$$x_m(j) = \sum_{n=1}^{N} a_{nm}(j) * s_n(j), \tag{1}$$

where $s_n(j)$ denotes the nth source signal and $a_{nm}(j)$ is the acoustic impulse response from the nth source to the mth sensor.

Transforming both sides of (1) into the short-time Fourier transform (STFT) domain and rearranging the results into a vector form gives

$$\begin{aligned} \mathbf{x}_{f,t} &= \sum_{n=1}^{N} \mathbf{a}_{n,f} S_{n,f,t} \\ &= \sum_{n=1}^{N} \mathbf{x}_{n,f,t}, \end{aligned} \tag{2}$$

where $S_{n,f,t}$ is the STFT of $s_n(j)$, $\mathbf{x}_{f,t} \triangleq \left[X_{1,f,t}, \cdots, X_{M,f,t}\right]^T \in \mathbb{C}^M$ with $X_{m,f,t}$ being the STFT of $x_m(j)$, $\mathbf{a}_{n,f} \triangleq \left[A_{n,1,f}, \cdots, A_{n,M,f}\right]^T$ with $A_{n,m,f}$ denoting the acoustic transfer function, the superscript ^{T} denotes the transpose operator, f and t denote, respectively, the frequency and frame indices, and $\mathbf{x}_{n,f,t} \triangleq \mathbf{a}_{n,f} S_{n,f,t}$, whose elements are often called the source images.

The signal model in (2) can be rearranged into a more compact form as

$$\mathbf{x}_{f,t} = \mathbf{A}_f \mathbf{s}_{f,t}, \tag{3}$$

where $\mathbf{A}_f \triangleq \left[\mathbf{a}_{1,f}, \cdots, \mathbf{a}_{n,f}\right] \in \mathbb{C}^{M \times N}$ is called the mixing matrix, and $\mathbf{s}_{f,t} \triangleq \left[S_{1,f,t}, \cdots, S_{N,f,t}\right]^T$ is a vector consisting of the N source signals. Now, the problem of BSS becomes one of identifying a demixing matrix such that

$$\mathbf{y}_{f,t} = \mathbf{D}_f \mathbf{x}_{f,t}, \tag{4}$$

where $\mathbf{D}_f = \left[\mathbf{d}_{1,f}, \cdots, \mathbf{d}_{N,f}\right] \in \mathbb{C}^{N\times M}$ denotes the demixing matrix, and $\mathbf{y}_{f,t}$ is an estimate of $\mathbf{s}_{f,t}$ (up to a scale and permutation). Note that if the mixing matrix $\mathbf{A}_f = \left[\mathbf{a}_{1,f}, \ldots, \mathbf{a}_{n,f}\right] \in \mathbb{C}^{M\times N}$ is not singular as assumed in such methods as ILRMA, the demixing matrix should be the inverse of the mixing matrix $\mathbf{A}_f$.

To achieve this identification, some source model has to be assumed. The so-called spherically invariant random processing (SIRP) model has been widely used in BSS for speech signals [19]. With this model, the multivariate probability density function can be derived from the corresponding univariate probability density function and the correlation matrices [18, 20]. As a particular case of SIRP, the local Gaussian model has gained much attention, in which the source spectrum in every time-frequency (TF) bin is modeled as a time-varying complex Gaussian distribution [17] and the spectral components from different frequency bins and time frames are assumed to be mutually independent, and as a result, $s_{n,f,t}$ follows a zero-mean complex Gaussian distribution with a time-varying variance $\lambda_{n,f,t}$, i.e.,

$$s_{n,f,t} \sim \mathcal{N}_{\mathbb{C}}\left(0, \lambda_{n,f,t}\right). \tag{5}$$

The critical parameter of this source model is the time-varying variance $\lambda_{n,f,t}$, which needs to be estimated. One way to achieve such estimation is through NMF, in which the variance matrix of every source is modeled as a low-rank approximation of the product of a basis matrix and an activation matrix. Given $\lambda_{n,f,t}$, the variance matrix is defined as

$$\boldsymbol{\lambda}_n \triangleq \begin{bmatrix} \lambda_{n,1,1} & \ldots & \lambda_{n,1,T} \\ \vdots & \ddots & \vdots \\ \lambda_{n,F,1} & \ldots & \lambda_{n,F,T} \end{bmatrix}, \tag{6}$$

which consists of the time-varying variance for all the time frames (the total number of frames is denoted as T) and frequencies bins (the number of frequency bins is denoted as F). The the low-rank approximation is then expressed as

$$\boldsymbol{\lambda}_n \approx \mathbf{W}_n\mathbf{H}_n, \tag{7}$$

where

$$\mathbf{W}_n = \begin{bmatrix} w_{n,1,1} & \ldots & w_{n,1,K} \\ \vdots & \ddots & \vdots \\ w_{n,F,1} & \ldots & w_{n,F,K} \end{bmatrix}, \tag{8}$$

$$\mathbf{H}_n = \begin{bmatrix} h_{n,1,1} & \ldots & h_{n,1,T} \\ \vdots & \ddots & \vdots \\ h_{n,K,1} & \ldots & h_{n,K,T} \end{bmatrix}, \tag{9}$$

are, respectively, the basis and activation matrices, and K denotes the number of basis vectors. With this approximation, the estimation of the time-varying variances, i.e., $\lambda_{n,f,t}$, for all the time frames and frequency bins is converted to a problem of estimating the basis and activation matrices, which will be discussed in the next section.

From (2) and (5), one can check that $\mathbf{x}_{n,f,t}$ follows a multivariate complex Gaussian distribution, i,e,.

$$\mathbf{x}_{n,f,t} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \quad \lambda_{n,f,t}\mathbf{R}_{n,f}\right), \tag{10}$$

where $\mathbf{0}$ is column vector with all its elements being 0, $\mathbf{R}_{n,f} \triangleq E\left[\mathbf{x}_{n,f,t}\mathbf{x}_{n,f,t}^H\right]$ is the spatial covariance matrix for the nth source. If one approximates this matrix as $\mathbf{R}_{n,f} = \mathbf{a}_{n,f}\mathbf{a}_{n,f}^H$, the model degenerates to a rank-1 spatial model. Given $\mathbf{R}_{n,f}$, one can be check that the observation signal vector $\mathbf{x}_{f,t}$ follows the following distribution:

$$\mathbf{x}_{f,t} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \quad \sum_{n=1}^{N}\lambda_{n,f,t}\mathbf{R}_{n,f}\right). \tag{11}$$

3 SINKHORN DIVERGENCE BASED MODEL PARAMETER ESTIMATION

Generally, the NMF based source model adopts the IS divergence as the cost function for optimization. For the ILRMA algorithm, the cost function, which is denoted as $\mathcal{L}_{\text{ILRMA}}$, is the sum of the logarithmic conditional probability $p\left(\mathbf{X}_{f,t}|\lambda_{n,f,t},\mathbf{D}_f\right)$, i.e.,

$$\begin{aligned}
\mathcal{L}_{\text{ILRMA}} &= \sum_{f=1}^{F}\sum_{t=1}^{T}\log\left[p\left(\mathbf{X}_{f,t}\middle|\lambda_{n,f,t},\mathbf{D}_f\right)\right] \\
&= \sum_{f=1}^{F}\sum_{t=1}^{T}\log\mathcal{N}_{\mathbb{C}}\left(\mathbf{X}_{f,t}\middle|\mathbf{0},\sum_{n=1}^{N}\lambda_{n,f,t}\mathbf{a}_{n,f}\mathbf{a}_{n,f}^{H}\right) \\
&= -\sum_{f=1}^{F}\sum_{t=1}^{T}\text{Tr}\left[\mathbf{y}_{f,t}^{H}\mathbf{D}_f^{-H}\left(\mathbf{D}_f^{H}\mathbf{\Lambda}_{f,t}^{-1}\mathbf{D}_f\right)\mathbf{D}_f^{-1}\mathbf{y}_{f,t}\right]+T\sum_{f=1}^{F}\log\left|\mathbf{D}_f\mathbf{D}_f^{H}\right|-\sum_{f=1}^{F}\sum_{t=1}^{T}\log\left|\mathbf{\Lambda}_{f,t}\right|+\text{Cst} \\
&= -\sum_{f=1}^{F}\sum_{t=1}^{T}\left[\sum_{n=1}^{N}\frac{\left|y_{n,f,t}\right|^2}{\sum_{k=1}^{K}w_{n,f,k}h_{n,k,t}}+\sum_{n=1}^{N}\log\left(\sum_{k=1}^{K}w_{n,f,k}h_{n,k,t}\right)\right]+2T\sum_{f=1}^{F}\log\left|\mathbf{D}_f\right|+\text{Cst},
\end{aligned} \tag{12}$$

where $\mathbf{\Lambda}_{f,t}=\text{Diag}\left(\lambda_{1,f,t},\ \ldots,\ \lambda_{N,f,t}\right)$ is a diagonal matrix. Note that the first term on the right-hand side of the last line in (12) denotes the source model, which can also be viewed as the IS divergence between the low-rank approximated spectra and the estimated source spectra for every source, and the second term denotes the spatial model.

It is seen from (12) that the spectra from different frequency bins are treated independently. In practice, the spectral components of the same source from different frequency bins may be correlated [6, 7]. In what follows, we introduce the Sinkhorn divergence based source model to replace the first term on the right-hand side of the last line in (12) so the cross-band information is used to estimate the model parameters. Specifically, the Sinkhorn divergence is expressed as

$$D_{\text{S}}\left(\mathbf{Y}_n\cdot\mathbf{Y}_n^{*}\ |\ \boldsymbol{\lambda}_n\right)=\sum_{t=1}^{T}\min_{\mathbf{P}_t}\ \langle\mathbf{P}_t,\mathbf{C}\rangle-\frac{1}{\mu}H\left(\mathbf{P}_t\right)\quad\text{s. t.}\quad\mathbf{P}_t\mathbf{1}=\mathbf{y}_{n,t}\cdot\mathbf{y}_{n,t}^{*},\quad\mathbf{P}_t\mathbf{1}^{T}=\boldsymbol{\lambda}_{n,t}, \tag{13}$$

where $\langle\rangle$ denotes the inner product between two matrices, $\cdot$ denotes the Hadamard Product (element-wise Multiplication), $\mathbf{P}_t\in U\left(\mathbf{y}_{n,t}\cdot\mathbf{y}_{n,t}^{*},\boldsymbol{\lambda}_{n,t}\right)$ denotes the transport matrix with $[\mathbf{P}_t]_{ij}$ describing the frequency component migrates from the ith frequency bin of $\mathbf{y}_{n,t}\cdot\mathbf{y}_{n,t}^{*}$ to the jth subband of $\boldsymbol{\lambda}_{n,t}$, $\mathbf{1}\in\mathbb{R}^{F}$ is the all-one vector, $U\left(\mathbf{y}_{n,t}\cdot\mathbf{y}_{n,t}^{*},\boldsymbol{\lambda}_{n,t}\right):=\left\{\mathbf{P}_t\in\mathbb{R}_{+}^{F\times F}\ |\ \mathbf{P}_t\mathbf{1}=\mathbf{y}_{n,t}\cdot\mathbf{y}_{n,t}^{*},\ \mathbf{P}_t^{T}\mathbf{1}=\boldsymbol{\lambda}_{n,t}\right\}$ denotes a transport polytope, which contains all paths from the estimated source $\mathbf{y}_{n,t}\cdot\mathbf{y}_{n,t}^{*}$ to the target parameter $\boldsymbol{\lambda}_n$, $\mathbf{C}\in\mathbb{R}^{F\times F}$ represents the cost of transporting one unit of the source vector to the target vector, and $H(\mathbf{P}_t)=-\sum_{i,j}[\mathbf{P}_t]_{ij}\log[\mathbf{P}_t]_{ij}$ denotes the entropic regularization term, which enables efficient approximation of the gradient of the Sinkhorn divergence.

Using the Lagrange multiplier method, one can express (13) as

$$D_{\text{S}}^{\mu,\gamma}\left(\mathbf{Y}_n\cdot\mathbf{Y}_n^{*}\ \middle|\ \boldsymbol{\lambda}_n\right)=\sum_{t=1}^{T}\left[\min_{\mathbf{P}_t}\ \langle\mathbf{P}_t,\ \mathbf{C}\rangle-\frac{1}{\mu}H(\mathbf{P}_t)+\gamma D_{\text{KL}}\left(\mathbf{P}_t\mathbf{1}\middle|\mathbf{y}_{n,t}\cdot\mathbf{y}_{n,t}^{*}\right)+\gamma D_{\text{KL}}\left(\mathbf{P}_t\mathbf{1}^{T}\middle|\boldsymbol{\lambda}_{n,t}\right)\right], \tag{14}$$

where $\boldsymbol{\lambda}_{n,t}=\sum_{k=1}^{K}\mathbf{w}_{n,k}h_{n,k,t}$, and $D_{\text{KL}}\left(x|y\right)=x\log\frac{x}{y}-x+y$. Note that only a single Lagrange multiplier is used in (14) to reduce the number of parameters.

The transport matrix $\mathbf{P}_t$ should satisfy $\mathbf{P}_t=\text{diag}(\mathbf{u})\ \mathbf{G}\ \text{diag}(\mathbf{v})$ when optimizing the cost function in (14), where $\mathbf{u}=\left(\frac{\mathbf{y}_{n,t}\cdot\mathbf{y}_{n,t}^{*}}{\mathbf{P}_t\mathbf{1}}\right)^{\gamma\mu}$, $\mathbf{v}=\left(\frac{\boldsymbol{\lambda}_{n,t}}{\mathbf{P}_t\mathbf{1}^{T}}\right)^{\gamma\mu}$ (note that here the fraction between two vectors denotes the element wise division), and $\mathbf{G}=\exp(-\mu\mathbf{C}-1)$. The optimal transport matrix $\mathbf{P}_t$ is estimated by a Sinkhorn-like iterative algorithm.

For the basis matrix $\mathbf{W}_n$ and the activation matrix $\mathbf{H}_n$, we construct an auxiliary function as

$$A\left(\mathbf{W}_n,\mathbf{W}_n^{\star}\right)=\sum_{t=1}^{T}\sum_{k_1,\ldots,k_F}\prod_{f}\alpha_{f,k_f}D_{\text{S}}^{\mu,\gamma}\left(\mathbf{y}_{n,t}\cdot\mathbf{y}_{n,t}^{*}\middle|\frac{\sum_{k=1}^{K}\mathbf{w}_{n,k}h_{n,k,t}}{\boldsymbol{\alpha}}\right), \tag{15}$$

$$A\left(\mathbf{H}_n, \mathbf{H}_n^{\star}\right) = \sum_{t=1}^{T} \sum_{k_1,\ldots,k_F} \prod_{f} \beta_{f,k_f} D_{\mathrm{S}}^{\mu,\gamma}\left(\mathbf{y}_{n,t} \cdot \mathbf{y}_{n,t}^{*} \middle| \frac{\sum_{k=1}^{K} \mathbf{w}_{n,k} h_{n,k,t}}{\boldsymbol{\beta}}\right) \tag{16}$$

where $\mathbf{H}_n^{\star}$ denotes an auxiliary matrix constructed from $\mathbf{H}$, $\alpha_{f,k_f} = \frac{w^{\star}_{n,f,k_f} h_{n,k_f,t}}{\sum_{k_f} w_{n,f,k_f} h^{\star}_{n,k_f,t}}$, and $\beta_{f,k_f} = \frac{w_{n,f,k_f} h^{\star}_{n,k_f,t}}{\sum_{k_f} w_{n,f,k_f} h^{\star}_{n,k_f,t}}$. Through evaluating the partial derivatives $\frac{\partial A(\mathbf{W}_n, \mathbf{W}_n^{\star})}{\partial w_{n,f,k}}$ and $\frac{\partial A(\mathbf{H}_n, \mathbf{H}_n^{\star})}{\partial h_{n,k,t}}$, we can obtain the algorithm to estimate the elements of the basis and activation matrices, i.e.,

$$w_{n,f,k} \leftarrow w_{n,f,k} \sqrt{\frac{\sum_t \left[\mathbf{P}_t \mathbf{1}\right]_f h_{n,k,t} \left(\sum_{k'} w_{n,f,k'} h_{n,k',t}\right)^{-2}}{\sum_t \left[\mathbf{P}_t \mathbf{1}\right]_f \left(\sum_{k'} w_{n,f,k'} h_{n,k',t}\right)^{-1}}}, \tag{17}$$

$$h_{n,k,t} \leftarrow h_{n,f,k} \sqrt{\frac{\sum_f \left[\mathbf{P}_t \mathbf{1}\right]_f w_{n,f,k} \left(\sum_{k'} w_{n,f,k'} h_{n,k',t}\right)^{-2}}{\sum_f \left[\mathbf{P}_t \mathbf{1}\right]_f \left(\sum_{k'} w_{n,f,k'} h_{n,k',t}\right)^{-1}}}. \tag{18}$$

The model parameters are optimized in a similar manner as ILRMA [5]. Note, however, computation of the transport matrix $\mathbf{P}_t$ in every frame for the nth source requires large memory and is computationally expensive. In the next section, we apply the Kronecker product tool to decompose the transport matrix $\mathbf{P}_t$ into a product of a number of matrices of much smaller dimensionality.

4 MODEL PARAMETER ESTIMATION BASED ON KRONECKER PRODUCT DECOMPOSITION

Property 1. *(sum of Kronecker product)[15]: Let two matrices be $\mathbf{A} \in \mathbb{R}^{m \times m}$ and $\mathbf{B} \in \mathbb{R}^{n \times n}$, their Kronecker sum can be expressed as*

$$\mathbf{A} \oplus \mathbf{B} = \mathbf{A} \otimes \mathbf{I}_n + \mathbf{I}_m \otimes \mathbf{B}, \tag{19}$$

where $\mathbf{I}_m$ and $\mathbf{I}_n$ are identity matrices of size $m \times m$ and $n \times n$, respectively, and $\otimes$ denotes the Kronecker product.

Since the above Kronecker product decomposition is based on two all-one matrices, we name it the all-one Kronecker product.

Let us decompose the cost matrix $\mathbf{C}$ as

$$\mathbf{C} = \oplus_{q=1}^{Q} \mathbf{C}_q = \mathbf{C}_1 \otimes \mathbf{C}_2 \otimes \cdots \otimes \mathbf{C}_Q, \tag{20}$$

where $\mathbf{C}_1 \in \mathbb{R}^{f_1 \times f_1}, \ldots, \mathbf{C}_Q \in \mathbb{R}^{f_Q \times f_Q}$, $F = f_1 \times \cdots \times f_Q$. The intermediate variable matrix $\mathbf{G}$ can then be written as

$$\mathbf{G} = \exp\left(-\mu \oplus_{q=1}^{Q} \mathbf{C}_q - 1\right) = e^{-1} \otimes_{q=1}^{Q} \exp\left(-\mu \mathbf{C}_q\right). \tag{21}$$

The product $\mathbf{P}_t \mathbf{1}$ in (17) and (18) can be calculated in another way:

$$\mathbf{P}_t \mathbf{1} = \mathrm{diag}\left(\mathbf{u}\right) \mathbf{G} \mathrm{diag}\left(\mathbf{v}\right) \mathbf{1} = \mathrm{diag}\left(\mathbf{u}\right) \mathbf{G} \mathbf{v} = \mathrm{diag}(\mathbf{u}) e^{-1} \otimes_{q=1}^{Q} \exp\left(-\mu \mathbf{C}_q\right) \mathbf{v}. \tag{22}$$

Now, let us use the relationship between vector-operator and Kronecker product, i.e., $\mathrm{vec}\left(\mathbf{ABC}\right) = \left(\mathbf{C}^T \otimes \mathbf{A}\right) \mathrm{vec}\left(\mathbf{B}\right)$. Then, we adopt a fold operator $\mathrm{fold}\left(\cdot\right)$ and a product operator $\times_{q=1}^{Q}$ to fold a vector into a tensor, thereby transforming the vector $\mathbf{v} \in \mathbb{R}^{F}$ into an Q order tensor $\mathcal{V} = \mathrm{fold}(\mathbf{v}) \in \mathbb{R}^{f_1 \times f_2 \times \cdots \times f_Q}$. This gives

$$P_t \mathbf{1} = \mathrm{diag}\left(\mathbf{u}\right) \mathrm{vec}\left(\mathcal{V} \times_{q=1}^{Q} \exp\left(-\mu \mathbf{C}_q\right)\right). \tag{23}$$

Note that (23) does not require to compute directly the transport matrix $\mathbf{P}_t$, which helps reduce the computational complexity by a magnitude. Now, the estimators in (17) and (18) can be updated as

$$w_{n,f,k} \leftarrow w_{n,f,k} \sqrt{\frac{\sum_t \left[\mathrm{diag}\left(\mathbf{u}\right) \mathrm{vec}\left(\mathcal{V} \times_{q=1}^{Q} \exp\left(-\mu \mathbf{C}_q\right)\right)\right]_f h_{n,k,t} \left(\sum_{k'} w_{n,f,k'} h_{n,k',t}\right)^{-2}}{\sum_t \left[\mathrm{diag}\left(\mathbf{u}\right) \mathrm{vec}\left(\mathcal{V} \times_{q=1}^{Q} \exp\left(-\mu \mathbf{C}_q\right)\right)\right]_f \left(\sum_{k'} w_{n,f,k'} h_{n,k',t}\right)^{-1}}}, \tag{24}$$

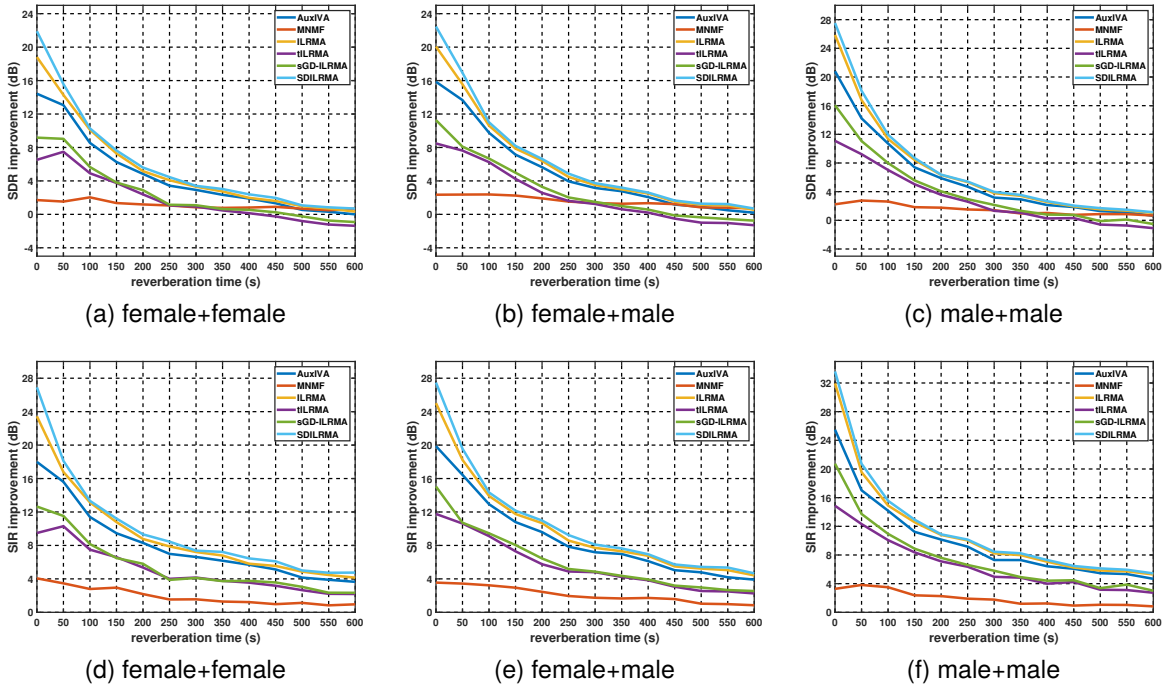


Figure 1. SDR and SIR improvement of the studied methods.

$$h_{n,k,t} \leftarrow h_{n,f,k} \sqrt{\frac{\sum_f \left[\operatorname{diag}(\mathbf{u}) \operatorname{vec}\left(\mathcal{V} \times_{q=1}^{Q} \exp\left(-\mu \mathbf{C}_q\right)\right)\right]_f w_{n,f,k} \left(\sum_{k'} w_{n,f,k'} h_{n,k',t}\right)^{-2}}{\sum_f \left[\operatorname{diag}(\mathbf{u}) \operatorname{vec}\left(\mathcal{V} \times_{q=1}^{Q} \exp\left(-\mu \mathbf{C}_q\right)\right)\right]_f \left(\sum_{k'} w_{n,f,k'} h_{n,k',t}\right)^{-1}}}. \tag{25}$$

5 SIMULATIONS

We used some speech signals from the Wall Street Journal (WSJ0) corpus [23] as the clean speech source signals and configured evaluation signals following the SISEC challenge [25] with $M = N = 2$, where the room size is $8 \times 8 \times 3$ m. The two sources are assumed to be 2 m away from the center of the two microphones and the microphone spacing is 5.66 cm. The incidence angles of the two sources are randomly selected from $[0°, 90°]$ and $[0°, -90°]$ respectively per mixture, where the direction normal to the line connecting two microphones is $0°$. The image source model [27] is used to generate the room impulse responses, where the sound absorption coefficients are calculated by Sabine's Formula [28] with the room aforementioned room size and reverberation time T_{60} changing from 0 to 600 ms with an interval of 50 ms. For each combination of sources (there are four combinations) and every value of T_{60}, 100 mixtures are generated for evaluation. The sampling rate is 16 kHz.

The parameters μ and γ of SDILRMA were set to 100, and 10, respectively. We compared SDILRMA with AuxIVA [10], MNMF [24], ILRMA [5], t-ILRMA and sGD-ILRMA [22]. The performance metrics used are the signal-to-distortion ration (SDR) and source-to-interferences ratio (SIR) [26].

Figure 1 presents the results in terms of the average SDR and SIR improvements. It is seen that SDILRMA outperforms MNMF, ILRMA, t-ILRMA and sGD-ILRMA, which demonstrates the effectiveness of SDILRMA for source separation.

Figure2 plots the spectrograms of the source signals as well as the signals estimated by ILRMA and SDILRMA. It is seen that both ILRMA and SDILRMA are effective. ILRMA suffers from a small number of permutations, which are not seen in SDILRMA. This, again, demonstrates the superiority of SDILRMA.

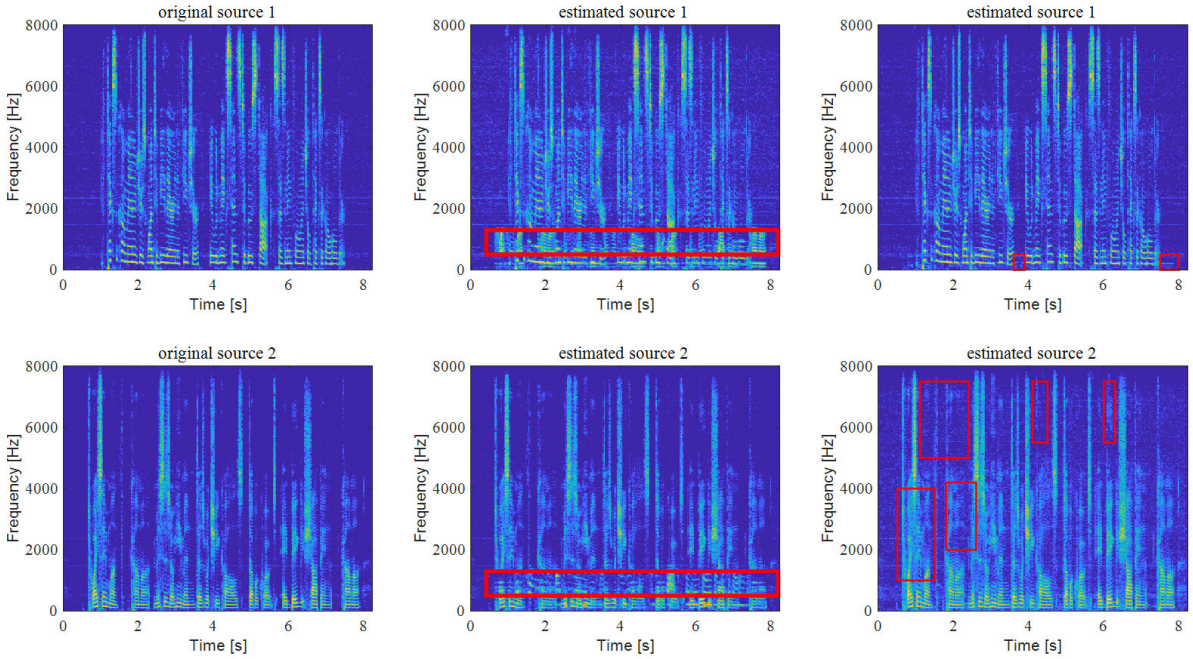


Figure 2. The spectrograms of the source and separated signals. Left panels: the original source signals. Middle panels: the separated signals by ILRMA. Right panels: the separated signals by SDILRMA.

6 CONCLUSION

This paper studied the determined BSS problem for audio and speech applications. We presented an improved version of ILRMA, which applies NMF to decompose the time-varying source model and Sinkhorn divergence as the cost function to optimize the model parameters. To simplify the algorithm to reduce its computational complexity, the Kronecker product tool was used to decompose the modeling matrix into the product of a number of matrices of much smaller dimensionality, resulting in a simplified SDILRMA algorithm. Simulation results verified that the simplified SDILRMA is able to achieve better BSS performance than ILRMA and is also computationally more efficient than its counterpart without Kronecker product decomposition.

REFERENCES

[1] Benesty J, Makino S, Chen J. Speech enhancement. New York, NY, USA: Springer, 2005.

[2] Makino S, Lee TW, Sawada H. Blind Speech Separation. New York, NY, USA: Springer, 2007.

[3] Li Y, Amari S, Cichocki A, Ho DWC, Xie S. Underdetermined blind source separation based on sparse representation. IEEE Trans. Signal Process., vol. 54, no. 2, pp. 423–437, Feb. 2006.

[4] Bofill P, Zibulevsky M. Underdetermined blind source separation using sparse representations. Signal Process. vol. 81, no. 11, pp. 2353–2362, Nov. 2001.

[5] Kitamura D, Ono N, Sawada H, Kameoka H, Saruwatari H. Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 14, no. 9, pp. 1626–1641, Sep. 2016.

[6] Benesty J, Chen J, Habets E. Speech Enhancement in the STFT Domain. Berlin: Springer-Verlag, 2011.

[7] Huang H, Zhao L, Benesty J, Chen J. A minimum-variance-distortionless-response filter based on the bifrequency spectrum for single-channel noise reduction. Digital Sig. Process., vol. 33, pp. 169–179, Oct. 2014.

[8] Comon P. Independent component analysis, a new concept. Signal Process., vol. 36, no. 3, pp. 287–314, Apri. 1994.

[9] Kim T, Eltoft T, Lee TW. Independent vector analysis: An extension of ICA to multivariate components. in Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation, Oct. 2006, pp. 165–172.

[10] Ono N. Stable and fast update rules for independent vector analysis based on auxiliary function technique. in Proc. IEEE Workshop Appl. Signal Process. Audio Acoust., 2011, pp. 189–192.

[11] Vallender S. Calculation of the wasserstein distance between probability distributions on the line. Theory of Probability& Its Applications, vol. 18, no. 4, pp. 784–786, 1974.

[12] Kantorovic LV, Rubinstein GS. On a functional space and certain extremum problems. in *Doklady Akademii Nauk*, Russian Academy of Sciences, vol. 115, pp. 1058–1061, 1957.

[13] Cuturi M. Sinkhorn distances: Light-speed computation of optimal transport. in Advances in Neural Information Processing Systems (NIPS). Red Hook, NY, USA: Curran, 2013, pp. 2292—2300.

[14] Rolet A, Cuturi M, Peyré G. Fast dictionary learning with a smoothed wasserstein loss. in Proc. Int. Conf. Artif. Intell. Stat., Cadiz, Spain, May 2016, pp. 630—638.

[15] Benzi M, Simoncini V. Approximation of functions of large matrices with Kronecker structure. Numerische Mathematik, vol. 135 no. 1, pp. 1—26, Jan. 2017.

[16] Motamed M. Hierarchical Low-Rank Approximation of Regularized Wasserstein Distance. arXiv preprint arXiv:2004.12511, 2020.

[17] Févotte C, Cardoso JF. Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models. in Proc. IEEE Workshop Appl. Signal Process. Audio Acoust., Oct. 2005, pp. 78–81.

[18] Vincent E, Jafari MG, Abdallah SA, Plumbley MD, Davies ME. Probabilistic modeling paradigms for audio source separation. In Machine Audition: Principles, Algorithms and Systems. IGI global, pp. 162–185, 2011.

[19] Brehm H, Stammler W. Description and generation of spherically invariant speech-model signals. Signal Process. vol. 12, no. 2, pp. 119–141, Mar. 1987.

[20] Buchner H, Aichner R, Kellermann W. Blind source separation for convolutive mixtures: A unified treatment. In Audio signal processing for next-generation multimedia communication systems (pp. 255–293). Springer, Boston, MA, 2004.

[21] Wang J, Guan S, Liu S, Zhang XL. Minimum-Volume Multichannel Nonnegative Matrix Factorization for Blind Audio Source Separation. IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, no. 3089–3103, Aug. 2021.

[22] Mogami S, Takamune N, Kitamura D, Saruwatari H, Takahashi Y, Kondo K, Ono N. Independent low-rank matrix analysis based on time-variant sub-Gaussian source model for determined blind source separation. IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 28, pp. 503–518, Dec. 2019.

[23] Garofolo J, Graff D, Paul D, Pallett D. Csr-i (wsj0) complete ldc93s6a. Web Download. Philadelphia: Linguistic Data Consortium, 83, 1993.

[24] Sawada H, Kameoka H, Araki S, Ueda N. Multichannel extensions of non-negative matrix factorization with complex-valued data. IEEE Trans. Audio Speech Lang. Process, vol. 21, no. 5, pp. 971–982, Jan. 2013.

[25] Araki S, Nesta F, Vincent E, Koldovský Z, Nolte G, Ziehe A, Benichoux A. The 2011 signal separation evaluation campaign (SiSEC2011):-audio source separation. LVA/ICA (pp. 414–422). Springer, Berlin, Heidelberg 2012.

[26] Vincent E, Gribonval R, Févotte C. Performance measurement in blind audio source separation. IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 4, pp. 1462–1469, June 2006.

[27] Allen JB, Berkley DA. Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Amer., vol. 65, no. 4, pp. 943–950, Apr. 1979.

[28] Young RW. Sabine reverberation equation and sound power calculations. The Journal of the Acoustical Society of America, vol. 31, no. 7, pp. 912—921, July 1959.

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0870

Analysis and Source Separation of Overlapping Speech Using Corpus of Everyday Japanese Conversation

Haruki NAMMOKU; Kouei YAMAOKA; Taishi NAKASHIMA; Yukoh WAKABAYASHI; Nobutaka ONO

Tokyo Metropolitan University, Japan, {nammoku-haruki, yamaoka-kouei, nakashima-taishi}@ed.tmu.ac.jp, {wakayuko, onono}@tmu.ac.jp

ABSTRACT

In this study, we evaluate the performance of blind source separation (BSS) for audio data in the Corpus of Everyday Japanese Conversation (CEJC). In everyday conversation, people's utterances may frequently overlap, which affects the transcription accuracy for linguistic analysis. First, we analyze how much the utterances overlap and how the overlap affects transcription using meta-data-like transcribed texts provided in CEJC. Then, we apply BSS methods to solve this problem. CEJC audio data were recorded by multiple voice recorders that were not synchronized, and speakers and recorders could move during the conversation. Both conditions make the direct application of BSS difficult. To overcome this difficulty, we apply blind signal synchronization and BSS in block processing. In particular, we compare four BSS methods, namely, two conventional BSS methods (batch and online auxiliary-function-based independent vector analysis (AuxIVA)) and two proposed methods (maximum channel selection-based time–frequency masking (MaxChTF) and the combination of AuxIVA and MaxChTF). We evaluate the proposed methods by subjective listening tests and confirm that BSS can support the transcription of everyday conversations.

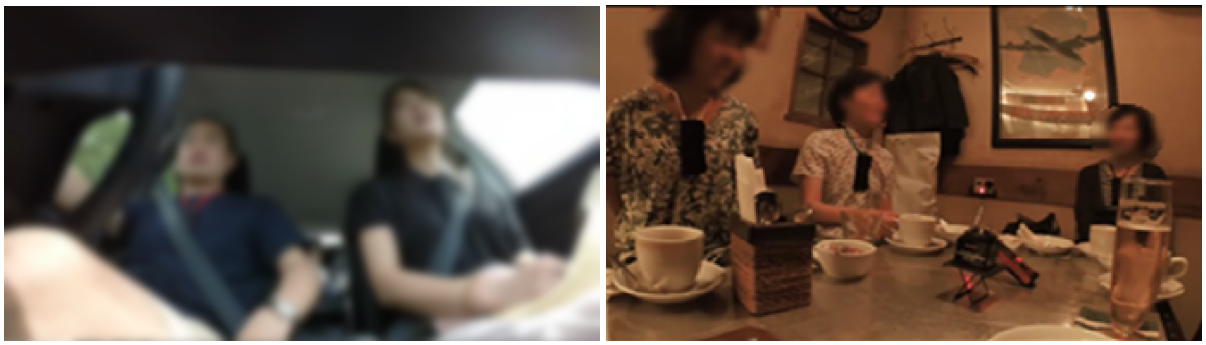
Keywords: Everyday conversation, Source separation, Independent vector analysis, Time–frequency masking

1 INTRODUCTION

The analysis of spoken language from everyday conversations is an essential topic in natural language processing. The language spoken in everyday conversations differs from that used in writing or lectures owing to verbal contractions and imprecise grammar. Collecting data from everyday conversations is challenging, and the lack of data has been one of the difficulties in research on everyday conversations. To address this issue, the National Institute for Japanese Language and Linguistics (NINJAL) has developed and published the Corpus of Everyday Japanese Conversation (CEJC) [1]. CEJC contains large-scale audio, video, and transcription data from everyday conversations in various situations. Audio and video data were collected using voice recorders and video cameras, respectively. CEJC is expected to develop Japanese language research through the use of collected data.

In everyday conversations, utterances sometimes overlap, making transcription difficult. We consider applying a blind source separation (BSS) method to extract each speaker's utterances from overlapped utterances. Various BSS methods that use spatial information from multichannel observation or assume the statistical independence of sound sources have been proposed [2, 3, 4, 5, 6]. In these BSS methods, it is assumed that sound sources and microphones do not move, and a microphone array is synchronized. However, this assumption is unrealistic in everyday conversations. Therefore, we do not expect these BSS methods would work well if applied directly to CECJ audio data.

In this study, we first statistically survey CEJC data to verify the effects of overlapped utterances on transcription. Next, we introduce a preprocessing technique to make the BSS methods applicable to overlapped utterances in CEJC. The preprocessing technique consists of two steps. First, to cope with the movement of sources, we divide the observation signals into blocks of short duration [7]. Second, we use the signal synchronization method developed by Miyabe *et al.* [8] for the time alignment of multiple asynchronized observation signals. After this preprocessing,



(a) Conversation scene while driving (b) Conversation scene in a restaurant

Figure 1. Examples of conversation situations recorded in CEJC

we apply BSS methods. Namely, auxiliary-function-based independent vector analysis (AuxIVA) [9], its online extension [10], and binary time–frequency (TF) masking that utilizes the characteristics of recordings in CEJC. We also propose an AuxIVA method combined with binary TF masking [11]. To evaluate the usefulness of these BSS methods for transcription, we conduct experiments by subjective listening tests to determine whether the BSS methods could be useful for transcribing everyday conversations.

2 ANALYSIS OF CEJC AUDIO DATA

2.1 Conversation recordings in CEJC

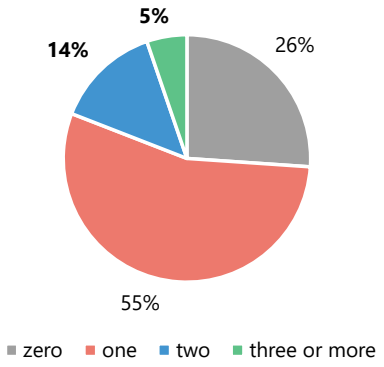
CEJC contains various conversation scenes with a balance of speaker ages and gender. More specifically, it consists of 200 hours of conversation data with 461 scenes in which 862 speakers participate. Figure 1 shows two scenes as examples: conversations during meals at a restaurant and during traveling in a private car. In recording CEJC audio data, each participant wears a voice recorder to transcribe his/her utterances accurately.

2.2 Statistical survey

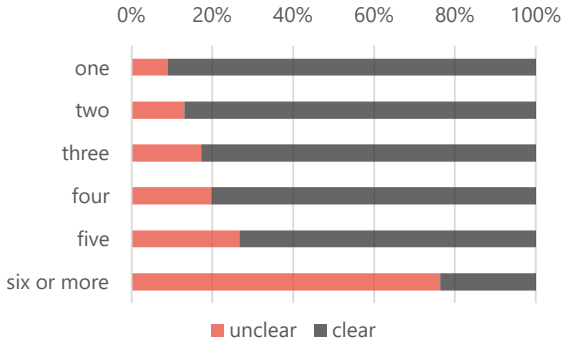
Although utterances were recorded with voice recorders, unclear segments occasionally exist in CEJC transcription data. Utterances in these segments have been transcribed with little reliability or have not been transcribed. Possible causes include overlapped utterances, individual differences in utterance volume, and background noise. In this study, we focus on overlapped utterances. To verify whether overlapped utterances affect the difficulty of transcribing, we conducted a statistical survey on the number of people who spoke simultaneously and the segments where the transcription was labeled as "unclear."

First, we counted the number of overlapped utterances from the transcription data, where the time stamps of all speakers' utterances are given. Figure 2a shows how many speakers spoke simultaneously and how long they spoke for the total recording time. As shown in this figure, two or more people spoke simultaneously in 19% of the total recorded data. Moreover, three or more people spoke simultaneously in 5% of the total recorded data. This result confirms that overlapped utterances sometimes occur in everyday conversations.

Next, we examined how many percentages of segments with overlapped utterances were labeled as "unclear," in CEJC transcription data and compared it for each number of simultaneous utterances. Figure 2b shows the percentage of unclear segments by the number of segments with overlapped utterances. As shown in Figure 2b, more than 10% of segments were labeled as "unclear" for overlapped utterances by two people. Additionally, the percentage of unclear segments increases with the number of simultaneous speakers in overlapped utterances. This result implies that the overlapped utterances make it difficult to understand each speaker's utterance, making the transcription difficult.



(a) Percentage of segments with overlapped utterances in CEJC audio data.



(b) Percentages of segments labeled as "unclear" in CEJC transcription data for each number of simultaneous utterances.

Figure 2. Results of statistical survey of data in CEJC

3 APPLIED FRAMEWORK

3.1 BSS for asynchronous recording of nonfixed sources

Overlapped utterances make transcription difficult as mentioned in Section 2.2. One possible solution to this problem is to apply BSS methods. Many BSS methods including AuxIVA require the following two assumptions:

1. Time invariance of the acoustic transfer function and
2. Time synchronization of each channel.

However, it is difficult to assume both in real environments, and CEJC audio data are an example for this case. In CEJC audio data, each speaker can move and microphones are not synchronized. To address the problem for the first assumption, it is necessary to adopt the separation filter for every moment as the acoustic transfer function changes.

Furthermore, there are discrepancies in the recording start time among recorders. There are also slight differences in sampling frequency among the recorders, even if the nominal sampling frequency of these devices is the same. The time information of a signal cannot be obtained correctly owing to these factors, and the separation performance is degraded [8, 12]. Therefore, we perform the following processes:

- Divide the entire audio recording into blocks and apply the blind synchronization method [8] in each block and
- Apply the BSS in shorter blocks (sub-blocks) that are expected to have relatively little change in the transmission system.

Conventionally, online ICA-based BSS methods have been studied to allow movement of speakers, and several pioneering works such as online versions of ICA [13] and TRINICON [14] are known. Taking the subsequent development of BSS methods into consideration, we employ newer BSS methods, AuxIVA, and its variants in this study. We also consider simple TF masking passing only the component with the largest amplitude. This design method is inspired by the CEJC recording condition where each speaker wears a recording device.

3.2 Problem formulation

Let K be the number of sound sources and I the number of voice recorders. In the short-time Fourier transform (STFT) domain, let the source signal of the kth speaker be $S_k(\tau,\omega)$ and the observed signal of the ith recorder be $X_i(\tau,\omega)$, where τ and ω are the time frame and frequency bin index, respectively. Because each speaker wears a single-channel voice recorder in the CEJC setting, we set that $K = I$ is necessary and sufficient. Their vector forms are denoted as $\boldsymbol{S}(\tau,\omega) = [S_1(\tau,\omega),\ldots,S_K(\tau,\omega)]^{\mathsf{T}}$, $\boldsymbol{X}(\tau,\omega) = [X_1(\tau,\omega),\ldots,X_I(\tau,\omega)]^{\mathsf{T}}$, where the superscript T denotes the transpose of a vector. In this paper, we aim to separate the mixed observations into individual sources

associated with each speaker. That is, we estimate the separation signals $\boldsymbol{Y}(\tau,\omega) = [Y_1(\tau,\omega),\ldots,Y_I(\tau,\omega)]^{\mathsf{T}}$ from $\boldsymbol{X}(\tau,\omega)$. We then obtain each of the persons' utterances even when they speak at the same time.

4 BSS

4.1 Auxiliary-function-based independent vector analysis

Ono proposed AuxIVA [9]. By applying the auxiliary function method to the optimization problem for the conventional gradient-descent-based IVA [3, 4], one can derive a fast and stable algorithm for the update of the separation matrix. AuxIVA assumes a generative model of the source vector that puts together the frequency components of each sound source. The separation filter $\boldsymbol{W}(\omega)$ is estimated by solving an optimization problem so that each sound source is statistically independent, and the separation sound is obtained as

$$\boldsymbol{Y}(\tau,\omega) = \boldsymbol{W}(\omega)\boldsymbol{X}(\tau,\omega). \tag{1}$$

The AuxIVA algorithm alternately updates the weighted covariance matrix $\boldsymbol{V}_k(\omega)$ and the separation matrix $\boldsymbol{W}(\omega)$.

$$\boldsymbol{V}_k(\omega) = \frac{1}{T}\sum_{\tau}\left(\frac{\boldsymbol{X}(\tau,\omega)\boldsymbol{X}^{\mathsf{H}}(\tau,\omega)}{\frac{1}{\Omega}\sum_{\omega}|Y_k(\tau,\omega)|^2}\right), \tag{2}$$

$$\boldsymbol{w}_k(\omega) \leftarrow (\boldsymbol{W}(\omega)\boldsymbol{V}_k(\omega))^{-1}\boldsymbol{e}_k, \tag{3}$$

$$\boldsymbol{w}_k(\omega) \leftarrow \frac{\boldsymbol{w}_k(\omega)}{\sqrt{\boldsymbol{w}_k^{\mathsf{H}}(\omega)\boldsymbol{V}_k(\omega)\boldsymbol{w}_k(\omega)}}, \tag{4}$$

where T is the total number of time frames, Ω is the total number of frequency bins, the superscript H is the complex conjugate transpose of the vector, and $\boldsymbol{e}_k$ and $\boldsymbol{w}_k(\omega)$ are the kth row vectors of the identity and separation matrices, respectively.

Because $\boldsymbol{W}(\omega)$ is assumed to be identical across time frames, AuxIVA may not separate long CEJC audio data well. Therefore, we divide synchronized blocks of signals into shorter time sub-blocks before applying AuxIVA. We expect that partitioning into sub-blocks will improve the separation performance of AuxIVA.

4.2 Online auxiliary-function-based independent vector analysis

AuxIVA described in the previous section is a batch algorithm, but an online version of AuxIVA has been proposed to realize real-time source separation [10]. Online AuxIVA estimates the separation filter $\boldsymbol{W}(\tau,\omega)$ and weighted covariance $\boldsymbol{V}_k(\tau,\omega)$ at each time frame. The forgetting factor α enables the estimation of $\boldsymbol{V}_k(\tau,\omega)$ while maintaining the information of the time frame before the current one.

$$\boldsymbol{V}_k(\tau,\omega) = \alpha\boldsymbol{V}_k(\tau-1,\omega) + (1-\alpha)\frac{\boldsymbol{X}(\tau,\omega)\boldsymbol{X}^{\mathsf{H}}(\tau,\omega)}{||\boldsymbol{Y}_k||_2}, \tag{5}$$

$$\boldsymbol{w}_k(\tau,\omega) \leftarrow (\boldsymbol{W}(\tau,\omega)\boldsymbol{V}_k(\tau,\omega))^{-1}\boldsymbol{e}_k, \tag{6}$$

$$\boldsymbol{w}_k(\tau,\omega) \leftarrow \frac{\boldsymbol{w}_k(\tau,\omega)}{\sqrt{\boldsymbol{w}_k^{\mathsf{H}}(\tau,\omega)\boldsymbol{V}_k(\tau,\omega)\boldsymbol{w}_k(\tau,\omega)}}. \tag{7}$$

We expect that the online AuxIVA cam achieve higher separation performance than the batch algorithm for real-world recording with nonfixed sources and microphones.

4.3 Maximum channel selection-based TF masking

Binary TF masking assumes at most one dominant source at each TF point (W-disjoint orthogonality [15]). On the basis of this assumption, we estimate the binary mask $M_k(\tau,\omega)$ that passes only the TF components of the target sound in $X_k(\tau,\omega)$. We then extract the target signal from the observed signal by taking the product of $M_k(\tau,\omega)$ and $X_k(\tau,\omega)$ as follows:

$$Y_k(\tau,\omega) = M_k(\tau,\omega)X_k(\tau,\omega). \tag{8}$$

There are many methods of estimating $M_k(\tau,\omega)$ such as clustering based on the time difference of arrival of the source [15] and classification based on deep learning [16]. Binary masking results in many zeros, and the output spectrum becomes discontinuous in many time frames. Therefore, $Y_k(\tau,\omega)$ may contain unpleasant noises called musical noise.

In this study, we consider a method that takes advantage of the CEJC recording environment, where there is a one-to-one correspondence between a speaker and a recorder, and the microphone is close to the speaker. We assume that when each speaker speaks, the signal is most significantly observed in each speaker's recorder, and we design a binary mask $M_k(\tau,\omega)$ on the basis of the maximum channel selection as follows:

$$M_k(\tau,\omega) = \begin{cases} 1, & (|X_k(\tau,\omega)| > |X_j(\tau,\omega)|,\ \forall j \neq k), \\ 0, & (\text{otherwise}). \end{cases} \tag{9}$$

Using $M_k(\tau,\omega)$ designed in this way, we can estimate $Y_k(\tau,\omega)$ for each speaker's voice by applying (8) to the mixed source signal.

4.4 Masking-based AuxIVA

Source separation by binary TF masking produces some artificial noises, including musical noise, which can affect human auditory perception. In contrast, AuxIVA separation does not produce musical noise. On the basis of these, we consider combining these advantages. Conventional AuxIVA uses a time-varying Gaussian distribution as the source model and assumes that the variance of the distribution is constant across frequencies [17]. More flexible source model with different variance for each time-frequency component would improve separation performance [11]. In this study, we use the binary masking separation result designed in Section 4.3, $\hat{S}_k(\tau,\omega) = M_k(\tau,\omega)X_k(\tau,\omega)$, as a source model to avoid musical noise and further improve the separation performance from conventional IVA. That is, we replace the denominator of (2) with the estimation result by binary TF masking designed in Section 4.3, $\hat{S}_k(\tau,\omega)$, as follows:

$$\boldsymbol{V}_k(\omega) = \frac{1}{T}\sum_{\tau}\left(\frac{\boldsymbol{X}(\tau,\omega)\boldsymbol{X}^{\mathsf{H}}(\tau,\omega)}{\left|\hat{S}_k(\tau,\omega)\right|^2}\right). \tag{10}$$

So as not to divide by zero, we replace zero value components of the TF mask with a minute value ε.

5 EXPERIMENTS

5.1 Conditions

In this study, we evaluated the performance of source separation processing by subjective measures using a listening test. Objective measures such as signal-to-distortion ratio (SDR) [18] are often used to evaluate the performance of source separation methods. Reference sound signals are required in the evaluation of performance with these indices. Because CEJC audio data were recorded in real environments, no reference sound source signal exists, so that SDR and SIR cannot be computed. Therefore, we need to evaluate them subjectively by a listening test. We conducted this evaluation with approval from the research ethics review committee of Tokyo Metropolitan University.

The audio data evaluated in this paper are the conversation utterances recorded in the public version of CEJC Monitor. We apply BSS processing to them, using the parameters shown in Table 1. The nominal sampling frequency of each voice recorder is 16 kHz, and we apply synchronization processing to compensate for minute errors due to individual microphone differences.

We recruited 13 normal-hearing subjects in the listening test. They evaluated 10-second audio data segments in each conversation scene. We instructed them to carry out two tasks. In the first task, we instructed them to transcribe by listening to all the audio data we gave them. In the second task, we instructed them to rank five sounds (one unprocessed observation signal and four separated signals) in terms of ease of listening to the utterance, assuming a detailed utterance transcription including fillers and dialects. We did not inform the subjects which method was applied for each audio data. We instructed the subjects to listen with earphones or headphones. In this section, the names of the compared signals are simplified, as shown in Table 2. For some of the compared audio data, it was difficult to judge their ease of listening. Therefore, we permitted the subjects to listen to the audio data for an

Table 1. Parameters for source separation

Parameter	Value
STFT frame length	2048
STFT frame shift	1024
STFT window function	Hamming
Block length for synchronization	30 s
Sub-block length for AuxIVA	10 s
Number of iterations in AuxIVA	20
Number of iterations in online AuxIVA	2
Forgetting factor α of online AuxIVA	0.94
Threshold for BM IVA ε	2^{-15}

Table 2. Methods compared in the experiment

Processing	Abbreviation
Unprocessed observation	Obs.
AuxIVA described in 4.1	IVA
Online AuxIVA described in 4.2	OIVA
TF masking proposed in 4.3	BM
TF-masking-based AuxIVA proposed in 4.4	BM IVA

Table 3. Conversation scenes used in the experiment

Scene ID	Number of speakers	Scene
`K004_008`	3	Over tea and sweets
`T003_021`	5	During lunchtime

unlimited number of times for the ranking. They evaluated all speaker's utterances in the two situations shown in Table 3. One scene was a conversation among three women, the other was a conversation among five women. Each conversation took place in a different place involving different people.

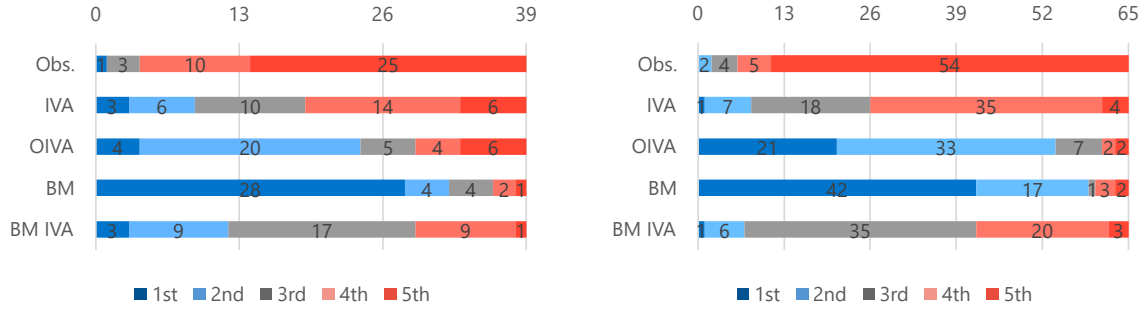
5.2 Results

Figure 3 shows the number of times subjects voted for each rank for each sound. The length of the horizontal axis is the product of the number of speakers in each scene and the number of subjects. We discuss the usefulness of each BSS method for transcription by focusing on the number of votes for each rank.

First, BM has the largest number of best ratings and a small number of poor ratings (fourth or fifth place) among all audios compared in both scenes. This result implies that BM has a high separation performance for CEJC audio data. We considered that the musical noise would make the BM's rating worse. However, for many subjects, the noise did not bother them enough to affect the ease of transcriptions.

Second, the number of second place ratings of online AuxIVA is the largest among all audios compared in both scenes. Moreover, online AuxIVA in `T003_021` (the scene of five speakers) has more good ratings (first or second place) than that in `T003_021` (the scene of three speakers), which implies that BSS methods with time-variant filters, such as online AuxIVA, are useful for transcribing audio data recorded in real environments. Figure 3a also shows that online AuxIVA has the largest number of worst ratings among all processed audios. This might be because of segments where someone does not speak for a few seconds. These segments interfere with the accurate update of the separation filter. There are many such segments during everyday conversations. Therefore, it is important to retain speakers' information when they are not speaking to further improve the performance.

Third, the number of third place ratings of BM IVA is slightly larger than that of AuxIVA. In addition, when we asked subjects to comment on their evaluation of the sound separation, many said that they felt the difference between AuxIVA and BM IVA was small. Although we expected BM IVA to be higher than AuxIVA and BM in separation performance, subjects rated BM IVA to be the same as AuxIVA and much worse than BM. Even when the TF masking distorted the target utterance or caused some musical noise, many subjects may have considered that the



(a) For three speakers in the conversation scene `K004_008`.

(b) For five speakers in the conversation scene `T003_021`.

Figure 3. Ranking results for each speaker's utterances in each conversation scene. The number on the band shows the number of times the corresponding rank was chosen.

more the interference was suppressed effectively, the easier to transcribe.

6 CONCLUSION

In this paper, we statistically investigated the frequency of overlapped utterances in CEJC and examined the performance of BSS for conversation recordings. For BSS, we focused on the recording environment of the conversation. We designed a BM that emphasizes the target speaker's utterance by comparing the amplitude spectrograms between the observed signals. We then proposed an AuxIVA method combined with simple binary TF masking. We conducted experiments to evaluate the separation performance through subjective listening tests and verified the usefulness of BSS methods in supporting transcription. Future work may include the combination of the online AuxIVA with a TF-masking-based source model and the evaluation of the separation performance in more variety of conversation scenes.

ACKNOWLEDGEMENTS

This research was supported by JSPS Grants-in-Aid for Scientific Research JP20H00613.

REFERENCES

[1] Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yasuyuki Usuda, and Yuka Watanabe. Design and evaluation of the corpus of everyday japanese conversation. In *the Language Resources and Evaluation Conference*, pages 5587–5594, Marseille, France, June 2022. European Language Resources Association.

[2] Paris Smaragdis. Blind separation of convolved mixtures in frequency domain. *Neurocomputing*, 22(1):21–34, 1998.

[3] Atsuo Hiroe. Solution of permutation problem in frequency domain ICA, using multivariate probability density functions. In *Independent Component Analysis and Blind Signal Separation*, pages 601–608, 2006.

[4] Taesu Kim, Hagai T. Attias, Soo young Lee, and Te won Lee. Blind source separation exploiting higher-order frequency dependencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):70–79, 2007.

[5] Alexey Ozerov and Cédric Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563, 2009.

[6] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari. Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1626–1641, 2016.

[7] Athanasios Koutras, Evangelos Dermatas, and Georgios Kokkinakis. Blind speech separation of moving speakers in real reverberant environments. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, volume 2, pages II1133–II1136, 2000.

[8] Shigeki Miyabe, Nobutaka Ono, and Shoji Makino. Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation. *Signal Processing*, 107:185–196, 2015.

[9] Nobutaka Ono. Stable and fast update rules for independent vector analysis based on auxiliary function technique. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 189–192, 2011.

[10] Toru Taniguchi, Nobutaka Ono, Akinori Kawamura, and Shigeki Sagayama. An auxiliary-function approach to online independent vector analysis for real-time blind source separation. In *2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pages 107–111, 2014.

[11] Ana Ramírez López, Nobutaka Ono, Ulpu Remes, Kalle Palomäki, and Mikko Kurimo. Designing multichannel source separation based on single-channel source separation. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 469–473, 2015.

[12] Lin Wang and Simon Doclo. Correlation maximization-based sampling rate offset estimation for distributed microphone arrays. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3):571–582, 2016.

[13] Robert Aichner, Herbert Buchner, Shoko Araki, and Shoji Makino. On-line time-domain blind source separation of nonstationary convolved signals. In *Int. Symposium on Independent Component Analysis and Blind Signal Separation*, 2003.

[14] Herbert Buchner, Robert Aichner, and Walter Kellermann. A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. *IEEE Transactions on Speech and Audio Processing*, 13(1):120–134, 2005.

[15] Özgür Yılmaz and Scott Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004.

[16] Yi Luo and Nima Mesgarani. TaSNet: Time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700, 2018.

[17] Nobutaka Ono. Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions. In *2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2012.

[18] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1462–1469, 2006.

PROCEEDINGS of the
24th International Congress on Acoustics
October 24 to 28, 2022 in Gyeongju, Korea

ABS-0242

Vocalic and consonantal identification in noisy and quiet environments

Joy Oluchi UGURU
Department of Linguistics, Igbo and other Nigerian Languages
University of Nigeria, Nsukka, Nigeria

ABSTRACT

In this paper, students were tested for vocalic and consonantal identification. English and Igbo words were dictated to seventy students, whose native language is Igbo, in noisy and quiet classrooms respectively. Both languages had words with varied consonants and vowels. Our comparative analysis of vocalic/consonantal perception by the students in both noise and quiet only showed significant difference for English vowels and English consonants in quiet environment; there was no significant difference observed in the rest. Conversely, there was significant difference in the isolated analyses carried out to reveal the students' perception for consonants in quiet and noise on one hand and vowels in quiet and noise on the other hand (for both languages). Our findings show that though vowels are more easily identified in speech than consonants, the perception of both vowels and consonants is hindered by noise.

Keywords: Sound perception, Noisy and quiet environments

1. INTRODUCTION

Igbo is a Benue Congo language of the Niger Congo family. It is one of the major Nigerian languages and is spoken in the South East of the country. Following colonization, English was adopted as the official language of Nigeria, a highly multilingual nation. English is mostly learnt in the school environment. Noise, is a major problem in the country (1) hence in the learning environments, learners have to cope with hearing in noise. It is therefore pertinent to understand how English and Igbo (the indigenous language of the immediate environment) are perceived in noise. To do this, we centred on the comparison of the perception of the consonants and vowels in the two languages.

At the syllable level, vowels are usually more intense than consonants and at the phrase level, syllables at the end of an utterance can become weaker in intensity (2). Vitevitch (3) shows that the initial part of a word is important for quick and accurate recognition of a spoken word. The back and round vowels /o/ and /u/ are the most visible, while the front unrounded vowels not involving lip rounding like /e/ and /i/ have been identified as the most audible (4).

More importantly, this experiment is conducted to help find out which of the two languages (Igbo or English) should be used more in very noisy environments. Ebem *et al* (5) discovered that a group of Igbo subjects listening to Igbo and American English speech, were more affected by noise than native American English speakers listening to their own native speech. Also, Uguru (6) shows that Igbo tone has a higher harmonic –to-noise ratio than English intonation. All these previous findings may point to the fact that the

joy.uguru@unn.edu.ng

perception of Igbo may be more affected by noise than English. The present study is yet another experiment to prove or disprove previous claims or findings.

2. METHODS

Twenty monosyllabic English words and ten Igbo words (five monosyllabic and five disyllabic) were read to seventy students in both quiet and noisy environments. They wrote down as perceived. In analyzing their written responses, attention was paid on the consonants and vowels they wrote. Their rates of perception were observed and analyzed with student 't' to see if there is any significant difference between perception of consonants and vowels. The words are as follows:

The words are as follows:

English Words

(1) Toy /tɔɪ/ (2) Boy /bɔɪ/ (3) Noise /nɔɪz/ (4) Girl /gɜl/ (5) Cry /kɹaɪ/ (6) Lip /lɪp/ (7) Light /laɪt/ (8) Like /laɪk/ (9) Lice /laɪs/ (10) Life /laɪf/ (11) Bread /bɹɛd/ (12) (12) Deep/dip/ (13) Gum /gʌm/ (14) Cat /kæt/ (15) Road /ɹəʊd/ (16) House /hɑʊs/ (17) Feet /fit/ (18) Pot /pɒt/ (19) Clay /kleɪ/ (20) Class /klɑs/

IGBO WORDS

Òké /oke/ 'rat' (2) Jí /ʤi/ 'yam' (3) Àlà /ala/ 'ground' (4) Nyé /ɲe/ 'give' (5) Ńrí /nri/ 'food' (6) Gbá /ɡba/ 'run' (7) Ókwú /okwu/ 'speech' (8) Tí /ti/ 'strike' (9) Ìgbà /ɪgba/ 'drum' (10)Gụ́ /gʊ/ 'read'

3. RESULTS AND FINDINGS

The results are shown in tables and also discussed.

Table 1- Recognition of English consonants versus vowels by respondents in noise.

Utterance	No. of respondents who identified vowel/diphthong correctly	No. of respondents who identified consonant correctly
Toy	63	42
Boy	70	70
Noise	70	70
Girl	70	70
Cry	42	56
Lip	42	56
Light	70	70
Like	63	49
Lies	70	49
Life	70	63
Bread	63	70
Deep	70	35

Gum	63	63
Cat	63	70
Road	70	56
House	70	70
Feet	56	56
Pot	56	56
Clay	70	70
Class	70	70
Total	65	58

Vowels – 91%; Consonants – 86%; t = 1.19; Probability = 0.244; Not significant (p > 0.05).

From Table 1, we see that the students' recognition of English vowels in noise is 91% while that of consonants is 86%.

t = 1.19; Probability = 0.244. This result is not significant (p > 0.05).

Table 2 - Recognition of Igbo consonants versus vowels by respondents in noise.

Utterance	No. of respondents who correctly identified vowel/diphthong	No. of respondents who correctly identified consonant
Òké	70	70
Jí	63	63
Àlà	63	63
Nyé	49	49
Ńrí	63	63
Gbá	42	42
Ókwú	63	56
Gụ́	56	35
Ịgbà	56	63
Tí	70	70

Vowels – 85%; Consonants – 82%; t = 0.43; Probability = 0.672; Not significant (p > 0.05).

A look at Table 2 reveals that the students' perception of Igbo vowels in noise is 85% while that of Igbo consonants is 82%.

t = 0.43. Probability = 0.672. The difference is not significant (p > 0.05).

Table 3-Recognition of English consonants versus vowels by respondents in quiet.

Utterance	No. of respondents who correctly identified vowel/diphthong	No. of respondents who correctly identified consonant
Toy	70	70
Boy	70	70
Noise	69	70

Girl	69	68
Cry	70	70
Lip	69	68
Light	68	68
Like	70	70
Lice	69	69
Life	69	67
Bread	70	70
Deep	70	69
Gum	69	65
Cat	70	69
Road	70	70
House	70	70
Feet	70	69
Pot	70	69
Clay	70	67
Class	70	70

Vowels – 99%; Consonants – 98%; t = −2.09; Probability = 0.046;
Significant ($p < 0.05$).

From Table 3, we see that the students' perception of English vowels in quiet is 99% while that of English consonants is 100%. t = −2.09.Probability = 0.046. The difference between the perception of English consonants and English vowels is therefore significant ($p < 0.05$).

Table 4-Recognition of Igbo consonants versus vowels by respondents in quiet.

Utterance	No. of respondents who correctly identified vowel/diphthong	No. of respondents who correctly identified consonant
Òké	70	70
Jí	70	70
Àlà	70	70
Nyé	70	70
Ńrí	70	70
Gbá	70	70
Ókwú	70	70
Gụ́	70	66
Ịgbà	70	70
Tí	70	70

Vowels – 100%; Consonants – 99%; t = 1.00; Probability = 0.331; Not significant (p > 0.05).

The recognition rates for Igbo consonants and vowels in quiet (Table 4) are 99% and 100%, respectively. The 't' test result for the analysis is t = 1.00; probability = 0.331 (not significant: (p > 0.05)).

Table 5-Recognition of English vowels in noise and quiet (values indicate the number of students that recognized the consonants/vowels).

	Noise	Quiet
Toy	63	70
Boy	70	70
Noise	70	69
Girl	70	69
Cry	42	70
Lip	42	69
Light	70	68
Like	63	70
Lies	70	69
Life	70	69
Bread	63	70
Deep	70	70
Gum	63	69
Cat	63	70
Road	70	70
House	70	70
Feet	56	70
Pot	56	70
Clay	70	70
Class	70	70

Noise – 91%; Quiet – 99%; t = 2.79; Probability = 0.012; Significant (p < 0.05)

Table 6-Recognition of English consonants in noise and quiet.

	Noise	Quiet
Toy	42	70
Boy	70	70
Noise	70	70
Girl	70	68
Cry	56	70
Lip	56	68
Light	70	68

Like	49	70
Lies	49	69
Life	63	67
Bread	70	70
Deep	35	69
Gum	63	65
Cat	70	69
Road	56	70
House	70	69
Feet	56	69
Pot	56	69
Clay	70	67
Class	70	70

Noise – 86%; Quiet – 98%; t = −3.44; Probability = 0.003; Significant ($p < 0.05$)

Table 7-Recognition of Igbo vowels in noise and quiet.

	Noise	Quiet
Òké	70	70
Jí	69	70
Àlà	63	70
Nyé	49	70
Ńrí	63	70
Gbá	42	70
Ókwú	63	70
Gụ́	56	70
Ị̀gbà	56	70
Tí	70	70

Noise – 85%; Quiet – 100%; t = 3.35; Probability = 0.004; Highly significant ($p < 0.05$)

Table 8-Recognition of Igbo consonants in noise and quiet.

	Noise	Quiet
Òké	70	70
Jí	69	70
Àlà	63	70
Nyé	49	70
Ńrí	63	70
Gbá	42	70
Ókwú	63	70
Gụ́	56	66
Ị̀gbà	56	70

Tí	70	70

Noise – 85%; Quiet – 99%; t = 3.19; Probability = 0.011;Significant ($p < 0.05$)

Table 5 shows the students' rate of recognition of English vowels in noise and quiet to be 91% and 99%, respectively. The difference is significant ($t = -2.79$; probability = 0.012 ($p< 0.05$)). Table 6 shows the analysis of the students' recognition of English consonants in noise and quiet to be 86% and 98%, respectively; 't' test result is significant ($t = -3.44$; probability = 0.003 ($p < 0.05$)). From Table 7, the students' recognition of Igbo vowels in noise and quiet is seen; 85% of the vowels were recognized in noise, while 100% were recognized in quiet. The 't' test result is: ($t = 3.35$, probability = 0.004 ($p < 0.05$)). Hence the difference is highly significant. Table 8 shows the recognition of the Igbo consonants in noise and quiet to be 85% and 99%, respectively. The difference is significant as revealed by the 't' test result ($t = 3.19$, probability =0.011 ($p < 0.05$)). From Tables 5 – 8, it can be seen that the perception and identification of both consonants and vowels were significantly worse in noise. From the foregoing, we conclude that noise is unfavourable for the recognition and the study of both English and Igbo languages.

4. Conclusion

Our findings show a significant difference between the students' perception for English vowels and English consonants in quiet environment (others did not show significant difference). Also, there were significant differences in the isolated analyses carried out to reveal the students' perception for consonants in quiet and noise on one hand and vowels in quiet and noise on the other hand (for both languages). Hence, we have discovered that though vowels are more easily identified in speech than consonants, the perception of both vowels and consonants is hindered by noise. These findings have implications for manufacturers of speech enhancement gadgets. Also, neither of the languages studied in this work should be studied in noisy environments if maximum learning is to be achieved. This has a great implication for other subjects which are studied in English throughout Nigeria. Noise should be limited as much as possible in our schools.

REFERENCES

1. Uguru, J.O. (2014). *Noise in Africa.* Enugu: Chenglo.
2. Klatt, D. (1987). Review of text-to-speech conversion for English. *Journal of Acoustical Society of America*, 82(3): 737-793.
3. Vitevitch, M.S. (2002). Naturalistic and experimental analyses of word frequency and neighbourhood Density effects in slips of the ear. *Language and Speech*, 45(4): 407–434.
4. Thirumalai, M.S., Gayathri, S.G. (1988). *The Speech Of The Hearing Impaired*. Manasagangotri: Central Institute.
5. Ebem, D.U.; Beerends, J.G.; Vugt, J.V.; Schmidmer, C.; Kooij, R.E. and Uguru, J.O. (2011) "The Impact of Tone Language and Non-Native Language Listening on Measuring Speech Quality", *Journal of Audio Engineering Society*. Vol. 59 (9) pp. 647 – 655.
6. Uguru, J.O. (2007) "Harmonics – to – Noise Ratio in Ika Igbo and English Utterances: Implications and Inferences", Proceedings of the 19th International Congress on Acoustics. Madrid, Spain. September 2nd – 7th.

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0466

AN ACOUSTIC ANALYSIS OF SPEECH PERCEPTION AND COMPREHENSION OF ONLINE LECTURES

Nkechi Mgbodichinma UKAEGBU[1]; **Kingsley K. UGWUAGBO**[2]

[1]**University of Nigeria, Nsukka, Nigeria.**

[2]**University of Port Harcourt, Rivers State**

ABSTRACT

Online lectures have become part of the 'new normal' with the increasing need to teach, learn, trade, and collaborate in the post-COVID era. However, the success of online interactions is hinged on effective communication which is dependent on speech and perception. This study seeks to examine the extent of students' perception and comprehension of English online lectures on pronunciation. Stimuli consists of six online phonetics and phonology lectures, taught by two British, American and Nigerian native speakers; one male and female each. Subjects comprising 134 language students between ages 18-30; who took a hearing test, listened to and rated tokens from lectures. The mean f0, duration, intensity and syllable-count of tokens are analyzed on Praat. The study observes that, stimuli from female speakers are perceived and understood easily than that of male speakers. Also, though the duration value are same for all tokens, the syllable count per duration for the male speakers are higher, showing a faster speech rate than that of the female speakers; which explains why there is low level of stimuli perception from subjects. The study recommends moderation of pitch and speakers' speech rate to accommodate the variety of English L2 speakers that listen to online lectures.

Keywords: Speech perception/comprehension, English Online lectures

1. INTRODUCTION

When listening to someone talking, the most important aspect of that communication is to comprehend the meaning of what is being said not necessarily based on the pronunciation of the speaker, though in many ways pronunciation plays a huge role in the understanding of that communication or interaction. Most times it is argued that the mispronunciation of certain sounds do not necessarily affect the comprehension of a conversation, however, instances abound when one makes a faulty pronunciation, makes use of a wrong intonation, stress or even tone pattern, thereby impeding meaning and the essence of that communication. The voice of a speaker is a dynamic and fundamental instrument of communication. This goes on to say that without voice, communication is possible, but it may be neither efficient nor timely. The voice carries a lot of perceptual and acoustic information that could be studied from different perspectives; which makes instrumental phonetics come in handy. The use of instrumental analysis can account for not only the production and perception of speech at the segmental level but also for the differences in speaking rates and other prosodic patterns which may play a role in listeners' comprehension (Munro, 1995 (1); Mitterer et al.,

[1] nkechi.ukaegbu@unn.edu.ng
[2] kachukwusicho@gmail.com

2016 (2); 2019 (3); Steffman, 2019 (4); Steffman, and Jun 2019 (5); 2021 (6)) . Although Holt, Lotto and Diehl (2004:1763) (7) argue that "communicative benefits arise not from acoustic distinctiveness among sounds but from distinctiveness within the human auditory system." This implies that comprehending spoken language is dependent more on how listeners extract fundamental linguistic elements from the perceptual signals of the voice or speech of speakers. This does not in any way undermine the use of acoustic signals from speakers' speech to see why certain listeners comprehend certain speakers and not others.

A normal voice can be convincing, motivating, and stimulating in any conversation, whether it be a casual conversation between two individuals or from the a lecturer at the lectern in front of the lecture hall to his/her students. It could pass across information about a particular person's speech pattern, emotions, speech variety and the environment in which s/he lives. That is why in Nigeria for example, the voice of a speaker can sometimes tell a listener whether the speaker is a native-speaker of Hausa, Igbo, Yoruba, Ibibio, Ijaw, Tiv or from any other language speaking group within Nigeria. The language or regional dialect that a person speaks could affect comprehension in many possible ways. Mattys, Davis, Bradlow and Scott (2012) (8) and Penga and Wang (2019) (9) support this assertion that speaker characteristics like foreign accent and listeners' individual characteristics such as language competence and performance, amongst other things can make it difficult and effortful to comprehend speech especially in complex realistic acoustic environments.

In this study, we argue that a calm, clear and pleasant voice draws listeners, maintains their attention and enhances comprehension of the message being communicated irrespective of speaker's accent or the listener's competence of the language used in the conversation. Imagine an instance when someone delivers a lecture or does a presentation in a clear, pleasant and resonant voice, as well as in a language you understand. Irrespective of culture, educational achievements or the variety of the language used, it does command much more attention and invariably, understanding the speech itself will not be not far-fetched. Therefore, the objectives of this study are to see if hearing ability affects perception and comprehension, to assess which of the speakers were more or least understood and finally, we use acoustic analysis to show whether intensity of voice and speaking rate contributes in any way to perception and comprehension of the online lectures.

2. MATERIALS AND METHODS

As a preliminary study, we use a population of 134 language students purposively sampled from the Department of Linguistics, Igbo and Other Nigerian Languages, University of Nigeria, Nsukka, who are between 18-30 years. The sample size was derived using Taro Yamane's formula for sampling technique which is:

$$n = \frac{N}{1 + N(e)^2} \tag{1}$$

n is the sample size

N is the total number of respondents or population

e is the error margin which 0.05

$$n = \frac{134}{1+134(0.05)^2} = 100.374 \quad (2)$$

Out of this sample size, 18 students representing 18% of the population did not return their responses, out of the returned responses 7 students representing 7% of the population were invalid bringing the number to 75 students which is 75% of the population. A hearing test (using Marcin Masalski's hearing test app, version 2.0.26) was first conducted to ensure that the subjects have an average hearing ability. Data (stimuli) consist of 6 online lectures on English sounds (and pronunciation) from 2 British speakers, 2 American speakers and 2 Nigerian speakers; one male, one female each. Tokens from the lectures within 13 – 14 seconds were segmented and analyzed on Praat at 44.1 KHz sampling rate. Subjects are made to listen to only the audio aspect of the lectures and thereafter asked to write down what they heard and understood from the lectures. Tokens were played between 2 to 6 times to enable subjects hear the tokens properly in order to write down what they heard. Descriptive statistics is used to get the mean f0, while the duration, intensity, Syllable count and speaking rate are based on the praat analysis.

3. Results and Discussion

3.1 Perception Analysis

The results from the figure below show the hearing level of the subjects for both ears.

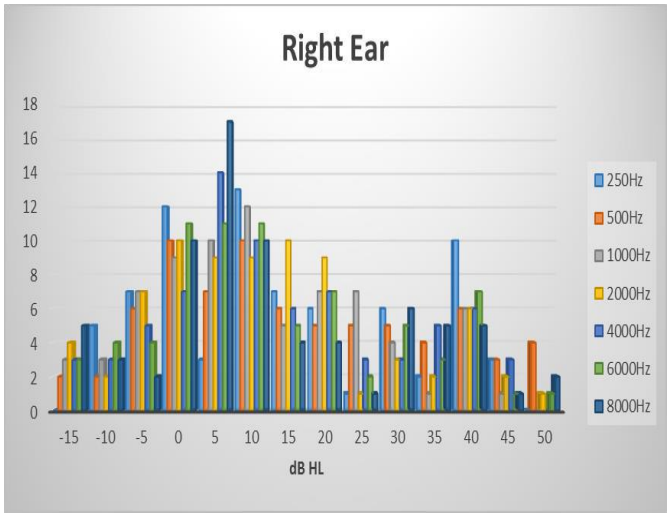


Figure 1 – Hearing level for the right ear

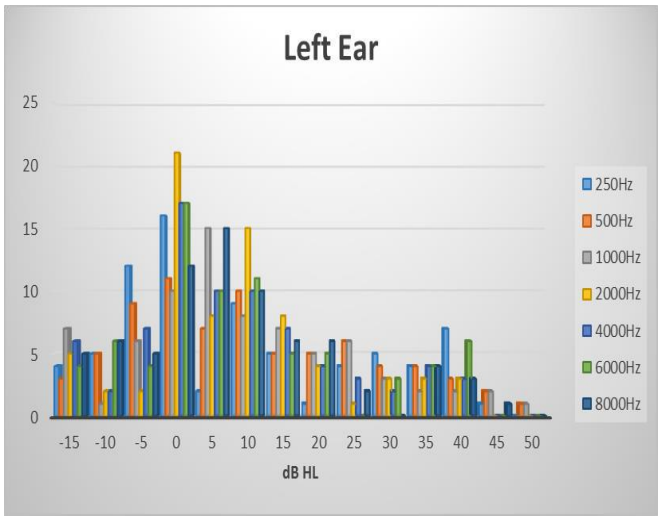


Figure 2 – Hearing level for the left ear

The results from the figures above show that majority of the subjects had good hearing level as many of them were within the hearing level threshold of -5dB to 20dB which is the normal hearing level as can be seen below.

Average hearing threshold at 500, 1000, 2000 and 4000 Hz

Right ear: -1 dB HL (Normal hearing)

Left ear: 3 dB HL (Normal hearing)

Normal hearing ≤ 20 dB HL

Mild impairment 21–40 dB HL

Moderate impairment 41–60 dB HL

Severe impairment 61–80 dB HL

Profound impairment including deafness > 80 dB HL

World Health Organization (WHO) grade of hearing loss

Better ear: -1 dB HL (Normal hearing)

Performance: No or very slight hearing problems. Able to hear whispers.

Recommendations: No recomendations.

Figure 3 – Average Hearing Threshold

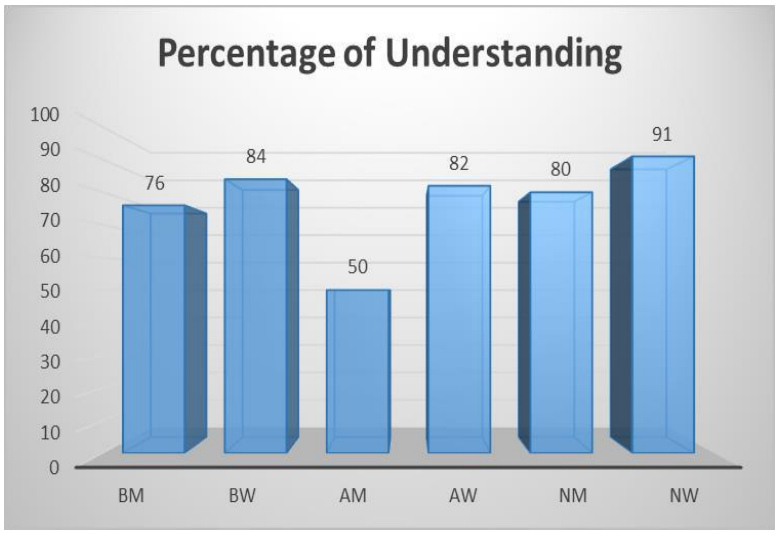


Figure 4 – Subjects ratings of level of comprehension of tokens

From figure 4 above, it can be observed based on subjects' ratings, the Nigerian woman (NW) has the highest level of understanding which is 91%. The next was the British Woman (BW) with 84% then, the American Woman (AW) with 82%. The Nigerian Man (NM) follows with 80%, the British Man (BM) with 76% and lastly, the American Man (AM) with 50%. This implies that the Nigerian Woman's lecture was most understood while the American Man's lecture was least understood.

3.2 Acoustic Analysis

These are the Praat spectrograms from the acoustic analysis of the tokens of the online lectures. See figures 5 – 10 below.

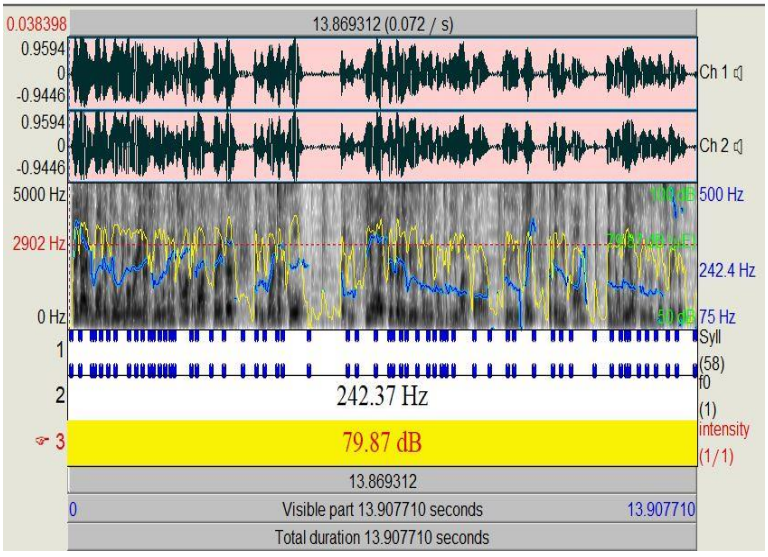


Figure 5 – Praat analysis of the token for the American Female speaker

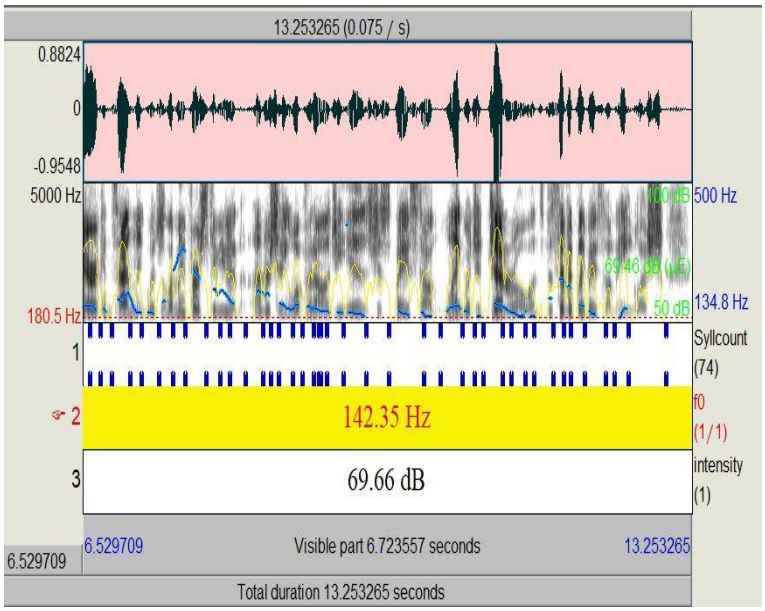


Figure 6 – Praat analysis of the token for the American Male speaker

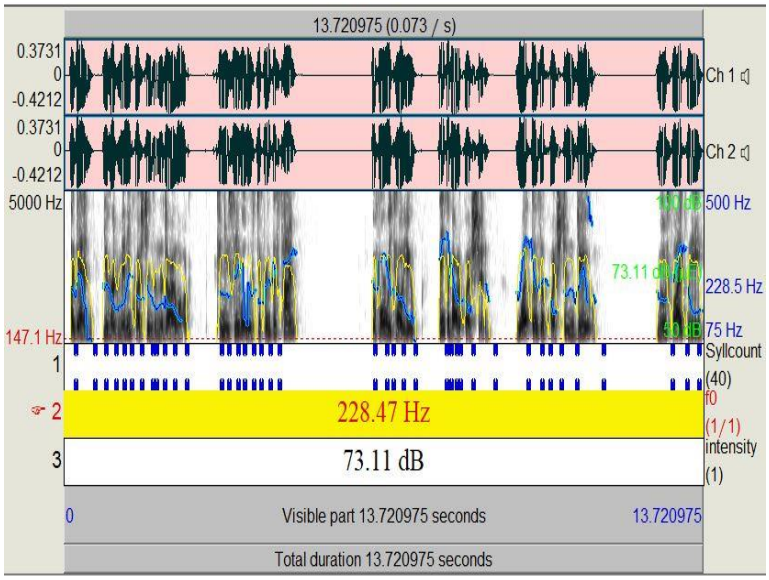


Figure 7 – Praat analysis of the token for the British Female speaker

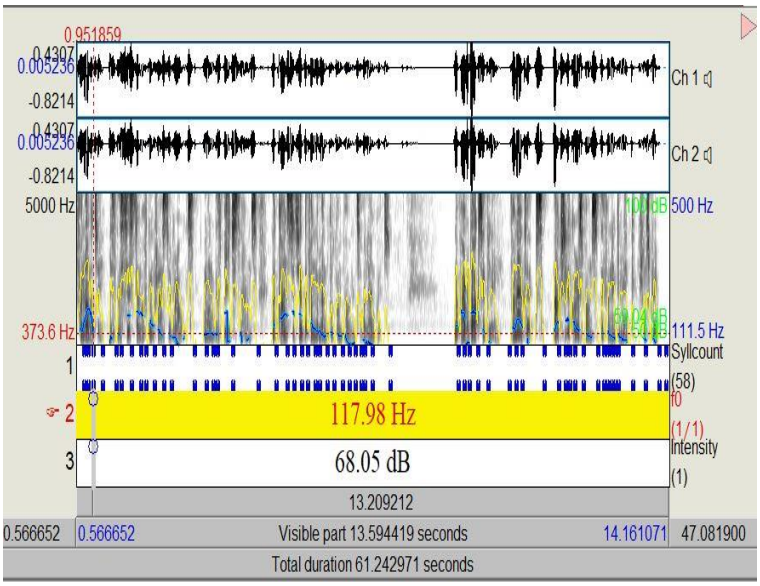


Figure 8 – Praat analysis of the token for the British Male speaker

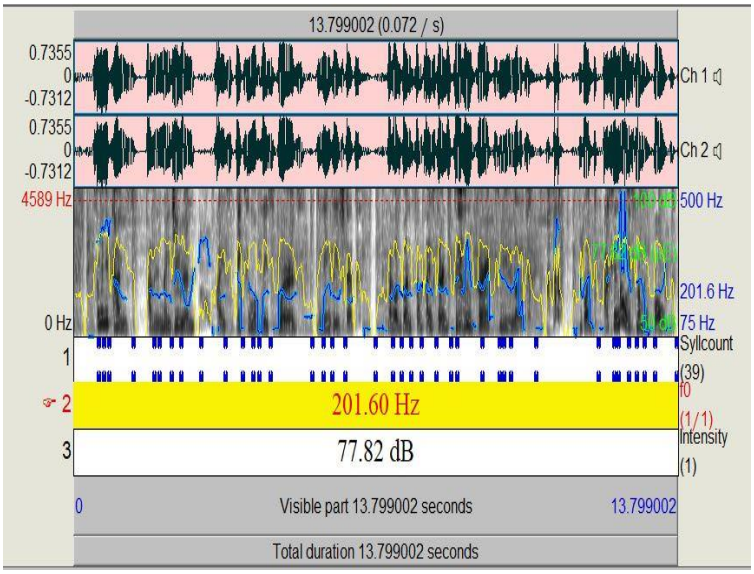


Figure 9 – Praat analysis of the token for the Nigerian Female speaker

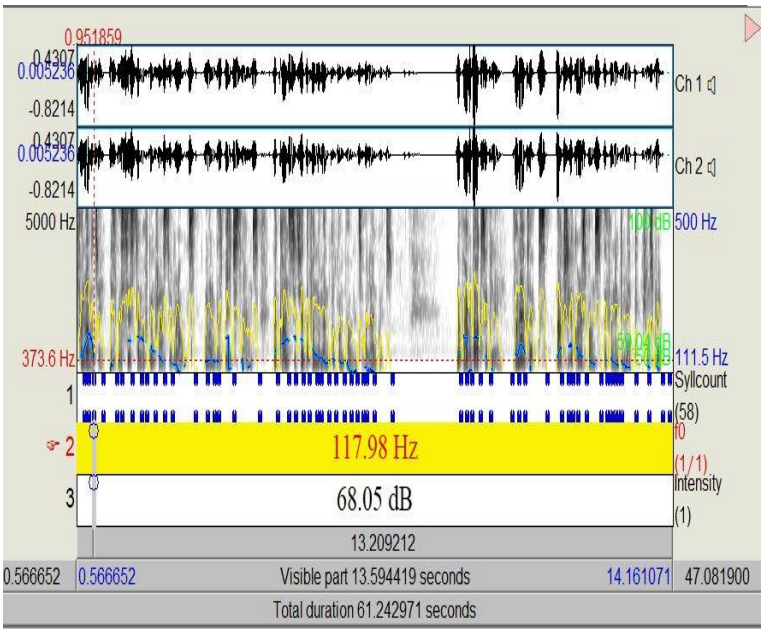


Figure 10 – Praat analysis of the token for the Nigerian Male speaker

Below is a table that collates the acoustic measurements of the different tokens as can be seen in the spectrograms above.

Table 1: Acoustic Analysis of Tokens

Speakers	Mean f_0 (Hz)	Intensity (dB)	Duration (sec)	Syllable Count	Speaking Rate (Syll/sec)
American Female	242.37	79.87	13.86	58	4.18
American Male	142.35	69.66	13.25	74	4.58
British Female	228.47	73.11	13.72	40	2.92
British Male	117.98	68.05	13.59	58	4.27
Nigerian Female	201.60	77.82	13.79	39	2.83
Nigerian Male	138.91	74.90	13.83	55	3.98

From the table above, it can be observed that the mean f_0 of the tokens were within the normal pitch range of male and female speakers generally but for the intensity, we see that the female speakers had higher intensity than their male counterparts which implies that the female speakers were more audible than the male speakers and so this may have contributed to why there was a higher percentage of subjects' understanding of the tokens of the female speakers than their male counterparts. Again, the syllable count per duration shows that the speaking rate of the Nigerian Woman is 2.83 syllables per seconds, followed by the British woman which is 2.92 syllables per seconds, then the Nigerian man which is 3.98 syllables per seconds, the American Woman which is 4.18 syllables per seconds, the British man which is 4.27 syllables per seconds and finally the American Man which is 4.58 syllables per seconds. Based on these results, it affirms that the Nigerian Woman's speaking rate is the slowest while the American Man speaking rate is the

fastest. By and large, we can also infer that the female speakers were generally slower in comparison to the male speakers contrary to popular opinion that females speaker rate are generally faster than that of males (though this is not one of our research objectives, it could open up avenue for further research).

4. SUMMARY OF FINDINGS

From the findings of the study, it can be seen that the subjects who rated the tokens have normal hearing ability, therefore, lack of perception or comprehension cannot be hinged on subjects' hearing ability.

Secondly, the subjects' ratings of understanding show that token got from the utterances of the Nigeria Female speaker was most understood than that of others while the American male speaker was the least understood.

Thirdly, the acoustic analysis of the speakers' utterances show that the intensity of the speakers' voice and the speaking rate of the speakers contributed largely to subjects' perception and comprehension of the tokens. This means that the louder and slower a person speaks, irrespective of the variety of English, it is easy for the audience to follow through with little effort. Put simply, clarity of speech, loudness and a moderate speaking rate plays more role than accent in the perception and comprehension of the online lectures. This agrees with Guyer, Fabrigar and Vaughan-Johnston (2019) (10) but disagrees with what has been reported in previous studies (c.f. Harding, 2008 (11); Lev-Ari, Van Heugten, and Peperkamp, 2017 (12); Wang, 2018 (13)).

5. CONCLUSIONS

Online lectures just like in-class or in-person lectures aim at not just teaching but effectively communicating in such a way that the audience understands with little effort. Since English has become one of the global languages, it is not surprising to see it being used in many online lectures. The study observes that if these lectures are not recorded, a lot of wealth of information may bypass the audience as they may have no way of replaying what has been said. Subsequently, in order to better communicate to an array of speakers; who speak English as their second language (L2), it is important that the intensity and speaking rate of speakers (teachers) be moderated to enable audience perception and comprehension of the online lectures.

ACKNOWLEDGEMENTS

Authors acknowledge that they received no financial support for this research.

REFERENCES

1. Munro M. Nonsegmental Factors in Foreign Accent: Ratings of Filtered Speech. Studies in Second Language Acquisition. 1995:17(1): 17-34.
2. Mitterer H, Cho T, Kim S. How does prosody influence speech categorization? Journal of Phonetics. 2016:*54*, 68-79.
3. Mitterer H, Kim S, Cho T. The glottal stop between segmental and suprasegmental processing: The case of Maltese. Journal of Memory and Language. 2019:108,104034.
4. Steffman J. Intonational structure mediates speech rate normalization in the perception of segmental

categories. *Journal of Phonetics*. 2019:*74*,114-129.
5. Steffman J, Jun SA. Effects of prosodic structure versus durational context on the perception of segmental categories: The case of focus realization. Proc 19th International Congress of Phonetic Sciences. Melboure, Australia; 2019
6. Steffman J, Jun, SA. Prosodic Prominence in Speech Perception: The Influence of Focus Prosody on the Perception of Durational and Spectral Cues. Proc 38th West Coast Conference on Formal Linguistics. Cascadilla Proceedings Project. 2021. p. 406-416.
7. Holt LL, Lotto AJ, Diehl RL. Auditory discontinuities interact with categorization: Implications for speech perception. *J. Acoust. Soc. Am*. 2004:116 (3):1763–1773.
8. Mattys SL, Davis MH, Bradlow AR, Scott SK. Speech recognition in adverse conditions: A review. Language and Cognitive Processes. 2012:27(7–8):953–978.
9. Penga ZE, Wang LM. Listening effort by native and nonnative listeners due to noise, reverberation, and talker foreign accent during English speech perception. Journal of Speech, Language, and Hearing Research. 2019: 62: 1068–1081.
10. Guyer JJ, Fabrigar, LR, Vaughan-Johnston TI. Speech rate, intonation, and pitch: Investigating the bias and cue effects of vocal confidence on persuasion. Personality and Social Psychology Bulletin. 2019:45(3): 389-405.
11. Harding L. Accent and academic listening assessment: A study of test-taker perceptions. Melbourne Papers in Language Testing. 2008:13(1):1-33.
12. Lev-Ari S, Van Heugten M, Peperkamp S. Relative difficulty of understanding foreign accents as a marker of proficiency. Cognitive Science, 2017:41(4):1106-1118.
13. Wang Z. The Effect of Accent on Listening Comprehension: Chinese L2 Learners' Perceptions and Attitudes. THAITESOL Journal. 2017:31 (2):47-71.

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022, in Gyeongju, Korea

Extended abstract

ABS-0642

Spatial unmasking in reverberation by children who use bilateral cochlear implants

Z. Ellen Peng[1]

[1] Boys Town National Research Hospital, USA

ABSTRACT

In addition to better speech perception, bilateral cochlear implants (BiCIs) provide some access to spatial hearing for individuals with profound hearing loss. One benefit of having access to auditory spatial cues is the ability to better attend to a target talker in competing background babbles, if the target is spatially separated from the speech masker. Such speech benefit, known as spatial unmasking, is measured as the difference in speech reception thresholds between two target-masker spatial conditions: co-located versus separated. Several studies have suggested that pediatric BiCI users have very little access to binaural cues, namely interaural time and level differences, but primarily rely on the monaural head shadow cue for spatial unmasking. In real-world listening with reverberation, the reflected sounds impose additional distortions on both binaural and monaural cues in the acoustic signals before CI processing. How spatial unmasking is affected by reverberant degradation among pediatric BiCI users is unknown. In this work, we use a novel measure (Peng & Litovsky, 2021) to assess spatial unmasking among pediatric BiCI users in simulated reverberant environments that mimic typical indoor learning environments. Results will be presented and provide further indications on the feasibility of individualized fitting strategies for pediatric patients with BiCIs.

Keywords: Spatial hearing, bilateral cochlear implant, children

1. INTRODUCTION

For children and adults with normal hearing, spatial hearing provides access to auditory cues that are critical for attending to target speech in the background of competing babbles in spatial unmasking (Brown et al., 2010; Buss et al., 2017; Corbin et al., 2017; Griffin et al., 2019; Litovsky, 2005). Spatial unmasking refers to the intelligibility benefit when the babble masker is spatially separated from the target from co-location. For children with severe to profound hearing loss, bilateral cochlear implants (BiCIs) provide some access to spatial hearing without complete restoration (Bennett & Litovsky, 2019; Grieco-Calub & Litovsky, 2010, 2012; Hess et al., 2018; Misurelli & Litovsky, 2012, 2015; Zheng et al., 2015). Recent work showed that while some children with BiCIs can receive an intelligibility benefit with only access to interaural timing and level difference cues relying on binaural processing, others needed additional monaural cues (Peng & Litovsky, 2021) – all demonstrating the ability to take advantage of spatial separation between the target and masker, as well as bilateral listening through two CIs.

In this work, we expand beyond the basic question of whether children with BiCIs have access to spatial unmasking in ideal anechoic auditory environments. Specifically, we measure spatial unmasking in reverberant environments similar to a standard classroom, where critical verbal communication occurs, among children with BiCIs. In indoor spaces, reverberation is the cascade of sound energy reflected from interior surfaces that arrives at the listener shortly after the direct sound from the sound source. For adults with NH, reverberation has been shown to reduce the effect of spatial unmasking (Kidd et al., 2005), with a smaller target intelligibility benefit when the masker is spatially separated. Reflected sounds not only reduce ILD, but also distorts the signal envelope that leads to more difficult extraction of ongoing ITD (Rennies & Kidd, 2018). The acoustic-side distortions from reverberation present additional challenges for CI processing to preserve envelope-based cues. This pilot work aims at understanding how reverberant distortions affect the access to

[1] Ellen.Peng@boystown.org

spatial unmasking among a group of children and young adults who experienced perilingual deafness and received BiCIs at an early age.

2. METHOD

2.1 Participants

Four children and two young adults who are early CI recipients (before three years of age) participated in this study. Table 1 shows their demographics, including ages of CI activation and bilateral experiences. All children had at least three years of regular daily BiCI use, as verified by data logging with their audiologists at the time of testing. All experimental procedures were approved by Institutional Review Board at Boys Town National Research Hospital.

Table 1. Demographics of bilateral cochlear implant users.

Subject ID	Age (yr; mo)	Age of First CI Activation (yr; mo)	Age of Second CI Activation (yr; mo)	Duration of Bilateral Experience (yr; mo)	Etiology	Speech Processor
ACI001	18; 2	2; 11	2; 11	15; 3	Unknown, Hereditary	AB, Naida
ACI002	19; 0	2; 2	4: 6	14; 6	Unknown	AB, Naida
CCI002	14; 8	1; 1	2; 9	11; 11	Meningitis	AB, Marvel
CCIAC	7; 9	1; 5	4; 9	3; 0	Unknown, Hereditary	AB, Marvel
CCIAF	7; 11	1; 0	1; 2	6; 9	Unknown, Hereditary	AB, Marvel
CCIAG	7; 11	1; 0	1; 0	6; 11	Waardenburg	AB, Marvel

Note: AB = Advanced Bionics.

2.2 Experimental Task and Procedure

Spatialized sounds were directly streamed into the speech processors for stereo playback. Spatial unmasking was assessed in two acoustic conditions, first in anechoic and again in reverberation. The reverberant condition was created by convolving the binaural room impulse responses simulated in ODEON for a virtual classroom with speech materials from virtual sound sources at various spatial locations. The virtual classroom had approximately 0.6 s reverberation time, based on ANSI S12.60 classroom acoustics recommendation. The default set of head-related transfer functions from ODEON was used in creating the room impulse responses to ensure all listeners had access to the same auditory cues for comparison.

All BiCI users began with testing speech in quiet with the target located at either -90° or +90° azimuth (to the left or right of the listener in the virtual environment) to determine if they had a better ear. The speech reception threshold (SRT) was measured by adaptively changing the target level using a one-down-one-up procedure (Levitt, 1971). We determined the target position for subsequent testing based on better SRT in quiet.

To measure spatial unmasking, we used a recently developed metric of minimum angular separation (MAS). The MAS is the smallest spatial separation needed between the target and masker for a 20% intelligibility increase. For each acoustic condition, we first measured SRT in babble noise at 50% word-level accuracy using the one-down-one-up adaptive procedure when the target and masker were co-located. Next, we displaced the masker at 180° separation with the target (i.e., on the opposite hemifield), and adaptively changed the separation using the two-down-one-up procedure for 70.7% accuracy. Both steps terminated after six reversals. The final angular separation from the second step

is the MAS.

For the older children, the target speech were open-set sentences spoken by a female (Dawson et al., 2013). The masker was same-sex two-talker (i.e., second female) babble of continuous discourse (e.g., science stories) presented at the fixed level of 55 dB SPL. On each trial, the listener was presented with a short sentence with three keywords and asked to repeat back words heard. Scoring was based on the number of keywords responded correctly. Three children were 7 years old at the time of testing and had limited experience completing open-set sentence recognition tasks. Hence, the same experiment was conducted using a word recognition task. For these younger children, the target was close-set rhyming words spoken by a male talker led a prompt "Show me the…", with four-talker babble (i.e., male and female) as masker at 55 dB SPL. On each trial, the child chose one of six pictures displayed on the screen that matched the spoken word.

3. RESULTS

Preliminary results are presented herein. Figure 1 shows SRTs at 50% accuracy for each child in anechoic and reverberation. Most children, particularly those who completed the task using open-set sentences, require a positive signal-to-noise ratio (SNR) for achieving 50% accuracy with co-located target and masker. The two children, Subject ID CCIAF and CCIAG, were able to complete the close-set word identification task using a negative SNR. However, when we proceeded to measure MAS, these two children was unable to produce measurable threshold using negative SNR in some conditions. Subsequently, we increased the target speech level to 0 dB SNR in anechoic for CCIAG, 3 dB SNR in reverberation for CCIAF. However, child CCIAF produced measurable MAS in anechoic with a negative SNR.

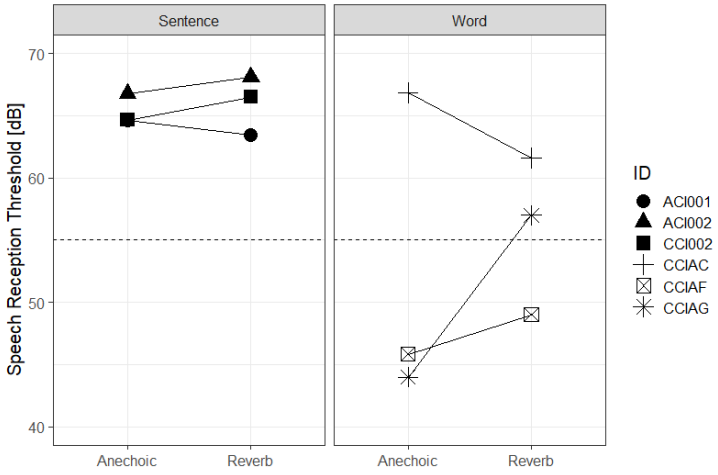


Figure 1. Speech reception thresholds (SRTs) as the target speech level for 50% word accuracy in anechoic and in reverberation for each child. Dashed line denotes the masker presentation level at 55 dB SPL.

In this task of measuring MAS with a target at either -90° or +90°, a final MAS ≤90° suggests the listener's ability to complete the task to gain 20% target intelligibility using only a combination of ITD and ILD cues for spatial unmasking. When MAS >90°, listeners need additional head shadow for spatial unmasking. Figure 2 illustrates MAS for these children with BiCIs. All participants demonstrate larger MAS (poorer spatial unmasking) when tested in reverberation than in anechoic. For the three participants tested using the sentence task, one adult (Subject ID ACI001) needed an additional head shadow cue when reverberation was introduced. The other two BiCI users (Subject ID ACI002 and CCI002) experienced the reverberant distortion in ITD/ILD cues with larger MAS, but were still able to complete the task using binaural cues only.

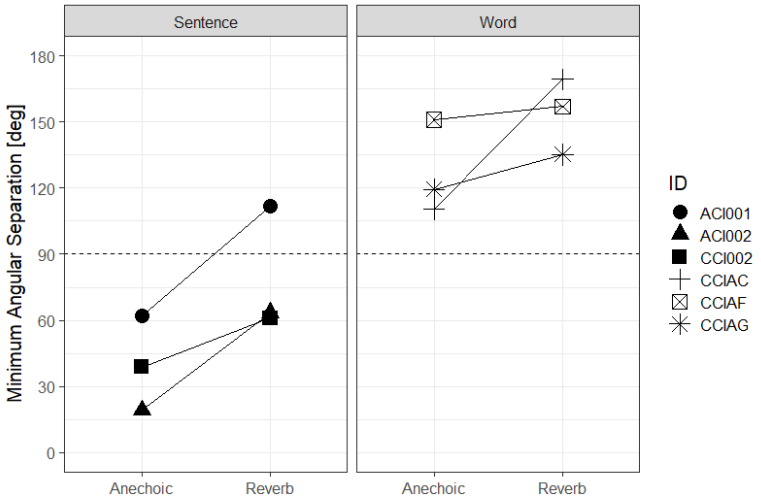


Figure 2. Minimum angular separation in anechoic and in low reverberation. Dashed line at 90° denotes the boundary distinguishing the use between binaural versus monaural cues for spatial unmasking.

4. DISCUSSION

Our preliminary results provided encouraging evidences that, even for young school-age children with BiCIs, they have access to spatial unmasking in both anechoic and low reverberant environments albeit needing a much larger MAS than older children and young adults. The two children, who produced negative SRT for co-located target and maskers but needed positive SNR for MAS, provided some insights on the limitation of close-set word recognition task. But their data are consistent with the idea that positive SNR is critical for speech-in-speech understanding by CI users. For younger children with BiCIs, at least around 7-8 years of age, are still developing spatial hearing abilities and primarily use head shadow cues for spatial unmasking. For the three oldest BiCI users, two participants completing the task solely relying on ITD/ILD cues and the third needing additional head shadow cues only in reverberation. Such individual differences in their MAS outcomes call for consideration for personalized fitting of speech processors for optimal listening outcomes.

ACKNOWLEDGEMENTS

The author is grateful to all families with children with cochlear implants who traveled to Omaha, Nebraska to participate in this study. We are grateful for assistance from Dr. Jeffrey Simmons for audiological assistance and recruitment effort, as well as Drake Hintz and Abigail Mollison for assistance during data collection. This work is in part supported by the Hearing Health Foundation and NIH NIDCD (R21DC020532).

REFERENCES

Bennett, E. E., & Litovsky, R. Y. (2019). Sound Localization in Toddlers with Normal Hearing and with Bilateral Cochlear Implants Revealed Through a Novel “Reaching for Sound” Task. *Journal of the American Academy of Audiology*, *14*(5), 1–14. https://doi.org/10.3766/jaaa18092

Brown, D. K., Cameron, S., Martin, J. S., Watson, C., & Dillon, H. (2010). The North American listening in spatialized noise - Sentences test (NA LiSN-S): Normative data and test-retest reliability studies for adolescents and young adults. *Journal of the American Academy of Audiology*, *21*(10), 629–641. https://doi.org/10.3766/jaaa.21.10.3

Buss, E., Leibold, L. J., Porter, H. L., & Grose, J. H. (2017). Speech recognition in one- and two-talker maskers in school-age children and adults: Development of perceptual masking and glimpsing. *The*

Journal of the Acoustical Society of America, *141*(4), 2650–2660. https://doi.org/10.1121/1.4979936

Corbin, N. E., Buss, E., & Leibold, L. J. (2017). Spatial Release From Masking in Children: Effects of Simulated Unilateral Hearing Loss. *Ear & Hearing*, *38*(2), 223–235. https://doi.org/10.1097/AUD.0000000000000376

Dawson, P. W., Hersbach, A. A., & Swanson, B. A. (2013). An adaptive Australian Sentence Test in Noise (AuSTIN). *Ear and Hearing*, *34*(5), 592–600. https://doi.org/10.1097/AUD.0b013e31828576fb

Grieco-Calub, T. M., & Litovsky, R. Y. (2010). Sound Localization Skills in Children Who Use Bilateral Cochlear Implants and in Children With Normal Acoustic Hearing. *Ear and Hearing*, *31*(5), 645–656. https://doi.org/10.1097/AUD.0b013e3181e50a1d

Grieco-Calub, T. M., & Litovsky, R. Y. (2012). Spatial Acuity in 2-to-3-Year-Old Children With Normal Acoustic Hearing, Unilateral Cochlear Implants, and Bilateral Cochlear Implants. *Ear and Hearing*, *33*(5), 561–572. https://doi.org/10.1097/AUD.0b013e31824c7801

Griffin, A. M., Poissant, S. F., & Freyman, R. L. (2019). Speech-in-Noise and Quality-of-Life Measures in School-Aged Children With Normal Hearing and With Unilateral Hearing Loss. *Ear & Hearing*, *40*(4), 887–904. https://doi.org/10.1097/AUD.0000000000000667

Hess, C. L., Misurelli, S. M., & Litovsky, R. Y. (2018). Spatial Release From Masking in 2-Year-Olds With Normal Hearing and With Bilateral Cochlear Implants. *Trends in Hearing*, *22*, 233121651877556. https://doi.org/10.1177/2331216518775567

Kidd, G., Mason, C. R., Brughera, A., & Hartmann, W. M. (2005). The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Acta Acustica United with Acustica*, *91*, 526–536.

Levitt, H. (1971). Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477. https://doi.org/10.1121/1.1912375

Litovsky, R. Y. (2005). Speech intelligibility and spatial release from masking in young children. *The Journal of the Acoustical Society of America*, *117*(5), 3091–3099. https://doi.org/10.1121/1.1873913

Misurelli, S. M., & Litovsky, R. Y. (2012). Spatial release from masking in children with normal hearing and with bilateral cochlear implants: Effect of interferer asymmetry. *The Journal of the Acoustical Society of America*, *132*(1), 380–391. https://doi.org/10.1121/1.4725760

Misurelli, S. M., & Litovsky, R. Y. (2015). Spatial release from masking in children with bilateral cochlear implants and with normal hearing: Effect of target-interferer similarity. *The Journal of the Acoustical Society of America*, *138*(1), 319–331. https://doi.org/10.1121/1.4922777

Peng, Z. E., & Litovsky, R. Y. (2021). Novel Approaches to Measure Spatial Release From Masking in Children With Bilateral Cochlear Implants. *Ear & Hearing*, *Publish Ah*. https://doi.org/10.1097/AUD.0000000000001080

Rennies, J., & Kidd, G. (2018). Benefit of binaural listening as revealed by speech intelligibility and listening effort. *The Journal of the Acoustical Society of America*, *144*(4), 2147–2159. https://doi.org/10.1121/1.5057114

Zheng, Y., Godar, S. P., & Litovsky, R. Y. (2015). Development of Sound Localization Strategies in Children with Bilateral Cochlear Implants. *PLOS ONE*, *10*(8), e0135790.

https://doi.org/10.1371/journal.pone.0135790

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022, in Gyeongju, Korea

Extended abstract

ABS-0823

Assessing the intelligibility of degraded speech through automatic speech recognition (ASR) systems

Wenhui Sun[1]; Tianyi Zhu[3]; Jiaxin Gao[2], Qingling Meng[4], Nai Ding[1,2*]

[1] Research Center for Applied Mathematics and Machine Intelligence, Research Institute of Basic Theories, Zhejiang Lab, Hangzhou, 311121, China

[2] Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrument Sciences, Zhejiang University, China

[3] College of Control Science and Engineering, Zhejiang University, China

[4] Acoustic Laboratory, School of Physics and Optoelectronics, South China University of Technology, China

ABSTRACT

The performance of automatic speech recognition (ASR) systems has been greatly improved in recent years, and has reached human-level performance for clean speech. The current study investigated whether state-of-the-art ASR systems can predict how well humans recognize degraded speech. In the first experiment, the ASR systems were trained on clean speech and tested using degraded speech. The degraded speech conditions included simulations of speech in common adverse listening conditions such as speech in babble noise, additive white noise and reverberated speech. The results showed that ASR speech recognition accuracy was lower than human performance in all tested degraded speech conditions. In the second experiment, the ASR systems were trained and tested using each kind of degraded speech, and the ASR accuracy on degraded speech was much higher in Experiment 2 than Experiment 1. ASR accuracy in Experiment 2 was comparable to human-level performance in most conditions. Therefore, when properly trained, ASR is a potential method to assess intelligibility of degraded speech.

Keywords: Automatic Speech Recognition, Speech Intelligibility, Degraded speech

1. INTRODUCTION

The performance of automatic speech recognition (ASR) systems has been greatly improved in recent years as the deep neural network (DNN) was adopted to ASR tasks a decade ago [1,2]. Recently, the end-to-end (E2E) techniques were employed on ASR tasks, which could directly translate a speech input into an output token [2]. It was with the advantage of simplifying the speech recognition pipeline compared with traditional hybrid ASR models. The E2E models reached the state-of-the-art results in most benchmarks in terms of recognition accuracies [2], which also achieved human-level performance in clean speech recognition tasks [3]. Whereas, the performance of ASR models was not robust and was sensitive to small perturbations in input sequences [4]. In most cases, human performance was less affected by variation of distorted speech. The current study was intended to investigate the performance of ASR models in recognizing distorted speeches compared with humans. If the state-of-the-art ASR model could reach human-level performance in degraded speech recognition, it indicated that the design of E2E ASR models was efficient and these models might also be a potential method for accessing intelligibility of degraded speech.

With the aim of evaluating the performance of ASR systems in degraded speech, two kinds of experiments were performed. In the first experiment, the ASR systems were trained with clean speeches and tested on degraded speeches. The degraded speech conditions included simulation of common adverse listening conditions such as speech in babble noise, additive white noise and reverberation. Results of the first experiment might help us to observe performance of ASR models with exposure to unseen samples. In the second experiment, the ASR systems were trained and tested

in each kind degraded speech. Human experiments corresponding to each kind of degrade speech were also conducted for comparison, in which the remoting web-based listening test were performed due to the ongoing impact of the COVID-19 pandemic. Our results demonstrated that there were still room for improvement of the ASR models in generalization ability. At the same time, the state-of-the-art ASR system might be a potential method for accessing the speech intelligibility of degrade speech when properly trained.

2. MATERIALS AND METHODS

Thirty-two (21 females; mean age, 23.2±2.8 years old) normal hearing native-Mandarin listeners were recruited mostly from Zhejiang university. Each subject was informed of the content of the experiments and received monetary payments after the experiment for their participation.

Speech materials in our experiments were selected from a popular open source Chinese Mandarin speech corpus AISHELL-1[3], which was consisted of 150-hour training set, a 10-hour development set and a 5-hour test set. This corpus was widely used for Mandarin speech recognition and building automatic speech recognition systems. Eighty-four utterances were randomly extracted without repeating from test set of AISHELL-1 to make a new test set for our experiments, which were employed in both human listening experiments and ASR model experiments. Specifically, the training and development sets were the same as AISHELL-1 corpus in our experiments. There were three kinds of degraded speech involved in our experiments, including white noise, babble noise and reverberation noise. The audio spectrograms of these conditions are illustrated in Figure 1.

In the human experiments, an average result of all subjects was used as the final accuracy corresponding to specific degraded speech condition. For ASR experiments, a popular state-of-the-art open-source ASR model, ESPnet combined with Conformer, was utilized in our experiments, which was consisted of a connectionist temporal classification (CTC)/attention architecture with conformer as the encoder [5]. In the first experiment, the ASR model was trained on the clean speech and tested using each kind of degraded speech; in the second ASR experiment, the ASR models were trained and tested on each kind of degraded speech.

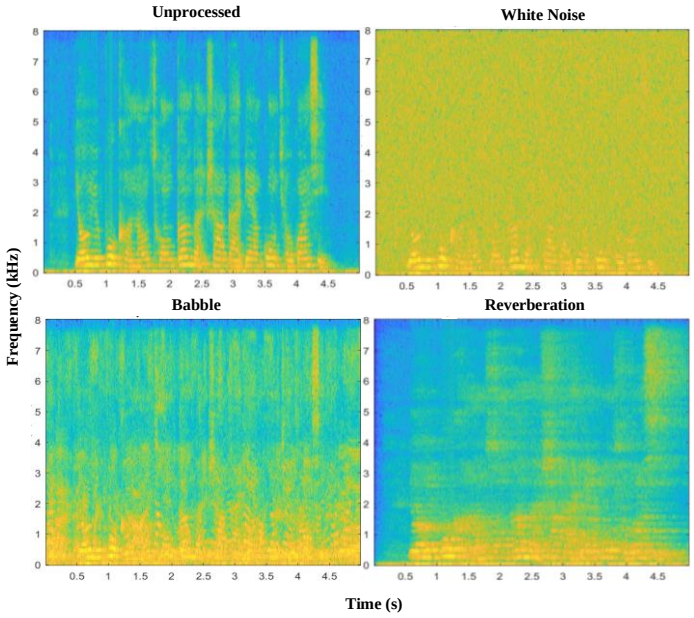


Figure 1: Spectrograms of unprocessed, white noise, babble and reverberation noise.

3. RESULTS AND CONCLUSION

We have tested performance of humans and ASR model in recognizing several distorted speeches, including white noise, babble and reverberation conditions. In the first experiments, performance of human is better than of models, as the models are trained on unprocessed speeches and tested on each kind of degraded speech. In the second experiments, when the models were trained and test on the same kind of degraded speeches, performances were improved and close to that of humans. Our results indicate that the current ASR systems may be a potential approach for achieving the speech intelligibility of degraded speeches when properly trained.

REFERENCES

1. Miao, Yajie, Mohammad Gowayyed, and Florian Metze. "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding." 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015.
2 Li, Jinyu. "Recent advances in end-to-end automatic speech recognition." APSIPA Transactions on Signal and Information Processing 11.1 (2022).
3 H. Bu, J. Du, X. Na, B. Wu and H. Zheng, "AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline," 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), 2017, pp. 1-5, doi: 10.1109/ICSDA.2017.8384449.
4 Ebrahimi, J.; Rao, A.; Lowd, D.; and Dou, D. 2017. Hotflip: White-box adversarial examples for text classification. arXiv preprint arXiv:1712.06751.
5 Guo, Pengcheng, et al. "Recent developments on espnet toolkit boosted by conformer." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

Extended abstract

ABS-0397

Performance of speech enhancement algorithms for native Mandarin listeners on English perception

Yunqi C. Zhang[1]*, C.T. Justine Hui[1], Yusuke Hioka[1]†, Catherine I. Watson[2]

[1]Acoustics Research Centre, Department of Mechanical and Mechatronics Engineering, University of Auckland, New Zealand

[2]Department of Electrical, Computer and Software Engineering, University of Auckland, New Zealand

ABSTRACT

Non-native listeners are usually more disadvantaged in listening to speech in noisy environments compared to native listeners. Many previous studies have investigated the differences in intelligibility between native and non-native listeners and the possible causes of the degradation due to language backgrounds. To improve the quality and intelligibility of speech, speech enhancement is widely used for signal processing for native listeners. However, no previous study has investigated how the technique would specifically improve the intelligibility of non-native listeners. This study evaluated noisy speech sentences that were processed by five commonly used single-channel speech enhancement algorithms: Wiener filter, Subspace, Nonnegative Matrix Factorization, Conv-TasNet, and Deep Complex U-Net. A listening test was carried out on native New Zealand English speakers and native Mandarin Chinese speakers. English sentences from the Bamford-Kowal-Bench corpus masked by either speech-shaped noise or babble noise were used. The enhanced speech sentences were played through headphones using an online questionnaire and the participants were asked to transcribe the sentences. The response of the native and non-native listeners was compared and analysed.

Keywords: Speech enhancement, Speech intelligibility, Non-native listeners, Mandarin, English

1 INTRODUCTION

Speech is the primary medium of communication in our daily life. The process of receiving information from a speech by the auditory system is called speech perception. Listening under adverse conditions such as under noise has been proved to be detrimental to speech perception and significantly reduce communication effectiveness [1]. It has been well studied that non-native listeners are disadvantaged more than native listeners when listening in adverse conditions [2]. The possible reasons for the degradation in perception of non-native listeners compared to native listeners under noisy environments are due to their language background [3], language proficiency, and the type of noise [4].

Speech enhancement (SE), is a technology that improves the speech quality or intelligibility of speech signals contaminated by noise. Although a variety of SE algorithms has been designed to improve the intelligibility of speech in noise, previous studies that evaluated the performance of SE algorithms were on native listeners of the speech sentences. There is no such method proposed for non-native listeners specifically. This prompts us to study how the existing SE algorithms would improve the speech perception in noise of non-native listeners. This study investigates the intelligibility of noisy English sentences that are processed by various existing SE algorithms between native and non-native English listeners. A subjective test was run to measure the intelligibility of the enhanced speech. Since the non-native listeners' first language would affect their perception phonetically and syntactically [3], this study only focuses on listeners with Mandarin as their first language. Participants of native listeners were recruited in New Zealand while those of non-native listeners were recruited in China.

2 TEST DESIGN

2.1 Speech enhancement algorithms

To examine the effectiveness of the technology with breadth, the SE algorithms tested need to be 1) widely used, 2) with a relatively simple concept and implementation, 3) easy to access the original implementation

*yzhb694@aucklanduni.ac.nz
†yusuke.hioka@ieee.org

(i.e. code is publicly available). Based on these criteria, five single-channel speech enhancement algorithms were selected in this study, which were: ***a priori* Wiener filter (WF)** [5], **generalised subspace (SS)** [6], **unsupervised Bayesian non-negative matrix factorisation (NMF)** [7], **Conv-TasNet (Conv)** [8], and **deep complex U-net (U-net)** [9] algorithms. The selected algorithms were implemented by programmes publicly available; where the MATLAB codes implementing WF and SS were acquired from the appendix in [10], the programmes implementing NMF [11] and U-net [12] were provided by the authors of the literature, and Conv was implemented through Asteroid [13].

Some parameters of the algorithms were adjusted to optimise the speech enhancement performance. For *a priori* WF, the smoothing factors of β and μ [5] were set to 0.96 and 0.99, respectively. As for the SS algorithm, the Lagrange multiplier μ [6] from the gain function was set to 0.99. For NMF, the size of the main buffer N_1 [7] was set to 50. The Conv and U-net methods are end-to-end trained methods that do not support tuning.

2.2 Stimuli

The Bamford-Kowal-Bench (BKB) sentences in the Speech Perception Assessment New Zealand (SPANZ) corpus [14] were used for the test. These sentences are semantically meaningful, have simple syntactic structures, and use words of a high frequency of occurrence, which is also suitable for testing non-native English listeners. The SPANZ corpus re-recorded the sentences in New Zealand accent and modified the words and sentences to expressions that are commonly used in New Zealand (e.g. "vacation" was replaced by "holiday"). Each sentence contains three to four keywords that are marked for measuring the speech intelligibility.

The current study simulated speech under stationary and non-stationary noise, which were represented by speech-shaped noise (SSN) and babble noise, respectively. A SSN was generated by shaping a white noise by the spectral shape of the sum of 288 SPANZ BKB sentences. For the babble noise, the NOISEX-92 babble noise [15] was used. For both noise types, a section with the same length as the target speech signal was truncated to mix with the clean speech at certain signal-to-noise (SNR) levels. The starting point of the noise was always the same regardless of its length. The length of each speech audio was between 3 - 4 seconds, where some audio files had a longer quiet section at the end of the signal. To avoid significant floor/ceiling effects in both participant groups (see Section 2.3), different sets of input SNR were chosen for each noise type: 0, -3, -6 dB for the babble noise, and -3, -6, -9 dB for the SSN.

The noise was added to the clean speech at specific SNR levels defined in Section 2.2 to generate the noisy speech, where the speech signals were normalised to have a root mean square of 1. The stimuli of enhanced speech were generated by feeding the noisy speech signal to each of the selected SE algorithms stated in Section 2.1 at a sampling rate of 16 kHz. Apart from the speech processed by five SE algorithms, the unprocessed noisy speech signals were also included as a baseline reference. For every participant, three sentences from a single condition were selected randomly for repetition purposes. Hence, a total of 108 BKB sentences (5 SE algorithms (+ 1 noisy speech) $\times$ 2 noise types $\times$ 3 SNR levels $\times$ 3 repetitions) were used in the test.

2.3 Participants

Forty-nine normal-hearing adults participated in the experiment, where 20 were native New Zealand English (NZE) listeners recruited in New Zealand (NZE group) and 19 were native Mandarin listeners recruited in Mainland China (CC group). All participants in both groups had been to tertiary level education.

Participants in the NZE group (mean age = 23.3, sd = 4.34, 6 female, 14 male) arrived in New Zealand before the age of 12 years old. One participant reported that they lived in Sweden until 15 years old and moved to New Zealand. The participant had lived in New Zealand since then and reported English as their mother tongue and home language. The response of this participant showed no abnormality compared to the others, hence the response was not removed.

Participants in the CC group (mean age = 23.5, sd = 1.63, 10 female, 8 male, 1 preferred not to say) were based in Mainland China and had never lived in an English-speaking country for more than one year. All participants learnt either British (n = 6) or/and American English (n = 13). One participant studied in Hong Kong from the age of 23 to 24. Most Chinese universities require students to pass at least one English certificate. English major students sit Test of English Majors (TEM) Band 4 and 8, and the others sit College English Test (CET) Band 4 and 6, with higher Band numbers indicating rising difficulty. Only two of the participants in the CC group had not received any relevant certificates. The rest had sat at least one of CET 4 (n = 2), CET 6 (n = 9), TEM 8 (n = 1), International English Language Testing System (IELTS) (n = 4, two scored 6.5 and two not mentioned), and Test of English as a Foreign Language (TOFEL) (n = 1). Therefore, participants from CC group should have learnt enough vocabularies to understand the BKB sentences.

2.4 Test Procedure

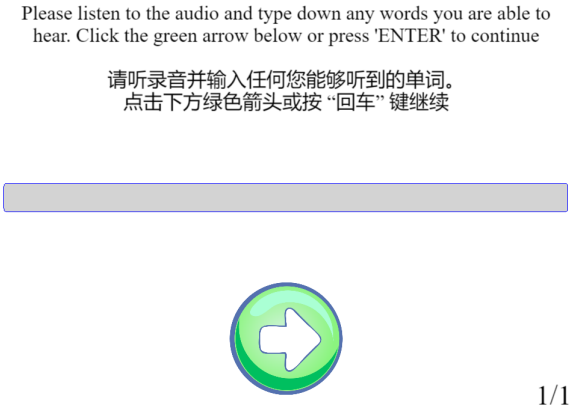


Figure 1. A screenshot of the GUI of the online test.

The test platform was developed on PsychoPy [16] which is available online through Pavlovia. The participants were asked to wear headphones throughout the experiment and transcribe any words they could hear from the automatically played audio via the GUI as shown in Figure 1 by a keyboard. A practice test with one unprocessed noisy signal from the BKB sentences at -9 dB SNR was given for the participants to get familiar with the test format. They were asked to maximise their volume to the highest they could tolerate and not to change the device's volume setting once the formal test started.

For the formal test, each participant responded to 108 sentences mentioned in Section 2.2. Each sentence was only played once. The sentences were identical for every participant but with random parameter combinations and were played in an arbitrarily randomised order. The test took around 30 minutes, and the participants were given a break after listening to 54 sentences.

The participants were also required to fill out a demographics form to collect their gender, age, mother tongue, type of English speaking, language background, proficient language, educational background, and English language certificates (for CC group).

The study was approved by the University of Auckland Human Participants Ethics Committee (UAH-PEC24202). The participants were rewarded *koha* (a New Zealand Māori custom which means gift, donation or contribution) as compensation.

2.5 Marking Rubic

To quantify speech intelligibility, participants' responses were marked manually following the suggestions in the SPANZ corpus [14]. The root of the word is marked rather than the whole word. Homonyms such as "buy" and "by", "two" and "to" were not penalised. For NZE group, words related to the diphthong merger /iə/ and /eə/ (e.g. "ear" and "air") [17] were given the same marks. The ratio of correct words (i.e. proportion correct) that indicates the speech intelligibility was calculated for each participant from the words across three sentences under the same condition. The proportion correct (ranging from 0 to 1) of each condition was calculated and treated as a single data point.

2.6 Statistical Analysis

The marked results were analysed by the linear mixed effect (LME) model in R by the lme4 package [18]. Interactions among multiple variables were hypothesised and tested by the lmerTest package [19] using the step function. Likelihood ratio test was carried out to check the significance of the fixed effect by comparing the p-values of models with and without each effect. Under the post-hoc pairwise comparisons by the emmeans package [20], the difference between every two effects was tested, and the p-values were adjusted by the Tukey HSD method. The pairs providing p-values being less than 0.05 were considered as significant.

Since the two noise types were tested under different SNR levels, separate models were developed for the results collected from each noise type. In summary, the fixed effects were SE algorithm (noisy, Conv, NMF, SS, U-net, WF), input SNR (0, -3, -6 or -3, -6, -9), and nativeness (NZE, CC). The random effect was the participant ID.

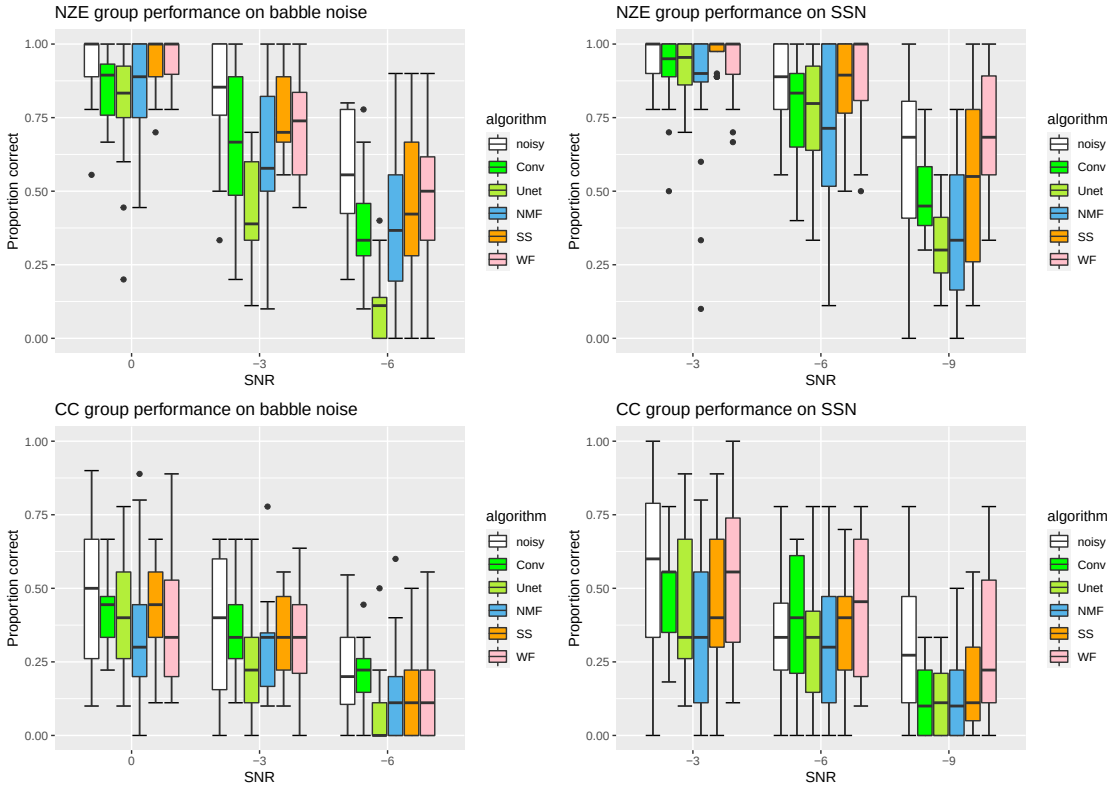


Figure 2. Proportion correct of responses of NZE group and CC group under babble and speech-shaped noise processed by 5 different speech enhancement algorithms.

3 Results

Figure 2 presents the proportion correct of NZE group and CC group under different noise types in boxplots. The median line indicates that the noisy signal outperforms most SE algorithms under every condition. Exceptions occurred for CC group when signals were enhanced by Conv-TasNet under both noise types at -6 dB SNR, and when processed by subspace and WF under SSN at -6 dB SNR. For NZE group, only signals enhanced by WF under SSN at -6 dB SNR had higher intelligibility than the noisy speech.

To ensure the speech enhancement algorithms have indeed performed as expectd, their performance was measured by the improvement of two objective metrics, namely SDR [21] and STOI [22], which are known to represent the quality and the intelligibility of the target speech, respectively. The average improvement of SDR and STOI for each SE method are reported in Table 1. It shows that NMF degraded the intelligibility of speech under both noise types, while SS and WF show degradation in intelligibility under babble noise.

The LME model for the SSN showed three-way interaction among SE algorithm, input SNR, and nativeness of participants, including the random effect of the participant ID. The effect of each variable was significant: algorithm ($\chi^2(30) = 128.58$, $p < .0001$), input SNR level ($\chi^2(24) = 394.29$, $p < .0001$), nativeness ($\chi^2(18) = 109.16$, $p < .0001$) from the likelihood ratio comparison. The model for the babble noise showed significant two way interactions between the algorithm and SNR level ($\chi^2(10) = 20.641$, $p = 0.02374$), algorithm and nativeness ($\chi^2(5) = 37.622$, $p < .0001$), and SNR and nativeness ($\chi^2(2) = 54.609$, $p < .0001$).

The post-hoc pairwise contrasts between the NZE and CC groups were always significant for all conditions, which indicates that the performance of CC group was always significantly worse than that of the NZE group regardless of the type of SE algorithm and SNR level under both noise types. The largest difference under babble noise was observed in WF at 0 dB SNR, where the proportion correct estimate was 0.54 ($t.ratio = 12.52$, $p < .0001$). The lowest estimate was observed in U-net at -6 dB SNR, with an estimate of 0.1 ($t.ratio = 2.27$, $p < .0001$). Under SSN, the highest proportion correct estimate was 0.54 ($t.ratio = 8.13$, $p < .0001$), which was observed in noisy speech under -6 dB SNR, and the lowest proportion correct estimate of 0.18 ($t.ratio = 2.76$, $p < .0001$) was observed in U-net but at -9 dB SNR.

The speech intelligibility in terms of the proportion correct decreases as the SNR level decreases. The contrasts between input SNR levels for NZE group were always significant under babble noise, for CC group, they were insignificant between SNR -3 dB and 0 dB for most algorithms apart from U-net and WF.

Table 1. Average improvements of SDR and STOI for each speech enhancement algorithm. SDR_impr = (output SDR) - (input SDR), STOI_impr = (output STOI) - (input STOI).

	Babble noise			SSN		
Algorithm	Input SNR	SDR_impr	STOI_impr	Input SNR	SDR_impr	STOI_impr
Conv	0	**7.92**	**0.13**	-3	**9.77**	**0.10**
	-3	**7.46**	**0.14**	-6	**6.82**	**0.11**
	-6	**5.59**	**0.10**	-9	**3.63**	**0.08**
NMF	0	**0.24**	-0.03	-3	-0.41	-0.03
	-3	**0.88**	-0.01	-6	**0.01**	-0.02
	-6	**1.45**	**0.002**	-9	**0.22**	-0.005
SS	0	**3.59**	**0.01**	-3	**4.79**	**0.03**
	-3	**3.13**	-0.001	-6	**4.50**	**0.02**
	-6	**2.52**	-0.02	-9	**3.98**	**0.007**
U-net	0	**2.54**	**0.07**	-3	**4.52**	**0.08**
	-3	**3.02**	**0.06**	-6	**4.83**	**0.08**
	-6	**2.65**	**0.002**	-9	**4.32**	**0.3**
WF	0	**2.29**	**0.001**	-3	**4.39**	**0.03**
	-3	**2.17**	-0.003	-6	**4.00**	**0.03**
	-6	**2.00**	-0.01	-9	**3.44**	**0.02**

As for SSN, both NZE and CC groups observed insignificant contrasts between SNR -3 and -6 dB for most conditions.

Post-hoc pairwise contrasts between SE algorithms suggest that different algorithms have little effect on performance for CC group under both noise types. According to the number of significant contrasts, the algorithms were more effective under babble noise (19 pairs for NZE group, 3 pairs for CC group) than under SSN (10 pairs for NZE group, 5 pairs for CC group) for NZE group, but hardly effective for CC group. Specifically, NMF and U-net showed significant degradation compared to the noisy speech for all SNR levels under babble noise for NZE group. It is evident that U-net performed consistently the worst among all of the algorithms, except when it is compared with Conv-TasNet under SSN at -9 dB for NZE group. Moreover, it was significant that the noisy signal always had higher intelligibility than the signals enhanced by Conv-TasNet, NMF, and U-net. WF always obtained a higher score than Conv-TasNet, NMF, and U-net under SSN, indicating a more effective performance of WF in terms of improving speech intelligibility.

The waveform of the enhanced speech found that NMF and SS reduce the magnitude of the enhanced speech compared to the noisy speech, leading to lower energy, i.e. lower volume signals. Conv amplified the magnitude of speech, resulting in "louder" signals which may favour the listeners' perception.

4 Conclusion

The present study investigated the performance of different existing speech enhancement algorithms on native New Zealand English listeners in New Zealand and native Mandarin listeners based in China. Speech signals contaminated by speech-shaped noise and babble noise under different SNR levels were processed by five speech enhancement algorithms and their speech intelligibility was tested subjectively. The result of the subjective test suggested that current algorithms showed little intelligibility improvement and even degradation over noisy speech under negative SNR levels.

5 Acknowledgements

We appreciate our participants for their participation and Dr. Suzanne Purdy for providing the SPANZ corpus.

REFERENCES

[1] Sven L. Mattys et al. “Speech recognition in adverse conditions: A review”. In: *Language and Cognitive Processes* 27.7 (Sept. 1, 2012), pp. 953–978.

[2] Martin Cooke, M. L. Garcia Lecumberri, and Jon Barker. “The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception”. In: *The Journal of the Acoustical Society of America* 123.1 (Jan. 1, 2008), pp. 414–427.

[3] Dollen Tabri, Kim Michelle Smith Abou Chacra, and Tim Pring. “Speech perception in noise by monolingual, bilingual and trilingual listeners”. In: *International Journal of Language & Communication Disorders* 46.4 (2011), pp. 411–422.

[4] Jin Zhang et al. “How Noise and Language Proficiency Influence Speech Recognition by Individual Non-Native Listeners”. In: *PLOS ONE* 9.11 (Nov. 19, 2014), e113386.

[5] P. Scalart and J.V. Filho. “Speech enhancement based on a priori signal to noise estimation”. In: *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. Vol. 2. May 1996, 629–632 vol. 2.

[6] Yi Hu and P.C. Loizou. “A generalized subspace approach for enhancing speech corrupted by colored noise”. In: *IEEE Transactions on Speech and Audio Processing* 11.4 (July 2003), pp. 334–341.

[7] Nasser Mohammadiha, Paris Smaragdis, and Arne Leijon. “Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.10 (Oct. 2013), pp. 2140–2151.

[8] Yi Luo and Nima Mesgarani. “Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.8 (Aug. 2019), pp. 1256–1266.

[9] Hyeong-Seok Choi et al. “PHASE-AWARE SPEECH ENHANCEMENT WITH DEEP COMPLEX U-NET”. In: *ICLR* (2019), p. 20.

[10] Philipos Loizou. *Speech Enhancement: Theory and Practice*. June 7, 2007. ISBN: 978-0-429-13373-2.

[11] mohammadiha. *bnmf*. original-date: 2018-08-23T11:48:02Z. Jan. 25, 2022. URL: https://github.com/mohammadiha/bnmf.

[12] ILJI CHOI. *Source Separation*. original-date: 2019-07-23T23:44:42Z. Aug. 1, 2022. URL: https://github.com/AppleHolic/source_separation.

[13] Manuel Pariente et al. “Asteroid: the PyTorch-based audio source separation toolkit for researchers”. In: *Proc. Interspeech*. 2020.

[14] Jae-Hyun Kim and Suzanne Purdy. “Speech Perception Assessments New Zealand (SPANZ)”. In: *New Zealand Audiological Society Bulletin* 24 (Jan. 1, 2014), pp. 9–16.

[15] Andrew Varga and Herman J. M. Steeneken. “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems”. In: *Speech Communication* 12.3 (July 1, 1993), pp. 247–251.

[16] Jonathan Peirce et al. “PsychoPy2: Experiments in behavior made easy”. In: *Behavior Research Methods* 51.1 (Feb. 1, 2019), pp. 195–203.

[17] Margaret A. Maclagan and Elizabeth Gordon. “Out of the AIR and into the EAR: Another view of the New Zealand diphthong merger”. In: *Language Variation and Change* 8.1 (Mar. 1996), pp. 125–147.

[18] Douglas Bates et al. “Fitting Linear Mixed-Effects Models Using lme4”. In: *Journal of Statistical Software* 67 (Oct. 7, 2015), pp. 1–48.

[19] Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. “lmerTest Package: Tests in Linear Mixed Effects Models”. In: *Journal of Statistical Software* 82 (Dec. 6, 2017), pp. 1–26.

[20] Russell V. Lenth et al. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. Version 1.7.5. June 22, 2022. URL: https://CRAN.R-project.org/package=emmeans.

[21] E. Vincent, R. Gribonval, and C. Fevotte. “Performance measurement in blind audio source separation”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.4 (July 2006), pp. 1462–1469.

[22] Cees H. Taal et al. “An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (Sept. 2011), pp. 2125–2136.

PROCEEDINGS of the
24th International Congress on Acoustics
October 24 to 28, 2022 in Gyeongju, Korea

ABS-0438

Effects of language development on sound symbolic associations

Tomomi WATANABE[1]; Momoko HISHITANI[1]; Sachi ITAGAKI[1]; Shota A. MURAI[1, 3]; Haruka SUZUKI[1]; Yuji SHIMA[2]; Kazuko SHINOHARA[4]; Ryoko UNO[4]; Kohta I. KOBAYASI[1*]

[1]Graduate School of Life and Medical Sciences, Doshisha University, Japan

[2]Faculty of Life and Medical Sciences, Doshisha University, Japan

[3]International Research Center for Neurointelligence, The University of Tokyo Institutes for Advanced Study, Japan

[4]Institute of Engineering, Tokyo University of Agriculture and Technology, Japan

ABSTRACT

The cross-modal mapping between sound and perceptual images (e.g., shape) is known as sound symbolism. A recent study suggested that cross-modal mappings between certain consonants and perception of hardness differ between different native languages in adults. Exploration regarding how language development affects these associations is required. This study tested adult Japanese and English speakers to see whether language-specific sound symbolism comes from language-specific pronunciation. Then, we tested Japanese infants, aged 9, 12, and 15 months to see whether the effect of sound symbolism changes in development.Hard and soft images were presented as video stimuli. As audio stimuli, two-mora nonsense words recorded with English- or Japanese-like pronunciations and with consonants /p/, /b/, /k/, /g/ and vowels /i/, /o/ (e.g., “bigi”) were presented. We confirmed that adult English speakers deemed voiceless consonants (e.g., /p/) harder than voiced consonants (e.g., /b/), contesting adult Japanese speakers’ perception. The hardness in sound symbolic associations when hearing Japanese was constant between 9 and 15 months old infants. However, the associations when listening to English differed from those of Japanese infants at 9 months and were the same at 15 months. This suggests that sound symbolic association develops in a native-language-specific manner between 9 and 15 months.

Keywords: Sound symbolism, Cross modal mappings, Language development

1. INTRODUCTION

Sound symbolism refers to the cross-modal correspondence between sounds and specific images (1), which is considered a universal phenomenon (2–4), as represented by the “takete-maluma” effect (5) and the “bouba-kiki” effect(6). Recently, native language-specific sound symbolism has been reported (7–9). Symbolic sound association between consonant voicing and hardness changes across languages. For example, according to a previous study (8), monolingual Japanese speakers would relate voiced consonants (/b/, /g/) to hardness and voiceless consonants (/k/, /p/) to softness, whereas monolingual English speakers would relate voiceless consonants (/k/, /p/) to hardness and voiced consonants (/b/, /g/) to softness. However, the effect of native language experience on sound symbolism is unclear. Previous studies on sound symbolism have examined the acoustic and articulatory bases (10–13). The frequency code hypothesis (14) has been considered one of the mechanisms to explain sound symbolism from acoustic bases. The articulation bias is also examined since articulation motion creates a sense of sound symbolism, such as the oral cavity’s size and vowel

backness (1,11,15).

This study confirmed whether the sound symbolic association between consonants and hardness differs between English and Japanese native speakers, and examined whether the sound symbolic association alters across English and Japanese pronunciations in speech stimuli. If hardness is evoked by acoustic properties specific to the speaker's native language, it would depend on the language-specific pronunciation of spoken sounds. In addition, a comparable study of Japanese infants was conducted to test whether the effect of language-specific pronunciation on sound symbolism changes in different stages of development. The results can provide insight into the acquisition of native language-specific sound symbolism.

2. EXPERIMENT1: A FORCED-CHOICE TASK IN ADULTS

We conducted an online experiment using the Gorilla Experiment Builder (www.gorilla.sc). Data from native Japanese speakers were collected between 23 October 2020 to 21 December 2020. Data from native English speakers were collected from 04 February 2022 to 27 February 2022.

2.1 Participants

Thirty native Japanese speakers participated in the experiment (Men = 15, women = 15; age = 19–25, mean = 21.9). Eleven native English speakers participated in the experiment (Men = 7, women = 4; age = 20-39, mean = 26; American English = 9, Australian English = 1, and Canadian English, 1). All participants provided written informed consent before the experiment started. The experiment was approved by the ethics committee of Doshisha University.

2.2 Audio Stimuli

Two-mora nonsense words comprising consonants (/b/, /g/, /k/, /p/) and vowels (/i/, /o/) (Table.1) were used. The set of stimuli contained eight nonsense words, as shown in Table 1, which were pronounced by two bilingual speakers of Japanese and English. Consequently, we had a total of 16 audio stimuli.

2.3 Visual Stimuli

Three visual stimuli were prepared along with the audio stimuli. Fig.1a shows a sample of the video images. Two white lines were moved and rotated on the black screen. Two different movements were designed to induce impressions of hardness and softness.

2.4 Procedure

In each trial, audio and video stimuli (Fig.1a) were simultaneously presented twice after fixation cross presentation for 0.25 seconds. The video stimuli contained two movements of white lines which were displayed on the left and right sides of the screen for four seconds. The left and right positions of the two movements were switched in each trial. Subsequently, the participants chose a video stimulus that matched each sound stimulus (Fig.1b). A total of 32 trials (16 trials × 2 sets) were conducted.

Table.1 Audio stimuli. All words were spoken with English- and Japanese-like pronunciation

Vowel	Voiced consonant	Voiceless consonant
/i/	bigi, gibi	kipi, piki
/o/	bogo, gobo	kopo, poko

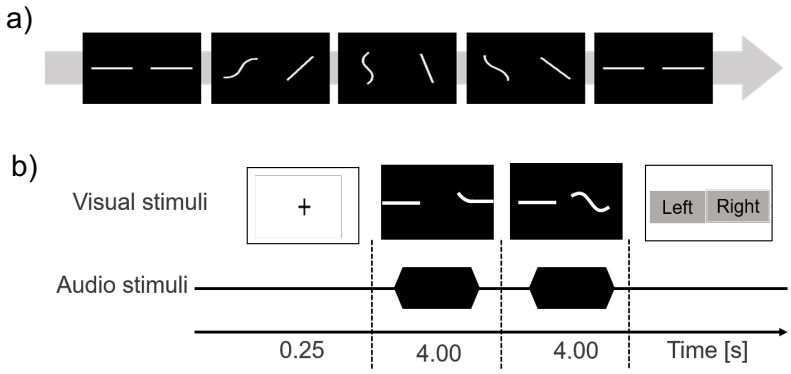


Fig. 1 (a) One of the visual stimuli. White lines moved and rotated. Two movements induced impressions of hardness and softness. (b) Sequence of a trial: A fixation cross was presented for 0.25 s, followed by the first presentation of audio and video stimuli for four seconds. The left and right positions of visual stimuli were then changed and presented again for four seconds with the audio stimulus. Participants chose videos that matched each sound stimulus.

2.5 Results

In the first analysis, we used a generalized linear model (GLM) to test the relationships between consonants (/b/and /g/ vs. /k/and /p/) and participants' native languages (Japanese vs. English). The results showed that there was an interaction between consonants and native language (Consonant × Native language: $p = 0.024$), indicating that native English listeners considered /b/ and /g/ as softer than native Japanese listeners did (Fig. 2).

In the second analysis, we investigated the relationships between pronunciation type (Japanese vs. English) and the effects of voicing consonants using GLM. The results showed no interaction between pronunciation and consonant types (Fig. 3 and 4). In other words, the symbolic sound association of hardness in adults was affected by the participants' native language, not by the pronunciation type that they listened to.

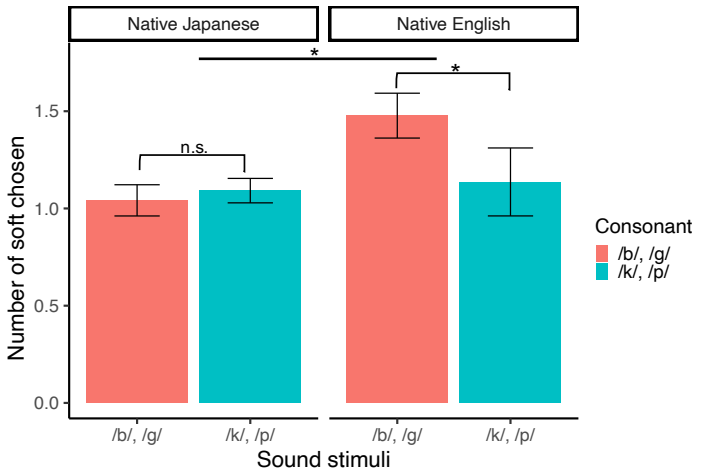


Fig.2 The average number of responses as soft by type of listener's native language. The vertical axis is the average number of times the participants selected "soft" for particular words (e.g., /bigi/) in two trials. The horizontal axis indicates the consonant codition: voiced (/b/, /g/) or voiceless (/k/, /p/). Error bars indicate standard error of the mean (SEM). Left: Native Japanese listener, Right：Native English listener. *p < .05

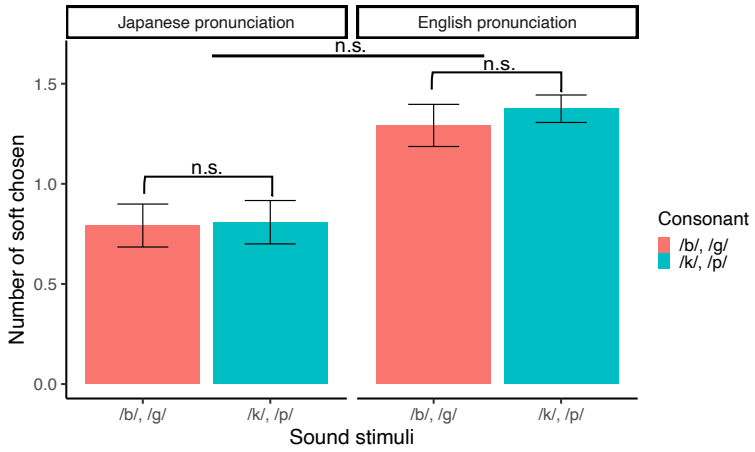


Fig.3 The average number of responses as soft when Japanese participants listened to the word (e.g., /bigi/) in Japanese-like or English-like pronunciation. The vertical axis indicates the average number of times the participants selected "soft" in two trials. The horizontal axis indicates the consonant condition: voiced (/b/, /g/) or voiceless (/k/, /p/). Left: Japanese-like pronunciation, Right: English-like pronunciation. Error bars indicate SEM.

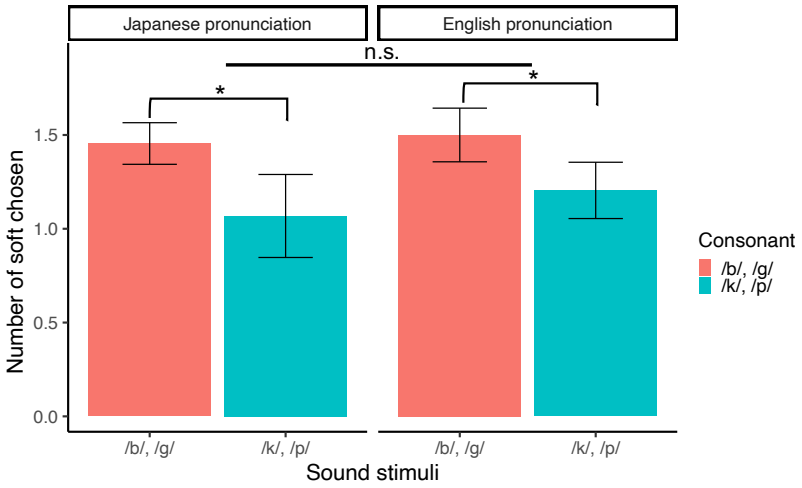


Fig.4 The average number of responses as soft when English participants listen to the word (e.g., /bigi/) in Japanese-like or English-like pronunciation. The vertical axis is the average number of times the participants selected "soft" in two trials. The horizontal axis indicates the consonant condition: voiced (/b/, /g/) or voiceless (/k/, /p/). Japanese-like pronunciation, Right: English-like pronunciation. Error bars indicate SEM. *p < .05

3. EXPERIMENT 2: PREFERENCE LOOKING EXPERIMENT IN INFANTS

3.1 Method

The same stimuli as in Experiment 1 were used. We measured the time each participant looking at each video stimulus. A total of 32 trials (16 trials × 2 sets) were conducted (Fig. 5).

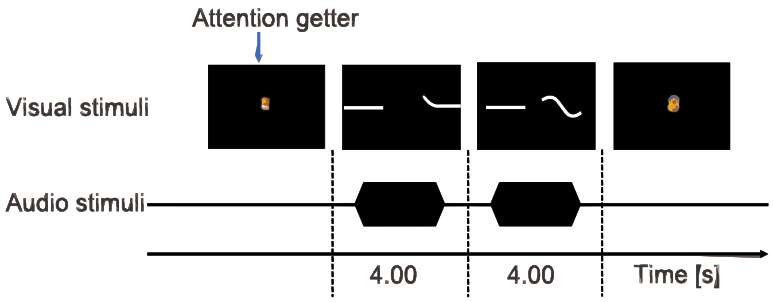


Fig.5 Sequence of a trial. We presented an animation called “Attention getter” to direct infants' attention to the display, followed by the first presentation of audio and video stimuli for four seconds. The left and right positions of visual stimuli were then changed and presented again for four seconds with the audio stimulus. We measured looking time.

3.2 Participants

Sixty-five infants participated, of which 57 were included in the study: 21 participants were 9 months old (ten male and eleven female participants), 18 participants were 12 months old (nine male and nine female participants), and 18 participants were 15 months old (eight male and ten female participants). Eight participants withdrew from the experiment because of fussiness. All the participants were Japanese monolinguals with monolingual Japanese parents. The experiment was approved by the Research Ethics Committee of Doshisha University, and the infants' parents provided written informed consent before the experiment.

3.3 Apparatus

A loudspeaker (Computer MusicMonitor, BOSE) was placed behind the display (1800FP,DELL). An eye tracking device (Tobii Pro X3-120, Tobii) was placed under the display. Each infant was positioned 65 cm from the display on their parents’ lap in a forward-facing position.

3.4 Data

The data of infants (34 out of 57) whose looking time was less than 30% of the total time, or who looked at only one side of the screen, were excluded. As a result, a total of 23 infants were included in the analysis: seven infants aged 9 months (three male and four female participants), six infants aged 12 months (two male four and female participants), and ten infants aged 15 months old (four male and six female participants).

3.5 Results

We analyzed the relationships between pronunciation type (Japanese vs. English) and consonant type based on voicing (/b, d/ vs. /p, k/) using GLM. The interaction between the consonant type and age was significant when the sound stimuli in English-like pronunciation were presented, whereas it was not significant when the sound stimuli in Japanese-like pronunciation were presented (Japanese: Consonant × Month $p = 0.163$, English: Consonant × Month $p = 0.003$). The results indicated that the effect of voicing in consonants differed between Japanese-like and English-like pronunciations in the case of infants aged 9 months, but the effect was not significantly different among 12 and 15 month old infants (Fig. 6).

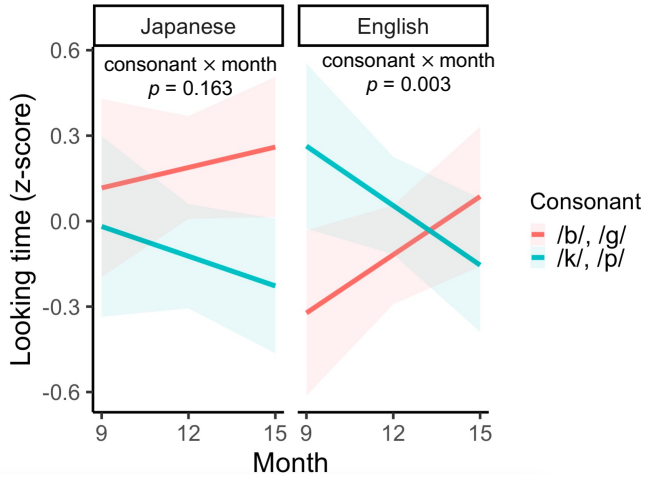


Fig. 6 The effects of consonants were compared among three age groups. The vertical axis shows looking time (z-score). The horizontal axis indicates ages in months. Shaded areas are confidence intervals based on standard errors. Left: Japanese-like pronunciation, Right: English-like pronunciation.

4. DISCUSSION

This study investigated whether Japanese and English adult speakers and Japanese infants at 9, 12, and 15 months of age differ in sensitivity to cross-modal mappings. Experiment 1 tested whether language-specific sound symbolism among adults comes from their linguistic experience (i.e., their first language) or from the acoustic features of pronunciation in speech stimuli. The results showed a significant interaction between consonant types and listeners' native language, supporting native language-specific sound symbolism, but not acoustically motivated pronunciation-based sound symbolism. This aligns with the findings of the previous study (8), which used disyllabic pseudowords in VCVC form (e.g., /apap/) and a four-point Likert scale from “very hard” to “very soft.” Since the current study employed a different task in which participants were asked to select a combination of audio stimuli in CVCV form (e.g., /piki/) and video stimuli (soft or hard movement), our findings provide further support for the native language-specific sound symbolism of hardness.

Surprisingly, no interaction between consonant and pronunciation types was observed. This suggests that adults’ symbolic sound intuition is possibly influenced by their native language rather than the physical features of the sounds that they hear. Although our initial hypothesis was that native language-specific sound symbolism may be motivated by acoustic information, as the frequency code hypothesis suggests, our results showed that images of hardness did not change depending on the pronunciation of the stimuli; rather, they were dependent on the native language of the participants. Thus, sound symbolism might be modulated by linguistic experiences or the acquired system of one’s native language rather than actual acoustic inputs on the spot.

In Experiment 2, a similar method was applied among 9–15-month-old Japanese infants to test the effect of language development on language-specific sound symbolism, and a significant effect of age was found. Although Japanese-like sounds did not affect symbolic associations for hardness, English-like pronunciation affected 9-month-old infants. This effect disappeared in the 15-month-old infants. These results suggest that sound symbolic intuitions in infants may be altered according to native language roughly between 9 and 15 months of age.

5. CONCLUSION

For adults, symbolic images of hardness associated with voicing consonants differed between English and Japanese. In addition, the effect of English-like sounds on Japanese infants’ looking behavior decreased from 9 to 15 months of age. Hence, language-specific sound symbolism might be formed in infants at around 12 months of age when their sensitivity to non-native phonemes decreases

(16). The process of infant language development is crucial for the formation of sound symbolic intuition. Thus, language-specific sound symbolism is influenced by language development.

ACKNOWLEDGMENTS

This study was supported by Japan Society for the Promotion of Science grant numbers 18H05089 and 22K18661.

REFERENCES

1. Sapir E. A study in phonetic symbolism. J Exp Psychol. 1929;12(3):225–39.
2. Noriko I, Vinson DP, Vigliocco G. What do English Speakers Know about gera-gera and yota-yota?:A Cross-linguistic Investigation of Mimetic Words of Laughing and Walking. Jpn Lang Educ Globe. 2007 Jun;17:53–78.
3. Lockwood G, Dingemanse M, Hagoort P. Sound-symbolism boosts novel word learning. J Exp Psychol Learn Mem Cogn. 2016;42(8):1274–81.
4. Wong LS, Kwon J, Zheng Z, Styles SJ, Sakamoto M, Kitada R. Japanese Sound-Symbolic Words for Representing the Hardness of an Object Are Judged Similarly by Japanese and English Speakers. Front Psychol. 2022 Mar 15;13:1–14.
5. Köhler W. Gestalt psychology, 2nd ed. Oxford, England: Liveright; 1947. (Gestalt psychology, 2nd ed).
6. Ramachandran VS, Hubbard EM. Synaesthesia--a window into perception, thought and language. J Conscious Stud. 2001;8(12):3–34.
7. Dingemanse M, Schuerman W, Reinisch E, Tufvesson S, Mitterer H. What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages. Language. 2016;92(2):e117–33.
8. Shinohara K, Uno R, Kobayashi F, Odake S. Sound symbolism of food texture: cross-linguistic differences in hardness. In University of Tartu, Tartu; 2017.
9. Saji N, Akita K, Kantartzis K, Kita S, Imai M. Cross-linguistically shared and language-specific sound symbolism in novel words elicited by locomotion videos in Japanese and English. Chen S, editor. PLOS ONE. 2019 Jul 10;14(7):e0218707.
10. Gallace A, Spence C. Multisensory synesthetic interactions in the speeded classification of visual size. Percept Psychophys. 2006 Oct;68(7):1191–203.
11. Shinohara K, Kawahara S. A Cross-linguistic Study of Sound Symbolism: The Images of Size. Annu Meet Berkeley Linguist Soc. 2010 Aug;36(1):396–410.
12. Knoeferle K, Li J, Maggioni E, Spence C. What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings. Sci Rep. 2017 Dec;7(1):1–11.
13. Akita K. Phonation Types Matter in Sound Symbolism. Cogn Sci [Internet]. 2021 May [cited 2022 Jul 6];45(5). Available from: https://onlinelibrary.wiley.com/doi/10.1111/cogs.12982
14. Ohala JJ. The frequency code underlies the sound-symbolic use of voice pitch. Sound Symb. 1994;2:325–47.

15. Newman SS. Further Experiments in Phonetic Symbolism. Am J Psychol. 1933 Jan;45(1):53–75.
16. Werker JF, Tees RC. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant Behav Dev. 1984;7(1):49–63.

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0462

Effect of frequency response equalization method on speech transmission index

Linda LIANG[(1)], Guangzheng YU[(2)]

[(1)]College of Civil Engineering and Architecture, Guangxi University, Nanning, China, scldliang@gmail.com

[(2)]Acoustic Laboratory, School of Physics and Optoelectronics, South China University of Technology, Guangzhou, China, scgzyu@scut.edu.cn

ABSTRACT

Frequency response of sound source may be one of the main causes of the inaccuracy in speech transmission index (STI) measurement. Thus, the frequency response equalization of sound source is important to satisfy the STI measured requirement. However, the performance of different equalization methods on STI measurement has rarely been discussed in previous research. This study therefore investigates the effect of equalization algorithms, including the Kirkeby algorithm and the minimum-phase reconstruction algorithm, and the effect of magnitude normalization methods of the frequency response used for equalization on the STI measurement. First, the impulse responses were measured in an anechoic room with three types of directional loudspeakers, and then used to calculate the corresponding STIs with impulse-based indirect method. Results show that compared with the Kirkeby algorithm, the equalization algorithm using minimum-phase reconstruction can obtain a flatter frequency response. The STI difference caused by equalization depends not only on the frequency response of the sound sources, but also on the magnitude normalization methods of the frequency response used for equalization. It recommends to use an energy-normalized frequency response for equalization, so as to avoid the introduction of additional energy.

Keywords: Speech Transmission Index, Frequency Response, Sound source Equalization

1 INTRODUCTION

The speech transmission index (STI) is one of the most common objective metrics for predicting speech intelligibility, and can be used to predict the loss of speech information between speaker and listener [1]. The STI is based on the fact that the interference on speech intelligibility due to the transmission system is related to the intensity reduction of modulations from the source signal to the receiver signal, which can be described by the modulation transfer function [2, 3]. By measuring the reduction in the modulation index, it is possible to calculate the apparent signal-to-noise ratio (SNR) of each frequency band, and then the transmission indexes and modulation transfer index can be obtained in turn. The STI is then the sum of the weighted contributions from each band [3]. The STI fully reflects the effects of background noise and the transfer function between speaker and listener on speech intelligibility; thus, it is better than the articulation index and the speech intelligibility index for assessing speech intelligibility in a steady acoustic environment that satisfies a linear time-invariant system.

In the light of the sound propagation path in a given space, in addition to background noise, binaural effect, and the acoustic conditions of the space [4, 5], STI measurement is considerably affected by the sound source characteristics including the frequency response and directivity. According to the specifications in IEC 60268-16 [3], the sound source used for STI measurement is better to have a similar directivity pattern as those of the average human mouth. Additionally, the frequency response of the sound source for the STI measurements should be within ± 1 dB in each 1/3 octave band over the range of 88–11300 Hz; otherwise, additional frequency response equalization is required. Previous studies have shown that sound source characteristics affect not only

the objective room acoustical parameters and subjective evaluation related to the sound quality [6, 7] but also the speech intelligibility in an actual room [8, 9, 10]. For instance, Zhu [10] measured the STI in different conditions using three different directional sound sources. They showed that both the directivity and frequency response of the sound sources may cause noticeable differences in the STI value, and the frequency response without equalization may afford large errors in the measured results. Additionally, Mapp [8] also showed that equalization in an unideal frequency response can significantly affect the speech intelligibility. There is no doubt that a careful frequency response equalization is always required when performing a STI measurement using a sound source with unideal frequency response.

The key technology of frequency response equalization is how to generate an appropriate inverse filter, which is always originated from the frequency response of sound source measured in an anechoic chamber [8, 10]. With regard to the generation of inverse filter, the choice of equalization algorithm and how to normalize the relative magnitude of the frequency response used for equalization are the main issues that need to be considered. However, the performance of different equalization algorithms and magnitude normalization methods of the frequency response used for equalization in STI measurement has not been discussed in previous research yet. This motivates the present systematic study of investigating the effect of equalization algorithms and magnitude normalization methods of the frequency response used for equalization on the STI measurement.

The impulse responses were measured in an anechoic room with three types of directional loudspeakers in the principal axis direction, and then used to calculate the corresponding STI with impulse-based indirect method according to IEC 60268-16 [3]. Two equalization algorithms, including the minimum-phase reconstruction algorithm [11] and the Kirkeby algorithm [12], and two magnitude normalization methods of the frequency response used for equalization, including the way aligning the relative magnitude in 1000 Hz to 0 dB following the treatment in reference [10] and the energy-normalized method which is first proposed in present study, were taken into account. Finally, the STI before being equalized and after being equalized by employing different equalization algorithms and magnitude normalization methods were analyzed.

2 METHODS

2.1 Impulse Response Measurements

Three directional loudspeakers with different frequency responses and directivity patterns were used as the sound sources in present study: (i) an artificial mouth Brüel & Kjær 4227A with a directivity and radiation pattern similar to those of an average human mouth; (ii) a monitor loudspeaker GENELEC 8010 with a frequency response within $\pm$2.5 dB in the range from 100 Hz to 20 kHz; and (iii) a domestic loudspeaker PHILIPS MCI500H with a low frequency limit of as low as 60 Hz, which is usually used for sound reproductions in traditional indoor environments.

The horizontal impulse responses in the principal axis direction of three sound sources were separately measured in an anechoic chamber. Excitation was induced using a logarithmic sweep signal with a sampling frequency of 44.1 kHz and 24-bit quantization for 3 s over a frequency range up to 22 kHz. This was fed to the sound sources after passing through a digital-to-analog converter in a sound card (Fireface UC; RME Audio, Haimhausen, Germany). Since the 4227A and MCI500H are passive sound sources, they need to be driven by a power amplifier (B&K 2716C). The sources were placed on a pallet fixed on a tripod and covered with sound-absorbing cotton to reduce reflection. The signals were recorded using an omnidirectional microphone (MicW M215), which was fixed at the source height (i.e., 1.2 m above the wire mesh of the anechoic chamber) and with a horizontal distance of 1 m relative to the sources. The partial photos of experimental environment in an anechoic room are shown in Figure 1. The recordings of the logarithmic sweep signals were deconvolved with the inverse of the original logarithmic sweep signal to yield impulse responses [13].

2.2 Equalization Methods

To smooth out the undesirable frequency response of the three sources, the inverse filters used for equalization were generated from the frequency responses at 1 m in front of the three sources that were measured in an anechoic chamber by employing the minimum-phase reconstruction algorithm and Kirkeby algorithm respectively.

If the frequency response used for generating inverse filtering is $H_0(f)$, the inverse filter $H^*(f)$ based on

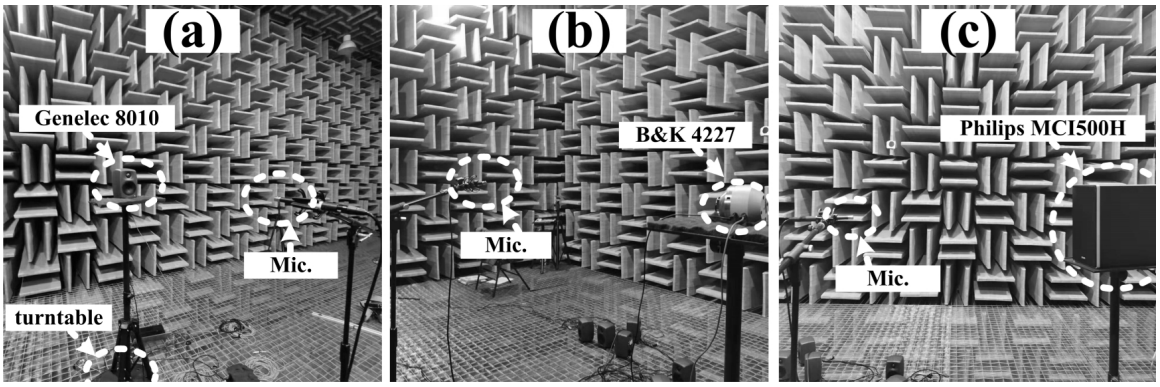


Figure 1. The partial photos of experimental environment in an anechoic chamber

the minimum-phase reconstruction [11] can be expressed as:

$$H^*(f) = \frac{1}{H_{\min}(f)} = \frac{1}{|H_0(f)| \exp\left(j\phi_{\min}(f)\right)}, \tag{1}$$

$$\phi_{\min}(f) = -\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\ln|H_0(f)|}{f-x} dx. \tag{2}$$

The Kirkeby method is based on the introducing of the regularization factor $\varepsilon(f)$ to design an inverse filter [12], which can be expression as:

$$H^*(f) = \frac{\mathrm{conj}\left[H_0(f)\right]}{\mathrm{conj}\left[H_0(f)\right] \cdot H_0\left(\psi_0, f\right) + \varepsilon(f)}, \tag{3}$$

where the regularization factor $\varepsilon(f)$ is set as an abnormally small value (1×10^{-100}) for frequencies within the target frequency range and is set to be abnormally large value (1×10^{100}) for frequencies outside the target frequency range.

2.3 STI Measurement Method

In the context of STI, the interference in speech intelligibility is related to a reduction in time modulation due to the transmission system, which can be described by the modulation transfer function. The modulation transfer function is distributed in 14 modulation frequencies, ranging from 0.63 to 12.5 Hz, and seven octave bands, with center frequencies ranging from 125 Hz to 8 kHz. The traditional method (direct method) for determining the modulation transfer function uses test signals modulated sinusoidally in terms of intensity. However, this method is time consuming because it must be repeated for each of the 98 combinations of modulation frequencies and octave band frequencies [3]. To reduce the number of measurements needed and enhance repeatability, a single-impulse response measurement (indirect) method was developed by Schroeder [14].

Usually, the sound pressure levels of the noise and binaural speech signal must be measured separately [1]. Here, we obtain the speech signals from the source by based on the Auralization technology [4, 5]. The speech sample was obtained by generating a pink noise signal and then filtering and adjusting the spectrum according to GB/T 7347 [15], the standard spectrum of Chinese speech. A stationary noise (i.e., pink noise) is regarded as background noise with the same sound pressure level in each frequency band. We then obtained the speech signals by convolving the speech sample with the corresponding impulse response measured in an anechoic chamber as Section 2.1. The STIs under different conditions were then calculated according to IEC 60268-16 [3].

3 RESULTS AND DISCUSSION

3.1 The Influence of Equalization Algorithm

The original frequency response in relation to 1000 Hz and the frequency response after being equalized by employing different equalization algorithms are shown in Figure 2.

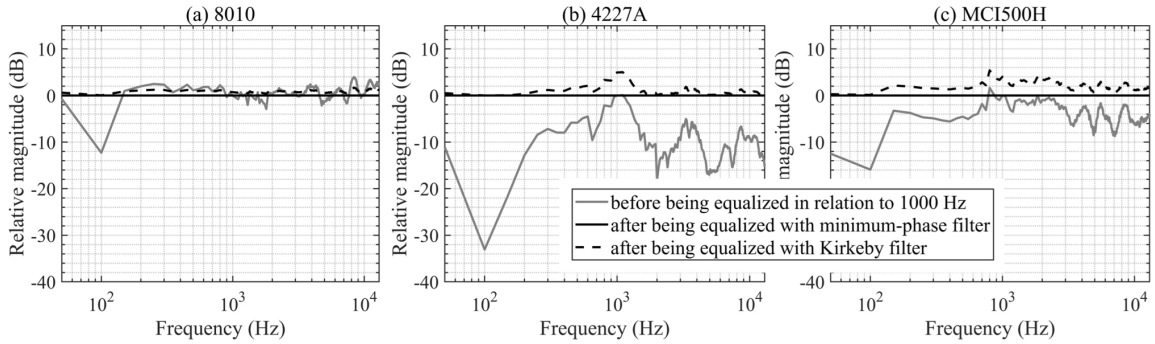


Figure 2. Frequency response before being equalized and after being equalized by employing different equalization algorithms in the principal axis direction for the (a) 4227A, (b) 8010, and (c) MCI500H sources, respectively

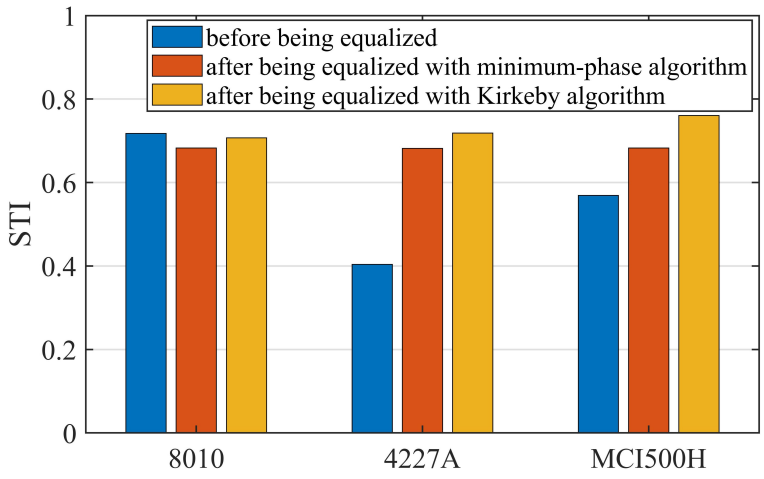


Figure 3. STI result before being equalized and after being equalized by employing different equalization algorithms for the 4227A, 8010, and MCI500H sources, respectively

It can be seen that the frequency response after being equalized with minimum-phase reconstruction algorithm is almost flat in the entire frequency range for each sound source. However, this is not the case for the Kirkeby algorithm, where the fluctuation range of the relative magnitude is even up to 5 dB for the 4227A and MCI500H sources. According to the specification in IEC 60268–16 [3], the frequency response of the sound source for the STI measurements should be within ± 1 dB in each 1/3 octave band over the range of 88–11300 Hz. Obviously, the performance of the Kirkeby algorithm does not meet the requirement.

Based on the original impulse responses with relative magnitude aligned to 0 dB in 1000 Hz and those after being equalized by employing different equalization algorithms, the corresponding STIs were calculated based on the impulse-based indirect method, as shown in Figure 3. Note that, except for the 8010 source, the STIs after being equalized obviously vary from those before being equalized, with difference of more than 0.3 in some cases, far exceeding the just-noticeable-difference (JND) of 0.03 [16]. On the one hand, the impulse

response used for equalization introduces additional energy to STI measurement to a large extent, and it means a large STI difference between before and after being equalized. On the other hand, the mismatch between the irregular frequency response of sound sources and the band-weighting coefficient at the STI calculation also contribute a portion of difference. It is worth mentioning that there is a negligible STI difference between before and after being equalized for the 8010 source (within 1 JND), due to its flat frequency response before being equalized as shown in Figure 2(a).

As shown in Figure 3, except for the 8010 source, there are observable difference between the STI with equalization employing the minimum-phase reconstruction algorithm and those employing the Kirkeby algorithm, especially for the MCI500H source. We can also find corresponding clues in Figure 2 for these results.

Overall, compared with the Kirkeby algorithm, the equalization using minimum-phase reconstruction algorithm can obtain a flatter frequency response, and there is not negligible difference between the STI result with two algorithms for the 4227A and MCI500H sources. Therefore, the inverse filtering algorithm employing the minimum phase reconstruction is adopted to the frequency response equalization mentioned below.

3.2 The Influence of Magnitude Normalization Method

As forementioned, if we align the relative magnitude of the frequency response used for equalization in 1000 Hz to 0 dB, the the impulse response used for equalization could introduce additional energy to the impulse responses to be equalized, thus there will be notable difference between the STI after being equalized and those before being equalized. To prevent the change in energy introduced by the equalization, the frequency response used for equalization should be processed by energy normalization. If the frequency response used for the inverse filtering is $H(f)$, the normalized frequency response can be expressed as:

$$\bar{H}(f) = \frac{H(f)}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}|H(f)|^2}}, \tag{4}$$

where N is the frequency counts. Then, the normalized frequency response $\bar{H}(f)$ was used to generate the inverse filter with minimum-phase reconstruction method as Eq. (1) and Eq. (2).

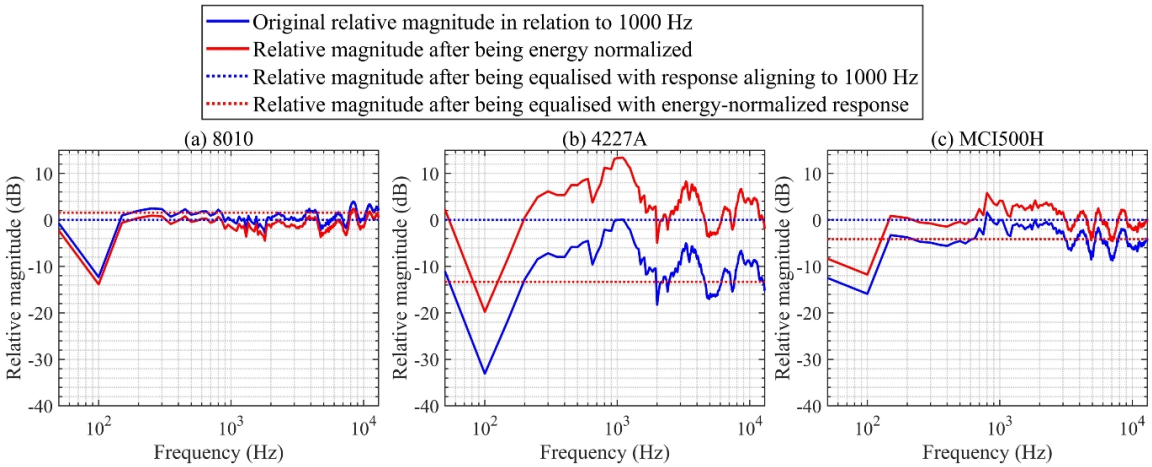


Figure 4. The frequency responses used for equalization by employing different magnitude normalization method and those after being equalized with them for the (a) 4227A, (b) 8010, and (c) MCI500H sources, respectively

The frequency responses used for equalization employing different magnitude normalization method and those after being equalized with them for three sound sources are shown in Figure 4. Note that, the differences in relative magnitude of the frequency response after being equalized employing different magnitude normalization method are about 2 dB, 14 dB, and 4 dB for 8010, 4227, and MCI500H sources respectively, which means

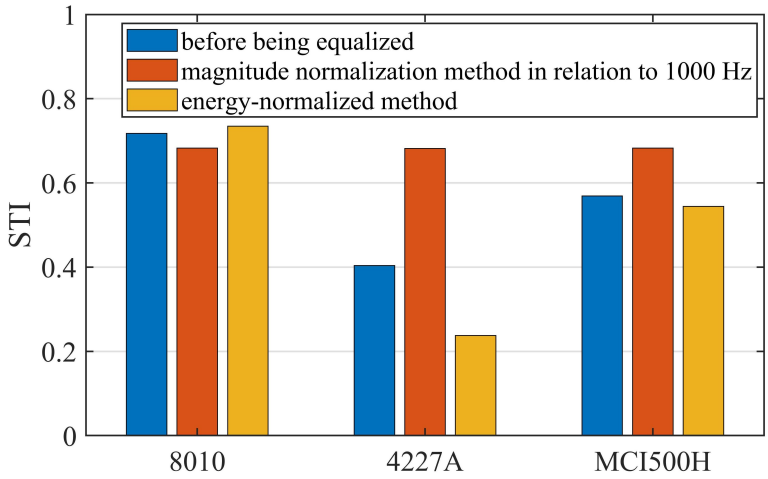


Figure 5. STI result before being equalized and after being equalized by employing different magnitude normalization methods for the 4227A, 8010, and MCI500H sources, respectively

STI difference of 0.07, 0.47, and 0.13 according to the definition of STI [3], i.e., the SNR from –15 dB to +15 dB is linearly correlated with the STI range from 0 to 1.

The corresponding STIs were also calculated based on the impulse-based indirect method, including the result before being equalized and those after being equalized with the frequency response under different magnitude normalization methods, as shown in Figure 5. There are similar phenomena as that observed in Figure 4. As can be seen in Figure 5, except for the 8010 source, there are considerably difference between the STIs after being equalized with frequency response by employing different magnitude normalization methods, even with difference of more than 0.4 for 4227A, far exceeding 1 JND of 0.03. This is principally because the frequency response used for equalization introduces additional energy to STI measurement, as shown in Figure 4. This recommends to use an energy-normalized frequency response for equalization in STI measurement, which can avoid the introduction of additional energy.

4 CONCLUSIONS

This study investigates the effect of different equalization algorithms and magnitude normalization methods of the frequency response used for equalization on the STI measurement. The impulse responses were measured in an anechoic chamber with three directional loudspeakers. Two equalization algorithms (i.e., the Kirkeby algorithm and minimum-phase reconstruction algorithm) and two magnitude normalization methods (i.e., the way aligning the relative magnitude in 1000 Hz to 0 dB and the energy-normalized method proposed in present study) were adopted in frequency response equalization. Then, STIs before being equalized and those after being equalized be employing different equalization algorithms and magnitude normalization methods were analyzed. Results show that the performance of the Kirkeby algorithm does not meet the requirement of STI measurement in term of the frequency response, thus causes a considerably error (far exceed 1 JND) except for the 8010 source. In contrast, the equalization algorithm using minimum-phase reconstruction can obtain a flatter frequency response. The comparison among the STI result before being equalized and those after being equalized employing different magnitude normalization methods shows that, the STI difference caused by equalization depends not only on the frequency response of the sound sources, but also on the magnitude normalization methods of the frequency response used for equalization. It recommends to use an energy-normalized frequency response for equalization in STI measurement, which can avoid the introduction of additional energy. This study provides reference for future evaluations of speech intelligibility using a sound source with unideal frequency response.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China through grant number 12074129.

REFERENCES

[1] Petra L, Hongistob V. Experimental comparison between speech transmission index, rapid speech transmission index, and speech intelligibility index. J. Acoust. Soc. Am. 2006;119(2):1106–1117.

[2] Houtgast T, Steeneken HJM. The modulation transfer function in room acoustics as a predictor of speech intelligibility. J. Acoust. Soc. Am. 1973;54:557.

[3] IEC 60268-16. Sound system equipment. Part 16: Objective rating of speech intelligibility by speech transmission index. International Electrotechnical Commission, Geneva; 2011.

[4] Liang LD, Yu L, Zhao T, Meng QL, Yu GZ. Speech intelligibility for various head orientations of a listener in an automobile using the speech transmission index. J. Acoust. Soc. Am. 2021; 149(4):2686–2697.

[5] Liang LD, Yu GZ. Binaural speech transmission index with spatialized virtual speaker in near field: distance and direction dependence. J. Acoust. Soc. Am. 2020;148(2):EL202–207.

[6] Dalenbäck BI, Kleiner M, Svensson P. Audibility of changes in geometric shape, source directivity, and absorptive treatment–experiments in auralization. J. Audio Eng. Soc. 1993;41(11):905–913.

[7] Otondo F, Rindel J. The influence of the directivity of musical instruments in a room. Acta Acust. united with Acust. 2004;90(6):1178–1184.

[8] Mapp P. Some effects of equalisation on sound system intelligibility and measurement. Proc 115th AES Convention; 10–13 October 2003; New York, USA, Paper 5986.

[9] Peng JX, Wang T, Wu S. Investigation on the effects of source directivity of Chinese speech intelligibility in real and virtual rooms. Appl. Acoust. 2013;74(8):1037–1043.

[10] Zhu P, Mo FS, Kang J. Influence of sound source characteristics in determining objective speech intelligibility metrics. Appl. Acoust. 2015;89:188–198.

[11] Xie B. Head-related transfer function and virtual auditory display. J Ross Publishing, New York, USA; 2013.

[12] Farina A. Advancements in impulse response measurements by sine sweeps. Proc 122nd AES Convention; 5–8 May 2007; Vienna, Austria.

[13] Majdak P, Balazs P, Labac B. Multiple exponential sweep method for fast measurement of head related transfer functions. J. Audio Eng. Soc. 2007;55:623–637;

[14] Schroeder MR. Modulation transfer functions: Definition and measurement. Acta Acust. United Acust. 1981;49(3):179–182.

[15] GB/T 7347-1987. The standard spectrum of Chinese speech. National Standard of China; Beijing, China; 1987.

[16] Bradley J, Reich R, Norcross SG. A just noticeable difference in C50 for speech. Appl. Acoust. 1999;58(2):99–108.

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0496

Indoor announcement system of emergency information based on human characteristics for sustainable use

Yoshifumi CHISAKI[(1)]; Ryo TAKAHASHI[(2)]

[(1)]Department of Advanced Media, Faculty of Advanced Engineering, Chiba Institute of Technology, Japan
[(2)]Graduate School of Advanced Engineering, Chiba Institute of Technology, Japan

ABSTRACT

After the 2011 Tōhoku earthquake and tsunami, infrastructure of radio wave and wireless internet is fulfilled rapidly, and alerts or notifications on emergency and warning from government can be obtained easily with outdoor loudspeaker, smartphone apps, e-mail and so on. Local government provides not only for emergency notification but also daily useful information over the infrastructure, and frequency of use of the system is increased. However, it is still difficult to catch the information in some cases indoors. This prevents behavior of listening carefully to the message from the government continuously.

This paper proposes an indoor announcement system for emergency information. The system is designed to encourage sustainable use considering human characteristics by relieving of missing important words in a sentence. A limitation of the number of words in a sentence in listening is examined using subjects in order to design an automatic optimized sentence generation considering human memory. Sentence composition method based on the limitation is discussed.

Keywords: indoor announcement, human characteristics, sustainable use

1 INTRODUCTION

The 2011 Tōhoku earthquake and tsunami caused huge damage. After the earthquake, the development of a system that can deliver appropriate information wherever has become active. System called disaster radio system or mass notification system built by the national and local governments play important roles, and those provide neighborhood fire information, missing person information and other useful information for residents. People receive that information from various ways. The information which requires rapid action is provided to television, radio, smartphone, household receiver, outdoor loudspeaker. Other daily announcement is always delivered by e-mail and sometimes household receiver, outdoor loudspeaker. Outdoor loudspeaker is useful to spread wide area all at once, however, residents annoyed loud sound. Long-path echo generated by buildings and others also makes it hard to hear. Since speech from outdoor loudspeaker is uttered with pauses to avoid long-path echo, a whole message requires longer time to understand the contents than time for a normal speed speech. Arrangements of outdoor loudspeaker location, characteristics of sound propagation and other related topics has been studied by references (1–5). It is also hard to hear from outdoor loudspeaker in a house due to a wall and a window of house. Ministry of Internal Affairs and Communications in Japan recommended to use a household receiver in a house, however, it is not widespread and to all rooms even if residents have. Due to the mentioned conditions above, delivery of e-mail which contains the same message is expected. The e-mail from local government is sent to a registered resident simultaneously. E-mail has an advantage of being able to be viewed as needed. Since we can read it as appropriate, if you are doing something, such as washing dishes, bathing, we may postpone reading the e-mail and forget that there was an announcement. These difficulties on listening from outdoor loudspeaker and getting a message by other methods hinders habit of getting information.

[1] yoshifumi.chisaki@p.chibakoudai.jp

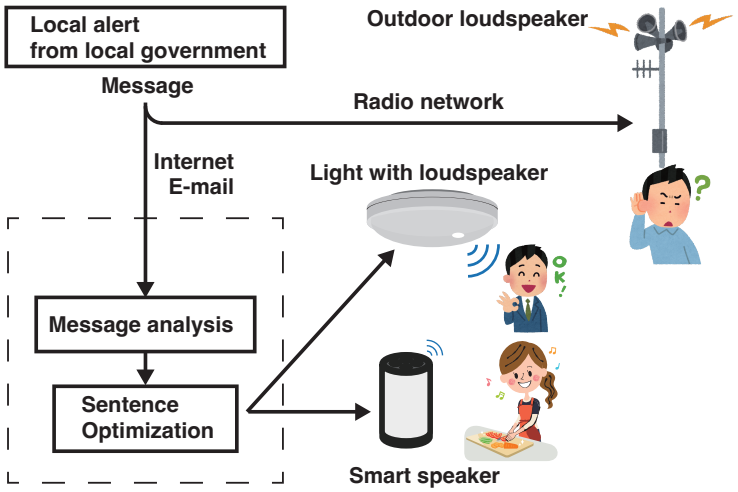


Figure 1. Message delivery flow from local government to residents.

In this paper, a push notification by speech is proposed. The notification method does not only read aloud e-mail but also include interactions based on human characteristics for sustainable use. In particular, the number of words in a sentence is investigated by listening tests, and a sentence is designed based on the findings of listening tests.

2 CONCEPT OF THE PROPOSED SYSTEM

After the 2011 Tōhoku earthquake and tsunami, information for safe and daily life from local government is delivered by several ways including e-mail. In this study, popular devices such as smartphone, smart speaker, or ceiling light with loudspeaker are supposed to read aloud a message in a room and the message from government is delivered by e-mail. The message contains a purpose, an incident, date and time. Since other warning sentences are included in the message, a total duration of message is over one minute when a whole message is read aloud. Since residents cannot keep paying attention to the speech during housework and so on, it is expected to summarize the e-mail contents for reducing the playback duration. This function is quite important for long-term use. FIgure 1 shows a flow of message delivery. The emergency information service is provided by local government. One of routes to residents is via outdoor loudspeaker. As mentioned above, sound propagation from outdoor loudspeaker is insufficient for residents in a room. On the other hand, residents can receive e-mail and get contents of message using smartphone or PC in a room. Since it is troublesome to pick up a smartphone to check e-mail while doing housework, the proposed system optimizes sentences of message in the e-mail and the optimized sentences are read aloud.

Category of message is indicated in Table 1, and category is not limited to emergency incident but also wide for daily life matters. Table 2 shows examples of message from local government. E-mail is consisted of the message with header and footer. Fig. 2 indicates an example of e-mail which includes message, and shows that e-mail has possibility of various sentences. As shown in Fig. 2, it is expected to take a minute to read aloud original contents of e-mail directly. It is considered that shorter sentence is preferred to avoid misunderstanding of message, and to encourage long-term use of the system.

Thus, functions of message analysis and sentence optimization are designed as follows:

1. specify a category from e-mail,
2. pick up essential items for a specified category from e-mail,
3. generate sentence for text-to-speech based on human's listening ability; the number of maximum words to catch in a sentence,

Table 1. Category of message from local government.

Topic	Details
Fire	Fire outbreak and suppression.
Disaster Information	J-alerts (Earthquake Early Warnings, Special Warnings of Weather, etc.) related to earthquakes and weather, as well as information on typhoons, heavy rain, sediment disasters, heavy snow, tornadoes, etc.
Photochemical Smog/PM 2.5	Precautions against photochemical smog and PM 2.5.
Suspicious person information	Witness information about suspicious persons, etc.
Lost or missing	Information about lost or missing persons.
City crime occurrence situation	Precautions regarding the occurrence of crimes in the city, etc.
City Emergency	J-Alert for public protection (Ballistic missile information, large-scale terrorist information, etc.), large-scale accidents, etc.
Broadcast contents of disaster prevention administration radio	Broadcast contents of disaster prevention administration radio (excluding regular evening broadcast).
Others	Other information (information on heat stroke, information on new coronavirus infection), etc.

Table 2. Examples of messages from local government.

Topic	Message
Fire	Building fire occurred, ○○○○ city ○○○ town, June 4th 2022, around one thirteen. Fire brigade is dispatched now.
Missing person	Mr. ○○○○, 74 years old has been missing since Tuesday, August 23. Height is 165 cm. The physique is medium meat and medium back. He wears a beige hat, navy blue short sleeve T-shirt, Gray long trousers. If you find, please call ○○○○ police station, 81-47-474-xxxx.
Warning	This is ○○○○ police station. we gets a lot of fraudulent calls recently in our city. The message is like ‘ There is a refund. The card has been used illegally and needs to update. I needed cash from my son and grandson.’ These calls are fraudulent. Be careful about fraudulent calls.

4. generate sentence based on personal characteristics; only what a user wants.

It is considered that a message part which are essentials to make notice could be formatted on each topic as shown in Table 2. Although those formatted sentences could be generated with a machine learning algorithm, a rule-based method to generate sentence is considered as a first trial in this paper. So, it is necessary to find out a category from e-mail as function 1. In function 2, some specific words are extracted from e-mail. For example, a message for missing person indicated in Table 2 includes some words; gender, ages, height, physique, wears, and police phone number to provide information. These feature words are extracted with dependency relationship of words, units and so on.

Function 3 and 4 are the key idea to encourage long-term use. In function 3, a user will stop using the system continuously if the frequency of not being able to hear the contents increases. Thus, an automatic

Example of e-mail

The number of suspicious persons is reported to the Youth Center in May.
[strange people talking: 3 incidents]ABC area, DEF area, GHI area
[perverted behavior : 4 incidents] JKL area, MNO area, PQR area
[violence, violent language : 1 incident] XYZ area

The Youth Center provided this information for the police, and carried out patrol.
For other suspicious person information, please see the XYZ Prefectural Police "Safety Map for Living".
Smartphone site
https://www2.xyz.jp/cp-gis-sp/

Click here to change user information or unsubscribe.
https://service.abcxyz.com/XYZ/m/u/i/9965a7a594b4
XYZ Youth Center
TEL 000-010-0203

Figure 2. Example of e-mail from local government to residents.

sentence generation function that limits the number of words that a person can hear in one sentence is required. In function 4, continuing to provide unnecessary information for a user does not encourage to long-term use. thus, it is preferred that playback message is selectable, whether playback or through based on the preset by a user.

In this paper, function 3 is focused to obtain findings for sentence generation based on human ability.

3 DESIGN OF SENTENCE STRUCTURE

From the viewpoint of human memory in hearing, sentence length for playback should be controlled. In this section, the appropriate number of words in a sentence is discussed.

3.1 Experiments on the number of words

Listening tests using subjects are conducted. The number of subjects is 6, and range of age is from 21 to 23. Their primary language is Japanese. All sentences are in Japanese. Experiments are performed with a headphone and 44.1kHz/16bits DA converter. Sound pressure level for playback is set to appropriate one by each subject at beginning of experiment. Four topics; lost or missing, fire, city crime occurrence situation, disaster Information are used.

The sentences blow are examples of sentence on each topic.

- Missing person
  Mr. ○○○○, 74 years old has been missing since Tuesday, August 23. Height is 165 cm. The physique is medium meat and medium back. He wears a beige hat, navy blue short sleeve T-shirt, Gray long trousers. If you find, please call ○○○○ police station, 81-47-474-xxxx.

- Fire
  Building fire occurred, ○○○○ city ○○○ town, 1 choume, 1 banchi, around one thirteen, June 4th, Thursday. Fire brigade is dispatched now.

- City crime occurrence situation
  The main crime that occurred was 5 snatching, around ○○○○ city ○○○ town, June 4th, Thursday, 2022. Please be careful.

- Disaster Information
  Emergency heavy rain warning was issued around ○○○○ city ○○○ town. The forecast will continue until around 23 on Thursday, March 2nd. Please be careful.

Although the above examples are in English to explain the experiments, all sentences are in Japanese in Listening tests. Underlined words are selected from the same category, e.g. red, black, gray and so on in case of color. The number of words in a sentence is also varied from 4 to 12. The number of generated sentences with varying all words is 2,700, and the sentences are presented to the subject randomly. Each session is 20 minutes and a break.

A subject answer a result from 5 choices. One is correct answer, other 3 choices are wrong words in the same category, and the last one is 'others or not included.'

3.2 Results

From the viewpoint of the number of words, percentage of correct answers is shown Table 3. According to the results, major trend was that percentage of correct answers becomes smaller as the number of words becomes larger. The maximum percentage of correct answers was 87 % even if the number of words was 4. Therefore, it is recommended shorter length of message for announcement is preferred.

Table 3. Results : percentage of correct answers on the number of words.

The number of words	4	5	6	7	8	9	10	11	12
% of correct answers	87	79	75	70	64	59	47	49	52

The average number of correct words was around 4 overall. In each trial, lower percentage was shown in category of 'Fire' with 7 to 12 words. In this case, a sentence included street address, such as 1 choume or 1 banchi. On the other hand, higher percentage was obtained in case that many numeric parameters; e.g. date, hour, street address were equal to or less than 2 words. Since it is considered that the numeric words make a user confused, the number of numeric words in a sentence is recommended up to 2 words.

3.3 Discussion

According to the results of experiments of the number of words, it is recommended that a sentence can include up to 4 words, and a numeric words is 2 words of 4. Therefore, only hour and minute can be indicated and two incident words in a sentence. Example sentence for missing person topic is below.

- Missing person
  sign sound, such as bell or beep
  + Mr. ○○○○, 74 years old has been missing for ELAPSED_TIME.
  + PAUSE
  + The physique is PHYSIQUE.
  + PAUSE
  + He wears CLOTHING_FEATURE1, CLOTHING_FEATURE2, CLOTHING_FEATURE3, CLOTHING_FEATURE4.
  + PAUSE
  + please call ○○○○ police station, PHONE_NUMBER.

The 'ELAPSED TIME' can be calculated as difference between the system time and date/hour/minute in e-mail. 'ELAPSED TIME' also can be rounded as only hours; e.g. 6 hours or 2 days to reduce the numeric words.

Example sentence for fire topic is below.

- Fire
  sign sound, such as bell or beep

+ Fire occurred ELAPSED_TIME ago.
+ PAUSE
+ It is DISTANCE_FROM_HERE in ○○○○ city ○○○ town.
+ PAUSE
+ Fire brigade is dispatched now.

The 'DISTANCE_FROM_HERE' can be calculated approximately using difference between GPS/internet access and address information in e-mail.

Example sentence for City crime occurrence situation topic is below.

- City crime occurrence situation
  sign sound, such as bell or beep
  + The main crime that occurred was NUMBER_OF_INCIDENTS CRIME these last ELAPSED_TIME.
  + PAUSE
  + It is DISTANCE_FROM_HERE in ○○○○ city ○○○ town.
  + PAUSE
  + Please be careful.

The 'DISTANCE_FROM_HERE' can be calculated approximately using difference between GPS/internet access and address information in e-mail.

As discussed above, words on date and time can be reduced with elapsed time. words on place also will be reduce the distance from a user. This method is expected to be useful to other topics.

4 CONCLUSIONS

In this paper, Indoor announcement system of emergency information based on human characteristics for sustainable use is proposed. Design of the system is proposed, and the ideas to use e-mail from L-alter is discussed. In particular, human characteristics on the number of words in a sentence in listening is investigated by listening tests. As a result, it is suggested that a sentence can include up to 4 words, and a numeric words is 2 words of 4. Some ideas of sentence generation based on the findings are discussed.

Listening tests based on ideas and field listening tests will be conducted in the future.

ACKNOWLEDGEMENTS

Part of this work was supported by JSPS KAKENHI Grant Number JP22K04638.

REFERENCES

(1) Onoguchi T, Murakami D, Chisaki Y, Emission timing control method for improving signal to interference ratio on public address system, Applied Acoustics 2015; 98, 70-78. https://doi.org/10.1016/j.apacoust.2015.04.019.

(2) Zhenglie Cui, Sakamoto S, Morimoto M, Suzuki Y, Sato H, Effect of word familiarity on word intelligibility of four continuous words under long-path echo conditions. Applied Acoustics 2017;124: 30-37. https://doi.org/10.1016/j.apacoust.2017.02.001.

(3) Sakamoto S, Zhenglie Cui, Miyashita T, Morimoto M, Suzuki Y, Sato H, Effects of inter-word pauses on speech intelligibility under long-path echo conditions. Applied Acoustics 2018; 140: 263-274. https://doi.org/10.1016/j.apacoust.2018.01.020.

(4) Nishimura R, Sakamoto S, Chisaki Y, Zhenglie Cui, Optimization of output level of outdoor loudspeakers for municipal radio systems in times of disaster. Journal of the Acoustical Society of Japan 2020; 76–9: 475–484. https://doi.org/10.20697/jasj.76.9_475. (in Japanese)

(5) Sato H, Kurisu K, Morimoto M, Maeda M, Effects of rainfall rate on physical characteristics of outdoor noise from the viewpoint of outdoor acoustic mass notification system Applied Acoustics 2021; 172. https://doi.org/10.1016/j.apacoust.2020.107616

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0519

Increasing speech intelligibility in noise based on concepts of modulation spectrum and voice conversion to professional announcer voice

Masato AKAGI[1]

[1] Japan Advanced Institute of Science and Technology, Japan

Abstract

This study introduces two different approaches to increase intelligibility of speech in adverse conditions. One is modifying portions of the modulation spectra of smeared voices showing high correlation with intelligibility scores from listening tests. The other is a speech enhancement using voice conversion (VQ-VAE) to modify the speaking styles of non-professional voices to that of professional announcers. The results of subjective evaluations confirm that the intelligibility of the enhanced voices is higher than that of the original voices.

Keywords: Speech intelligibility, Modulation transfer function, Voice conversion, Clear speech

1. Introduction

Japan is a disaster-prone country. There are many earthquakes, and in recent years, damage from wind and rain has also increased. To minimize the potential for these life-threatening emergencies to do damage and harm, it is important to present appropriate evacuation guidance according to the situation. Voiced evacuation guidance is used in various places because it is effective even if the visual signs for the evacuation guidance cannot be confirmed, and many people can be given guidance at one time. However, speech in environments with large amounts of noise and long reverberation times may be difficult to hear and understand.

To combat this problem, simply increasing the volume cannot improve the audibility of the voice. Also, even if the guidance can be heard, it may not convey a sense of danger due to the "normalcy bias", and as a result, evacuations may be delayed. Evacuation guidance must strongly convey the potential danger and strongly urge people to evacuate; that is, voice announcements must account for various environmental factors and convey the degree of danger.

Since the Great East Japan Earthquake on March 11, 2011, various disaster-prevention proposals have been made to the Cabinet Office. In particular, they have pointed out problems related to the difficulty of hearing outdoor loudspeakers belonging to the disaster administration wireless communication system, called bousai-musen (1) and understanding the content of the broadcast calling for evacuation. To convey the necessary information by voice in the event of a disaster, it is believed that speech-information technology can be used to clearly and reliably present the dangers and to strongly encourage evacuation.

Our research group has been studying a speech announcement system that can provide evacuation guidance with high intelligibility in order to provide the "necessary information by speech" capability called for by the SCOPE Program of the Ministry of Internal Affairs and Communications, Japan. The main issues in building a system for presenting evacuation guidance by voice are as follows.

1) Evacuation guidance must be heard clearly regardless of the acoustical environment of the disaster space and situations of evacuees.

2) Evacuation guidance must give the urge to evacuees to escape depending on the risk of disaster.

In this report, we introduce the research that our group has conducted on methods of appropriately presenting speech evacuation guidance.

[1] akagi@jaist.ac.jp

2. Research concept

The sound environment of the disaster space affects the effectiveness of spoken evacuation guidance. In noisy and reverberant environments, the presented speech reaches evacuees through the path illustrated in Fig. 1: guidance speech (a); presentation by audio equipment (b); distortion of speech by noisy and reverberant environment (c); hearing the distorted speech (d) (Fig. 1). Accordingly, the following three methods can be implemented as measures to improve intelligibility.

[1] To reduce the distortion of the sound by the space (c), the sound space should be prepared by taking measures against noise and reverberation that absorb sound from walls and ceilings.

[2] The characteristics of the audio equipment used to present the guidance (b) should be as good as possible by taking measures such as proper placement of speakers.

[3] The voice of the speaker giving the guidance (a) should be converted to ensure that is clear and presentable.

There are many studies on [1] and [2]. However, their application entails expensive new constructions. Our research focuses on [3] above.

The author is considering humans' excellent behavior performed consciously or unknowingly in speech communication in a noisy reverberant environment, like uttering Lombard speech or clear speech and controlling para-linguistic information. An automatic system mimicking this aspect can be constructed to generate voices that are clear even in such an environment and to present appropriate evacuation guidance according to the degree of risk and urgency. There are two points to focus on regarding the humans' behaviors.

(A) Adaptive control of the presented speech based on feedback of sound environment situations and knowledge of clear speech:

As in the case of the Lombard effect, by constantly monitoring the presented voices while measuring the noisy reverberant environment in which the listeners are present, we can generate an announcement voice that is natural and has the highest intelligibility in that environment. In addition, it is well known that the voices of professional announcers are clear and have a high degree of intelligibility in a noisy environment.

(B) Control of linguistic and para-linguistic information according to the situation:

By selecting linguistic information according to the situation and adding para-linguistic information adaptively, we can generate an announcement voice to alert listeners. We can determine what characteristics the evacuation guidance voice should have by considering the characteristics of speech perception in an emergency and generating a highly understandable natural voice based on the restrictions of the human vocal system.

Figure 2 shows a conceptual diagram of the proposed system. Due to the limitation on the number of pages, this paper focuses on the issues of (A).

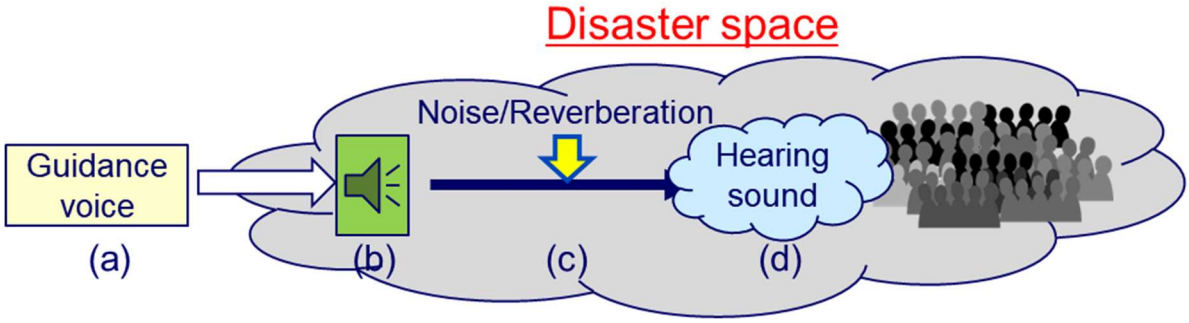


Figure 1 – Sound presentation in noisy and reverberant sound environment.

3. Adaptive control of presented speech based on feedback of sound environment situations and knowledge of clear speech

We implement two types of adaptive control:

- Speech modification based on the modulation transfer function (MTF) concept, and
- Mimicry of clear speech uttered by professional announcers by using voice conversion techniques.

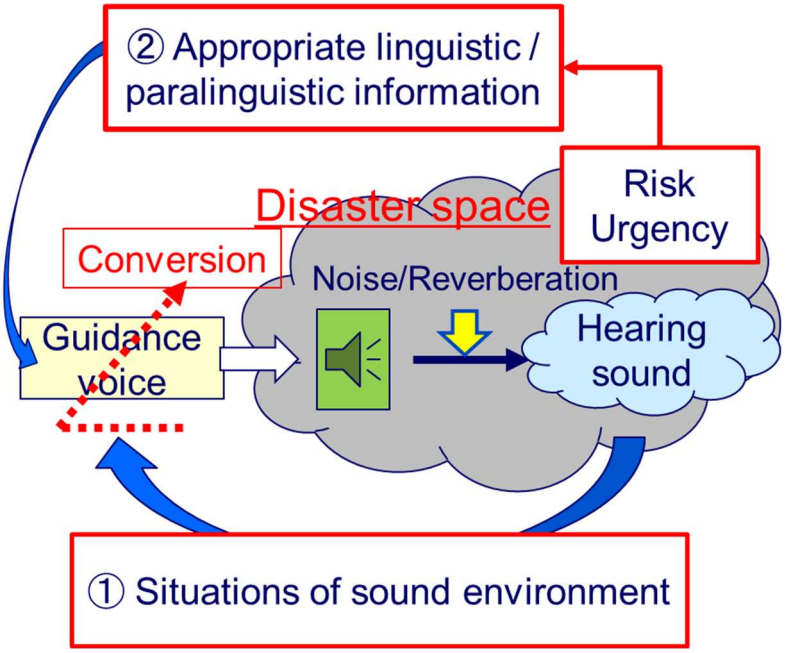


Figure 2 – Design of proposed system. The guidance voices are converted controllably by feedbacking situations of sound environment and appropriate linguistic and para-linguistic information.

3.1 Speech modification based on MTF concept (2)

The Lombard effect is simulated by converting speech by feeding back information of sound environments. At first, we proposed a method to convert speech by feeding back the noise level (3). In this paper, a method using a modulation transfer function (MTF) as the information of noisy and reverberant environment is proposed. The proposed method controls the speech modulation spectrum (MS) based on a room acoustic model, MTF, and demonstrates a more systematic and explicit derivation to enhance intelligibility of speech against environmentally caused smearing.

As proposed by Houtgast and Steeneken (4), the MTF reduces fluctuations in the envelope of the output signal relative to the envelope of the input signal during its transmission in a room. The MTF has been used in the calculation of the speech transmission index (STI) (5), which is an important objective index for the intelligibility of speech in noisy reverberant environments. In the MS concept, the speech MS is produced by spectral analysis of the temporal amplitude envelope of the frequency spectra. The dominant MS component of continuous speech lies between modulation frequencies of 1 and 16 Hz, with a peak around 4 Hz (4). Recent studies (6) have reported that higher intelligibility is obtained when the MS index is high as in the case of Lombard speech. In the other words, the higher the MS index is at these modulation frequencies, the greater the intelligibility becomes.

As shown in Fig. 3, suppose a ‘‘smeared MS’’ (MS_S) is given by

$$MS_S = MS_O \times MTF\,, \tag{1}$$

where MS_O is the MS of the original speech. Then, an ‘‘optimally resistant MS’’ (MS_R) can be calculated using

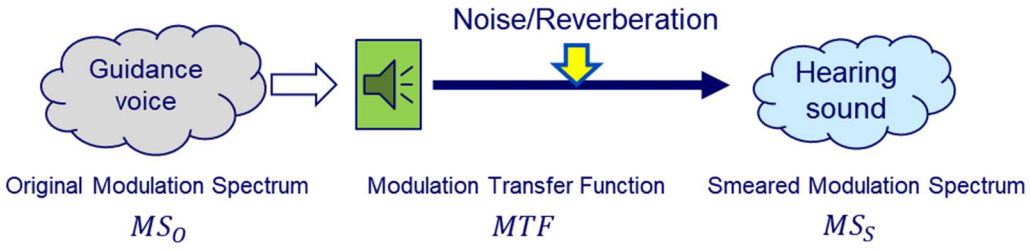
$$MS_R = MS_O \times MTF^{-1}. \tag{2}$$

If MS_R is presented in an adverse environment with such an MTF, the MS of the speech reaching the listeners should be MS_O as follows,

$$MS_O = MS_R \times MTF\,, \tag{3}$$

which has the original intelligible MS.

Directly obtaining the inverse MTF, MTF^{-1}, which requires an estimate of the MTF, is complicated; the assumption is that the estimation is usually done with the provided noise and room impulse response. Though MTF^{-1} can be obtained under this assumption, a critical problem is how it can be used efficiently. If we use the inverse MTF strictly according to the theory to modify speech MS (7)(8), it is sometimes ineffective, or even dangerous in our case, because the modification might destroy important speech features and it takes a lot of energy to boost all the acoustic and modulation frequency regions of smeared MS.



$$MS_O \times MTF = MS_S \quad \Rightarrow \quad \frac{(MS_O \times MTF^{-1})}{MS_R} \times MTF = MS_O$$

Figure 3 – Speech modification based on MTF concept. MS_O : Modulation spectrum (MS) of original speech, MS_S : MS of smeared speech, MS_R : MS of optimally resistant speech.

We consider that the basic concepts of the MS and MTF are central to an efficient way to modify the speech MS. To this end, we concentrated on significant acoustic and modulation frequency regions and their appropriate amplitude levels for improving the intelligibility and naturalness of speech in adverse conditions. Furthermore, our concept is not constrained by noise level or SNR; the conversion controls are based on the MTF which can be estimated at any time.

3.1.1 Monitoring of the sound environment (9)

In order to monitor sound environments that change from moment to moment depending on the situation, the STI and five other room-acoustic parameters can be estimated promptly and simultaneously in almost real time. However, it is difficult to obtain such parameters in spaces occupied by people, since the room impulse response (RIR) cannot be measured. In such cases, the room acoustic parameters have to be blindly estimated from the observed signals without measuring the RIR. To this end, we proposed a method for estimating the STI, MTF, reverberation time (T60), Deutlichkeit (D50), clarity index (C80), and early decay time (EDT) simultaneously (9). In this method, the temporal amplitude envelope of a reverberant speech signal is mapped to the parameters of an RIR model for each sub-band and the six parameters excluding the STI are calculated from the estimated temporal amplitude envelope.

We evaluate the proposed method in unseen reverberant environments. The root-mean-square errors between the ground-truths and estimated parameters suggest that the accuracy of the proposed scheme is close to the standardized measurements. This method is now being implemented in mobile equipment. Figure 4 shows an example of an operation screen of the sound environment monitoring system on a laptop computer.

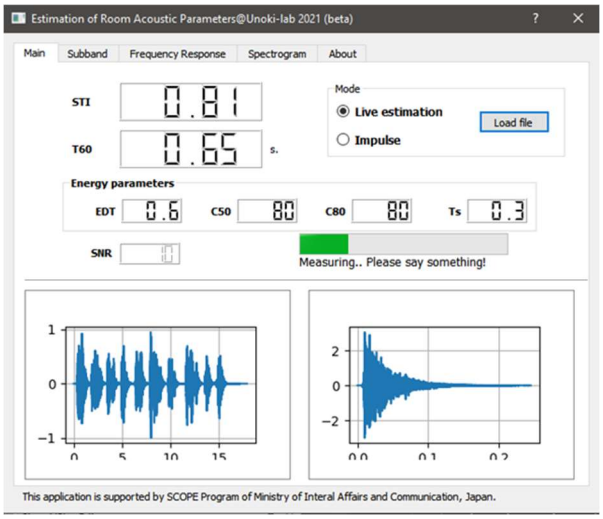


Figure 4 – Operation screen of sound environment monitoring system on a laptop computer.

3.1.2 Speech modification algorithm based on MTF concept

The basic concept of the algorithm is that speech is intelligible if its MS resists smearing of the MTF by the environment. Our resistant MS is calculated for multiple narrow acoustic frequency ranges; it is thus a two-dimensional spectrum on the two axes of acoustic frequency (AF) and modulation frequency (MF). Speech is analyzed by passing a speech wave through a band-pass (BP) filterbank on the AF; each wave output by the BP filterbank is transformed into an amplitude envelope and carrier, and the amplitude envelope is transformed into an MS on the MF at a certain AF. Thus, the MS for the outputs of the BP filterbank is a two-dimensional spectrum. The essential of this concept is to gain AF and MF regions of the smeared MS, MS_S under the effect of the environment that relate to the intelligibility and naturalness of speech.

Two-dimensional filters (2-D filters) are used to modify the plain speech MS efficiently by using both AF and MF filtering. The 2-D filter is not used the inverse MTF, MTF^{-1} simply but designed based on the estimated MTF by considering the intelligibility and naturalness of the modified speech having the resistant MS.

The problems of identifying acoustic and modulation frequency regions and their tuning gains are tackled separately. Here,

[1] we aim to obtain correlated acoustic and modulation frequency regions, which are called MS features by analyzing relationships between the estimated MS_S and the intelligibility and naturalness scores by conducting listening tests on enhanced and plain speech.

[2] The tuning amplitude levels or the gain of the 2-D filters for each feature are estimated from MTF^{-1} and are limited to within ranges with small gaps between them in order to make a fair comparison among the MS features. Listening tests were conducted to evaluate the intelligibility and naturalness of these modified feature combinations. The significant MS features were identified.

[3] The gains of the 2-D filters in the regions of the identified significant MS features were tuned by performing trials in different ranges to identify the optimal filter design for modifying the speech MS.

The detailed methodology can be found in the literature (2). Figure 5 illustrates the steps to convert plain speech into intelligible and natural speech by modifying the plain-speech MS. Figure 6 shows examples of an audio spectrum before and after the processing.

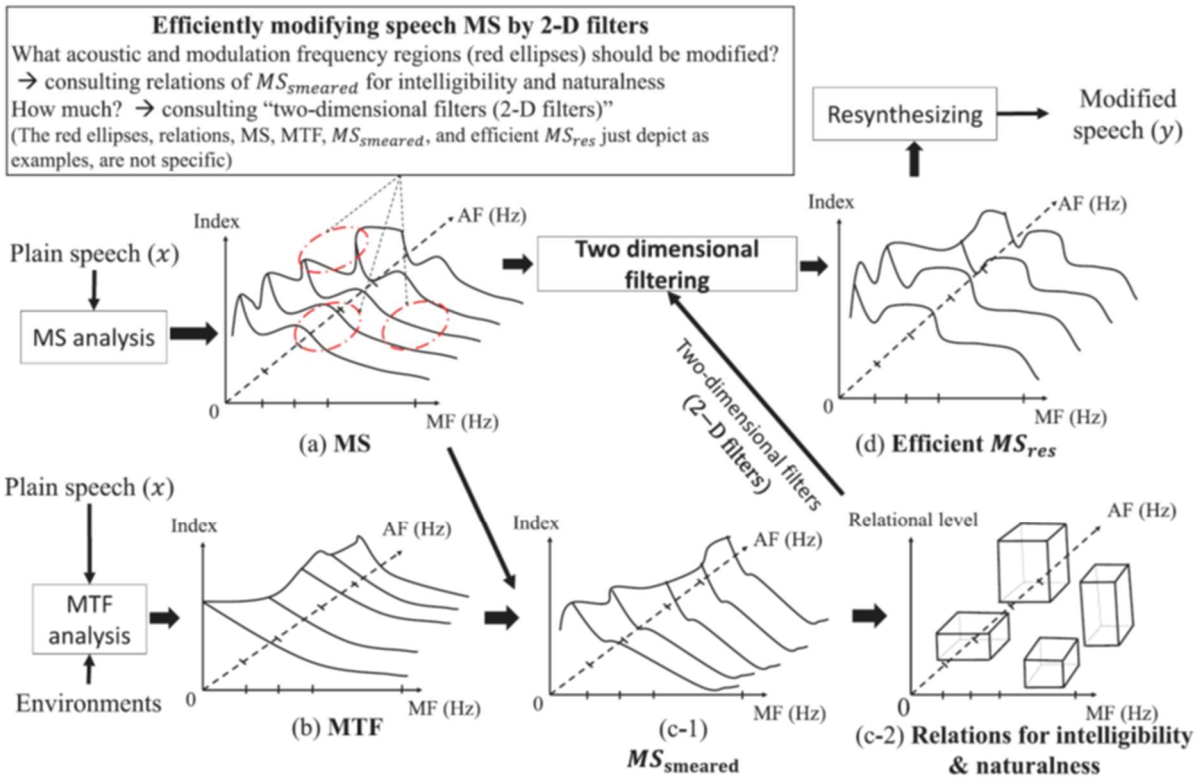


Figure 5 – Procedure of speech modification. AF and MF stand for acoustic frequency and modulation frequency respectively. Environments might contain both noise and reverberation (After Fig. 1 in (2)).

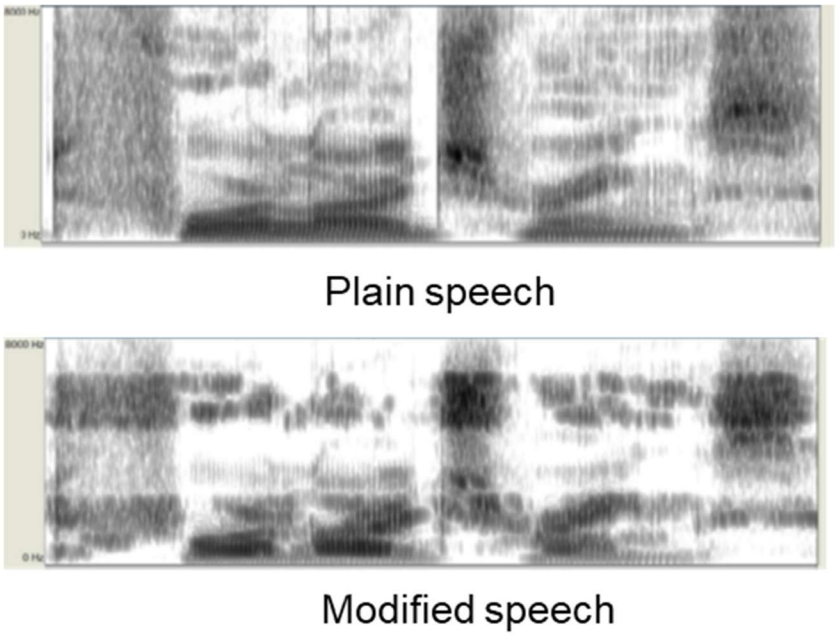


Figure 6 – Spectrograms of plain speech (top) and modified speech (bottom) for the English words ‘‘four large rings” (After Fig. 13 in (2)).

To evaluate the performance of the proposed algorithm, we participated in the Hurricane challenge 2.0 (HC 2.0) (10) and tried to improve the intelligibility of the provided speech data. The name of our algorithm in reference (11) which describes the results of the competition is MS500.

The speech material was in German and Spanish (100 sentences each) and English (90 sentences). It was recorded by male native speakers and was used as plain speech. The evaluation was performed by HC 2.0 with about 180 listeners. MS500 obtained an improvement in intelligibility for English and Spanish in all conditions but for German only in some conditions. An average improvement of about 4–18% in recognition rate compared with plain speech was obtained.

3.2 Mimicry of clear speech (12)

In most practical scenarios, the announcement system must deliver speech messages in a noisy environment in which background noise cannot be cancelled out. The noise reduces the intelligibility of the speech and increases listening effort; hence, it hampers the effectiveness of the announcement system. Recent studies have shown that the speech of professional announcers is more intelligible than speech of non-professionals in very noisy environments (13). This finding suggests that speech intelligibility might be related to the speaking style, which can be adapted by using a voice conversion method. Motivated by this idea, we devised a speech intelligibility enhancement for noisy environments by applying voice conversion to the voices of people who are not professional announcers.

The voice conversion method is based on StarGANv2 (14). An overview of the voice conversion model is shown in Fig. 7.

The training data consisted of utterances from 20 professional announcers from ATR dataset A-set and 20 non-expert speakers from ATR dataset C-set. All the utterances were preprocessed by resampling to 24 kHz. We followed the training strategy described in (15) with the same objective functions and hyper-parameters.

The speaker embedding encodes the speaker individuality conveyed in the input mel-spectrogram into a compact vector. By analyzing the speaker embedding by using principal component analysis (PCA), we can factorize out the dominant features of speaker individuality. Figure 8 plots the first and second principal components of the speaker embedding after training. The first component corresponds to the gender of the speakers, while the second component corresponds to the voice type, which is non-expert voice or professional announcer voice. This result suggests that the style of voice, i.e., non-expert or professional announcer, can be controlled independently from other voice attributes.

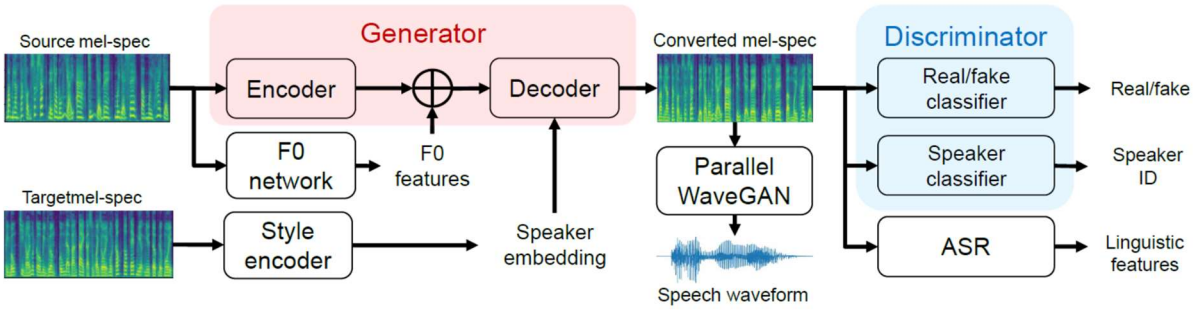


Figure 7 – Overview of StarGAN-v2 voice conversion model. The model consists of a generator network to convert the input mel-spectrogram, a discriminator network for adversarial training, a style encoder to extract the speaker embedding, a pretrained F0 network for extracting the F0 feature, a pretrained speech recognition to extract linguistic features, and a pretrained Parallel WaveGAN vocoder to generate the waveform from the mel-spectrogram.

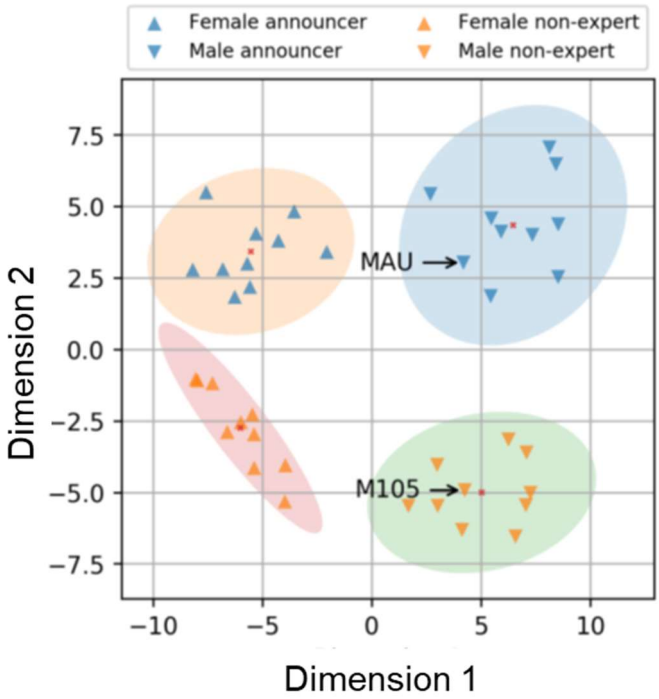


Figure 8 – 2D visualization of first and second principal components of speaker embedding. The red dots and shaded ellipses denote the centroid and covariance of each cluster. The M105 and MAU are the non-expert speaker and the target announcer used for the conversion.

To evaluate the proposed method, we carried out objective and subjective evaluations. As experiment speech stimuli, we selected 520 Japanese words, each containing 1 to 4 morae, from the ATR Digital Voice Database A-set (ATR-A) and ATR Digital Voice Database C-set (ATR-C) as clean stimuli for the target and source speaker. All of the speech waveforms were resampled at a 16 kHz sampling rate. There were four types of speech stimuli in the experiments, as follows:

Non-expert: Natural speech of a non-professional speaker, which was collected from speaker M105 in ATR-C.

Announcer: Natural speech of a professional announcer, which was collected from speaker MAU in ATR-A.

VC-1: Speech converted from speaker M105 to speaker MAU by using the voice conversion model. To check whether the announcer-adapted voice still possessed the properties of the natural announcer voice, we transformed the speaker individuality of speaker M105 to that of the target speaker MAU by using the voice conversion model. The speaker embedding of the MAU speaker was used to

synthesize the converted stimuli.

VC-2: Speech converted by shifting the second principal component of the speaker embedding of speaker M105 by using the voice conversion model. It was expected that the second principal component can be used to increase the intelligibility of the non-expert voice. To clarify this point, we replaced the second principal component of the M105 speaker embeddings with the average value calculated from the second principal component of all male professional announcers. Then, the obtained speaker embedding was used to synthesize converted stimuli.

Non-expert and Announcer were the reference stimuli.

To create the noisy stimuli, we masked the clean stimuli with pink noise at five different SNR levels: -9 dB, -6 dB, -3 dB, 0 dB, and ∞ (no noise).

3.2.1 Objective evaluation

Two objective metrics were used to evaluate the intelligibility of the converted speech: 1) average vowel space and 2) extended short-time objective intelligibility (eSTOI).

We compared the areas of the average vowel spaces derived from the different stimuli. The formant frequencies of five Japanese vowels (/a/, /e/, /i/, /o/, and /u/) were extracted using Praat. The locations of the vowels in each utterance were determined using the provided text transcription. The average frequencies of the first and second formants of the vowels were calculated across all speech utterances. The vowel space is defined as the smallest polygon that fits all the vowels. As can be seen from Fig. 9, the average vowel space of the professional announcer has the largest area. Interestingly, the converted voices from the non-expert speaker (VC-1 and VC-2) show an expansion of the vowel space from the non-expert vowel space.

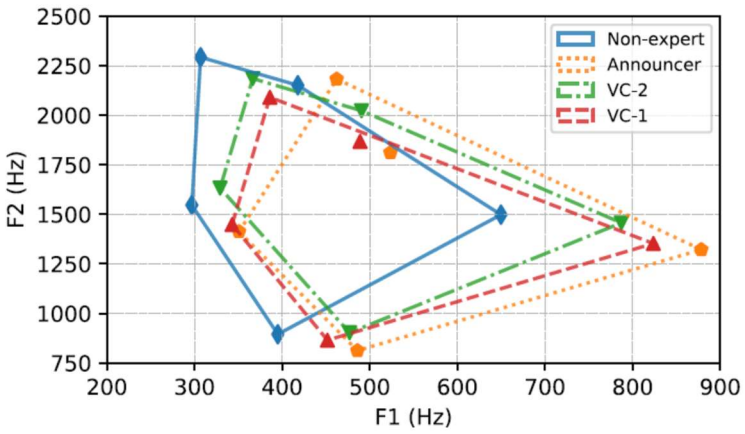


Figure 9 – Average vowel space of non-expert, announcer, VC-1, and VC-2 stimuli.

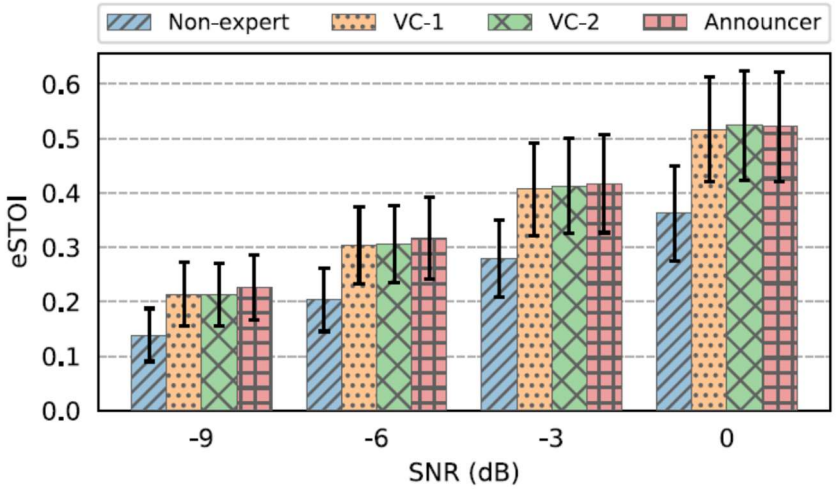


Figure 10 – Mean and standard deviation of eSTOI score of non-expert, announcer, VC-1, and VC-2 stimuli across all utterances. The horizontal axis is SNR in dB. The eSTOI score is in the range [0, 1]: a higher score indicates better intelligibility.

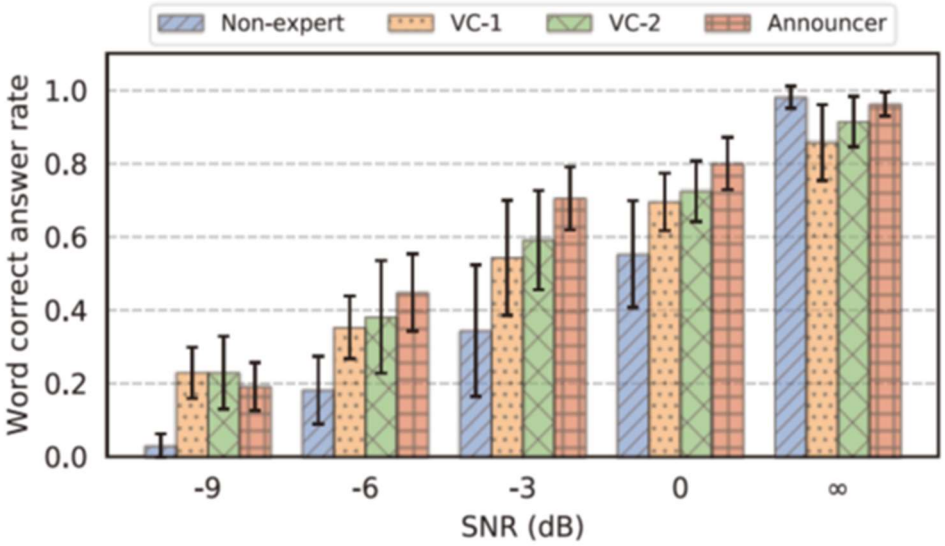


Figure 11 – Mean and standard deviation of word correct answer rates across participants. The horizontal axis is SNR in dB.

To objectively measure the intelligibility of speech in the presence of noise, we calculated the eSTOI of the speech stimuli at four SNR levels, -9 dB, -6 dB, -3 dB and 0 dB, using the pySTOI python package. The clean speech was used as the reference signal for the eSTOI calculation.

As can be seen from Fig. 10, the announcer voice resisted the noisy environment better than the non-expert voice, as expected. Moreover, the VC-1 and VC-2 stimuli showed comparable performance to that of the announcer voice.

3.2.2 Subjective evaluation

We conducted a listening test comparing the intelligibility of four types of speech stimuli. There were seven native Japanese participants. Each participant listened to a set of 300 different random words, which were equally distributed into five SNR levels and included four types of speech stimuli. Each stimulus was presented only once in each trial in a diotic fashion and the order of the presented stimuli was randomized for each participant.

Figure 11 reports the average word correction rate across participants. A one-way ANOVA test indicated statistical differences between the four types of stimuli. A post-hoc pairwise analysis using a Tukey HSD test ($p < 0.05$) was carried out to determine the statistical differences between pairs of stimulus types in different SNR conditions. The results indicated that the converted and announcer stimuli are significantly different from non-expert stimuli in noisy conditions. In addition, no statistical difference between the four types of stimuli was found in the clean condition.

The results of the objective measurements and subjective evaluation confirm that adapting the speech of a non-expert speaker to that of a professional announcer can increase its intelligibility. By modifying the second principal component of the speaker embedding, we can manually control the extent to which the announcer speaking style is used and hence increase the intelligibility of the speech in a noisy environment.

4. Conclusion

This paper described two different approaches for adaptive control of presented speech to increase its intelligibility in adverse conditions. One was modifying portions of the modulation spectra of the smeared voices. The other was a speech enhancement in which a voice conversion method is used to modify the speaking styles of non-professional voices to that of professional announcers. The results of objective and subjective evaluations confirmed that the intelligibility of the enhanced voices is higher than that of the original voices.

We did not discuss the control of linguistic and para-linguistic information according to the situation in this paper, due to the limitation on the number of pages. A detailed explanation is given

in reference (1). In future, by combining these technologies, we will construct a speech announcement system that can give evacuation guidance with high intelligibility.

ACKNOWLEDGEMENTS

This work was supported by the SCOPE Program of Ministry of Internal Affairs and Communications (Grant number: 201605002) and Grant-in-Aid for Scientific Research (Grant number: 20H04207). The research issues in this paper were mainly carried out by the project members, Dr. Masashi Unoki, Dr. Maori Kobayashi, Dr. Shunsuke Kidani, Dr. Suradej Duangpummet, Dr. Thuanvan Ngo, Dr. Tuan Vu Ho, and Dr. Masato Akagi.

REFERENCES

1. M. Kobayashi, Y. Hamada, and M. Akagi, “Acoustic features correlated to perceived urgency in evacuation announcements,” Speech Communication 139, 22-34 (2022).
2. T. Ngo, R. Kubo, and M. Akagi, “Increasing speech intelligibility and naturalness in noise based on concepts of modulation spectrum and modulation transfer function,” Speech Communication 135, 11-24 (2021).
3. T. Ngo, R. Kubo, and M. Akagi, “Mimicking Lombard effect: An analysis and reconstruction,” IEICE Trans. Information and Systems, E103-D, 5, 1108-1117 (2020).
4. T. Houtgast and H. J. M. Steeneken, “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria,” J. Acoust. Soc. Am. 77 (3), 1069–1077 (1985).
5. T. Houtgast and H. J. M. Steeneken, “The modulation transfer function in room acoustics as a predictor of speech intelligibility,” Acta Acust. United Acust. 28 (1), 66–73 (1973).
6. H. R. Bosker and M. Cooke, “Enhanced amplitude modulations contribute to the Lombard intelligibility benefit: Evidence from the Nijmegen Corpus of Lombard Speech,” J. Acoust. Soc. Am., 147 (2), 721-730 (2020).
7. M. Koutsogiannaki and Y. Stylianou, “Modulation enhancement of temporal envelopes for increasing speech intelligibility in noise,” Interspeech2016. 2508–2512 (2016).
8. A. Kusumoto, T. Arai, K. Kinoshita, N. Hodoshima, Nao, and N. Vaughan, 2005. “Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments,” Speech Communication 45 (2), 101–113 (2005).
9. S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki, “Blind estimation of speech transmission index and room acoustic parameters based on the extended model of room impulse response,” Applied Acoustics, 185, 108372 (2022).
10. J. Rennies-Hochmuth, M. Cooke, and C. Valentini-Botinhao, “The hurricane challenge 2020.” URL https://hurricane-challenge.inf.ed.ac.uk/
11. J. Rennies, H. Schepker, C. Valentini-Botinhao, and M. Cooke, “Intelligibility-enhancing speech modifications–the hurricane challenge 2.0,” Proc. Interspeech2020, Shanghai, China, 1341-1345 (2020).
12. T. Ho, M. Kobayashi, and M. Akagi, “Speak Like a Professional: Increasing Speech Intelligibility by Mimicking Professional Announcer Voice with Voice Conversion,” Interspeech2022, Inchon, Korea (Accepted).
13. M. Kobayashi and M. Akagi, “Intelligibility of announcer’s speech in noisy environments,” IEICE Technical Report, vol. 119, pp. 95–99, 2020.
14. Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, “Stargan v2: Diverse image synthesis for multiple domains,” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8185–8194 (2020).
15. Y. A. Li, A. Zare, and N. Mesgarani, “Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion,” Interspeech2021, 1349–1353 (2021).

PROCEEDINGS of the
24th International Congress on Acoustics
October 24 to 28, 2022 in Gyeongju, Korea

ABS-0547

Study on the modulation frequency range that contributes to the perception of urgency

Shunsuke KIDANI[1]; Xiaoting LIU[1]; Taiyang GUO[1]; Takuto ISOYAMA[1];
Junfeng LI[2]; Masashi UNOKI[1]

[1] School of Information Science, Japan Advanced Institute Science and Technology, Japan

[2] Institute of Acoustics, Chinese Academy of China, China

ABSTRACT

Non-linguistic information plays an important role in speech communication. Our previous study on noise-vocoded speech (NVS) showed that the temporal modulation cue in the temporal amplitude envelope (TAE) affects the perception of non-linguistic information such as urgency. We determined the upper or lower limitation of temporal modulation frequencies required for a sense of urgency perception using low-pass or high-pass filtering of TAEs. However, the frequency range of temporal modulation required for urgency perception has not yet been clarified. This study established an analysis-synthesis system of NVS with a modulation filterbank to identify the modulation frequency range that contributes to the perception of urgency. We compared NVS in which the TAEs were identical to those of the original speech with a band-limited NVS for the modulation frequencies. Urgent scales were derived from Scheffe's paired comparison of the results. The derived scales determined the relationship between the temporal modulation frequencies and urgency perception. The results show that the temporal modulation component contributes to the perception of urgency. Our new finding was that the temporal modulation frequencies important for urgency perception were 4 – 16 Hz, indicating that there must be a wide bandwidth. Therefore, the perception of urgency might be able to be manipulated by controlling the modulation frequencies.

Keywords: Modulation frequency, Urgency perception, Noise-vocoded speech, Temporal amplitude envelope, Non-linguistic information

1. INTRODUCTION

Speech communication plays an important role in human-to-human information communication. Non-linguistic information is important to convey the intent of speech to the dialoguer. If you hear a foreign language you do not know, you will still be able to determine the speaker's emotions, gender, etc. because non-linguistic information is not associated with language or culture (1). This example indicates that humans receive a lot of information from non-linguistic information.

There are many studies on what (acoustical) features of speech contain non-linguistic information perception (2-4). Previous studies showed a kind of static feature used in machine recognition (3, 4) or the fundamental frequency (F0) (2). These studies indicate the relationship between non-linguistic information and static feature. However, the relationship between non-linguistic information and temporal cues is not clear.

The temporal cue of speech perception by the temporal amplitude envelope (TAE) was demonstrated by experiments using noise-vocoded speech (NVS). NVS was proposed to reveal the primary cue of the recognition of linguistic information in speech perception (5). Studies using NVS showed that TAE of speech, as amplitude modulation of speech, is an important cue for speech perception (5, 6). Moreover, the temporal modulation frequency (TMF) band required for speech recognition is 4 – 16 Hz (7). These studies suggest that TAE might also contain a factor of perception on non-linguistic information.

[1] {kidani, s1810441, guotaiyang, isoyama-t, unoki}@jaist.ac.jp
[2] lijunfeng@hccl.ioa.ac.cn

Zhu et al. conducted psychoacoustic experiments using NVS and showed that the temporal cues involved speaker individuality (8, 9) and vocal emotion recognition (9, 10) as the non-linguistic information. They used a low-pass filter (LPF) to control the variability of TAEs by varying with cut-off frequency, revealing an upper limit on the TMF that is important for the perception of non-linguistic information. The results revealed that both speaker individuality and vocal emotion recognition have a TMF between 4 – 16 Hz. Additionally, the modulation spectral feature concerning to vocal emotion recognition is also shown (10).

Our previous study, which was based on studies by Zhu et al., showed that temporal cues are also involved in the perception of urgency as para-linguistic information (11). During a disaster, a voice with a sense of urgency is essential to encourage people to evacuate. The TMF band that causes the perception of urgency needs to be identified to manipulate the vocal urgency. Our previous study used the high-pass filter (HPF) in addition to the LPF to identify the upper and lower limits of the TMF involved in the perception of urgency. The results showed that the TMF band that contributes to the perception of urgency is at 6 – 8 Hz. However, it has not been possible to determine by the previous method whether 6 – 8 Hz of the TMF band is all that is needed to perceive urgency, or whether the band before and after are also needed.

This study aims to investigate the TMF band(s) that are important for perceiving urgency. Therefore, we construct a synthesis system for NVS using a modulation filterbank (MFB) to control speech in each modulation frequency band. This paper (I) determines whether NVS created using MFB and NVS created using previous methods are perceptually equivalent and (II) investigates the TMF bands that contribute to the perception of urgency.

2. NOISE-VOCODED SPEECH

2.1 Speech data

The input signal were real speech data: the evacuation call voices of a male announcer used in Kobayashi & Akagi (12) to investigate perception of urgency. The content of all the voices was "Please run away now" in Japanese ("*Ima sugu nigete kudasai*"). The speech data have four different types that have been shown to have different perceived urgency. These NVS stimuli were labeled "a," "b," "c," and "d."

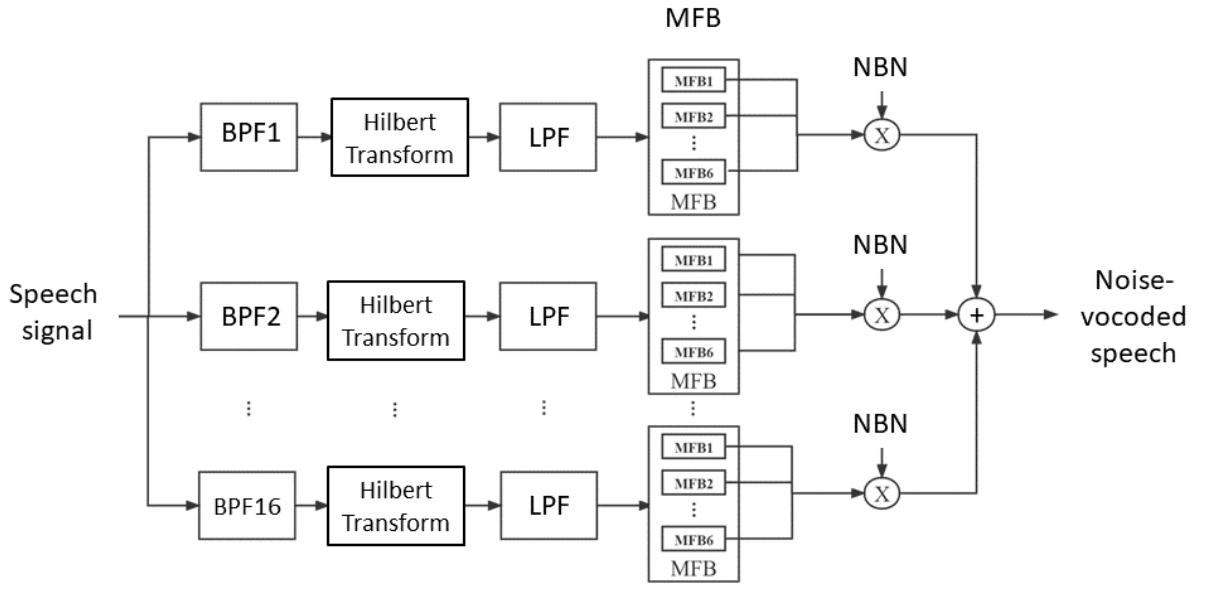


Figure 1 – Schematic diagram of noise-vocoder method with modulation filterbank used to generate stimuli.

2.2 Speech synthesis for generating NVS

Figure 1 schematically illustrates the signal processing to generate NVS using MFB. The procedure for making NVS is shown below. First, to reduce the effect of the average intensity, the active speech levels of all speech signals were normalized to -26 dBov by the same method as (9). Next, the input signal was divided into 16 frequency bands using an auditory filterbank. The 6th-order Butterworth band-pass filters (BPFs) of the Infinite Impulse Response (IIR) type were used as the auditory filterbank. The bandwidth of each filter was defined to be the bandwidth of the auditory filter, and the order of the filters was along the Equivalent Rectangular Bandwidth (ERB_N) and ERB_N-number scale (13). The relationship between ERB_N-number and acoustic frequency is defined as follows:

$$\mathrm{ERB_N\text{-}number} = 21.4\log_{10}\left(\frac{4.37f}{1000}+1\right) \tag{1}$$

where f is acoustic frequency in Hz. The boundary frequencies of the BPFs were defined from 3 to 35 ERB_N-number with bandwidth as 2 ERB_N.

Then, the amplitude envelope of the signal in each band was extracted using the Hilbert transform and a 2nd-order IIR-type Butterworth LPF, in each frequency band. The cut-off frequency was 64 Hz.

The TAE was divided into six bands for each modulation frequency band using MFBs. The MFBs utilized a group of IIR-type Butterworth BPFs. The MFBs have consisted of one LPF (0 to 2 Hz) and five LPF (a bandwidth of nth filter was 2^{n-1} Hz between 2 and 64 Hz).

Finally, the band signal is generated by multiplying the narrow band-limited noise (NBN) carriers restricted to each frequency band by the TAE controlled above. The NVS was generated by summing the bandwidth signals over the entire frequency range.

3. EXPERIMENT 1: PERCEPTUAL EQUIVALANCE OF DIFFERENT METHODS OF GENERATING NVS

Perceptual experiments were conducted to determine whether NVSs (stimuli No. a_1 – d_1) were generated using MFB, and whether NVSs (stimuli No. a – d) generated without MFB by using the same method used in previous studies (11) are perceptually equivalent. Stimuli with the same letters represent the same original voice.

3.1 Stimuli, Participants, and Procedure

Stimulus pairs for the pairwise comparison method were randomly made from the eight stimuli, "a – d" and "a_1 - d_1". The interval between the two stimuli was 0.5 s. The total number of paired stimuli was 28.

Eleven native Japanese speakers (three females and eight males, aged 22 to 25 years old) participated in experiment 1. All participants had normal hearing (hearing losses of the participants were below the hearing level of 12 dB in the frequency range from 125 to 8,000 Hz).

The experiment was conducted while the participants were in a soundproof room. The stimuli were simultaneously presented to both ears of a participant through a PC, audio interface (RME, Fireface UCX), and a set of headphones (Sennheiser HDA 200). The sound pressure levels were calibrated to be the same for all participants by using a head and torso simulator (B&K, type 4128) and sound level meter (B&K type 2231).

The stimuli were presented to the experimental participants. The experiment carried out using Scheffe's method of paired comparison (Ura variation) to evaluate the degree of urgency of stimuli. Participants were asked to evaluate whether or not the first stimulus was more urgent than the second one by using a five-grade evaluation measure (from "very strained (+2)" to "not quite strained (−2)").

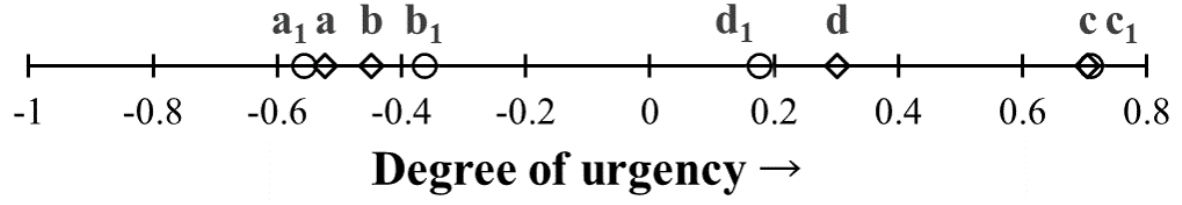


Figure 2 – Urgency scale of NVS. Diamonds indicate the previous method of generating NVS (without MFB), and circles indicate the proposed method of generating NVS (with MFB).

Table 1 – Corresponding table of modulation frequencies and stimulus conditions used in experiment 2. Fill indicates the number of MFBs used.

Stimulus condition	Channel number and frequency of modulation filterbank					
	1 0 – 2 Hz	2 2 – 4 Hz	3 4 – 8 Hz	4 8 – 16 Hz	5 16 – 32 Hz	6 32 – 64 Hz
Condition 1: a1 – d1	■	■	■	■	■	■
Condition 2: a2 – d2			■			
Condition 3: a3 – d3	■	■		■	■	■
Condition 4: a4 – d4				■		
Condition 5: a5 – d5	■	■	■		■	■
Condition 6: a6 – d6			■	■		
Condition 7: a7 – d7	■	■			■	■

3.2 Results

The psychological scale of urgency perception (named the urgency scale) obtained in experiment 1 is shown in Fig. 2. The horizontal axis represents the degree of urgency perception, with larger positive values indicating a higher perception of urgency. Diamonds indicate the generating method of NVS without MFB, and circles indicate the proposed method of generating NVS using MFB.

The result is that the same symbol is appended to the same position on the urgency scale. An analysis of variance (ANOVA) was conducted with differences in generating methods of NVS as a factor. According to the ANOVA, the main effect on generating methods of NVS ($F(7,507) = 171.5$, ($p < 0.01$)) was determined. This indicates that if the original voices are the same, they are perceived with the same degree of urgency, regardless of the NVS generating method. The results of experiment 1 show that NVS generated with MFB can be used to study the perception of urgency.

4. EXPERIMENT 2: CONTRIBUTION OF TMF BANDS TO PERCEPTION OF URGENCY

This experiment was conducted to investigate the TMF bands that contribute to the perception of urgency.

4.1 Stimuli, Participants, and Procedure

In experiment 2, the NVS was generated by controlling the TMF band. Table 1 shows the corresponding table of modulation frequency bands and stimulus conditions used to generate the NVS. The filled part in Table 1 indicates the number of MFBs used. Stimuli determined the frequency band to be controlled in this experiment because we considered that there is a core of urgency perception at 4 – 8 Hz in TMF based on the results of our previous study (11). Stimulus pairs for the pairwise comparison method were randomly made from the twenty-eight stimuli. The total number of paired stimuli was 756.

Eleven native Japanese speakers (four females and seven males, aged 22 to 25 years old) participated in experiment 2. All participants had normal hearing (hearing losses of the participants were below the hearing level of 12 dB in the frequency range from 125 to 8000 Hz).

In experiment 2, 28 stimulus pairs were used as one unit and a break was given for each unit. The stimuli presentation method, equipment, and procedure were the same as in experiment 1.

4.2 Results

Figure 3 shows that the results obtained in experiment 2. The results could be presented on a single urgency scale, but since the number of stimulus conditions is large and difficult to understand, urgency scales are presented for each stimulus condition being compared. The scale is the same as that in experiment 1.

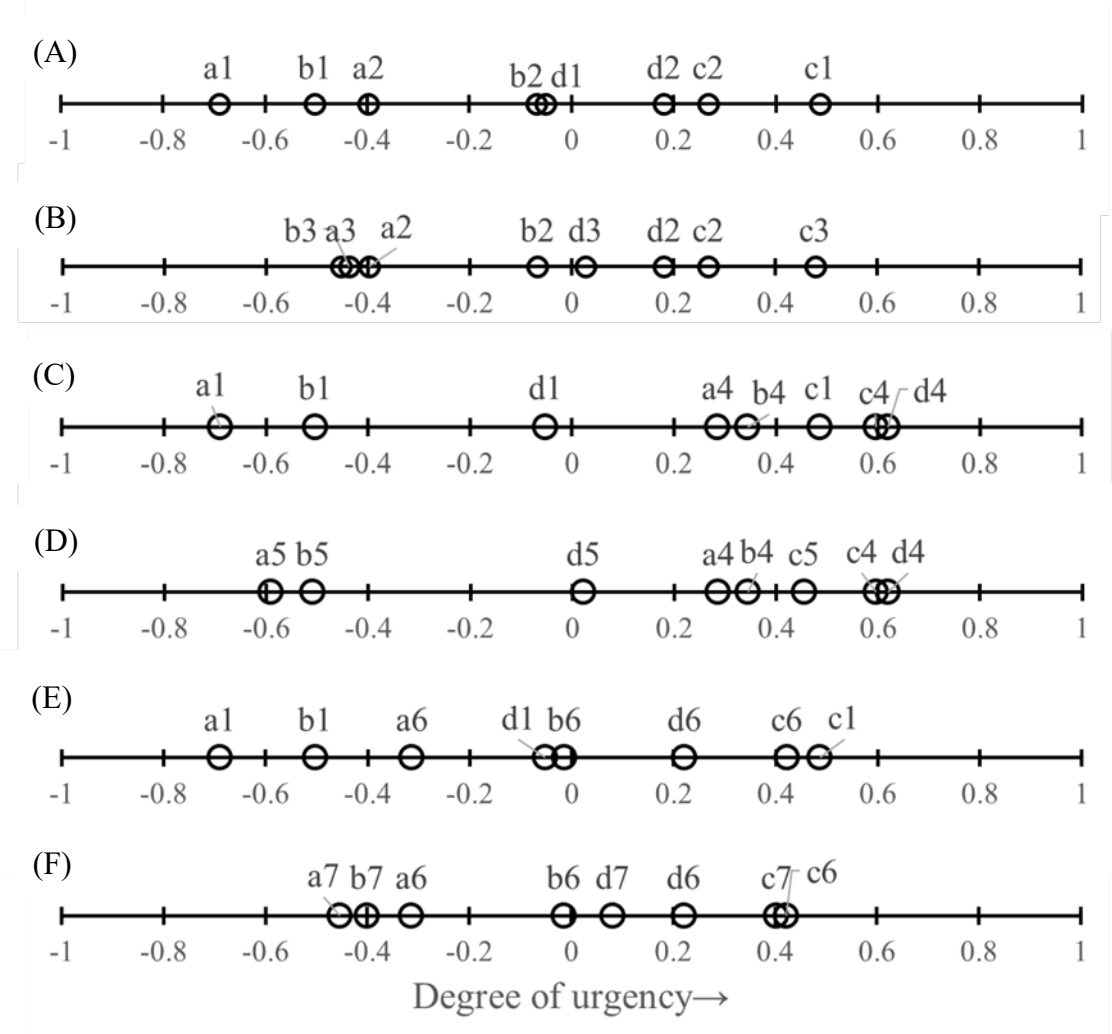


Figure 3 – Urgency scale of NVS which urgency perception with controlled the TMF bandwidth.

Fig. 3(A) shows that the perception of urgency decreases when 4 – 8 Hz is excluded, as predicted from our previous study (11). Fig. 3(B), however, shows that 4 – 8 Hz of the modulation frequency does not have a higher urgency perception compared with condition 1, which includes all modulation frequencies.

Figs. 3(C) and 3(D) show that the highest urgency is found in condition 4, which has only modulation frequencies between 8 and 16 Hz. However, the different order of urgency perception compared with condition 1 suggests that factors other than urgency might be involved.

In Fig. 3(E), condition 6, with the modulation frequencies between 4 and 16 Hz, is found to be comparable in urgency perception to condition 1. Fig. 3(F) shows that condition 7, in which this band is excluded, clusters toward 0 on the urgency scale.

5. DISCUSSION

The results of experiment 1 show that the contribution of TAE to non-linguistic information can be revealed using the analytical synthesis system of NVS with MFB. It is shown that the instantaneous modulation frequency of speech can be manipulated using an analytical synthesis system with MFB. Therefore, the effect of the instantaneous modulation frequency of speech on speech perception can be considered by using the proposed method of generating NVS in this paper.

The results of experiment 2 indicated that the upper and lower TMF limits determined using the LPF and HPF are not only modulation frequency bands that contribute to urgency perception. Method for the limiting of modulation frequency using LPF or HPF includes frequencies on the higher or lower side of the limited band in addition to the in-band frequencies. Therefore, it is considered that the range of modulation frequencies (4 – 8 Hz) shown in our previous studies could not explain all of the urgency perception.

To summarize and discuss the results of two experiments, it was suggested that the temporal cue for the

perception of urgency is not a single modulation frequency, but a simultaneous variation with several TMFs.

6. CONCLUSION

This paper generated the noise-vocoded speech (NVS) with modulation filterbank (MFB) and conducted perception of urgency experiments to evaluate the methods of generating NVS with/without MFB and investigate the temporal modulation frequency (TMF) band that contributes to the urgency perception. The results showed that (I) there is no difference in the methods of generating NVS, and (II) the TMF band between 4 and 16 Hz was important for the perception of urgency. Our future work is to investigate the contribution of temporal variation of modulation frequency to the perception of urgency.

ACKNOWLEDGEMENTS

This work was supported by the SCOPE Program of Ministry of Internal Affairs and Communications (Grant number: 201605002), JSPS-NSFC Bilateral Programs (Grant number: JSJSBP120197416), and Grant-in-Aid for Scientific Research (Grant number: 20KK0233, 21H03463).

REFERENCES

1. Laukka1 P, Elfenbein HA, Söder1 N, Nordström1 H., Althoff J., Chui W., Iraki FK., Rockstuhl T., and Thingujam NS. Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations. Front. Psychol., 30; 2013. doi: 10.3389/fpsyg.2013.00353.
2. Latinus M and Belin P. Human voice perception. Curr. Biol., vol. 21(4); 2011. pp. R143-R145.
3. Swain M, Routray A, and Kabisatpathy P. Databases, features and classifiers for speech emotion recognition: a review. Intl. J. Speech Technol., vol. 21; 2018. pp. 93 – 120.
4. Akçay MB and Oğuz K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun., vol. 116; 2020. pp. 56 – 76.
5. Shannon RV, Zeng FG, Kamath V, Wygonski J, and Ekelid M. Speech recognition with primarily temporal cues. Science, vol. 270; 1995. pp. 303 – 304.
6. Loizou PC, Dorman M, and Tu Z. On the number of channels needed to understand speech. J. Acoust. Soc. Am., vol. 106; 1999. pp. 2097 – 2103.
7. Drullman R, Festen J, and Plomp R. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Am., vol. 95(5); 1994. pp. 2670 – 2680.
8. Zhu Z, Nishino Y, Miyauchi R, and Unoki M. Study on linguistic information and speaker individuality contained in temporal envelope of speech. Acoust. Sci. & Tech., vol. 37(5); 2016. pp. 258 – 261.
9. Zhu Z, Miyauchi R, Araki Y, and Unoki M. Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech. Acoust. Sci. & Tech., vol. 39(3); 2018. pp. 234 – 242.
10. Zhu Z, Miyauchi R, Araki Y, and Unoki M. Contribution of modulation spectral features on the perception of vocal-emotion using noise-vocoded speech. Acoust. Sci. & Tech., vol. 39(6); 2018. pp. 379 – 386.
11. Unoki M, Kawamura M, Kobayashi M, Kidani S, and Akagi M. How the temporal amplitude envelope of speech contributes to urgency perception. Proc. ICA2019; 2019. pp. 1739 – 1744.
12. Kobayashi M and Akagi M. Psychological evaluation of evacuation announcements. J. Acoust. Soc. Jpn. vol. 74(12); 2018. pp. 633 – 640 (in Japanese).
13. Moore BCJ. An Introduction to the Psychology of Hearing (sixth edition). Brill Academic Pub. 2013.

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0721

Improving the accuracy of non-intrusive intelligibility estimation for reverberant speech using speech enhancement by optimizing the speech feature parameters

Kazushi Nakazawa[(1)], Kazuhiro Kondo[(2)]

[(1)]Yamagata University Graduate School of Science and Technology, Japan, ttk00758@st.yamagata-u.ac.jp

[(2)]Yamagata University Graduate School of Science and Technology, Japan, kkondo@yz.yamagata-u.ac.jp

ABSTRACT

Objective evaluation of speech intelligibility is a viable alternative to a time-consuming subjective evaluation in evaluating the performance of systems that process speech. Of these, non-intrusive objective evaluation is more practical than the intrusive methods, which require a clean speech signal corresponding to the speech under test. We previously proposed a non-intrusively intelligibility method that generates pseudo-clean speech using speech enhancement, mainly for reverberant speech. The frequency-weighted segmental (fwSNRseg) is calculated from the degraded speech and the pseudo-clean speech and is input to a neural network to estimate intelligibility. The fwSNRseg is calculated using the number of subbands and the SNR weight intensity optimized for this purpose. In the previous study, the effect of changing these parameters on the accuracy of intelligibility estimation was not evaluated in detail. Thus in this study, we aimed to improve the accuracy of intelligibility estimation by optimizing these parameters. We found that the linear correlation coefficient (LCC) between estimated intelligibility and subjective intelligibility was highest when the number of subbands was 50 and the SNR weight intensity was 2. The analysis between these parameters and LCC suggests a strong influence of the SNR weight intensity on the intelligibility accuracy.

Keywords: Speech intelligibility, Speech enhancement, Neural networks, Reverberation

1 INTRODUCTION

The development of communication systems and devices has allowed telephone conversations to take place in a variety of environments. However, in environments where noise and reverberation are mixed in, it can be difficult to use the system comfortably. Therefore, it is necessary to improve or maintain the quality of the system using intelligibility as an indicator. The simplest method to obtain intelligibility is the subjective evaluation method, in which a listening test is conducted on a human being. The intelligibility obtained by the subjective evaluation method is the most reliable way to evaluate a system used by humans, but it is expensive because it requires time-consuming testing on a large number of subjects. To solve this problem, objective evaluation methods that estimate intelligibility using an estimator have been developed. There are two types of objective evaluation methods: the intrusive method, which requires the degraded speech (target speech) and the corresponding original speech as input to the estimator, and the non-intrusive method, which requires only the target speech as input. The intrusive method is characterized by its ability to calculate the difference between the degraded voice and the original voice, and thus can estimate more accurately than the non-intrusive method. On the other hand, the non-intrusive method can be used even when the original voice is unavailable, because it can be estimated only from the degraded voice. Therefore, a non-intrusive model that can estimate intelligibility with high accuracy is needed.

We proposed a non-intrusive intelligibility estimation method using enhanced speech in our previous study and showed that intelligibility estimation is possible for additive noise degraded speech[1] and raververant speech[2]. In these studies, intelligibility estimation was performed using the feature fwSNRseg calculated from

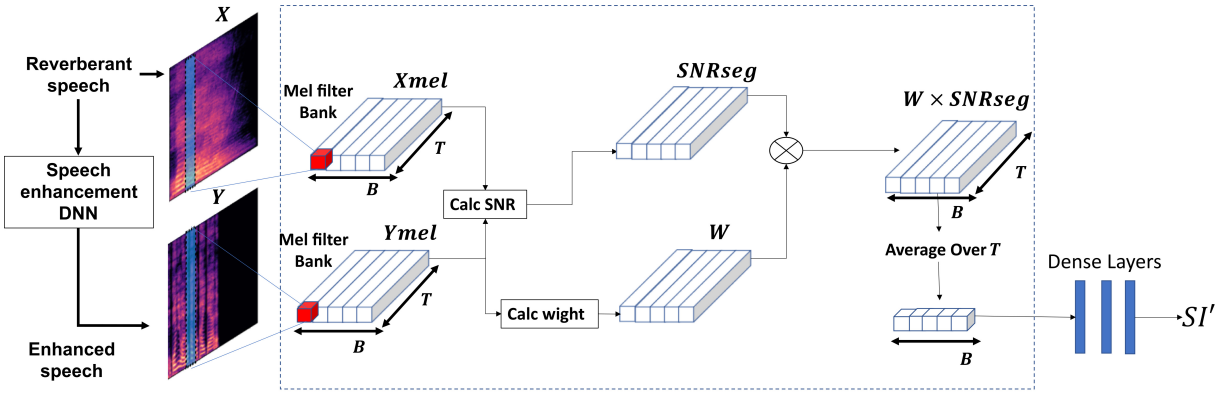


Figure 1. Overview of speech intelligibility estimation method. The input is reverberation-degraded speech and emphasized speech, and feature calculation is performed according to fwsnrseg to output estimated speech intelligibility SI' from the fully-connected DNN.

speech. The calculation involves parameters such as the number of filter bank bands and the strength of the weights, but in the previous study, only one parameter set was examined, and the effect of the parameters on the accuracy of intelligibility estimation was not confirmed. Therefore, this study aims to improve the intelligibility estimation accuracy of non-intrusive speech intelligibility method using enhanced speech by searching for the optimal parameters for intelligibility estimation for reverberant speech.

2 SPEECH AND SUBJECTIVE INTELLIGIBILITY DATA

The speech used in this study is the Japanese version of the DRT speech set[3] .We generated reverberated speech by convolving Room ImpulseResponse (RIR) with the above speech. RIR is simulated using the RIR Generator[4]. The RIRs were generated with reverberation times of 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, and 1.2. These RIRs were convolved with the JDRT speech data of three speakers (two males and one female) to generate the reverberant speech. The sampling frequency of the speech data was 16 kHz and the number of quantization bits was 16 bits. We conducted a listening test according to the DRT on subjects using the above reverberated speech to measure their subjective intelligibility. The intelligibility used in this study is the mean value obtained from seven healthy hearing subjects in their 20s.

3 SPEECH ENHANCEMENTMODEL

Spectral mapping based on BiLSTM[5] is used for speech enhancement. The input is a spectrogram of reverberation-degraded speech, with 20 frames of spectrum input and a clean spectrum of the target frame output. The configuration consists of two 512-unit BiLSTM layers and a subsequent fully connected layer. The phase obtained from the reverberation degraded speech is used to recover the speech from the estimated spectrogram.

4 FEATURE CALCULATION

In this study, fwSNRseg is used as a feature. This feature has been conventionally used as a intrusive objective intelligibility index[6, 7]. It has also been shown to be effective in non-intrusive intelligibility estimation using emphasized speech. The calculation procedure is shown in Fig. 2, with the input spectrogram X of the reververant speech and the input spectrogram Y of the enhanced speech. A mel filter bank with B subbands is applied to these spectrograms to generate melspectrograms $Xmel$ and $Ymel$. The SNR for each frame and subband of these melspectrograms is calculated according to Equation 1.

$$SNRseg(b,t) = 10\log_{10}\frac{Y_{mel}^2(b,t)}{\{X_{mel}(b,t) - Y_{mel}(b,t)\}^2} \tag{1}$$

Calculate the weights W for the SNRs according to Equation 2.

$$W(b,t) = |Y_{mel}(b,t)|^p \tag{2}$$

Multiply *SNRseg* by the weights W, average them in the time direction, and use them as the input to the fully-connected DNN.

$$fwSNRseg(b) = \frac{1}{T}\sum_t \frac{W(b,t)SNRseg(b,t)}{\sum_b W(b,t)} \tag{3}$$

5 EXPERIMENT

We use reverberated speech of reverberation durations 0.2, 0.4, 0.6, 0.8, and 1.0 seconds and the corresponding intelligibility as training data, and reverberated speech of reverberation durations 0.3, 0.5, 0.7, 0.9, and 1.2 seconds and intelligibility as test data. The enhanced speech DNN and intelligibility estimation DNN are trained using the above training data, and the estimation accuracy is evaluated on test data. In this study, the spectrogram is calculated by a 512-point FFT with 75% overlap using a Hanning window.

5.1 SPEECH ENHANCEMENT DNN

In training the enhanced speech DNN, supervised learning is used, where reverberated speech is used as input and the corresponding clean speech is used as the teacher data. The loss function is the mean squared error between the estimated spectrum and the true clean spectrum, and the optimization method is Adam with 1000 epochs of training.

5.2 FEATURE CALCULATION AND INTELLIGIBILITY ESTIMATION DNN

The parameters to be explored in this study are the number of bandwidth divisions B of the mel filter bank and the SNR weight intensity p. The search ranges are set to 16, 32, and 50 for B and 0, 1, and 2 for p, respectively. The fully-connected layer consists of 5 layers of 128 units. The loss function is the mean squared error between the subjective intelligibility SI and the estimated intelligibility SI', and the optimization method is Adam, which is used to train 1000 epochs. We cross-validated the training data by dividing it into 10 parts.

5.3 EVALUATION

We evaluate estimation accuracy using the linear correlation coefficient and RMSE between subjective intelligibility and estimated intelligibility for the test data.

6 RESULT

Fig.42 and Fig.3 show the LCC and RMSE between estimated intelligibility and subjective intelligibility in the test data. The horizontal axis represents the SNR weight intensity p, and the legend represents the number of bandwidth divisions B. It can be seen that LCC tends to increase and RMSE tends to decrease as p and B increase. Based on these results, the accuracy of intelligibility estimation can be improved by adjusting the parameters of feature calculation.

7 DISCUSSION

The reasons for these results are discussed in terms of the number of subbands B in the Melfilter bank and the SNR weight intensity p. The accuracy of intelligibility estimation as the number of subbands B increases, and we speculate that this is because a finer analysis of frequency components makes it possible to extract acoustic features of consonants that affect intelligibility.

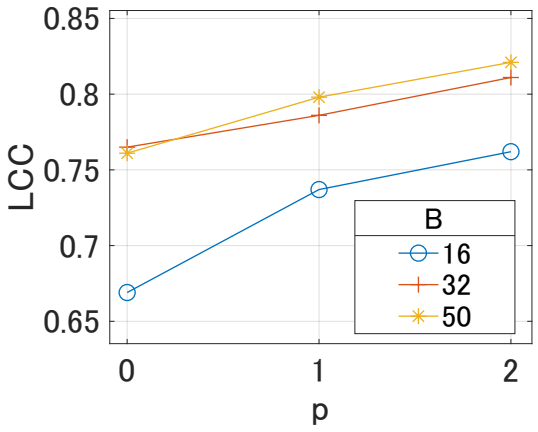


Figure 2. Linear correlation coefficient (LCC) between estimated intelligibility and subjective intelligibility. B is the number of subbands in the mel filter bank and p is the SNR weight intensity.

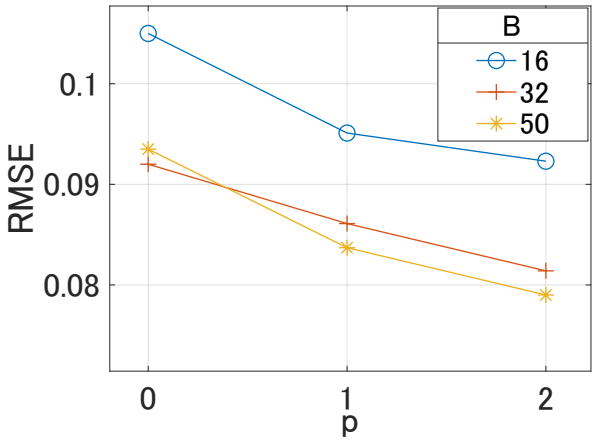


Figure 3. RMSE between estimated intelligibility and subjective intelligibility. B is the number of subbands in the mel filter bank and p is the SNR weight intensity.

As for p, it represents the intensity of the mask W that extracts the location of enhanced speech from the SNR from Equations(2) and (3), and the intensity of W varies with p. Therefore, it is assumed that by selecting an appropriate p, it is possible to effectively extract the SNR of the location where the speech is present.

8 CONCLUSION

In this study, we attempted to improve intelligibility estimation in non-intrusive speech enhanced speech by adjusting parameters during feature computation. As a result, the intelligibility estimation accuracy can be improved by increasing the number of subbands in the Melfilter bank used for bandwidth segmentation and by increasing the SNR weight intensity , and the highest accuracy was obtained with a LCC of 0.821 when the number of subbands was 50 and the SNR weight intensity was 2. However, the filter bank used in this study was not used. Future plans include verifying the improvement in estimation accuracy by using a filter bank other than the Mel filter bank used in this study. It is also necessary to verify that the system works well with speech degraded by both reverberation and additive noise.

REFERENCES

[1] Hiroto Takahashi and Kazuhiro Kondo. On Non-Reference speech intelligibility estimation using DNN noise reduction. In *Proceedings of the 23rd International Congress on Acoustics*, pages 3103–3108, September 2019.

[2] Kazushi Nakazawa and Kazuhiro Kondo. On non-reference speech intelligibility estimation using dnn dereverberation. *Proc. IEEE Global Conference on Consumer Electronics*, 10 2020.

[3] Kazuhiro Kondo, Ryo Izumi, Masaya Fujimori, Rui Kaga, and Kiyoshi Nakagawa. Two-to-one selection-based japanese speech intelligibility test. *Journal of the Acoustical Society of Japan*, 63(4):196–205, 4 2007.

[4] E.A.P.Habets. Room impulse response (RIR) generator. https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator, 2008.

[5] M Schuster and K K Paliwal. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, 45(11):2673–2681, November 1997.

[6] Yi Hu and Philipos C Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio Speech Lang. Processing*, 16(1):229–238, January 2008.

[7] Jianfen Ma, Yi Hu, and Philipos C Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *J. Acoust. Soc. Am.*, 125(5):3387–3405, May 2009.

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0760

Subjective evaluation regarding mixing ratio of bone-conducted to air-conducted speech for own-voice perception

Teruki TOYA[(1)], Peter BIRKHOLZ[(2)], Masashi UNOKI[(1)]

[(1)]Japan Advanced Institute of Science and Technology, Japan, {t-toya, unoki}@jaist.ac.jp

[(2)]Technische Universität Dresden, Germany, peter.birkholz@tu-dresden.de

ABSTRACT

To clarify human speech perception and production mechanisms, perceptual properties of speakers' own voices transmitted via bone conduction should be further understood. In our previous studies, transfer functions of ear-canal sound pressure and regio temporalis vibration relative to oral-cavity sound pressure were physically measured. The former and latter were assumed to represent perceptual properties of the outer-ear and middle/inner-ear part of bone-conducted (BC) speech transmission during voicing, respectively. This paper investigated the contribution of BC speech reaching the middle/inner ear to one's own voice perception, as well as that reaching the outer ear, by making two types of filtered speech signals based on the measured transfer functions. The mixing ratio of speech signals filtered by each transfer function (i.e., mimicked BC speech signals) relative to the original air-conducted (AC) speech signals were determined through subjective evaluations. It was found that two filtered speech signals were mixed at almost the same level, and a combination of two filtered speech signals were mixed at almost the same level as the AC speech. These findings suggest that the middle/inner-ear part of transmission contributes to BC speech perception at almost the same extent as the outer-ear part.

Keywords: Own voice, air-conducted speech, bone-conducted speech, transmission characteristics, mixing ratio

1 INTRODUCTION

During speaking, humans perceive their own voices to control their speech production systems (1). There are two different types of sound transmission of one's own voice: air-conducted (AC) and bone-conducted (BC) speech. While the voice is generated by the larynx, modified by the vocal tract, then transmitted as AC speech to the auditory system, the sound inside the vocal tract is also transmitted to the auditory system through the soft tissue and the skull bone as BC speech. To further explore human speaking/hearing mechanisms, both AC and BC speech transmission processes during the perception of one's own voice need to be clarified.

The effect of BC speech perception on one's speech production has been investigated using a delayed auditory feedback technique (2), showing that BC speech perception affects one's speech production similarly to AC speech perception. However, the transmission process of BC speech has not been well understood.

Stenfelt previously proposed a model of BC speech transmission pathways for one's own voice (3). This model shows the relationships between the soft tissue/skull bone vibration and each part of the auditory system on the basis of the physiological aspects of BC hearing. Vibrations of the soft tissue and that of the skull bone were hypothesized to produce the ear-canal (EC) sound pressure, respectively. The skull-bone vibration was also hypothesized to cause inertial forces in the middle ear ossicles and the cochlea fluid.

One of the largest issues in investigating the BC speech transmission process is to identify the contribution of each pathway to one's own voice perception (hereafter, own-voice perception). Békésy showed in his perceptual investigation that BC speech transmission has almost the same order of magnitude as AC speech transmission (4). Reinfeldt *et al.* clarified the contribution of BC speech reaching the outer ear to own-voice perception for multiple phonemes, showing that the magnitude of BC speech dominated at frequencies between

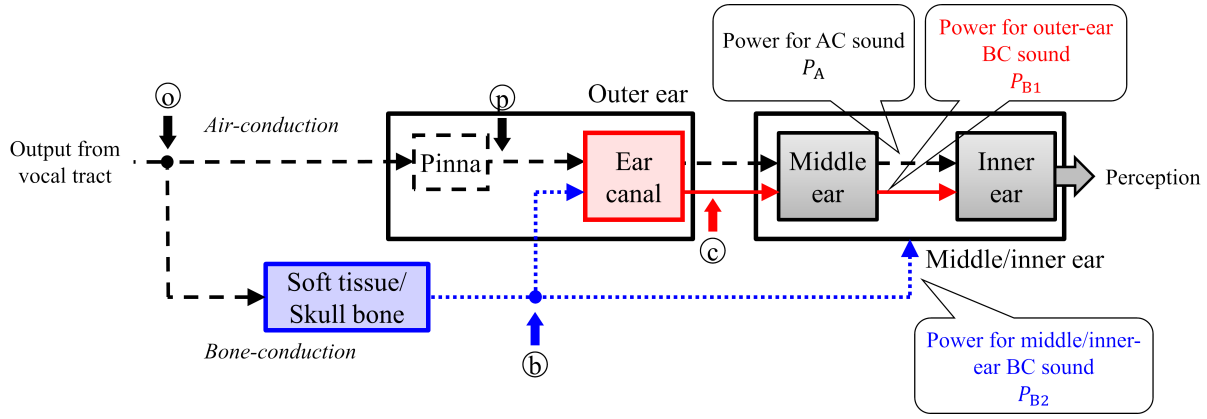


Figure 1. Simple model of AC/BC speech transmission pathways for own voice. AC/BC speech transmission process is assumed as summation of power of signal reached from each pathway to inner ear.

1 and 2 kHz (5). Pörschmann (6) investigated the perceptual relationship between AC and BC speech as a function of frequency, concluding that BC speech has a perceptually greater magnitude at frequencies between 0.7 and 1.2 kHz than AC speech, but the contribution of the outer ear component to BC speech transmission is very small (6). Currently, it is still not clear whether outer- or middle/inner-ear components of BC speech transmission dominate own-voice perception.

This paper aims to identify the contribution of BC speech reaching the middle/inner ear to own-voice perception, as well as that reaching the outer ear. Here, we focus on two types of observable BC speech components: the sound radiated into the ear canal (EC) and the vibration of the regio temporalis (RT). This study attempts to mimic the spectra of the EC sound radiation and RT vibration due to the oral-cavity (OC) sound pressure. The voice timbre for mixed stimuli with AC and two types of mimicked BC speech is evaluated to determine the perceptually optimal mixing ratio of those speech signals.

2 CONCEPTS FOR INVESTIGATING PERCEPTUAL CONTRIBUTION

On the basis of Stenfelt's model (3), this paper hypothesizes a simple model of AC and BC speech transmission pathways for one's own voice. Figure 1 shows an overview of our hypothesized transmission pathways. Although the transmission characteristics for the middle/inner ear parts of bone-conduction cannot be directly measured, the transmission characteristics from the speech production system to the outer ear through the soft tissue and skull bone can be measured physically. Here, the middle-ear ossicular vibration relative to the RT vibration does not drastically fluctuate across frequencies (7). Skull-bone vibration is also reported to cause an excitation in the cochlea (8). Considering these facts, it is assumed that the spectral shape of the RT vibration itself represents the contribution of the middle/inner ear parts of bone-conduction to own-voice perception.

The authors previously measured the transmission characteristics of bone conduction focusing on EC sound pressure and RT vibration due to acoustic excitation in the OC (9, 10). Figure 2 shows the amplitude characteristics of the transfer functions of EC sound pressure relative to OC sound pressure ($|H_{oc}(f)|$) and that of RT vibration relative to OC sound pressure ($|H_{ob}(f)|$) averaged across five participants, which was measured in the authors' previous study (10). The characters "o", "b" and "c" correspond to the observation positions ⓞ, ⓑ, and ⓒ in Fig. 1. On the basis of the assumption stated above, the measured transfer functions $|H_{oc}(f)|$ and $|H_{ob}(f)|$ are regarded as the transmission characteristics of the outer-ear BC and middle/inner-ear BC speech pathways, respectively.

Here, the total power of the own-voice signal reaching the inner ear P_{OV} can be represented as follows:

$$P_{OV} = P_A + P_{B1} + P_{B2}, \tag{1}$$

where P_A, P_{B1}, and P_{B2} denote the power for AC speech signals, BC speech signals reaching the outer ear, and

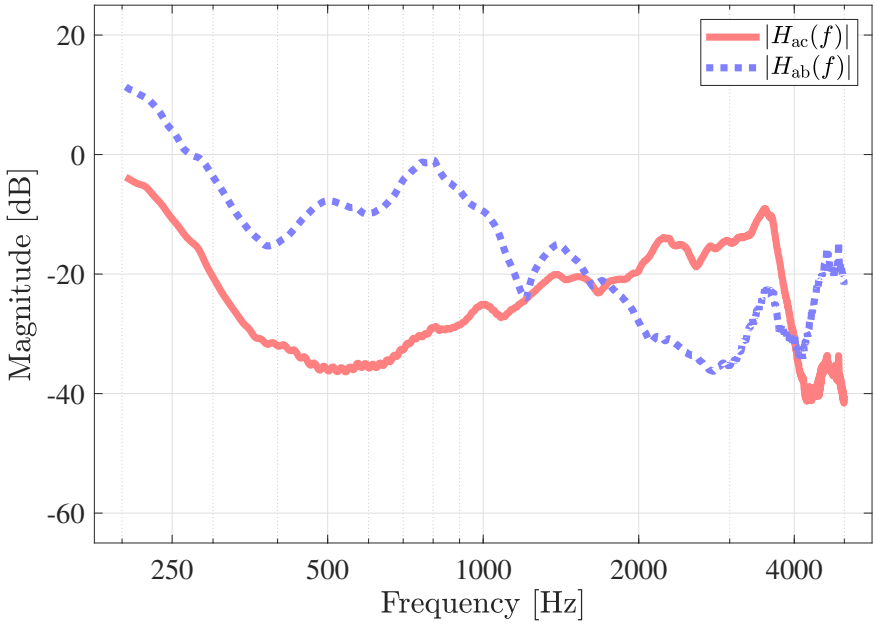


Figure 2. Transfer functions of EC sound pressure ($|H_{oc}(f)|$) and RT vibration ($|H_{ob}(f)|$) relative to OC sound pressure measured by Toya *et al.* (10)

BC speech signals reaching the middle/inner-ear, respectively. Note that the phase characteristics of the different transmission pathways were ignored since those could not be measured directly.

In this paper, the mixing ratio between P_A, P_{B1} and P_{B2} was investigated subjectively using recorded and filtered AC speech signals.

3 SPECTRALLY-MIMICKED BC SPEECH

3.1 Vocal recording

Seven students (five males and two females, aged 22 to 29) participated in the production tasks for vocal recording. All were native Japanese speakers with normal hearing, and none had a speaking disorder.

Production tasks for recording were conducted in a soundproof room. Speakers' voices were recorded through a microphone (Rode NT1-A), located 10 cm from their lips. The recorded signals were routed through an amplifier and an A/D converter (Steinberg UR44) to a PC (Windows 10, with MATLAB 2020a). The sampling frequency was 44.1 kHz and the number of quantizing bits was 16. The speakers were asked to utter a Japanese vowel /a/ with as constant vocal intensity as possible. The utterance duration was around 2.5 sec.

3.2 FIR filter design and filtering for spectral mimicking

Two types of FIR filters were designed to mimic the spectrum of BC speech signals from the recorded AC speech signals. In Eq. (1), the total power P_{OV} is defined as the summation of power of the multiple signals at the inner ear, while mixed signals of recorded AC speech and simulated BC speech can just be presented as AC sound at the EC entrance (shown as ⓟ in Fig.1). To compensate the measured transfer functions ($|H_{oc}(f)|$ and $|H_{ob}(f)|$) for the transfer functions between the middle/inner ear and EC entrance, the desired transfer functions for mimicking the spectrum of BC speech reaching the outer ear $|H_{ocp}(f)|$ and that reaching the middle/inner ear $|H_{obp}(f)|$ were defined as follows:

$$|H_{ocp}(f)| = |O^{-1}(f)||H_{oc}(f)|, \tag{2}$$

$$|H_{obp}(f)| = |O^{-1}(f)||M^{-1}(f)||H_{ob}(f)|, \tag{3}$$

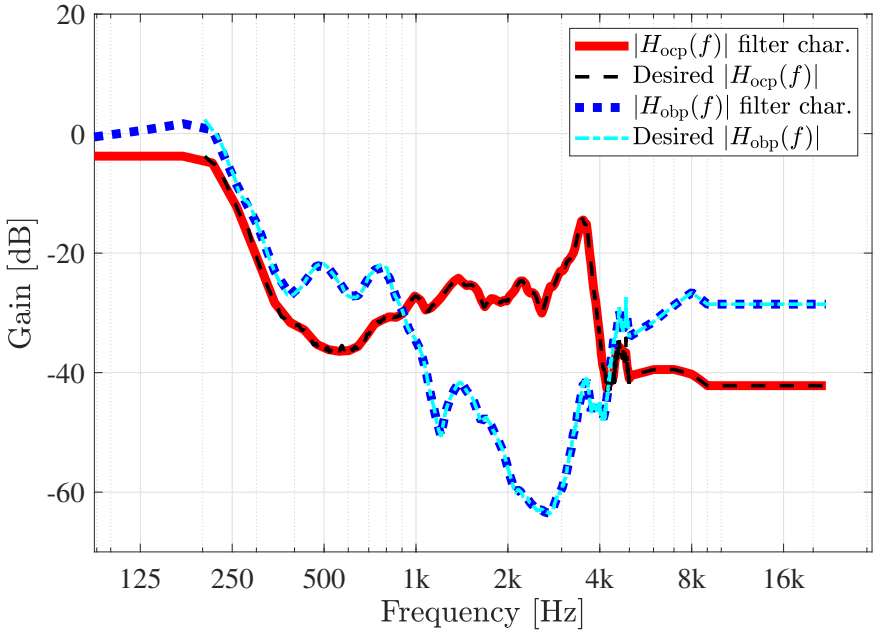


Figure 3. Amplitude characteristics of filters for mimicking spectra of BC speech reaching outer ear ($|H_{\mathrm{ocp}}(f)|$) and middle/inner ear ($|H_{\mathrm{obp}}(f)|$)

where $|O^{-1}(f)|$ denotes the inverse transfer function of EC and eardrum during AC sound transmission reported by Shaw (11), and $|M^{-1}(f)|$ denotes the inverse transfer function of the middle ear ossicles reported by Aibara *et al.* (12). On the basis of Eqs. (2) and (3), FIR filters for mimicking the spectra of BC speech reaching the outer- and middle/inner-ear were designed, so that the squared error between the filter characteristics and each desired transfer function ($|H_{\mathrm{ocp}}(f)|$ and $|H_{\mathrm{obp}}(f)|$) was minimized. The sampling frequency was set to 44.1 kHz and the filter order was set to 8192. Here, phase characteristics of the desired $|H_{\mathrm{ocp}}(f)|$ and $|H_{\mathrm{obp}}(f)|$ were unknown. Therefore, the FIR filters were constructed with a linear phase characteristics.

Figure 3 shows the amplitude characteristics of the filters for mimicking the spectra of the BC speech reaching the outer ear ($|H_{\mathrm{ocp}}(f)|$) and the middle/inner ear ($|H_{\mathrm{obp}}(f)|$). The recorded speech signals for each speaker were filtered by those filters to make two types of spectrally-mimicked BC speech signals.

4 EXPERIMENT

Subjective evaluation of own-voice timbre was conducted to determine the mixing relationship among P_{A}, P_{B1} and P_{B2}. The seven speakers mentioned in Sec. 3.1 participated in this experiment.

4.1 Stimuli

For mimicking each participant's own voice, their own recorded AC speech and the spectrally-mimicked BC speech reaching the outer and middle/inner ear were mixed under a certain power combination (P_{A}, P_{B1}, and P_{B2}). P_{A} was determined so that the A-weighted sound pressure level in the headphone was 60 dB. The mixing ratio of total BC speech to AC speech (L_{BA}) and that of BC speech reaching the middle/inner ear to BC speech reaching the outer ear (L_{B2B1}) were defined as follows:

$$L_{\mathrm{BA}} = 10\log_{10}\frac{(P_{\mathrm{B1}}+P_{\mathrm{B2}})}{P_{\mathrm{A}}}, \tag{4}$$

$$L_{\mathrm{B2B1}} = 10\log_{10}\frac{P_{\mathrm{B2}}}{P_{\mathrm{B1}}}. \tag{5}$$

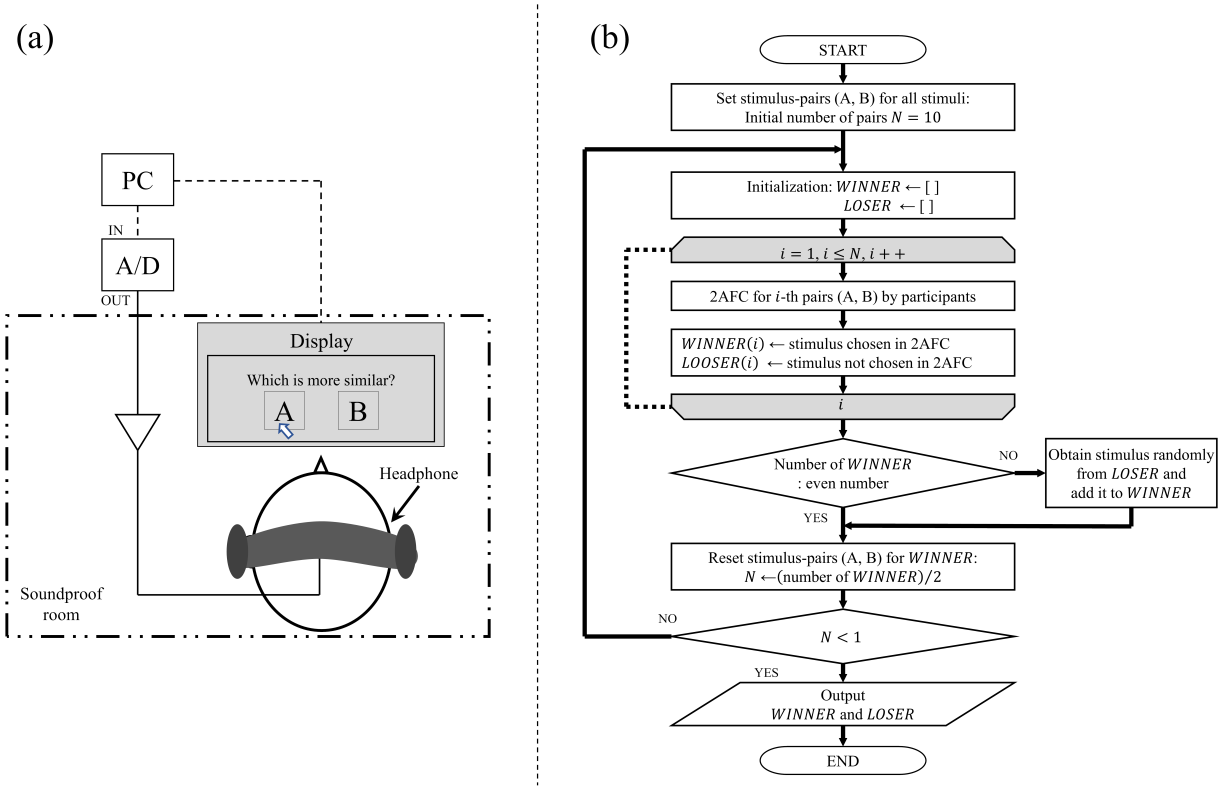


Figure 4. Schematic illustration of experimental setup and flow chart of stimulus presentation for subjective evaluation of own-voice quality

L_{BA} was set to $L_{\mathrm{BA}} = -10, 0$, and 10 dB, considering the fact that BC speech transmission has almost the same order of magnitude as AC speech transmission (4). L_{B2B1} was set to $L_{\mathrm{B2B1}} = -6, -3, 0, 3$, and 6 dB. Additionally, following five stimuli were also prepared as extra conditions:

Original The recorded AC speech signal itself,
Low-1 The low-pass filtered AC speech signal with a cut-off frequency of 1 kHz,
Low-4 The low-pass filtered AC speech signal with a cut-off frequency of 4 kHz,
High-1 The high-pass filtered AC speech signal with a cut-off frequency of 1 kHz,
High-4 The high-pass filtered AC speech signal with a cut-off frequency of 4 kHz.

Those stimuli for the extra conditions were assumed to be dissimilar perceptually to the participants' own produced voice. Therefore, those are prepared to confirm whether they could successfully compare the similarity of the stimuli to their own produced voice.

The total number of conditions was 20 (3 variants of L_{BA} $\times$ 5 variants of L_{B2B1} + 5 extra conditions).

4.2 Apparatus and procedure

Figure 4 shows a schematic illustration of the experimental setup and a flow chart of stimulus presentation. The experiment was conducted in the same soundproof room mentioned in Sec. 3.1. An open-air headphone (Stax SR-L700) was used for presenting experimental stimuli. For controlling the stimulus presentation, the same A/D converter and the PC mentioned in Sec. 3.1 were used. A GUI on MATLAB 2021a were used for obtaining participants' answers.

A tournament-style listening test with two alternative forced choice (2AFC) trials was adopted in the experiment. The participants were asked to listen to a pair of stimuli (A and B), and choose the one which is

Table 1. Number of times stimulus was chosen as first and second places by all participants for tournament-style listening test

L_{BA} \ L_{B2B1}		−6 dB	−3 dB	0 dB	3 dB	6 dB
10 dB	(1st)	0	0	2	2	0
	(2nd)	3	1	1	0	2
0 dB	(1st)	0	7	5	2	4
	(2nd)	2	4	6	4	4
−10 dB	(1st)	3	5	0	3	1
	(2nd)	0	0	4	1	1
Extra		Low-1	Low-4	Original	High-1	High-4
	(1st)	0	0	1	0	0
	(2nd)	0	0	2	0	0

perceptually more similar to their own produced voice. Before answering each trial, they were allowed to utter the vowel /a/, which enabled them to compare the voice quality between the stimuli and their own produced voice. For the paired stimuli (A and B) in each trial, the stimulus chosen in 2AFC is regarded as the *WINNER*, while the stimulus not chosen is regarded as the *LOSER*.

As shown in Fig. 4(b), 10 initial stimulus pairs were determined in a random sequence of all 20 stimuli. After completion of the 2AFC trials for all 10 pairs, the next stimulus pairs were determined in the *WINNER* stimuli to continue 2AFC trials. This procedure was repeated until only one *WINNER* (i.e., the first-place stimuli) and one *LOSER* (i.e., the second-place stimuli) were determined.

A total of 21 trials were included in a tournament. For each participant, the tournament was performed five times. The total number of tournaments was 35 (7 participants $\times$ 5 tournaments), and for each condition, it was counted how often it reached the first or second place in the tournaments.

5 RESULTS

Table 1 shows how often a condition was chosen as the first and second place across all participants in the tournament-style listening test. Overall, the conditions under which $L_{BA} = 0$ tended to be chosen most often as the first and second places. Under the conditions $(L_{BA}, L_{B2B1}) = (0, -3)$ dB, the number of times chosen as the first places was seven, which is the best in all conditions. The conditions $(L_{BA}, L_{B2B1}) = (0, 0)$ dB was chosen for the first place five times (the second best), and the number of times it was chosen as the second place was six (the best). In the extra conditions, only the original recorded stimuli were chosen as the first and second places three times in total.

6 DISCUSSION

The tendency of the greater number of times chosen as the first and second places when $L_{BA} = 0$ is assumed to support the fact reported by Békésy (4) that AC and BC speech transmissions have almost the same order of magnitude. Under $L_{BA} = 0$, the number of times a condition was chosen as the first and second places was greater when $L_{B2B1} = -3$ and 0 dB, which means that P_{B2} almost equals, or is at least half as great as P_{B1}. This finding suggests that the contribution of BC speech transmission reaching the middle/inner ear to own-voice perception cannot be ignored, as well as that reaching the outer ear. Pörschmann suggests in his psychoacoustical study that the contribution of the outer ear component to BC speech transmission is very small (6). The results in this study may partly support Pörschmann's suggestion using a different method.

Here, the averaged transfer functions $|H_{oc}(f)|$ and $|H_{ob}(f)|$ across five participants were used for making

spectrally-mimicked BC speech. Although the authors' previous study found that global spectral shapes of the measured transfer functions were similar among all participants (9, 10), individual differences of the transfer functions should be further considered for more precise evaluation.

In this paper, a tournament-style listening test was carried out. Such methods have been adopted mainly for investigating perceptual best-fitting conditions (e.g., individualization for virtual audio display (13)). The tournament-style method is assumed to be more effective if the transfer functions for mimicking BC speech were well-customized individually.

7 CONCLUSION

This paper investigated the contribution of BC speech reaching the middle/inner ear to own-voice perception, as well as that reaching the outer ear, by making two types (the outer- and middle/inner-ear components) of spectrally-mimicked BC speech signals based on the measured transfer functions. The subjective evaluation of own-voice quality for the mixed signal of recorded AC speech and two types of spectrally-mimicked BC speech found that two mimicked BC speech signals were mixed at almost the same level, and a combination of two mimicked BC speech signals were mixed at almost the same level as the AC speech. This suggests that the middle/inner-ear part of transmission contributes to BC speech perception at almost the same extent as the outer-ear part of transmission.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant No. 20KK0233, 21H03463, and 21K21314.

REFERENCES

(1) Denes PB, Pinson EN. The speech chain. 2nd ed. NY: Freeman; 1993.

(2) Toya T, Ishikawa D, Miyauchi R, Nishimoto K, Unoki M. Study on effects of speech production during delayed auditory feedback for air-conducted and bone-conducted speech. J Signal Processing, 2016;20(4):197–200. https://doi.org/10.2299/jsp.20.197.

(3) Stenfelt S. Acoustic and physiological aspects of bone conduction hearing. Advances in Oto-Rhino-Laryngology, 2011;71:10–21. https://doi.org/10.1159/000323574.

(4) Békésy GV. The structure of middle ear and the hearing of one's own voice by bone conduction. J Acoust Soc Am. 1949;21:217–232. https://doi.org/10.1121/1.1906501.

(5) Reinfeldt S, Ostli P, Håkansson B, Stenfelt S. Hearing one's own voice during phoneme vocalization — Transmission by air and bone conduction. J Acoust Soc Am. 2010;128(2):751–762. https://doi.org/10.1121/1.3458855.

(6) Pörschmann C. Influences of bone conduction and air conduction on the sound of one's own voice. Acta Acustica unitid with Acustica. 2000;86(6):1038–1045.

(7) Stenfelt S, Hato N, Goode RL. Factors contributing to bone conduction: The middle ear. J Acoust Soc Am. 2002;111(2):947–959. https://doi.org/10.1121/1.1432977

(8) Stenfelt, S. Inner ear contribution to bone conduction hearing in the human. Hearing Research, 2015;329:41–51. https://doi.org/10.1016/j.heares.2014.12.003.

(9) Toya T, Birkholz P, Unoki M. Estimates of transmission characteristics related to perception of bone-conducted speech using real utterances and transcutaneous vibration on larynx. In: Salar AA, Karpov A, Potapova R, editors. Speech and computer. Switzweland, Springer; 2019;11658:491–500. https://doi.org/10.1007/978-3-030-26061-3_50.

(10) Toya T, Birkholz P, Unoki M. Measurements of transmission characteristics related to bone-conducted speech using excitation signals in the oral cavity. J Speech Lang & Hear Res. 2020;63:4252–64. https://doi.org/10.1044/2020_JSLHR-20-00097.

(11) Shaw EAG. The external ear. In: Kaidel WD, Neff WD, editors. Handbook of Sensory Physiology. Berlin, Springer; 1974;5:455–490. https://doi.org/10.1007/978-3-642-65829-7_14.

(12) Aibara R, Welsh JT, Puria S, Goode RL. Human middle-ear sound transfer function and cochlea input impedance. Hear Res. 2001;152:100–109. https://doi.org/10.1016/S0378-5955(00)00240-9.

(13) Iwaya Y. Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears. Acoust Sci & Tech. 2006;27(6):340–343. https://doi.org/10.1250/ast.27.340.

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022, in Gyeongju, Korea

Extended abstract

ABS-0792

Methods to improve production and perception skills of foreign sounds

Takayuki ARAI[1]

[1] Sophia University, Japan

ABSTRACT

I am currently supervising a TV program to help children in Japan acquire English sounds. Through discussing the contents of the show, we often come across that some foreign sounds are difficult for non-native speakers. As an example, native Japanese speakers tend to have difficulty producing and perceiving the English /r/ and /l/ sounds. In this study, we discuss several methods to overcome such barriers for foreign sounds. The first is a simple but effective emphasis of a set of important components of a target sound. For example, lowering the third formant of the English /r/ is essential, and to train the ears to detect this component, listening to sounds with the emphasized component is effective. The second method is to give a non-native speaker some feedback on such important component. One example of this is an auditory or a visual feedback while producing a speech sound. The third method is to use physical models of the human vocal tract. For example, we have developed several vocal-tract models, such as one for the retroflex /r/ sound and the other for the bunched /r/ sound. These models can be utilized to help language learners acquire the pronunciations of English /r/ sound.

Keywords: Speech Production, Speech Perception, Foreign Sounds

1. INTRODUCTION

Six-month-old infants are known to make a distinction between two sounds, A and B, but when they get to 10–12 months old, they are unable to make such distinction if they are raised in a language environment in which A and B sounds belong to the same phoneme (1). Therefore, native Japanese speakers find it difficult to discriminate English /r/ and /l/ sounds, leading to many studies on this "challenge," such as (2). The difficulty is not only speech perception but also speech production, where Japanese native speakers usually cannot articulate the English /r/ and /l/ sounds well. Thus, we will discuss several approaches to overcoming this difficulty.

2. KIDS' TV SHOW ON ENGLISH SOUNDS

In 2017, the Japan Broadcasting Corporation, or NHK, started an English TV show for children, "Eigo-de Asobo with Orton." I am the sound supervisor for this TV show, including pronunciation of English vowels and consonants. For example, in an episode related to the English sound /f/, we introduced a so-called "super machine" made of a string and a rubber plate. In the episode, when a child actor placed the rubber plate between the upper teeth and the lower lip, pulled the string, and released it, the machine fished up a dummy fish. By doing this, viewers can learn and even feel that the /f/ in "fish" is articulated by the upper teeth and the lower lip.

The two most difficult sounds in English for Japanese children are /r/ and /l/. In the same TV show, we invented ways to practice these two sounds. For /r/, we wanted the children to train the retroflex shape of the tongue. Therefore, we introduced the "tongue trainer" shown in Fig. 1 (a), which is three beads threaded together on a string. The children placed the beads on their tongue and scooped them up to shape their tongue like the retroflex configuration. For /l/, we wanted the children to make sure there are lateral pathways of the air on the left and right sides of their tongue. Therefore, we introduced the "lion mask" shown in Fig. 1 (b), where a straw is placed between the tongue and the palate when the child makes a tongue configuration of this sound.

[1] arai@sophia.ac.jp

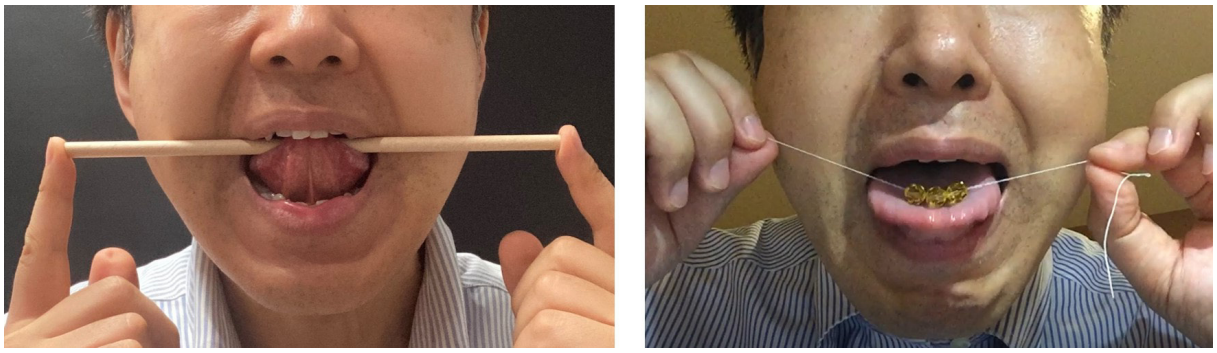


Figure 1 – Tools and their use for producing English sounds (left: /l/, right: /r/).

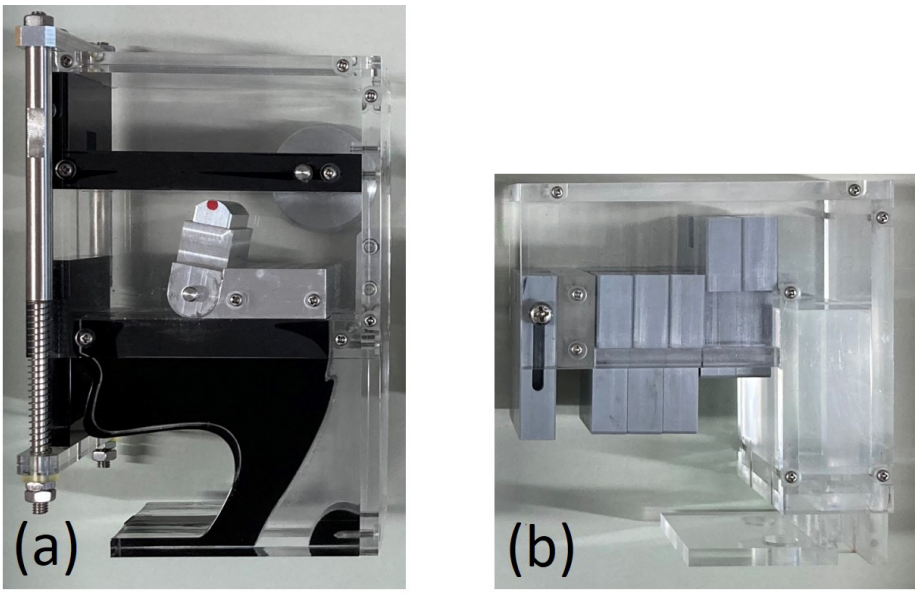


Figure 2 – Vocal-tract models. (a) BMW-RL model, (b) bunched /r/ model.

3. EMPHASIZING CUES FOR PERCEPTION

The most important cue to distinguish /r/ and /l/ sounds is the movement of the third formant (3). One way to train the perception of such a cue is by emphasizing it and letting listeners focus on the cue. For English /r/, the third formant frequency goes as low as 2 kHz. To train the ears to detect this frequency range, the frequency components between 1500 and 3000 Hz are emphasized for listening.

4. VOCAL-TRACT MODELS FOR PRODUCTION

We have developed a set of vocal-tract models for several purposes, including pronunciation training. For instance, our “BMW-RL” model (4) can mechanically produce /b/, /m/, /w/, /r/, and /l/ English consonants (Fig. 2 (a)). For the /l/ sound, the first half of the tongue can be rotated. A lateral sound is produced when the tongue tip touches the alveolar ridge because there are still spaces on the left and right sides of the tongue. For /r/, the tongue rotation does not touch anywhere on the palate and forms its retroflexion when the tongue is shortened in the same model, mimicking the retroflex /r/. Model (5) can change each part of the tongue height (Fig. 2 (b)), enabling us to bunch the tongue to form the bunched /r/. With these models, learners can listen to sounds, touch the models, and watch the movements of speech organs to help them acquire such sounds.

ACKNOWLEDGEMENTS

This work was partially supported by JSPS KAKENHI Grant Number 21K02889.

REFERENCES

1. Werker, J. F. and Tees, R. C. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behav Dev*. 1984; 7: 49–63.
2. Akahane-Yamada, R. The new tide of the development of ICT-enhanced learning content for foreign language learning. *J Acoust Soc Jpn*. 2014; 70 (8): 446–451.
3. Tomaru, K. and Arai, T. Perception of multiple series of English /ra/-/la/ continuum having different end frequencies of formant transitions. *Acoust Sci Technol*. 2014; 35 (3): 166–169.
4. Arai, T. Integrated mechanical model of [r]-[l] and [b]-[m]-[w] producing consonant cluster [br]. *Proc. of INTERSPEECH*. 2017; 979–983.
5. Arai, T. Retroflex and bunched English /r/ with physical models of the human vocal tract. *Proc*. *of INTERSPEECH*. 2014; 706–710.

PROCEEDINGS of the
24th International Congress on Acoustics
October 24 to 28, 2022, in Gyeongju, Korea

Extended abstract

ABS-0814

Evaluating speech intelligibility degradation under different orders of Ambisonics

Zhenyu GUO[1]; Huali ZHOU[2]; Yuezhe ZHAO[3]

[1] School of Architecture, South China University of Technology, Guangzhou 510640, China

[2] College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China

[3] School of Architecture, South China University of Technology, Guangzhou 510640, China

ABSTRACT

The state-of-the-art virtual audio display technologies reproducing complex sound scenes show considerable auditory research prospects. The authentic sound field recreated by multi-channel loudspeakers makes it possible to evaluate speech perception for both normal-hearing and hearing-impaired listeners under some critical scenarios. Among the loudspeaker-based sound reproduction system, the Ambisonics system is considered to have a good trade-off between authenticity and hardware cost. However, speech intelligibility could be damaged by the crosstalk of multiple loudspeakers, especially when the lower-order Ambisonics system with a large energy spread is used. To evaluate the degradation of speech intelligibility under various orders of Ambisonics, the speech reception thresholds (SRTs) of twelve participants under different orders of Ambisonics and the one-channel playback (reference) conditions were measured. Results show that SRTs under the 1st and 3rd order Ambisonics were significantly higher (worse intelligibility) than those under the reference condition. The SRTs of higher-order (7th and 9th) Ambisonics are slightly higher than those under the reference condition, but it is not statistically significant. The variation range of SRTs under first-order Ambisonics is more extensive than other conditions, which may be related to the limited sweet area of first-order Ambisonics.

Keywords: Speech intelligibility, Ambisonics, Sound reproduction

1. INTRODUCTION

Conventional auditory research predominantly adopts simple audio (e.g., free-field recording) stimuli in experiments. Albeit using simplified stimuli may be capable of revealing the intrinsic auditory mechanism, realistic sound scenes may fetch conclusions with higher ecological validity in some venues. In other cases, realistic sound scenes may be desired due to the research purpose, e.g., speech perception in a specific reverberation environment[1]. Hence, the state-of-the-art virtual audio display technologies reproducing the elaborate sound scenes confirm considerable prospects for auditory research.

Binaural virtual audio display technology with earphones meets its massive application in auditory research. It is still regarded as the most accurate method up to now, even when several apparent drawbacks, including front-back confusion and in-head localization, exist[2]. However, the inherent limitation of binaural reproduction due to the earphones restricts the usage for hearing-assisted individuals.

The authentic sound field recreated by multi-channel loudspeakers makes it possible to evaluate speech perception for normal-hearing and hearing-impaired listeners under critical scenarios[3]. Among the loudspeaker-based sound reproduction system, the Ambisonics system is considered to have a good trade-off between authenticity and hardware cost. However, the Ambisonics system cannot reproduce the ideal sound field without distortions due to the limitation of the actual system. Hence, understanding how these distortions impact specific aspects of auditory perception may be one of the essential premises for applications of the Ambisonics system.

[1] zhenyuguo404@qq.com
[2] tuobamao@qq.com
[3] arzhyzh@scut.edu.cn

The present study investigated the speech intelligibility damaged by the multiple-loudspeaker reproduction system, especially when the lower-order Ambisonics system with a large energy spread is used. The hypothesis of this study is that both the distortions from crosstalk of multiple loudspeakers and spatial masking release from the energy spread of Ambisonics may influence speech intelligibility. To evaluate the degradation of speech intelligibility, the speech reception thresholds (SRTs) under different orders of Ambisonics and the one-channel playback (reference) conditions were measured.

2. Methods

During the SRT measurement, the masking noise was speech shape noise, and the sentences from Mandarin Hearing In Noise Test (MHINT) corpus were used as the speech stimuli. The noise was played with a single front loudspeaker directly while the speech was played in different methods. Six types of speech stimuli were incorporated in the experiment, i.e., played with a single front loudspeaker (reference) and synthesized with 0-, 1-, 3-, 7-, and 9-order of Ambisonics. Note that the 0-order Ambisonics implies the signal was distributed to all loudspeakers equally; hence the speech had no definite direction. In other conditions, the locations of noise and speech were always spatially coincident. The sound level differences among 6 conditions were measured with a monaural microphone placed in the center of the loudspeaker array and compensated correspondingly.

The Ambisonics system used in the present study comprises 192 spherically distributed loudspeakers. The virtual stimuli were reproduction with the model-match methods.

Twelve participants (aged from 21 to 29) were recruited with payments. Each participant underwent 12 SRT measurements of 6 different conditions (1 reference and 5 different orders of Ambisonics) with 2 repetitions. The sequence of tests was balanced with a 12×12 Latin square. Before the formal experiment, a training phase including 2 measurements of SRT was assigned to participants.

3. Results and discussions

Totally 144 thresholds of speech recognition were gathered in the experiment. The SRTs of all conditions were illustrated in Fig. 1. It clearly shows that the SRTs decreased and approached the results of the reference conditions when the order was promoted except for the 0-order condition, which means higher speech intelligibility. SRTs of the 0-order condition were inferior to other conditions using Ambisonics.

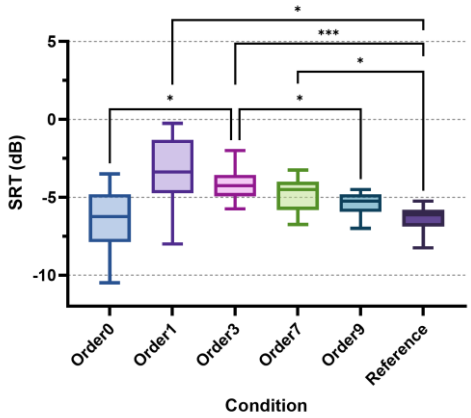


Fig 1. Tukey box and whisker plot of SRTs. The upper, middle, and lower lines of the box indicate the 75th, middle, and 25th of the SRTs, respectively. The * and ** caps denote significant differences at a level of 0.05 and 0.001, respectively.

The SRTs were statistically analyzed with the one-way repeated measures ANOVA, given the SRTs conform to the normality distribution. It reveals that conditions make a significant influence on the SRTs [$F(2.651, 29.16) = 7.652$, $p = 0.0009$]. According to the Tukey multiple comparisons, SRTs of the 1-, 3-, and 7-order conditions were significantly superior to those of the reference condition ($p = 0.0112$, 0.0001, and 0.0394, respectively). No significant difference between 9-order and reference conditions was observed. SRTs of 3-order were significantly higher than SRTs of 9-order condition ($p = 0.0409$). Besides, the variations of SRTs were also much larger for 0- and 1-order conditions. The

degraded speech intelligibility under lower order of Ambisonics probably comes from the crosstalk of multiple loudspeakers, which was verified for stereo[4] and Ambisonics system[5]. However, these results were disparate from those of Ahrens *et al*.[6], which show the SRTs barely not change with the order of Ambisonics when noise and speech are spatially coincident. Considering the differences between the two experimental configurations, further study may be needed.

While it seems SRTs of 0-order condition were below other conditions, a significant difference was only observed compared to 3-order condition (p = 0.0201). Given the noise was concentrated in the front loudspeaker in the present study, the spatial masking release may contribute to the speech intelligibility with the enlarged energy spread of speech in the lower order of Ambisonics. It may account for the phenomenon that the 0-order condition was completely contrary to the trend of other conditions.

4. Conclusion

The present study examined SRTs under different orders of Ambisonics. The results reveal that speech intelligibility was deteriorated by Ambisonics under the 9 order compared to playing with a single loudspeaker. Besides, the energy spread under the lower order of Ambisonics may induce perceivable spatial masking release. The study emphasizes the importance of considering the limitation of the Ambisonics system when using it to conduct auditory research.

ACKNOWLEDGEMENTS

Authors express gratitude to all participants involved in the experiment.

REFERENCES

1 Badajoz-Davila J, Buchholz JM, Van-Hoesel R. Effect of noise and reverberation on speech intelligibility for cochlear implant recipients in realistic sound environments. J Acoust Soc Am. 2020;147:3538.
2 Xie B. Head-related transfer function and virtual auditory display. USA: J. Ross Publishing; 2013.
3 Oreinos C, Buchholz JM. Evaluation of Loudspeaker-Based Virtual Sound Environments for Testing Directional Hearing Aids. J Am Acad Audiol. 2016;27:541–56.
4 Shirley B, Kendrick P, Churchill C. The effect of stereo crosstalk on intelligibility: comparison of a phantom stereo image and a central loudspeaker source. Journal of the Audio Engineering Society. 2007;55:852–63.
5 Solvang A. Spectral Impairment for Two-Dimensional Higher Order Ambisonics. Journal of the Audio Engineering Society. 2008;56:267–79.
6 Ahrens A, Marschall M, Dau T. Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments. Hear Res. 2019;377:307–17.

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0824

Applying Lombard speech improves speech intelligibility of outdoor public address systems under aircraft noise

Nao HODOSHIMA

Tokai University, Japan

ABSTRACT

An outdoor public address (PA) system is a popular method of rapid information propagation. This is especially important in events such as an emergency or disaster evacuation. However, spoken announcements broadcast from the outdoor PA system are sometimes unintelligible due to multiple echoes and aircraft noises. This study investigated whether adapting the Lombard speech (i.e., humans modify their speech to make it more intelligible against noise) to spoken announcements of the outdoor PA system improved speech intelligibility under aircraft noise. A young adult recorded words in a carrier sentence under a quiet condition (Q) and under an aircraft noise and multiple echo condition (NE), where the aircraft noise and speaker outputs convolved by the multiple echoes were fed back to the speaker through headphones. With multiple echoes and aircraft noise, listening tests were conducted with 24 young adults under a simulated outdoor PA system. The result of the listening tests showed that NE had significantly higher speech intelligibility than Q. The result suggests that introducing the Lombard speech as an input to spoken announcements of outdoor PA systems might yield higher speech intelligibility than conventional spoken announcements.

Keywords: Lombard speech, speech intelligibility, aircraft noise, multiple echoes, outdoor PA system

1. INTRODUCTION :

Outdoor public address (PA) systems, which are among the most popular methods for rapidly broadcasting information to numerous people simultaneously, are used in a wide variety of residential areas for purposes such as disseminating disaster evacuation instructions and air pollution warnings. However, spoken announcements broadcast from outdoor PA systems are often insufficiently intelligible. Indeed, in response to a survey conducted after the Great East Japan earthquake in 2011, 57.1% of 303 residents in a targeted disaster-struck area reported that they could not understand information broadcast from the outdoor PA system loudspeakers[1].

There are various reasons why the speech intelligibility of outdoor PA systems commonly becomes unintelligible. First, multiple or long-path echoes (consisting of delays longer than 50 ms) caused by sound reflections from neighboring buildings or mountains, as well as interference from the announcements broadcast from adjacent outdoor PA loudspeakers, can decrease listener comprehension. In fact, several studies have shown that speech with multiple simulated echoes was considerably less intelligible than speech without echoes[2].

Aircraft noise, which is a major problem in residential areas close to airports, also decreases the speech intelligibility of outdoor PA systems. However, it is typically difficult to restrict aircraft noise levels because air traffic control laws govern aviation routes and aviation safety regulations have priority over aircraft noise abatement considerations. As a result, according to monthly monitoring results produced by the Tokyo Metropolitan Government at seven locations (five schools, one library, and a central wholesale market) within six to 24 km of the Tokyo International Airport in June 2022, the number of high aircraft noise periods measured were from 34 to 73 times a day, maximum noise levels were from 63 to 74 dB, and day-evening-night equivalent noise levels (Lden) were from 40 to 53 dB at five of those seven locations[3].

This is significant because the World Health Organization (WHO) strongly recommends that Lden aviation noise exposure be kept below 45 dB[4], which means that three of the five abovementioned monitoring points failed to meet the WHO guideline. Furthermore, unlike in Europe, where there have been serious attempts to regulate aircraft noise levels near the main airports[5], no strict noise regulation attempts have been implemented at the Tokyo International Airport, at least during the

daytime. On the other hand, the average sound pressure levels (SPLs) measured at 13 fishing ports in the Tokyo area exceeded 60 dB within a 300 m radius of the area's outdoor PA system loudspeakers[6], thus indicating they had a lower signal-to-noise ratio (SNR) than the 0 dB level measured in many residential areas near the Tokyo International Airport.

In a previous study that simulated an outdoor PA system to evaluate the effects of aircraft noise and multiple echoes on speech intelligibility[7], the results obtained showed that when an aircraft noise was added at an SNR of −5 dB, the correct word recognition rate when multiple echoes were present was significantly lower than when the noises were absent. Those results also showed that the correct word recognition rate with aircraft noise present at an SPL of 70 dB was significantly higher than that at an SPL of 60 dB. However, that rate did not increase significantly between SPLs of 70 and 80 dB, regardless of whether or not multiple echoes were present. Based on these results, we determined that increasing the SPLs of spoken announcements could increase the speech intelligibility of outdoor PA systems but that they might also end up causing further noise-related problems.

During speech communications, we modify our speech to make it more intelligible against the surrounding noise. This is known as the Lombard effect[8]. Furthermore, when compared with speech spoken under a quiet condition, the speech spoken under noisy condition shows different acoustical characteristics (fundamental frequency, intensity, and duration) as well as increased speech intelligibility[8-10]. Similar results, both in acoustical characteristics and speech intelligibility, have been observed in reverberant environments[11,12], even though the masking patterns from noise and reverberation are spectrally and temporally different. For example, noise masks speech simultaneously, while overlap-masking results in reverberation[13].

With the above points in mind, this study aims to determine whether speech uttered in high aircraft noise and multiple echo surroundings is more intelligible for outdoor PA systems than spoken announcements in quiet when those announcements were heard in simulated residential areas near an airport. To accomplish this, 24 young adults were subjected to listening tests during which spoken sentences were presented with and without aircraft noise (SNR of 0 dB) and multiple artificial echoes.

2. LISTENING TEST

2.1 Participants

As stated above, 24 native speakers of Japanese (ages 21 to 24 years) with normal hearing participated in this study. All of the research methods used in this study were approved by the Ethics Committee of Tokai University, and written informed consent was obtained from all participants before the listening tests began.

2.2 Stimuli

The speech materials were target words embedded in the carrier sentence, "This is a [target word] announcement from the local government office." (English translation from Japanese). In total, 52 target words consisting of four morae (Japanese phonological syllable-like units) were selected from word lists, and each target word consisted of four consonant-vowel sequences[14]. To prevent the participants from using context and semantic cues, all of the words were selected with familiarity levels between 2.5 and 4.0 on a seven-point scale ranging from (1) unfamiliar to (7) very familiar.

Table 1 shows two speaking conditions used in the recording: quiet (Q) and under aircraft noise and multiple echoes (NE). The aircraft noise was recorded using a condenser microphone (Shure KSM141) connected to a portable SD recorder (Marantz PMD661) at the Jonanjima Seaside Park in Tokyo's Ota Ward, which is under an airport flight path approximately 3 km away from the Tokyo International Airport. After analyzing all the recorded aircraft noise, a period when the background noise level (e.g., ship noises, people talking, bird calls, etc.) was low was selected for use. Next, multiple echoes were created at delay times of 84 with an amplitude of 1, 189 with an amplitude of 0.5, and 400 ms with an amplitude of 0.25 using MATLAB software. These delay times were selected to simulate the impulse responses of an outdoor PA system[7].

Figure 1 shows the recording setup. Speech materials were recorded on a computer through the abovementioned microphone and a digital audio interface (Tascam US-144mkII) at a sampling rate of 44.1 kHz in a soundproof room. The speaker was a 22-year-old male native Japanese speaker. Under NE conditions, utterances were added to the aircraft noise, convolved by the impulse response, and then fed to the speaker through dynamic, closed circumaural type headphones (Sennheiser HDA200). The playback level of the noise was set at 85 dB. The playback level of the impulse response was set

at −10 dB relative to the speaking level at the speaker's ears. The recording was controlled by Adobe Audition 11.0. The speaker was asked to imagine that his speech was being broadcast from an outdoor PA system under the same acoustic conditions he was experiencing through the headphones.

After the speech materials were recorded, one carrier sentence was chosen for each speaking condition, and target words were embedded within the carrier sentence (with 150 ms before and after pauses) were produced for each speaking condition. In other words, each of the 52 target words spoken under either Q or NE condition was paired with a carrier phrase spoken under either Q or NE condition, respectively. The intensity ratio of the carrier sentence relative to the target word was normalized.

Finally, the concatenated speech sounds produced by the aircraft noise were added at SNR of 0 dB and convolved with the same multiple echoes that were used in the recording. After the overall intensity of the stimuli was normalized across speaking conditions, 104 stimuli were created (two speaking conditions × 52 sentences).

Table 1 – Speaking conditions

Condition	Presented through headphones during recordings
Q	
NE	aircraft noise and multiple echoes

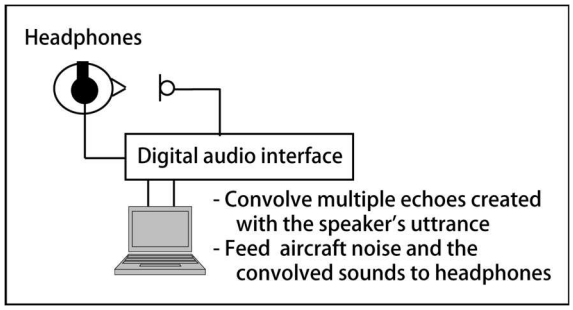


Figure 1 – Recording setup

2.3 Procedures

During the listening tests, which were also carried out in a soundproof room, stimuli were diotically presented to each participant via the abovementioned digital audio interfaces and headphones, which were connected to a computer. The playback level was set to 70 dB to simulate the SPL near an airport[3]. Two practice trials were held with each participant to familiarize them with the experimental procedure.

In each trial, a stimulus was presented once, after which the participants were instructed to write down the target word they heard on their answer sheets. The participants then were then asked to rate the impression of the stimulus according to their listening difficulty on a four-point scale where (1) indicated easy to comprehend and (4) corresponded to difficult to understand. For each participant, 100 stimuli (two speaking conditions × 50 sentences) were presented randomly. The target word and speaking condition combinations were counterbalanced across the participants.

3. RESULTS and DISCUSSION

Figure 2 shows the mean percentage of correct mora rates of the target words for each condition. A paired *t*-test was carried out for both speaking conditions (Q and NE) as the repeated variables and with the correct mora rate as the dependent variable. The results, which showed that NE had significantly higher correct rates than Q ($p < 0.01$), were consistent with previous studies that used white noise-induced[9] and reverberation-induced speech[12]. This result indicates that speech spoken under an actual condition with aircraft noise and multiple echoes induced was more intelligible than speech spoken under a quiet condition when the listeners were presented with the same noise and multiple echoes.

Figure 3 shows the four-point scale mean ratings of the stimuli. Here, we can see that Q was rated significantly more difficult to understand than NE ($p < 0.01$) based on Wilcoxon signed-rank test. As

for intelligibility, we can see that listeners found NE condition easier to listen to than Q condition.

Acoustical analysis showed that fundamental frequency was higher under NE (111 Hz) than under Q (90 Hz) conditions, which is significant because that higher fundamental frequency is one of the characteristics of speech spoken in noise or reverberation surroundings[8-11]. Therefore, the higher fundamental frequency in speech spoken in the presence of aircraft noise and multiple echoes may be one of the reasons for higher speech intelligibility and the impression that such speech is easier to understand than speech spoken in quiet locations.

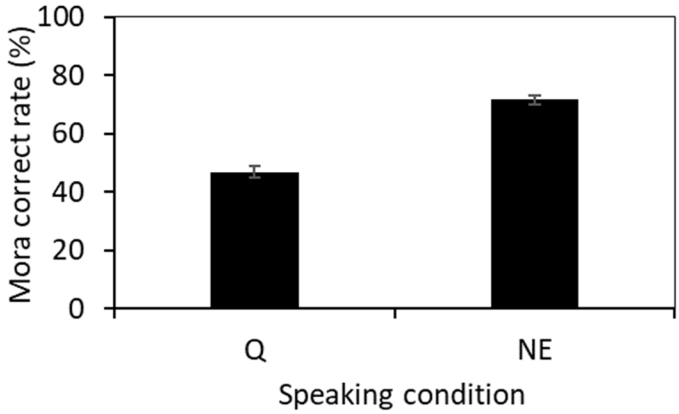


Figure 2 – Mean correct mora rate and standard error of target words for each speaking condition.

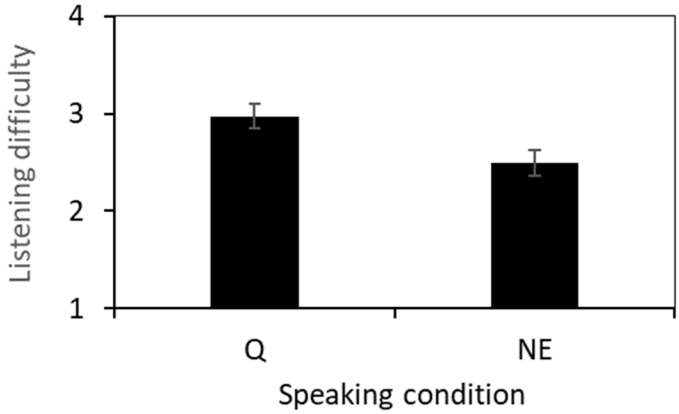


Figure 3 – Mean listening difficulty (1: easy to comprehend, 4: difficult to comprehend) and the standard error of target words for each speaking condition.

4. CONCLUSIONS

In this study, we carried out listening tests to investigate whether speech spoken under high aircraft noise and multiple echo conditions improved the intelligibility of speech uttered via a simulated residential area outdoor PA system near an airport when compared with the currently spoken PA announcements. The results of listening tests conducted on 24 participants subjected to speech recordings that included aircraft noise and multiple echoes showed that such speech was significantly more intelligible than the same speech spoken under a quiet condition. The results also showed that, on a four-point impression scale, speech spoken under quiet condition was significantly more difficult to comprehend than that spoken under aircraft noise and multiple echo conditions.

However, since this study used a single speaker, aircraft noise, and multiple echoes at a fixed SNR, additional research with more speakers and a wider range of noise and echo conditions will be needed in the future. It is also unclear how aircraft noise and multiple echoes separately degrade speech intelligibility of outdoor PA announcements. Further study could investigate the effect of multiple echoes and aircraft noise on speech intelligibility of outdoor PA systems, respectively.

Furthermore, the present study results revealed that, compared with current spoken announcements, speech spoken under aircraft noise and multiple echo conditions might increase the intelligibility of outdoor PA announcements near an airport where the SNR of PA announcements are often below 0 dB, and thus highly unintelligible. Accordingly, additional research will be needed to determine the

most appropriate conditions, such as acoustic settings and outdoor speaker surroundings, for further improving the intelligibility of outdoor PA announcements, especially in emergency situations.

ACKNOWLEDGEMENTS

This work was supported by a Grant-in-Aid for Scientific Research (B) from the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant No. 21H01596. We are very grateful to Hideki Tachibana, Kanako Ueno, and Sakae Yokoyama for providing the impulse response data, and to Miharu Kuniyoshi for conducting the listening tests.

REFERENCES

1. Ministry of Internal Affairs and Communications. Report on an ideal state of Telecommunications during disaster. 7 March 2012 (In Japanese).
2. Cui Z, Sakamoto S, Morimoto M, Suzuki Y, Sato H. Effect of word familiarity on word intelligibility of four continuous words under long-path echo conditions. Applied Acoustics; 2017; 124: 30-37.
3. Tokyo Metropolitan Government. Aircraft noise monitoring results according to new flight routes at Tokyo International Airport. June 2022 (In Japanese).
4. World Health Organization. Environmental Noise Guidelines for the European Region. 2018.
5. Policy Department for Citizens' Rights and Constitutional Affairs. Impact of aircraft noise pollution on residents of large cities. 2021.
6. Gotoh H, Takezawa M. Study on disaster announcement systems in coastal area. Research papers on the Japanese Institute of Fisheries Infrastructure and Communities; 2009; 21: 141-146 (In Japanese).
7. Takanashi K, Hodoshima N. The effects of the aircraft noise and multiple echoes on speech intelligibility of outdoor public address system. Proc. Internoise; 2014.
8. Lane H, Tranel B. The Lombard sign and the role of hearing in speech. J. Speech Hear. Res.; 1971: 14: 677-709.
9. Van Summers W, Pisoni D B, Bernacki R H, Pedlow R I, Stokes M A. Effects of noise on speech production: Acoustics and perceptual analysis. J. Acoust. Soc. Am.; 1988; 84: 917-928.
10. Junqua J C. The Lombard reflex and its role on human listeners and automatic speech recognizers. J. Acoust. Soc. Am.; 1993; 93: 510-524.
11. Hodoshima N, Arai T, Kurisu K. Speaker variabilities of speech in noise and reverberation. IEICE Technical Report; 2009; SP2009-69: 43-48, 2009 (In Japanese).
12. Hodoshima N, Arai T, Kurisu K. Intelligibility of speech spoken in noise and reverberation. Proc. International Congress on Acoustics; 2010 (paper ID: 663).
13. Nabelek A K, Letowski T R, Tucker F M. Reverberant overlap- and self-masking in consonant identification. J. Acoust. Soc. Am.; 1989; 86: 1259–1265.
14. Amano S, Kondo T, Sakamoto S, Suzuki Y. Familiarity-controlled word lists 2007 (FW07). The Speech Resources Consortium, National Institute of Informatics in Japan, 2007.

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0019

A practical method for generating whisper voices: Improvements in phantom silhouette method and application to multiple languages

Teruhisa UCHIDA[1]; Masanori MORISE[2]

[1] Research Division, National Center for University Entrance Examinations, Japan

[2] School of Interdisciplinary Mathematical Sciences, Meiji University, Japan

ABSTRACT

This research aims to make "expression via voice" richer and freer. Our immediate aim is to develop a system for converting standard utterances into attractive whispering. So far, we have devised a practical method for converting regular speech into whispered speech by using WORLD, a high-quality vocoder. With this "phantom silhouette method," the spectral envelope is first extracted from regular speech. Then, the envelope is manipulated, so the speech sounds like a whisper. Finally, a pseudo-whisper is created by driving the manipulated envelope using white noise instead of a vocal cord sound source signal. The spectral envelope is transformed using three operations to manipulate the timbre: (1) upward shifting of the spectrum in the F_1-F_2 formant frequency bands; (2) compensation of the breathy sound component in the high-frequency range; (3) suppression of the low-frequency band in the spectral envelope. Since the timbre is directly manipulated, the desired timbre can be obtained when synthesizing speech using WORLD. This report describes the latest improvements: implementation of multi-language support (i.e., English, Chinese, and Korean), conversion of the speaking rate, and expansion and contraction of the vocal tract length.

Keywords: Noise-vocoded speech, spectral envelopment, voice conversion, speech synthesis, WORLD

1. INTRODUCTION

1.1 Whisper voice generation from standard utterance speech

We previously presented the phantom silhouette method, a practical method for converting standard utterance speech into whispered speech to enrich "expression via voice." It is a simple parametric method using vocoder-type speech analysis and synthesis. The use of a high-quality vocoder, WORLD, makes it easy to understand which parts of the speech sound are manipulated intuitively.

The WORLD speech analysis/synthesis system is used in various applications, such as voice conversion and statistical parametric speech synthesis (1, 2). It is a high-quality vocoder-based system that accurately decomposes a speech waveform into the fundamental frequency (f_o), spectral envelope, and aperiodicity and synthesizes a new voice by integrating the transformed f_o, spectral envelope, and aperiodicity.

1.2 Multilingual support and new features

The method was originally tuned for Japanese speech (3–6). We have now tuned it for non-Japanese speech, aiming for multilingual support, and have added new functions for converting the speech rate and voice timbre. The results of whispered voice generation experiments demonstrated that the improved phantom silhouette method had enhanced performance.

[1] uchida@rd.dnc.ac.jp
[2] mmorise@meiji.ac.jp

2. PREVIOUS VERSION OF THE PHANTOM SILHOUETTE METHOD

We first give an overview of the previous version of the phantom silhouette method (3). In version two of the method (PS-2), the spectral envelope of the standard speech is first extracted using WORLD. The envelope is then transformed so that the voice sounds like a whisper. Finally, a whole devoiced pseudo-whisper is created by driving the manipulated spectral envelope using white noise instead of the vocal cord sound source signal (7). The spectral envelope is transformed by using three operations to manipulate the timbre:

(1) Upward shifting of the spectrum in the F_1-F_2 formant frequency bands,
(2) Compensation of the breathy sound component in the high-frequency range,
(3) Suppression of the low-frequency band in the spectral envelope.

This timbre manipulation is not signal processing performed after voice conversion but a rather direct manipulation of the spectrum. The desired timbre can thus be obtained when synthesizing speech.

The conversion process can be likened to the image of a Halloween ghost (Figure 1). The core of the method is noise-driven devoicing transformation and suppression of the low-frequency spectrum, which are referred to as the "phantomization of standard voice." The procedure for adding the characteristics of whispering to the spectrum of standard speech is called "spectral silhouette compensation." Specifically, it is the upward shifting of F_1-F_2 and compensating for the high-frequency component.

2.1 Upshifting of F_1-F_2 frequency bands

Matsuda et al. (8) reported that formants below 1200 Hz rise when whispering. Therefore, in the conversion, the standard speech's spectral frequency axis is partially expanded or contracted on the equivalent-rectangular-band-width (ERB_{RATE}) scale, which corresponds to the critical bandwidth of hearing.

A certain amount of shift is required for male voices, and the increase in F_1-F_2 is not as large for female whisper voices as it is for male voices. The quality of the converted voice is better for female voices when the shift amount is small. Therefore, the shift amount is controlled in accordance with the median value of f_0 so that the shift amount is large for low-f_0 male voices and small for high-f_0 female voices (Figure 2).

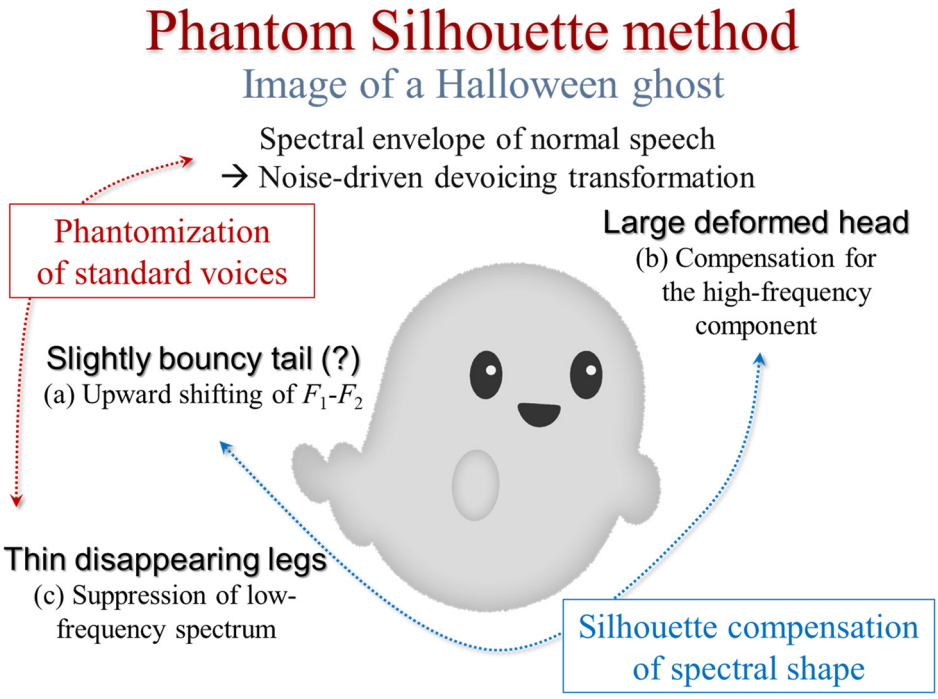


Figure 1 – Image of phantom silhouette method

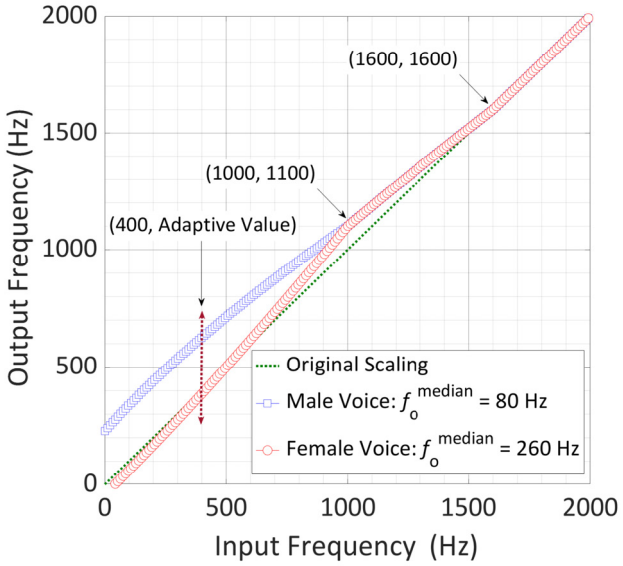


Figure 2 – Expanding and contracting spectral frequency scales

2.2 Compensation for high-frequency component

Compared with that of actual whispered speech, the standard speech spectrum may lack high-frequency components, as shown in Figure 3. The figure shows the time-averaged spectra of a female speaker uttering the same sentence in standard and whisper voices as recorded in a narration booth. The spectral gradient difference in the high-frequency range (1.6 kHz – 10 kHz) can be considered the amount of compensation required when generating a whisper.

The difference in spectral gradient corresponding to the f_o median for male and female voices is plotted in Figure 4. The figure shows that more high-frequency compensation was required for male whisper voices. Furthermore, the higher the f_o, the more compensation was required for both male and female voices. Therefore, after classification of male and female voices on the basis of the f_o median, the weighting of high-frequency emphasis is corrected in accordance with the f_o value (Figure 5).

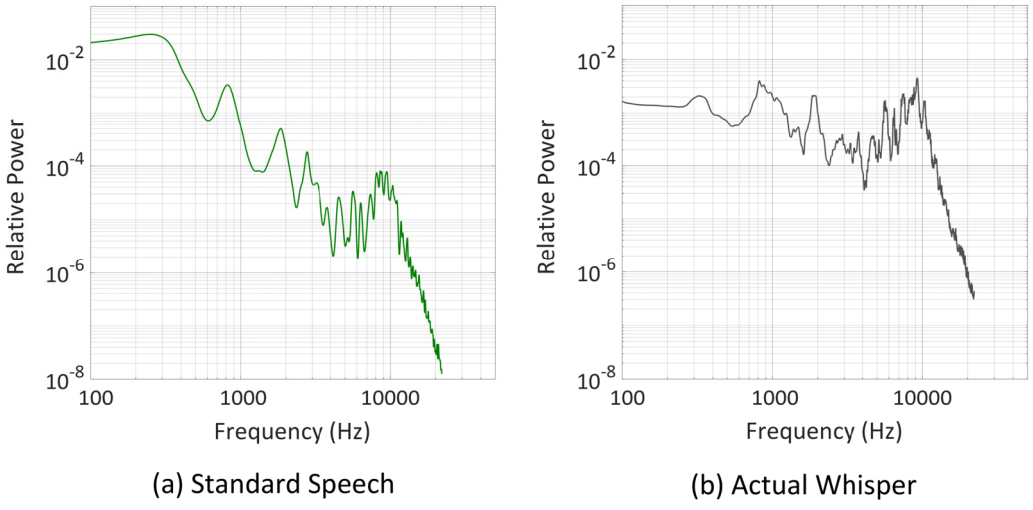


Figure 3 – Time-averaged spectra of standard speech and whispered speech by female speaker.

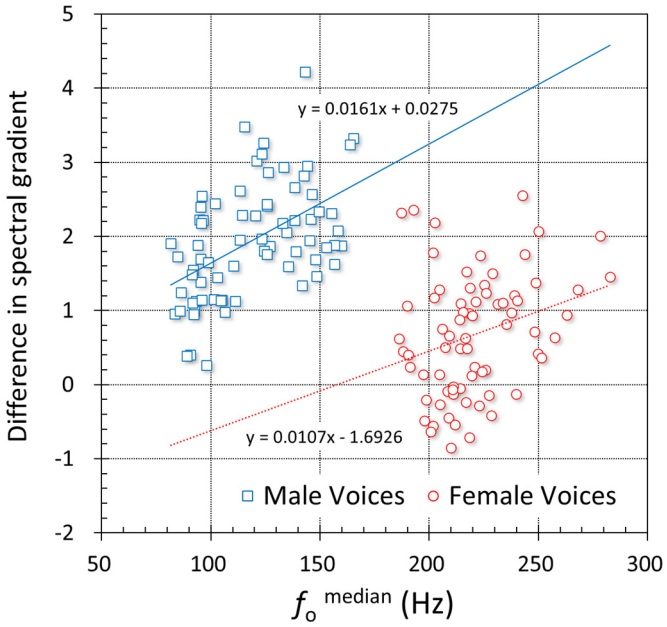


Figure 4 –Difference in spectral gradient (1.6 kHz – 10 kHz) between whispered and standard speech at f_o median for each standard speech sample

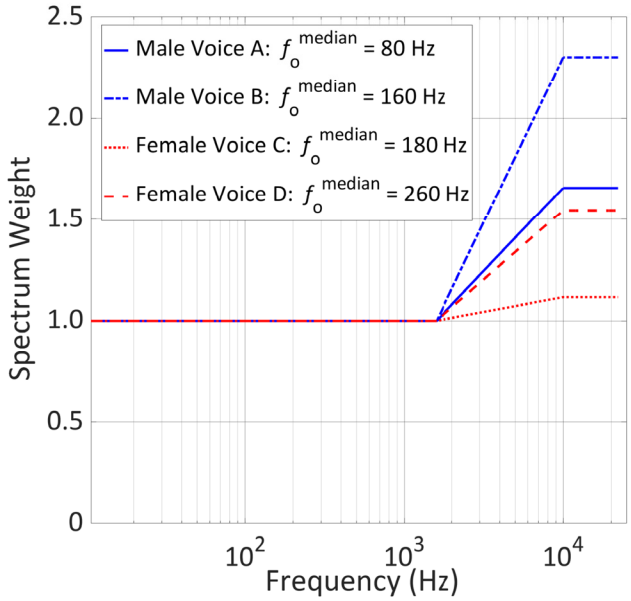


Figure 5 – Compensation functions used for high-frequency range based on f_o median

2.3 Low-frequency spectrum suppression

Matsuda et al. (8) observed that the sound pressure level in a whisper is in the range where the frequency is below 1 kHz. Kishida et al. (9) discovered that the amplitude information from 570 Hz to 1370 Hz plays an essential role in phonological comprehension in noise-vocoded speech. We thus set the transition range to 550 Hz – 1350 Hz, and the low-frequency spectrum was suppressed (Figure 6).

After the three spectral transformations described above, the transformed spectrum is used by WORLD as the basis for synthesizing a pseudo-whispering voice by driving it with white noise.

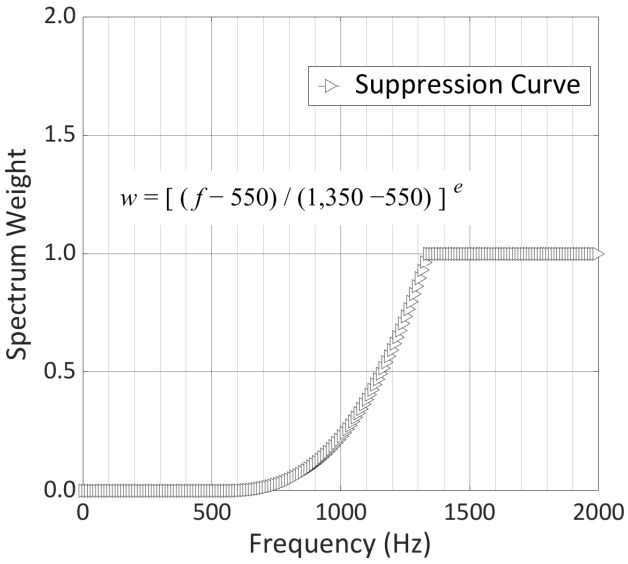


Figure 6 – Suppression curve for low-frequency spectrum

3. IMPROVED VERSION

The improved version of the Phantom Silhouette Method (PS-3) incorporated three significant enhancements. First, we have worked on multilingual support. The phantom silhouette method (PS-2), which previously supported only Japanese speech, has been expanded to multiple languages. Second, we added a function to adjust the speech rate when converting standard speeches to whispers. Third, we also added a voice timbre conversion function that mimics manipulating the speaker's physique. Then, we compared the parameters of the previous and improved methods and auditioned the generated whispers.

3.1 Multi-language support

We implemented multi-language support using speech data (24-kHz sampling, 16-bit quantization) of a female speaker in tri-jek, a corpus of Japanese, English, and Korean trilingual speech (10), as multilingual speech by the same speaker. For multiple speakers, we used male and female speech data (44.1-kHz sampling, 16-bit quantization) from the English listening comprehension test for the National Center Test for University Admissions in Japan. For Chinese speakers, we used male and female speech data (44.1-kHz sampling, 16-bit quantization) obtained with the publisher's permission from a learning material CD (11).

3.2 Problems and solutions for multilingual support

In PS-2, the shift amount of F_1-F_2 and high-frequency compensation ratio were adjusted following the median f_o of the original voice. The tuning was based on Japanese speech. However, with multilingualization, differences in the speaker's physique and speech style must be considered. In fact, when speech samples from the English listening test of the National Center Test were used, the shift amount of F_1-F_2 and high-frequency compensation ratio frequently were wrongly estimated due to errors in discriminating between male and female voices. Therefore, we tentatively set the initial values of the shift of F_1-F_2 and high-frequency compensation ratio at moderate levels and left the fine-tuning to the user.

Furthermore, the cutoff frequency for suppression of the low-frequency spectrum was made adjustable. These changes enable the user to manipulate the parameters while searching for the desired tone freely.

3.3 Additional features for proactive sound creation

In PS-3, the policy for specification settings was changed to leave sound creation to the user. Therefore, we added more optional functions to enable users to have the pleasure of creating voices that they find suitable.

One added function enables the user to adjust the speech rate. Since the phantom silhouette uses WORLD, it is relatively easy to adjust the time axis. In addition, the quality degradation caused by the speech rate adjustment is negligible for pseudo-whispers compared with standard speech. Therefore, we explicitly added it as a new feature.

Another added function enables the user to adjust the voice timbre to mimic the expansion and contraction of the speaker's vocal organs. This is done by stretching and contracting the spectral frequency axis. This function has also been called “vocal tract length transformation.” This new function is not intended to accurately reproduce the original speaker's voice. Instead, it is intended to create special effects, such as a whispering voice reminiscent of a small fairy or the whispering voice of a giant demon, which do not exist in reality.

By adding these new functions, we have expanded the degree of freedom in sound creation intending to make sound creation pleasurable.

3.4 Comparison of conversion parameters between PS-2 and PS-3

In PS-2, the parameters for conversion to whispered speech are automatically calculated from the f_o of the original standard speech but with the limitation that the estimation is based on the features of Japanese speakers' speech. In PS-3, the user searches for the optimal parameters for creating the desired whisper by repeatedly generating pseudo-whispering voices.

To evaluate whether there is any systematic difference between the parameters used in the two versions of the phantom silhouette method, we compared the English speech produced by English speakers using the two versions. The physiques of the speakers were assumed to differ greatly from those of Japanese speakers. Because the size of the speech organs corresponds to the speaker's physique, the physical characteristics of the vocalizations as well as the language-specific characteristics were assumed to differ systematically.

A total of 24 speech samples uttered by three male and three female speakers in the English listening portion of the National Center Test for University Admissions were used as the original voices. The mean and standard deviation of the parameters used in the generation of pseudo-whispering voices were compared between the two versions of the phantom silhouette method (Table 1). The parameters of the f_o search range during speech analysis in WORLD, the first conversion stage, are also listed.

There were systematic differences in the parameter values between the two versions. First, the F_1-F_2 formant shift was larger for PS-3. In English speech, it is assumed that the formants, which are considerably lower in frequency, must be raised during the conversion to remain even in whispered speech. Second, the amount of high-frequency compensation was lower with PS-3. It is thought that English speech does not require much compensation because of the power of high frequencies, even in standard speech. Furthermore, the pitch search range for f_o with PS-3 was lower. This lowering may be because English speakers are larger in stature than Japanese speakers, so f_o of their voices is necessarily lower.

These comparisons suggest that the improved version is more effective because the parameter search is more adaptive to the characteristics of the targeted language and speakers.

Table 1 – Comparison of conversion parameters between previous and improved versions
(24 voices; uttered by three male and three female English speakers)

	F_1-F_2 formant shift amount (Magnification at 400 Hz)	Compensation ratio of high-frequency component	Cutoff frequency for suppression of low-frequency spectrum (Offset in Hz)	WORLD: minimal pitch (Hz)	WORLD: maximal pitch (Hz)
Previous Version: PS-2	1.29 † (0.20) ††	1.61 (0.35)	0.0 (0.00)	71.0 (0.00)	800.0 (0.00)
Improved Version: PS-3	1.37 (0.22)	1.53 (0.38)	−4.39 (21.57)	58.3 (15.03)	433.3 (74.54)
Paired t-test: $t_{(23)}$	27.18 ***	−3.41 **	−0.86	−3.88 ***	−23.59 ***

†: mean, ††: *SD*

+: p < .10; *: p < .05; **: p < .01, ***: p < .001

3.5 Generated whispered speech and its evaluation

The PS-3 version of the phantom silhouette method was used to generate a pseudo-whisper from a standard speech in each language. Figure 7(a) shows the waveforms and spectrograms of the standard male speech samples taken from the English listening comprehension portion of the National Center Test for University Admissions. Figure 7(b) shows the generated whispered speech. The whispered speech was transformed by reducing the vocal tract length by a factor of 0.78, i.e., by stretching the spectral frequency axis by a factor of 1.28 (1/0.78), to produce a small speaker's timbre reminiscent of a child.

Figure 8(a) shows the waveforms and spectrograms of the standard female speech samples taken from the same test, and Figure 8(b) shows the generated whispered speech. The whispered speech was extended by a factor of 1.25 on the time scale to produce slower speech.

The generated pseudo-whispers were subjectively evaluated by a native speaker of each language working at a Japanese university. The speech was recognized as a whisper in each case, and the linguistic content was clearly understood.

4. DISCUSSION

4.1 Multilingual support and user-centered voice timbre adjustment

The phantom silhouette method is based on WORLD. Since WORLD is currently being used worldwide, it is clearly language-independent. In addition, whispering is a universal speech utterance method that does not involve vocal cord vibration and is derived from turbulence in the vocal tract.

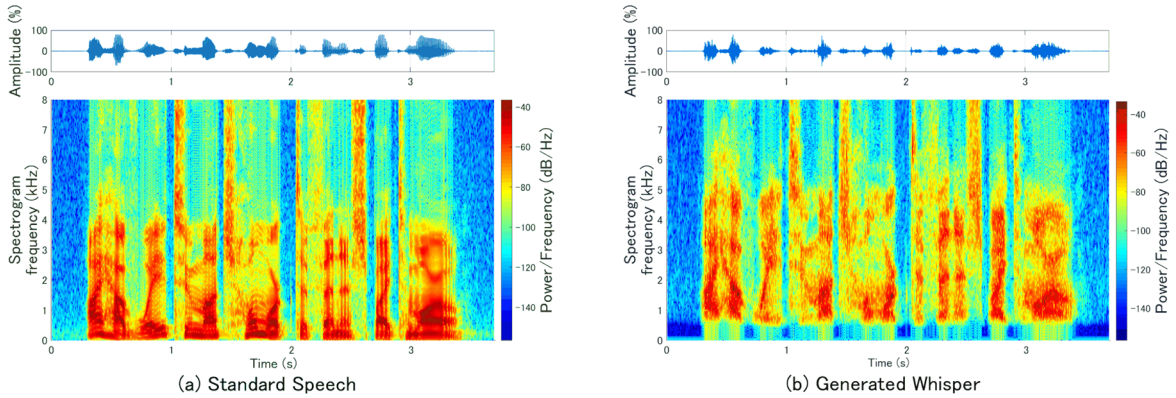


Figure 7 – Waveforms and spectrograms of standard speech and generated whispered speech for which spectral frequency axis was stretched by a factor of 1.28 (male voice: “I have way too much homework to finish by tomorrow.”)

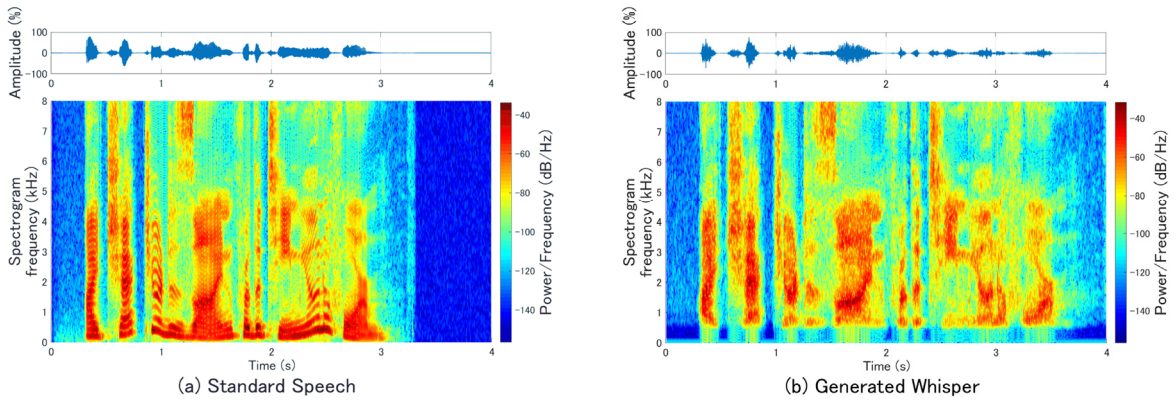


Figure 8 – Waveforms and spectrograms of standard speech and generated whispered speech for which time axis was stretched by a factor of 1.25 (female voice: “I checked your design for the team uniform.”)

Multilingual support, therefore, did not inherently require any special treatment. However, the previous version of our phantom silhouette method (PS-2) was tuned based on the speech of Japanese speakers, which hindered multilingual support.

Our improved version (PS-3) has naively supported multiple languages by leaving parameter adjustment to the user. This change enables the user to explore parameters more suitable for individual original voices than by using the parameters calculated from the median of f_o, as is done in the previous version. In short, we have paved the way toward producing higher-quality whispered speech.

The improved version supports manipulating the F_1-F_2 formant shift and the compensation ratio of the high-frequency component. It also supports the controllability of the cutoff frequency for suppression of the low-frequency spectrum. These changes enable the user to actively participate in the sound creation process by gradually changing the parameters and searching for the desired tone. The addition of functions for adjusting the speech rate and voice timbre has also expanded the freedom of vocal expression.

4.2 Symbiosis with neural vocoder using DNN and challenges

Continuing research on voice generation using deep neural networks (DNNs) and deep learning is expected to lead to the development of methods for generating higher quality "AI whispers" using a neural vocoder and to the development of other promising methods (12, 13).

One advantage of the phantom silhouette method is that it does not require prior training. Moreover, it easily runs in computer environments that do not include a GPU. Furthermore, it can be used to create unreal voices, such as fairy whispers so that they can be separated from AI whispers.

Future work includes developing a more efficient way to search for parameters and a method for numerically evaluating the generated whispered speech.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number JP21H04900.

REFERENCES

1. Morise M, Yokomori F, Ozawa K. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. IEICE Trans. Inf. Syst. 2016;E99-D(7):1877–1884.
2. Morise M. D4C, a band-aperiodicity estimator for high-quality speech synthesis. Speech Commun. 2016;84:57–65.
3. Uchida T, Morise M. A practical method of generating whisper voice: Development of phantom silhouette method and its improvement. Acoust. Sci. & Tech. 2021;42(4):214-217.
4. Uchida T. As we speak: Pure culture of the voice timbre. Proc. 85th Annu. Convention of the Jpn. Psych. Assoc.; 1–8 September 2021, Online, Japan 2021. p. 114. (in Japanese)
5. Uchida T, Morise M. Investigation of voice pitch illusion using quasi singing voice and quasi whisper. IPSJ Journal. 2020;61(4):807–816. (in Japanese)
6. Uchida T. Reversal of relationship between impression of voice pitch and height of fundamental frequency: Its appearance and disappearance. Acoust. Sci. & Tech. 2019:40(3):198–208
7. Shannon RV, Zeng F-G, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. Science 1995;270(5234):303–304.
8. Matsuda M, Mori H, Kasuya H. Formant structure of whispered vowels. J. Acoust. Soc. Jpn. 2000;56(7):477–487. (in Japanese)
9. Kishida T, Nakajima Y, Ueda K, Remijn GB. Effects of factor elimination on intelligibility of noise-vocoded Japanese speech. Proc. 31st Int'l Cong. of Psych; 24–29 July 2016; Yokohama, Japan 2016. PS28A-01-19.
10. Takamichi S. tri-jek: Japanese-English-Korean tri-lingual speech corpus. https://sites.google.com/site/shinnosuketakamichi/research-topics/tri-jek_corpus
11. Yamada R, He N, Yu. M, Nagano Y. Business Chinese through Stories (with CD). Tokyo, Japan: Surugadai-shuppansha; 2016. (in Japanese)
12. Wang X, Takaki S, Yamagishi J. Neural source-filter-based waveform model for statistical parametric speech synthesis. Proc. Int'l. Conf. Acoust. Speech Signal Process. (ICASSP) 2019; 12–17 May 2019; Brighton, UK 2019. p. 5916–5920.
13. Cotescu M, Drugman T, Huybrechts G, Lorenzo-Trueba J, Moinet A. Voice conversion for whispered speech synthesis. IEEE Signal Processing Letters 2019;27:186–190

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0196

A neural encoder-decoder framework for dysarthric speech reconstruction

Disong WANG[1]; Songxiang LIU[1]; Xixin WU[1]; Hui LU[1]; Lifa SUN[2]; Xunying LIU[1]; Helen MENG[1]

[1] The Chinese University of Hong Kong, Hong Kong SAR, China

[2] SpeechX Limited, Shenzhen, China

ABSTRACT

Dysarthric speech reconstruction (DSR) aims to improve the intelligibility and naturalness of dysarthric speech, while maintaining the speaker identity. In this paper, we propose to decompose the problem of DSR into three subproblems, i.e., content restoration, prosody correction and speaker identity preservation. To properly tackle each subproblems, we present a neural encoder-decoder framework for DSR. The framework includes: (1) a content encoder that extracts accurate content representations from the dysarthric speech; (2) a prosody encoder that infers normal prosodic features, i.e., pitch and duration, to replace their abnormal counterparts; (3) a speaker encoder that extracts an effective speaker representation; and (4) a decoder that aggregates the above representations to generate the reconstructed speech with improved quality. Subjective listening evaluation results verify the effectiveness of proposed framework to enhance both articulation and prosody of dysarthric speech. Besides, objective evaluation is conducted through automatic speech recognition, compared with original dysarthric speech, reconstructed speech achieves 23.2% and 31.8% absolute word error rate reduction for speakers with moderate and moderate-to-severe dysarthria respectively.

Keywords: Dysarthric speech reconstruction, Neural encoder-decoder

1. INTRODUCTION

Dysarthria arises from various neurological disorders including Parkinson's disease or amyotrophic lateral sclerosis, leading to weak regulation of articulators such as jaw, tongue, and lips (1). Therefore, the resulting dysarthric speech may be perceived as harsh or breathy with abnormal prosody and inaccurate pronunciation, which degrades the efficiency of vocal communication for dysarthric patients.

To improve the quality of dysarthric speech, various dysarthric speech reconstruction (DSR) approaches have been investigated. Generally, these approaches can be divided into rule-based DSR and statistical DSR. Rule-based DSR tends to apply manually designed, speaker-dependent rules to correct phoneme errors or modify temporal and frequency features to improve intelligibility (2, 3), while such rules are not flexible and restrict their applications in practice. Statistical DSR automatically maps the features of dysarthric speech to those of normal speech, where typical approaches contain Gaussian mixture model (4), non-negative matrix factorization (5, 6), partial least squares (7), and deep learning methods including sequence-to-sequence (seq2seq) models (8-11) and gated convolutional networks (12). Though significant progress has been achieved, most approaches rely on the parallel speech data from dysarthric speaker and healthy speaker to train the mapping models, while such data is generally unavailable in reality.

In this paper, we decompose the problem of DSR into three subproblems, i.e., content restoration, prosody correction and speaker identity preservation, which are tackled by our proposed neural encoder-decoder (NED) framework that contains four modules: (1) a content encoder extracting accurate phoneme embeddings from dysarthric speech to restore the linguistic content; (2) a prosody encoder inferring normal prosody features that are treated as canonical values for correction; (3) a speaker encoder producing a single vector as speaker representation used to preserve the speaker identity; and (4) a decoder mapping phoneme embeddings, prosody features and speaker representation to reconstructed mel-spectrograms. The proposed NED framework has high degrees of

[1] dswang@se.cuhk.edu.hk

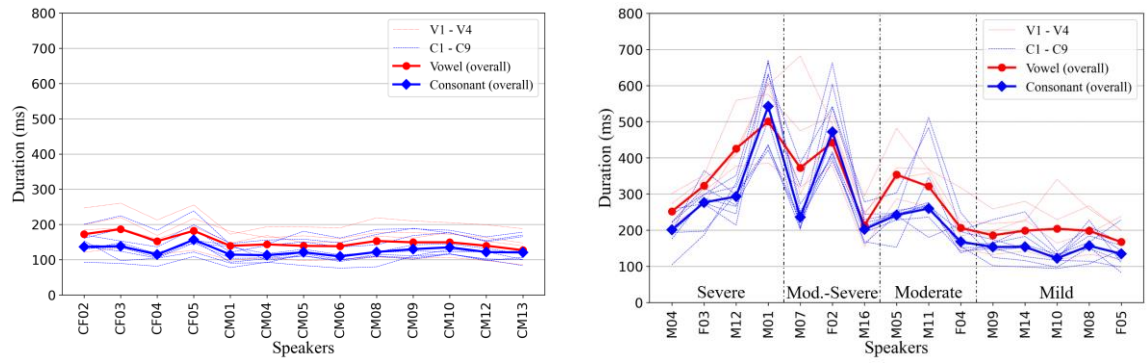
interpretability and flexibility, since each encoder produces meaningful intermediate representation and each factor of dysarthric speech can be flexibly modified. Besides, each module of NED framework is independently trained by using carefully-designed training strategies, which remove the requirements of parallel speech data.

The rest of the paper is organized as follows: Section 2 presents the acoustic characteristics of dysarthric speech. Section 3 describes the proposed NED framework for DSR. Section 4 gives the experimental results and analysis, and Section 5 concludes the paper.

2. ACOUSTIC CHARACTERISTICS OF DYSARTHRIC SPEECH

To demonstrate the acoustic characteristics of dysarthric speech, we select UASPEECH dataset (13) that contains 15 dysarthric speakers and 13 normal speakers with three blocks of speech data per speaker for prosodic analysis of different phonemes. We categorize all phonemes into four vowel groups and nine consonant groups following (14). Vowels are divided into short vowels (V1), medium vowels (V2), long vowels (V3) and diphthongs (V4). Consonants are divided into glides (C1), voiced stops (C2), nasals (C3), voiced fricatives (C4), voiced affricates (C5), aspirates (C6), unvoiced stops (C7), unvoiced fricatives (C8) and unvoiced affricates (C9).

To obtain the speech segments corresponding to different phonemes, the speech signals and phoneme sequence are aligned via forced-alignment by using speaker-dependent automatic speech recognition (ASR) models. Then we compute the values of phoneme duration and pitch for analysis.



(a) Normal speech (b) Dysarthric speech

Figure 1 – Average duration of different phoneme groups across all utterances from: (a) Healthy speakers; and (b) Dysarthric speakers

2.1 Duration Analysis

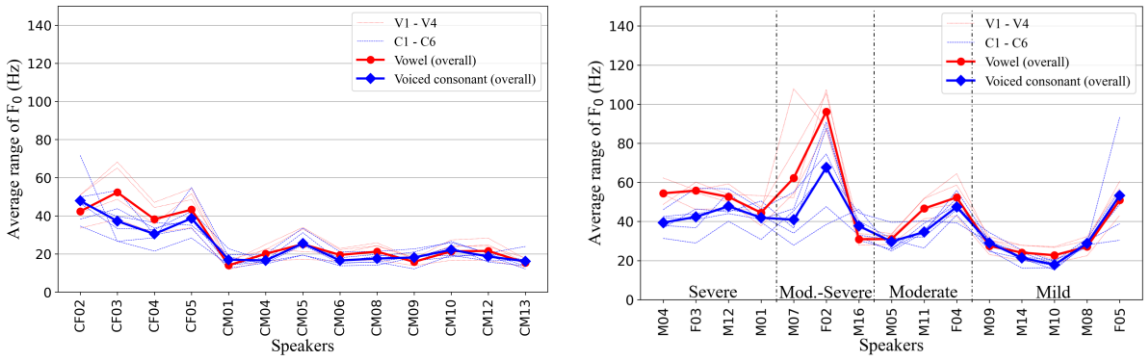
With the segment-phoneme alignment information, we can compute the average duration of different phoneme groups across all utterances of each speaker as shown in Figure 1, and we have the following observations:

Normal speech: From the results of normal speech in Figure 1 (a), we observe that the average durations of vowels and consonants mainly lie in the interval [100ms, 200ms], the duration of vowels exceeds that of consonants with around 26ms in average, and durations of different phoneme groups from different healthy speakers are relatively consistent in general, which shows that normal speakers tend to have similar and stable speaking rate.

Dysarthric speech: From the results of dysarthric speech in Figure 1 (b), we see that the average durations of different phonemes vary dramatically across different dysarthric speakers. For those with relatively severe dysarthria (Severe and Moderate-Severe), the durations of different patients are significantly different and their durations are longer than those of normal speech, especially for speakers M12, M01 and F02. For those with relatively mild dysarthria (Moderate and Mild), the durations tend to be shorter and approximate those of normal speech. Roughly speaking, it seems that the duration is proportional to the dysarthria severity, i.e., the higher the severity, the longer the duration. However, it does not hold for patients within the severe group that is diagnosed as spastic dysarthria, where the duration is in reverse ratio to the dysarthria severity, we reckon that it may be

labored for patients with severe dysarthria to pronounce standard phonemes of words in a clear manner.

Through above observations and analysis, we find that different dysarthric speakers have different speaking rates which depend on the severity of dysarthria and other speaker-dependent causes, e.g., etiologies. Generally, due to the weak control of muscles responsible for speaking, the phoneme duration of dysarthric speech is longer than that of normal speech, which lowers the speech naturalness and intelligibility. Therefore, the abnormal duration is a key factor that requires amendment during the voice reconstruction of dysarthria.



(a) Normal speech (b) Dysarthric speech

Figure 2 – Average range of F_0 of different phoneme groups across all utterances from: (a) Healthy speakers; and (b) Dysarthric speakers

2.2 Pitch Analysis

From the acoustic perspective, pitch can be described by fundamental frequency (F_0). The typical characteristics of dysarthric speech, such as hypernasality, mono-pitch, breathiness and harsh voice, are strongly associated with the abnormal F_0 contour. To demonstrate these issues, we illustrate the average range of F_0 that is used to measure the variations of F_0 within each phoneme group, it can be employed to indicate whether the F_0 variation is excessive or limited. For each phoneme, we first compute the F_0 sequence for the corresponding segment, calculate the range of F_0 as $Max(F_0)$-$Min(F_0)$ that is the difference between the maximum and minimum values of F_0 sequence, then the average range of F_0 is calculated by using all ranges of that phoneme. The statistical results are illustrated in Figure 2, where we only show the results for vowels and voiced consonants with the source excitation signals produced by the vibration of vocal folds with specific F_0 values. We have the following observations:

Normal speech: From the results of normal speech in Figure 2 (a), we observe that the average ranges of F_0 of female speakers (with "CF" as prefix) are larger than those of male speakers (with "CM" as prefix) in general, the female and male speakers also have the following differences: (1) For female speakers, the average ranges of F_0 of vowels and voiced consonants roughly lie in the intervals [38Hz, 52Hz] and [30Hz, 48Hz] respectively, and the average F_0 ranges of vowels tend to be larger than those of voiced consonants; (2) For male speakers, the average ranges of F_0 of vowels and voiced consonants lie in the intervals [14Hz, 25Hz] and [16Hz, 25Hz] respectively, and the average ranges of F_0 are nearly the same for both vowels and voiced consonants.

Dysarthric speech: From the results of dysarthric speech in Figure 2 (b), we observe that most female dysarthric speakers (with "F" as prefix) have the average ranges of F_0 matching those of healthy female speakers, except that the average ranges of F_0 of vowels and voiced consonants for the speaker F02 are 96Hz and 70Hz respectively, which greatly exceed the normal average ranges of F_0, indicating that F02 cannot well control the intonations with excessive F_0 variations. Besides, we can see that for male dysarthric speakers (with "M" as prefix) with mild and moderate dysarthria, their average ranges of F_0 match those of healthy male speakers. However, when the dysarthria becomes severe, male dysarthric speakers with severe and moderate-severe dysarthria have excessive F_0 variations as their average ranges of F_0 of vowels and voiced consonants significantly deviate from the normal average ranges of F_0 of male healthy speakers, which shows that those patients also have weak control of intonations with excessive F_0 variations.

Through above observations and analysis, we find that the speech of different dysarthric speakers

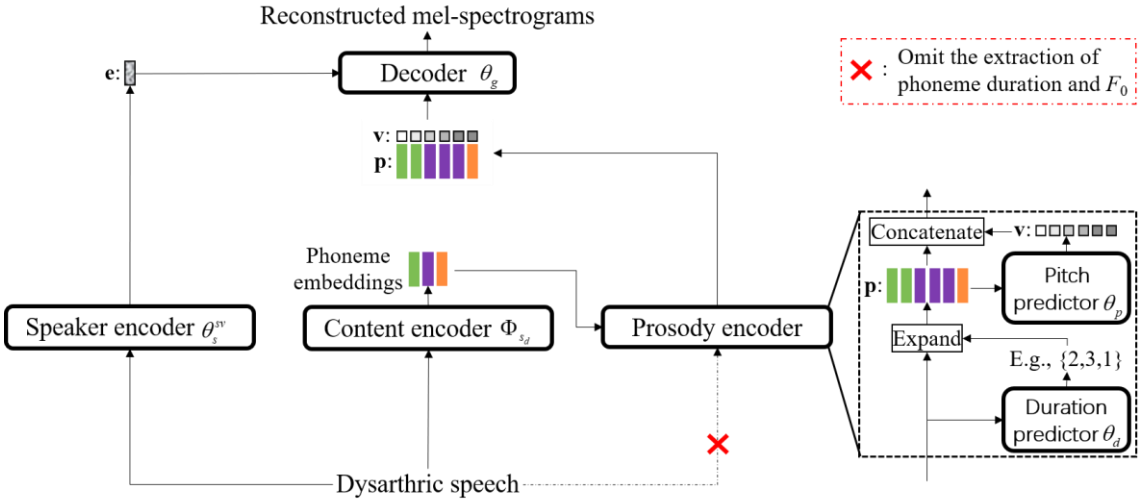


Figure 3 – The proposed neural encoder-decoder (NED) framework for dysarthric speech reconstruction

has different F_0 variations which generally depend on the dysarthria severity of patients. For dysarthric speakers with relatively severe dysarthria, the average ranges of F_0 of their speech tend to be excessive, which corresponds to dramatic or sudden changes of intonations from the perspective of perception. As a result, the speech naturalness and intelligibility are degraded significantly. Therefore, the abnormal F_0 is also a key factor that requires amendment during the voice reconstruction of dysarthria.

3. PROPOSED APPROACH

The proposed NED framework for DSR is shown in Figure 3, it contains four modules: content encoder, prosody encoder, speaker encoder and decoder. The first three modules respectively produce phoneme embeddings, prosody values and speaker representation; and the fourth module, the decoder, maps these features to reconstructed mel-spectrograms

Content encoder: We adopt a sequence-to-sequence (seq2seq) based content encoder to restore the content, it consists of 6-layer VGG extractor and 5-layer bidirectional long short-term memory (BLSTM) with 320 units per direction, location-aware attention, one-layer LSTM with 320 units and a 75-dim fully connected (FC) layer. The content encoder is only optimized to predict the phoneme sequence (73 phonemes + 1 start token + 1 end token) from the input speech by minimizing the sum of cross-entropy (CE) and connectionist temporal classification (CTC) losses between the predicted and ground-truth phoneme sequence. To improve the phoneme prediction accuracy on dysarthric speech, the content encoder is trained in two stages: (1) Pre-training on large-scale normal speech data to obtain an initialization model that has powerful generalization capacity; (2) Fine-tuning on the dysarthric speech of a certain patient s_d to boost the phoneme prediction performance. The outputs of pre-trained content encoder Φ_p or fine-tuned content encoder Φ_{s_d} are used as content representations (phoneme embeddings) that denote posterior phoneme probability distributions. In the following, Φ_p and Φ_{s_d} are used to extract content representations from the normal speech and dysarthric speech, respectively.

Prosody encoder: the prosody encoder contains two predictors to respectively infer normal phoneme duration and fundamental frequency (F_0). As we do not need the original prosodic features, we omit the extraction of abnormal phoneme duration and F_0 from the dysarthric speech. The prosody encoder is trained by a healthy speaker's speech with normal prosodic patterns: (1) Given the phoneme embeddings extracted by the speech encoder Φ_p as inputs, the phoneme duration predictor θ_d is trained to infer the normal phoneme durations that are obtained from force-alignment via Montreal Forced Aligner toolkit (15); (2) The ground-truth phoneme durations are used to align phoneme embeddings and F_0 as shown in Figure 3, the expanded phoneme embeddings are denoted as **p** and fed into the pitch predictor θ_p to infer normal F_0 that is denoted by **v**. The prosody encoder is expected to take in phoneme embeddings extracted from dysarthric speech to infer normal values of phoneme duration and F_0, which can be used as canonical values to replace their abnormal counterparts for generating the speech with normal

prosodic patterns.

Speaker encoder: The speaker encoder, θ_s^{sv}, is trained on a speaker verification (SV) task to capture speaker characteristics. θ_s^{sv} takes in mel-spectrograms $\mathbf{m}$ of one utterance with arbitrary length to produce a single vector as speaker representation: $\mathbf{e} = f_s(\mathbf{m}; \theta_s^{sv})$. Following the training scheme in (16), θ_s^{sv} is optimized to minimize a generalized end-to-end loss (17) by using normal speech data that is easily acquired from thousands of healthy speakers.

Decoder: The decoder aims to generate mel-spectrograms, it consists of two 512-dimensional FC layers, 4-layer BLSTM with 512 units per direction and one 80-dimensional FC layer to predict mel-spectrograms. The speaker representation extracted by the speaker encoder is repeated and concatenated with the expanded phoneme embeddings and F_0 sequence as the input of decoder. The decoder with parameters θ_g predicts mel-spectrograms as: $\mathbf{z} = f_g(\mathbf{p}, \mathbf{v}, \mathbf{e}; \theta_g)$. The decoder is also trained by using normal speech data from a set of healthy speakers $\mathcal{S}$. Each speaker $s_i \sim \mathcal{S}$ has the training data set $\mathcal{T}_{s_i} = \{(\mathbf{m}_j, \mathbf{p}_j, \mathbf{v}_j)\}$, where each sample corresponds to one utterance and contains mel-spectrograms $\mathbf{m}_j$, expanded phoneme embeddings $\mathbf{p}_j$ and pitch features $\mathbf{v}_j$. Then decoder is optimized by minimizing the generation loss, i.e., the L2-norm between the predicted mel-spectrograms $\mathbf{z}_j^{sv}$ and $\mathbf{m}_j$:

$$\mathcal{L}_{gen}^{sv} = \mathbb{E}_{s_i \sim \mathcal{S}, (\mathbf{m}_j, \mathbf{p}_j, \mathbf{v}_j) \sim \mathcal{T}_{s_i}} \| \mathbf{z}_j^{sv} - \mathbf{m}_j \|_2 \quad (1)$$

$$\mathbf{z}_j^{sv} = f_g\left(\mathbf{p}_j, \mathbf{v}_j, \mathbf{e}_j^{sv}; \theta_g\right), \mathbf{e}_j^{sv} = f_s\left(\mathbf{m}_j; \theta_s^{sv}\right) \quad (2)$$

During the reconstruction phase, the DSR system takes in the dysarthric speech of speaker s_d to generate reconstructed mel-spectrograms as $f_g(\tilde{\mathbf{p}}, \tilde{\mathbf{v}}, \mathbf{e}^{sv}; \theta_g)$, where $\tilde{\mathbf{p}}$ are phoneme embeddings extracted by fine-tuned content encoder Φ_{sd} and expanded with predicted normal duration, $\tilde{\mathbf{v}}$ is predicted normal F_0 contour, and $\mathbf{e}^{sv}$ is the speaker representation. Then Parallel WaveGAN (PWG) (18) is adopted as the neural vocoder to transform $f_g(\tilde{\mathbf{p}}, \tilde{\mathbf{v}}, \mathbf{e}^{sv}; \theta_g)$ into speech waveform.

4. EXPERIMENTAL RESULTS AND ANALYSIS

The datasets used in our experiments contain LibriSpeech (19), LJSpeech (20), VCTK (21) and UASPEECH (13). Different modules of proposed SED-based DSR system are independently trained by using different datasets. Content encoder Φ_p is pre-trained by 960h training data of LibriSpeech, prosody encoder including the phoneme duration predictor and F_0 predictor is trained by the data of a healthy female speaker from LJSpeech, decoder θ_g and PWG vocoder are trained by VCTK. For dysarthric patients, we choose UASPEECH for experiments, where two male speakers (M05 and M07) and two female speakers (F04 and F02) are selected from UASPEECH, M05/F04 and M07/F02 have moderate and moderate-severe dysarthria respectively. We use the speech data of blocks 1 and 3 of each dysarthric speaker for fine-tuning content encoder, and block 2 for testing. The inputs of content encoder are 40-dim mel-spectrograms appended with deltas and delta-deltas which results in 120-dim vectors, the targets of decoder are 80-dim mel-spectrograms, all mel-spectrograms are computed with 400-point Fourier transform, 25ms Hanning window and 10ms hop length. All acoustic features including mel-spectrograms and log-scale F_0 are normalized to have zero mean and unit variance.

The speaker encoder training follows (16). The pre-training and fine-tuning of content encoder are performed by Adadelta optimizer (22) with 1M and 2K steps respectively by using learning rate of 1 and batch size of 8. For the phoneme duration predictor and F_0 predictor, 3-layers with 256 units per direction are used to form the stacked Bi-GRU, which is followed by a 1-layer 256-dim Bi-GRU and the filter-bank that contains 3 convolution layers with kernel size of 5, 9 and 19 respectively, and a 1-dim fully-connected (FC) layer to predict the duration or F_0 value. Both duration and F_0 predictors are trained by Adam optimizer (23) with 30K steps by using learning rate of 1e-3 and batch size of 16. Decoder is optimized in a similar way except that the training steps are set to 50K.

We compare the proposed approach with an end-to-end DSR (E2E-DSR) system (8). Subjective listening tests have been conducted to measure the speech naturalness and speaker similarity, in terms of 5-point mean opinion score (MOS, 1-bad, 2-poor, 3-fair, 4-good, 5-excellent) rated by 20 subjects for 20 utterances randomly selected from each of four dysarthric speakers[2]. Objective evaluation based on ASR is used to measure the speech intelligibility in terms of word error rate (WER).

[2] Some audio samples used for subjective evaluation: https://wendison.github.io/ASA-DSR-demo/

Table 1 – Comparison results of MOS with 95% confidence intervals for speech naturalness

Approaches	M05	F04	M07	F02
Original	2.37±0.08	2.49±0.09	1.95±0.10	1.79±0.09
E2E-DSR	3.64±0.11	3.40±0.13	3.58±0.12	3.35±0.12
Proposed	3.88±0.11	3.92±0.10	3.80±0.10	3.79±0.09

Table 2 – Comparison results of MOS with 95% confidence intervals for speaker similarity

Approaches	M05	F04	M07	F02
Original	4.93±0.01	4.89±0.02	4.95±0.01	4.96±0.01
E2E-DSR	2.66±0.12	2.50±0.13	2.47±0.16	2.27±0.14
Proposed	2.70±0.14	2.27±0.10	2.55±0.14	1.88±0.13

Table 3 – WER (Δ) (%) results comparison, where Δ denotes the WER reduction of different approaches compared with original dysarthric Speech.

Approaches	M05	F04	M07	F02
Original	91.0	81.7	95.6	95.9
E2E-DSR	69.8 (21.2)	69.3 (12.4)	73.1 (22.5)	72.0 (23.9)
Proposed	61.7 (29.3)	64.6 (17.1)	62.7 (32.9)	65.3 (30.6)

4.1 Comparison Based on Speech Naturalness

Table 1 gives the MOS results of naturalness of original or reconstructed speech from different systems. We can see that the original dysarthric speech has lowest naturalness, due to non-standard articulation with abnormal phoneme duration and F_0 contour that degrade the perceptual quality. However, it is encouraging to see that both E2E-DSR and proposed system improve the naturalness of original dysarthric speech, where the proposed DSR system consistently achieves higher speech naturalness scores for all speakers. On one hand, E2E-DSR system models the phoneme duration implicitly by using the attention mechanism to align the length between input and output, but the attention module tends to make alignment errors, e.g., the stop token may not be predicted accurately, which degrades the speech naturalness. On the other hand, the proposed phoneme duration predictor can accurately infer the duration of each phoneme for expansion, together with the explicit modelling of pitch by using the F_0 predictor, the generated speech has higher quality with stable and accurate prosodic patterns.

4.2 Comparison Based on Speaker Similarity

Table 2 gives the MOS results of speaker similarity. We observe that the original speech has high scores of speaker similarity as expected, while both E2E-DSR and proposed system achieve low scores of speaker similarity, as the speaker encoder is trained on large-scale normal speech data, it may not fully capture the characteristics of previously unseen dysarthric speaker. Through our listening tests, the gender of reconstructed speech by E2E-DSR and proposed system may be changed especially for female speakers, this shows the limited generalization ability of the speaker encoder to extract effective speaker representations from the dysarthric speech.

4.3 Comparison Based on Speech Intelligibility

Objective evaluation of speech intelligibility is conducted by using a publicly released speech recognition model, i.e., Jasper (24), to compute WER with greedy decoding, and the results are shown in Table 3. We observe that the original dysarthric speech has high WER that is larger than 80%, which shows that an ASR system trained by large-scale normal speech data of healthy speakers nearly cannot

recognize the spoken content of dysarthric speech, indicating that dysarthric speaker has significant deviations of pronunciation patterns from those of healthy speakers. However, compared with original dysarthric speech, it is promising to see that both E2E-DSR and proposed systems achieve WER reduction, where the proposed system consistently achieves larger WER reduction for all dysarthric speakers, showing the effectiveness of proposed explicit prosody correction to improve the speech intelligibility, leading to 23.2% and 31.8% absolute WER reduction on average for speakers M05/F04 and M07/F02 that have moderate and moderate-severe dysarthria respectively.

5. CONCLUSIONS

This paper presents a DSR system based on NED framework, which employs a bank of encoders that have the capacity of modelling different factors of speech signals, i.e., content, prosody and speaker, which correspond to three subproblems of DSR, i.e., content restoration, prosody correction and speaker identity preservation, respectively. Then a decoder combines the representations of restored content, corrected prosody and original speaker identity to generate normal sounding speech. Experimental results show that the proposed DSR system achieves significant improvements of speech intelligibility and naturalness. However, the reconstructed speech still has low speaker similarity. One simple way to tackle this issue is to fine-tune the speaker encoder by using dysarthric speech as proposed in (25), which improves the speaker similarity to some extent. However, there is still room for improvement, we leave this as our future work.

ACKNOWLEDGEMENTS

This research is supported partially by the HKSAR Research Grants Council's General Research Fund (Ref Number 14208817) and also partially by the Centre for Perceptual and Interactive Intelligence, a CUHK InnoCentre.

REFERENCES

1. Yunusova Y, Weismer G, Westbury JR, Lindstrom MJ. Articulatory movements during vowels in speakers with dysarthria and healthy controls. 2008.
2. Rudzicz F, editor Acoustic transformations to improve the intelligibility of dysarthric speech. Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies; 2011.
3. Kumar SA, Kumar CS, editors. Improving the intelligibility of dysarthric speech towards enhancing the effectiveness of speech therapy. 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI); 2016: IEEE.
4. Kain AB, Hosom J-P, Niu X, Van Santen JP, Fried-Oken M, Staehely J. Improving the intelligibility of dysarthric speech. Speech communication. 2007;49(9):743-59.
5. Aihara R, Takashima R, Takiguchi T, Ariki Y, editors. Consonant enhancement for articulation disorders based on non-negative matrix factorization. Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference; 2012: IEEE.
6. Aihara R, Takashima R, Takiguchi T, Ariki Y, editors. Individuality-preserving voice conversion for articulation disorders based on non-negative matrix factorization. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; 2013: IEEE.
7. Aihara R, Takiguchi T, Ariki Y, editors. Phoneme-Discriminative Features for Dysarthric Speech Conversion. Interspeech; 2017.
8. Wang D, Yu J, Wu X, Liu S, Sun L, Liu X, et al., editors. End-to-end voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020: IEEE.

9. Doshi R, Chen Y, Jiang L, Zhang X, Biadsy F, Ramabhadran B, et al., editors. Extending parrotron: An end-to-end, speech conversion and speech recognition model for atypical speech. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2021: IEEE.
10. Chen Z, Ramabhadran B, Biadsy F, Zhang X, Chen Y, Jiang L, et al. Conformer Parrotron: a Faster and Stronger End-to-end SpeechConversion and Recognition Model for Atypical Speech. 2021.
11. Huang W-C, Kobayashi K, Peng Y-H, Liu C-F, Tsao Y, Wang H-M, et al. A preliminary study of a two-stage paradigm for preserving speaker identity in dysarthric voice conversion. arXiv preprint arXiv:210601415. 2021.
12. Chen C-Y, Zheng W-Z, Wang S-S, Tsao Y, Li P-C, Lai Y-H, editors. Enhancing Intelligibility of Dysarthric Speech Using Gated Convolutional-Based Voice Conversion System. INTERSPEECH; 2020.
13. Kim H, Hasegawa-Johnson M, Perlman A, Gunderson J, Huang TS, Watkin K, et al., editors. Dysarthric speech database for universal access research. Ninth Annual Conference of the International Speech Communication Association; 2008.
14. Xiong F, Barker J, Christensen H, editors. Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019: IEEE.
15. McAuliffe M, Socolof M, Mihuc S, Wagner M, Sonderegger M, editors. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. Interspeech; 2017.
16. Liu S, Wang D, Cao Y, Sun L, Wu X, Kang S, et al., editors. End-to-end accent conversion without using native utterances. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020: IEEE.
17. Wan L, Wang Q, Papir A, Moreno IL, editors. Generalized end-to-end loss for speaker verification. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2018: IEEE.
18. Yamamoto R, Song E, Kim J-M, editors. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020: IEEE.
19. Panayotov V, Chen G, Povey D, Khudanpur S, editors. Librispeech: an asr corpus based on public domain audio books. 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP); 2015: IEEE.
20. Ito K, Johnson L. The lj speech dataset. 2017.
21. Veaux C, Yamagishi J, MacDonald K. Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. 2016.
22. Zeiler MD. Adadelta: an adaptive learning rate method. arXiv preprint arXiv:12125701. 2012.
23. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.
24. Li J, Lavrukhin V, Ginsburg B, Leary R, Kuchaiev O, Cohen JM, et al. Jasper: An end-to-end convolutional neural acoustic model. arXiv preprint arXiv:190403288. 2019.
25. Wang D, Liu S, Wu X, Lu H, Sun L, Liu X, et al., editors. Speaker Identity Preservation in Dysarthric Speech Reconstruction by Adversarial Speaker Adaptation. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022: IEEE.

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0198

Towards robust one-shot voice conversion with cycle phonetic posteriorgrams and multi-scale speaker representations

Yannian CHEN[(1)], Lijuan LIU[(2)], Yajun HU[(2)] and Zhenhua LING[(1)]

[(1)]National Engineering Research Center of Speech and Language Information Processing,
University of Science and Technology of China, China
[(2)]iFLYTEK Research, iFLYTEK Co., Ltd., China

ABSTRACT

One-shot voice conversion (VC) aims to convert the voice across arbitrary speakers even unseen during training, with only one reference utterance from the target speaker. It is still a challenging task as both content and speaker representations estimated from speech are required to be reliable. In this paper, we propose a novel method which combines phonetic posteriorgrams (PPGs) and multi-scale speaker representations to achieve robust one-shot VC. PPGs are extracted by a pretrained automatic speech recognition (ASR) model and contain robust linguistic information. Cycle PPGs which are generated from a cycle conversion process are used for training to eliminate the influence of residual speaker information in PPGs. Furthermore, multi-scale speaker representations composed of global and local ones are utilized. Global speaker representations are modeled by an advanced speaker embedding network which integrates squeeze-excitation blocks and attentive statistics pooling to get utterance-level vectors. In order to extract time-varying and content-dependent local speaker representations, an attention mechanism is adopted to select the most suitable features depending on each content frame, which is expected to refine the coarse speaker information given by utterance-level speaker representations. Experimental results showed that the proposed method outperformed baseline methods on one-shot VC.

Keywords: Voice Conversion, Multi-scale, Speaker Representations, Phonetic Posteriorgrams

1 INTRODUCTION

Voice conversion (VC) aims to convert certain speech characteristic from one speaker to make it sound like spoken by another speaker, without changing linguistic contents [1]. Early VC methods are built on parallel training corpus, where source and target speakers speak the same linguistic contents. Thus, the mapping function between source and target speeches can be learned directly. Considering that it is difficult and time-consuming to collect parallel utterances, developing VC methods free of parallel data is necessary. In recent years non-parallel VC methods have been explored, such as phonetic posteriorgrams (PPGs) based ones [2, 3, 4, 5], variational auto-encoder (VAE) [6, 7] based ones, generative adversarial network (GAN) based ones [8, 9] and disentangled representations based ones [10].

Recently, more researches have been attracted to one-shot VC, which deal with the conversion across arbitrary speakers which are even unseen during training, with only one reference utterance from the target speaker. Many approaches have been explored for one-shot VC task. In particular, autoencoders with disentangled linguistic and speaker representations has been proved to be an effective way. The linguistic and speaker representations are extracted by two encoders respectively. And it is assumed that the linguistic information is dynamic and time-varying while the speaker information is static and time-invarant. Thus, the speaker representation is often modeled as utterance-lever vectors. Disentangling linguistic and speaker information can be achieved by carefully tuned bottleneck [11], adaptive instance normalization (IN) [12], vector quantization (VQ) [13, 14] and mutual information (MI) [15]. Despite recent progress, it is still a challenging task as the current one-shot VC

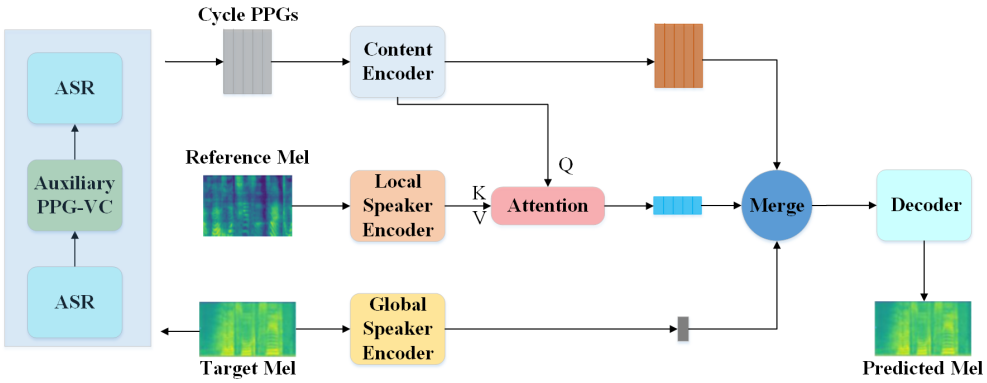


Figure 1. The structure of the proposed method.

methods are not robust for unseen speakers [16]. This is mainly because either linguistic or speaker representations estimated from speech is not accurate and reliable. Disentanglement methods like IN,VQ and MI may cause the loss of linguistic information, and the training is sometimes unstable. Representing speaker characteristics with utterance-level vectors is also insufficient, as some local characteristics like pronunciation variations depend on linguistic contents.

In this paper, we contribute to propose a robust one-shot VC method by modeling linguistic and speaker representations in a more effective way. Phonetic posteriorgrams (PPGs) are adopted as the linguistic representations, which are extracted by a pretrained automatic speech recognition (ASR) model and contain robust linguistic information. To eliminate the influence of residual speaker information in PPGs, previous work [17] proposed a cyclic training methods in any-to-one VC. But in one-shot VC, it has not been explored. We modify the cyclic training methods and adopt it to one-shot VC. The initial PPGs are consumed by an auxiliary PPG-VC model to produce pseudo speech with different speaker characteristics, and then PPGs are extracted from the pseudo speech, called Cycle PPGs, as the final linguistic representations. Furthermore, multi-scale speaker representations composed of global and local ones are utilized. Global speaker representations are modeled by an advanced speaker embedding network which integrates squeeze-excitation blocks and attentive statistics pooling to get utterance-level vectors. In order to extract time-varying and content-dependent local speaker representation, an attention mechanism is adopted to select the most suitable features depending on each content frame, which is expected to refine the coarse speaker information given by utterance-level speaker representations.

2 METHODS

The overall framework of the proposed method is illustrated in Figure 1, which consists of an ASR model, an auxiliary PPG-VC model, a content encoder, two speaker encoders, a merge module and a decoder. The ASR model extract PPGs from mel-spectrograms. The content encoder captures linguistic information from PPGs. The speaker encoders consist of a global one which models global speaker representations and a local one which models content-dependent frame-level local speaker representations from mel-spectrograms. The linguistic and multi-scale speaker representations are concatenated and passed into the merge module to produce merged representations, and a Tacotron2-like autoregressive LSTM based decoder predicts mel-spectrograms from it. In training, linguistic and multi-scale speaker representaions are all obtained from the mel-spectrograms of the same target speaker, and the training goal is to reconstruct the target mel-spectrograms. Instead of using normal PPGs, Cycle PPGs, which are extracted by converting the target mel-spectrograms towards sampled speakers to generate pseudo mel-spectrograms and then extracting PPGs, are used. At the conversion stage, PPGs are extracted from the source speech and passed into the content encoder to obtain linguistic representations while multi-scale speaker representations are obtained from the reference speech of the target speaker. Then a HiFiGAN vocoder is used to generate waveforms from the converted mel-spectrograms.

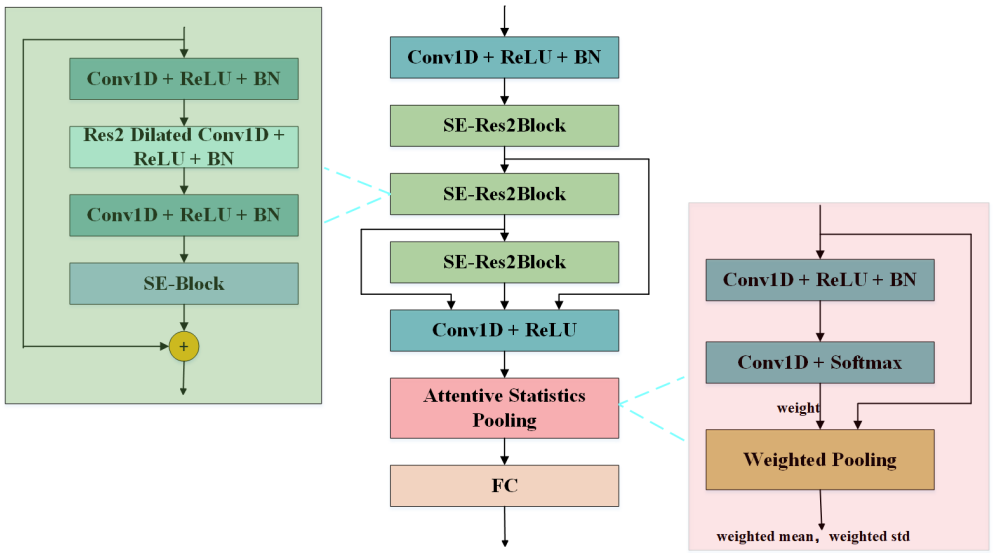


Figure 2. The structure for modeling global speaker representations.

2.1 Extracting Cycle PPGs

PPGs extracted from pretrained ASR model contain rich content-related features to ensure content correctness in the conversion. Traditional PPGs-based VC treats PPGs as speaker-independent content features, which takes PPGs of the target speaker as input for training while takes that of the source speaker for conversion. However, some residual speaker information also remains in PPGs, and the source speaker information in PPGs degrades the similarity of converted speech to the target speaker at the conversion stage. Previous work [17] proposed a cyclic training method to deal with this problem. But it focuses on training on the task with one specific target speaker to achieve any-to-one VC, and can not be applied to unseen target speakers. We modify this method and expand it to one-shot VC. An auxiliary VC model which takes PPGs as input and produces mel-spectrograms is introduced. It is trained with a multi-speaker speech corpus, and we treat these speakers as pseudo-source speakers. In training, the auxiliary VC model converts the target mel-spectrograms towards a randomly sampled pseudo-source speaker, and produces pseudo mel-spectrograms which have the same temporal structure but different speaker characteristics with the target mel-spectrograms. Then we extract PPGs from the pseudo mel-spectrograms as the content features for following modeling, and name them Cycle PPGs. Since the global and local speaker representations are all from the target speakers, the training procedure directly optimizes the conversion flow of pseudo-source to target, which is consistent with the conversion flow of source to target at the conversion stage.

2.2 Modeling global speaker representations

As shown in Figure 2, the global speaker encoder is based on a ResNet architecture, with squeeze-excitation blocks (SE-block) [18] and attentive statistics pooling, to get utterance-level vectors. The model structure is similar to the ECAPA-TDNN model[19] for speaker verification area.

The SE-block is first proposed for computer vision tasks to model global channel interdependence of feature maps and is also proven successful in speaker verification. The SE-block consists of two operations, and the first operation is squeeze which generates channel-wise descriptor by calculating the mean vector $\boldsymbol{z}$ of frame-level features across the time-axis as

$$\boldsymbol{z} = \frac{1}{T}\sum_{t}^{T}\boldsymbol{h}_t. \tag{1}$$

$\boldsymbol{z}$ is then used in the second operation of excitation, which calculates weights for each channel, formulated as

$$\boldsymbol{s} = \sigma(\boldsymbol{W}_2 f(\boldsymbol{W}_1 \boldsymbol{z} + \boldsymbol{b}_1) + \boldsymbol{b}_2), \tag{2}$$

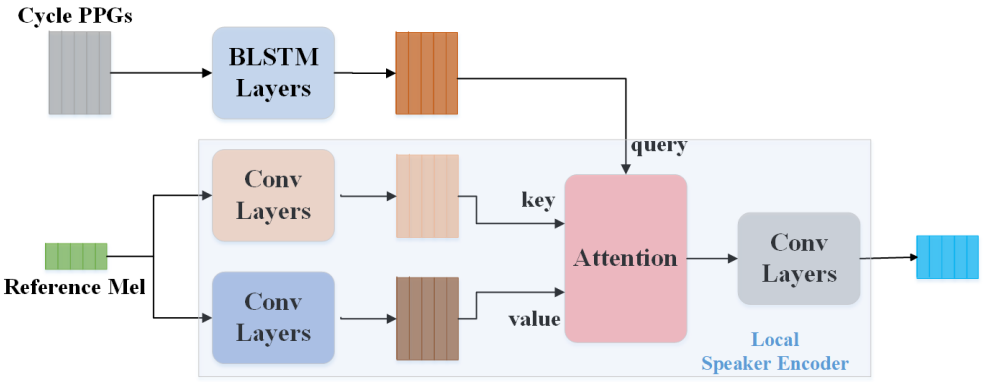


Figure 3. The structure for modeling local speaker representations.

where $\sigma(\cdot)$ and $f(\cdot)$ denote the sigmoid function and a non-linear function respectively, $\boldsymbol{W}_1 \in \mathbb{R}^{R\times C}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{C\times R}$, and C and R are the numbers of input channels and reduced dimensionality respectively. The resulting vector $\boldsymbol{s}$ contains the channel-wise weight s_c for each channel, which are multiplicated with the original input to obtain the final output of each channel c

$$\tilde{\boldsymbol{h}}_c = s_c \boldsymbol{h}_c. \tag{3}$$

Then attentive statistics pooling is adopted to generate utterance-level vectors. A attention module calculates a channel-dependent self-attention score $e_{t,c}$ for each frame-level feature and these scores are then normalized by a softmax function as

$$e_{t,c} = \boldsymbol{v}_c^T f(\boldsymbol{W}\boldsymbol{h}_t + \boldsymbol{b}) + k_c, \tag{4}$$

$$\alpha_{t,c} = \frac{\exp(e_{t,c})}{\sum_{\tau}^{T} \exp(e_{\tau,c})}. \tag{5}$$

$\alpha_{t,c}$ represents the importance of each frame given the channel and is used to calculate the weighted statistics of channel c. For each utterance, the channel components $\tilde{\mu}_c$ of the weighted mean vector $\tilde{\boldsymbol{\mu}}$ and the weighted standard deviation vector $\tilde{\boldsymbol{\sigma}}$ are estimated as

$$\tilde{\mu}_c = \sum_{t}^{T} \alpha_{t,c} h_{t,c}, \tag{6}$$

$$\tilde{\sigma}_c = \sqrt{\sum_{t}^{T} \alpha_{t,c} h_{t,c}^2 - \tilde{\mu}_c^2}. \tag{7}$$

$\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\sigma}}$ are concatenated as the final output. Unlike the global average pooling usually used in speaker embeddings in which each frame of features contributes equally, the attentive statistics pooling further analyses the feature maps and obtains more speaker related information.

2.3 Modeling local speaker representations

As shown in Figure 3, the local speaker encoder captures local speaker information from the reference mel-spectrograms spoken by the same target speaker. An attention module is introduced to attend to the most content-relavant frames, and the multi-head attention is adopted in our implementation for sufficient model capacity. The reference mel-spectrograms are passed into a convolutional block which consists of three 1-D convolutional layers with ReLU activations to give attention keys, and are then sent into another convolutional block to get attention values. The attention query is given by the hidden features of the content encoder, which is constructed by two bidirectional LSTM layers. The attention output is passed into a post-processing convolutional block to form the local speaker representations. In this manner, the generated local speaker representations have the same sequence length as the linguistic representations, and each content frame of linguistic

Table 1. MCDs (dB) and F_0 RMSEs (Hz) of baseline and proposed methods.

Method	TMM1-to-TMF1		TMF1-to-TMM1	
	MCD	F_0 RMSE	MCD	F_0 RMSE
AdaIN-VC	4.085	41.390	3.943	36.863
VQMIVC	3.810	48.769	3.838	42.995
Proposed	**3.531**	**24.939**	**3.595**	**29.754**

Table 2. Naturalness and similarity of baseline and proposed methods.

Method	TMM1-to-TMF1		TMF1-to-TMM1	
	Naturalness	Similarity	Naturalness	Similarity
AdaIN-VC	2.045±0.093	1.860±0.059	2.245±0.092	1.910±0.049
VQMIVC	2.796±0.093	2.520±0.075	2.940±0.101	2.895±0.087
Proposed	**3.765±0.091**	**3.860±0.078**	**3.940±0.083**	**3.900±0.077**

representations obtains a related frame of local speaker representations, ie., fine-grained speaker representations. Designing the fine-grained local speaker representations aims to refine the coarse speaker information given by the utterance-level global speaker embeddings.

2.4 Training details

Only reconstruction loss, i.e., the L1 loss between the predicted and the target mel-spectrograms, is used for training. As only linguistic information in Cycle PPGs is useful for reconstructing the target mel-spectrograms, the speaker encoders automatically learn global and local speaker-dependent representations for better reconstruction. The input mel-spectrograms of global speaker encoder are the reconstruction target during the entile training stage, for obtaining accurate global information. While, content-dependent local speaker encoder must be trained to handle the situation that the speaking content of the reference utterance given by the target speaker differs from the source utterance given by the source speaker at the conversion stage. Thus, we first give the same utterance of the reconstruction target for training some epochs to stablize the attention. Then the reference mel-spectrograms are chosen between the reconstruction target and the randomly sampled mel-spectrograms from the same speaker, where the probability of choosing the randomly sampled mel-spectrograms increases from 0 to 0.95 during training procedure. The ASR model and the auxiliary VC model are pretrained, and are not optimized during training.

3 EXPERIMENTS

3.1 Experiment setup

Open-source multi-speaker Mandarin corpora, AISHELL-3 [20] which contained 218 speakers with 85 hours of recordings and DiDiSpeech [21] which contained 500 speakers with 60 hours of recordings, were used for training the VC models. For one-shot testing, two Mandarin speakers in Voice Conversion Challenge 2020 (one male and one female, denoted as TMM1 and TMF1) were used as the test set. They had 70 parallel Mandarin utterances. We converted each one of the two speakers towards another, resulting in two conversion pairs TMM1-to-TMF1 and TMF1-to-TMM1.

We used 80-dimensional mel-spectrograms with 10 ms frame shift as acoustic features. The joint CTC-attention based approach similar to previous work [22] was employed to build the ASR model, which was composed of 2 VGG-like CNN layers, 5 BLSTM layers with 512 dimensions per direction, a fully-connected bottleneck layer with 256 dimensions, location-sensitive attention [23] and LSTM-based decoder with 512 dimensions. 2200 hours of Chinese recordings were used for training the ASR model. The 256-dimensional

Table 3. MCDs (dB) and F_0 RMSEs (Hz) of proposed and ablated methods.

Method	TMM1-to-TMF1		TMF1-to-TMM1	
	MCD	F_0 RMSE	MCD	F_0 RMSE
Proposed	**3.531**	**24.939**	**3.595**	**29.754**
-cycle	3.601	29.295	3.671	32.317
-local	3.587	27.519	3.621	31.846
-global	3.695	36.236	3.727	36.022

bottleneck features were taken as PPGs. The merge module were built with a simple two-layer BLSTM with 256 dimensions per direction.

Two state-of-the-art one-shot VC methods, AdaIN-VC [12] and VQMIVC [15], were adopted as baselines for performance comparision. We used their official open source implementations and modified some settings to our dataset. We followed the official open-source implementation of HiFiGAN [24] to train the universal vocoder using LibriTTS [25] and AISHELL-3 datasets, for generating 24kHz wavforms [1].

3.2 Objective evaluation

We randomly sampled 50 utterances from the 70 parallel recordings as source speech and randomly sampled 4 utterances from the remaining 20 utterances as reference speech, resulting in 200 converted speech of each conversion pairs for objective evaluation. Mel-cepstral coefficients (MCCs) and F_0 were extract from the natural and converted speeches. And we used Mel-cepstrum distortion (MCD) and root mean square error of F_0 (F_0 RMSE) as objective metrics. The results are reported in Table 1. For both conversion pairs, our proposed method achieved the best performance on all metrics.

3.3 Subjective evaluation

We generated 20 converted speech for each conversion pair, by randomly sampling 20 utterances as source speeches and 1 utterance from the remaining as reference speech, for subjective evaluation. For each conversion pair, all methods were grouped into a mean opinion score (MOS) listening test to compare their naturalness and similarity. 12 native Chinese listeners were involve in the listening test, and they were asked to give a 5-scale opinion score (5: excellent, 4: good, 3: fair, 2: poor, 1: bad) on both naturalness and similarity of each sample. For scoring the similarity, natural speech of the target speaker were also given for reference. The results are reported in Table 2. Our proposed method achieved significantly higher naturalness and similarity MOS for both TMM1-to-TMF1 and TMF1-to-TMM1 pairs.

3.4 Ablation studies

In this section, we conducted ablation studies to validate the effectiveness of several components in the proposed methods. For investigating the effects of Cycle PPGs, we removed this strategy by replacing the Cycle PPGs with normal PPGs extracted from the target mel-spectrograms, denoted as "-cycle". For investigating the effects of local speaker representations, we removed the local speaker encoder and the attention module, with only global speaker representations providing speaker information, denoted as "-local". For investigating the effects of global speaker representations, we removed the global speaker encoder and only used local speaker representations, denoted as "-global". Table 3 demostrates the objective evaluation results of ablation studies. As we can see, the proposed method outperformed all abalted methods, which confirmed the effectiveness of different components.

[1] Audio samples are available at https://nian2932491631.github.io/CyclePPGMSSROneShotVC/.

4 CONCLUSION

In this paper we propose a novel method for robust voice conversion, which combines cycle phonetic posteriorgrams and multi-scale speaker representations. We extract cycle phonetic posteriorgrams using a pretrained automatic speech recognition model from a cycle conversion process, and adopt them as linguistic representations. And we propose multi-scale speaker representations composed of global and local ones. The global speaker representations are represented as utterance-level vectors and are modeled by a network which integrates squeeze-excitation blocks and attentive statistics pooling. While modeling local speaker representations, an attention mechanism is adopted to select the most suitable features depending on each content frame. The proposed method models linguistic and speaker representations in a more effective way. Experimental results showed that it outperformed baseline methods in terms of natrualness and similarity.

REFERENCES

[1] D. G. Childers, K. Wu, D. M. Hicks, and B. Yegnanarayana, "Voice conversion," Speech Communication, vol. 8, no. 2, pp. 147–158, 1989.

[2] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in ICME, 2016, pp. 1–6.

[3] L. J. Liu, Z. H. Ling, Y. Jiang, M. Zhou, and L. R. Dai, "WaveNet vocoder with limited training data for voice conversion," in INTERSPEECH, 2018, pp. 1983–1987.

[4] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in ICASSP, 2019, pp. 6790–6794.

[5] L. J. Liu, Y. N. Chen, J. X. Zhang, Y. Jiang, Z. H. Ling, and L. R. Dai, "Non-parallel voice conversion with autoregressive conversion model and duration adjustment," in Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, 2020.

[6] C. C. Hsu, H. T. Hwang, Y. C. Wu, Y. Tsao, and H. M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in APSIPA, 2016, pp. 1–6.

[7] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in ICASSP, 2018, pp. 5274–5278

[8] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in European Signal Processing Conference (EUSIPCO), 2018, pp. 2100–2104.

[9] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 266–273.

[10] J. X. Zhang, Z. H. Ling, and L. R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019.

[11] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa Johnson,"Autovc: Zero-shot voice style transfer with only autoencoder loss," in International Conference on Machine Learning. PMLR, 2019.

[12] J. C. Chou, and H. Y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," in INTERSPEECH, 2019, pp. 664–668.

[13] D. Y. Wu and H. Y. Lee, "One-shot voice conversion by vector quantization," in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7734–7738.

[14] D. Y. Wu, Y. H. Chen, and H. Y. Lee, "Vqvc+: One-shot voice conversion by vector quantization and u-net architecture," in INTERSPEECH, 2020, pp. 4691–4695

[15] D. S. Wang, L. Q. Deng, Y. T. Yeung, X. Chen, X. Y. Liu, and H. Meng. "VQMIVC: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion," in INTERSPEECH,2021, pp. 1344–1348.

[16] T. H. Huang, J. H. Lin, and H. Y. Lee, "How far are we from robust voice conversion: A survey," in IEEE Spoken Language Technology Workshop, 2021, pp. 514–521.

[17] Y. N. Chen, L. J. Liu, Y. J. Hu, Y. Jiang, and Z. H. Ling, "Improving recognition-synthesis based any-to-one voice conversion with cyclic training," in ICASSP, 2022, pp 7007-7011.

[18] J. Hu, L. Shen, and G. Sun,"Squeeze-and-excitation networks," in IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141

[19] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in INTERSPEECH, 2020.

[20] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," in INTERSPEECH, 2021, pp. 2756–2760

[21] T. W. Guo, C. Wen, D. W. Jiang, and et al., "Didispeech: A large scale mandarin speech corpus," in ICASSP, 2021, pp. 6968–6972.

[22] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in ICASSP, 2017, pp. 4835–4839.

[23] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in Advances in neural information processing systems, 2015, pp. 577–585.

[24] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in Advances in Neural Information Processing Systems, vol. 33, 2020.

[25] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in INTERSPEECH,2019, pp. 1526-1530.

PROCEEDINGS of the
24th International Congress on Acoustics
October 24 to 28, 2022 in Gyeongju, Korea

ABS-0265

Towards laughter synthesis using voice conversion

Noritake SUTO; Kazuhiro KONDO

Yamagata University, Japan

ABSTRACT

The purpose of this study is to synthesize laughter by transforming normal speech. Laughter frequently appears in human daily conversations, making communication rich and enjoyable.

A variety of synthetic voices already exist in everyday life, such as voice assistants and video narration, and is expected that they will be widely used in dialogue systems in the future. However, while emotion-rich speech synthesis has been actively studied, studies on speech with no conventional linguistic information, such as laughter, are still in their developmental stages. In this study, we first extract the features of read-sentence speech, convert them into features corresponding to laughter using a conversion model learned by adversarial learning, and then synthesize laughter. Since nonparallel data is used to train the model, this method is considered effective for laughter synthesis since high-quality speech data of laughter can be considered difficult to record.

The converted laughter was evaluated by subjects for its likeliness as laughter on a 5-point scale, resulting in a score of 3.66, higher than the score of 1.23 for the voice before conversion. This indicates that the proposed voice conversion is effective in synthesizing naturally sounding laughter.

Keywords: Laughter, Voice Conversion

1. INTRODUCTION

Laughter is a common part of everyday conversation. For example, a 2011 study by Tanaka and Campbell (1) showed that approximately 11% of 30-minute conversations are composed of laughter. Laughter occurs frequently in conversation and enriches our communication. On the other hand, recent speech synthesis technology has been developing rapidly along with machine learning to produce synthetic speech that is indistinguishable from human speech. However, this is mostly in the form of read-sentence speech, and emotional speech, especially in the form of laughter, is still in the development stage. Rich communication using synthetic speech requires improvement in its naturalness and expression.

The purpose of this study is to propose an effective method to convert emotionless read voice into laughter by extending recent voice quality conversion methods using deep learning.

2. VOICE CONVERSION AND EVALUATION

2.1 Audio data conditions

The audio data used in this study was recorded by us and is a sample of spontaneous laughter when watching some comedy videos. The recorded audio included 297 laughter sounds, from which 96 sounds that were perceived as typical laughter, such as “ha-ha-ha” and “hu-hu-hu”, were selected.

The read-sentence speech was recorded by listening to those sounds and reading the transcribed text including laughter. The speaker was a 22-year-old male, and the transcription and selection of the laughter were done manually. The recordings were done using a headset microphone (SENNHEISER HMD300PRO). The sampling frequency of the original sample was 44.1 kHz and was down-sampled to 24 kHz.

Table 1 – summary of the recorded speech data

Categories	Number of data	Average time [s]	Average F_0 [Hz]
Natural laughter	96	2.1	208
Read laughter	96	2.1	142

The read laughter is based on the transcription of the natural laughter, and is linguistically identical. However, the fundamental frequency, which is one of the features that defines the quality of the speech, was about 66 Hz lower on average than natural laughter in an emotionless voice. An example sequence of these two is presented in Figure 1.

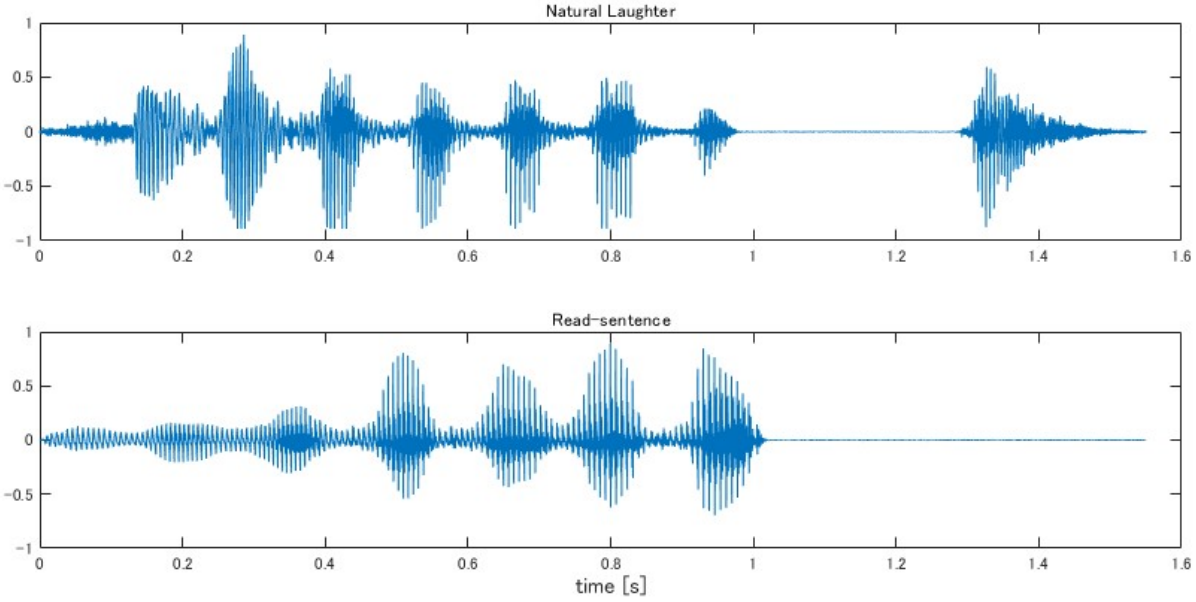


Figure 1 – Natural laughter and a reading laughter
(Laughter includes an inhalation sound beyond 1.3 s)

2.2 Introduction of the voice conversion model

The conversion experiments used CycleGAN-VC2 (2) (Kaneko and Kameoka, 2018), which is known to work with non-parallel data. CycleGAN learns the mapping between domains by alternating learning between two generators and two classifiers. One generator translates from domain A to domain B (in this case, from “read” to “natural laughter”) and the other from domain B to domain A (in this case, from “natural laughter” to “read”), each of which is learned by adversarial learning. In voice conversion using this, vocal tract features are extracted from speech, and the learning described above is carried out to create a generator capable of converting vocal tract features, and converted speech is obtained by resynthesizing speech using the converted vocal tract features.

First, Mel-cepstral coefficients (MCEPs), logarithmic fundamental frequency (log F_0), and aperiodicities were extracted from speech using the WORLD analysis system (3) (M. Morise, et al., 2016). Of these, MCEPs were fed to a Neural Network to map between speech and laughter, and log F_0 is converted using logarithm Gaussian normalized transformation. Finally, the aperiodicities were used as is to synthesize using the converted features. The frame length during analysis was 5ms and the order of the MCEPs was 35.

In the learning of CycleGAN, 96 voice data were divided into 12 groups of 8 each, of which one group was used as verification data for a 12-way cross-validation using the remaining data.

The above learning conditions are summarized in Table 2.

Table 2 – Condition of the converter by CycleGAN

Input-output size	36×128×1
Optimizer	Adam
Loss	MSE
Number of voice data	192 (read:96, laughter:96)

2.3 Subjective evaluation

We conducted a subjective evaluation based on two indicators:

(1) Naturalness: The subjective evaluation experiment on the naturalness of the speech samples was conducted for two types of sound: original laughter and converted laughter. Subjects listened to a total of 24 sounds, 12 each in random order, and rated their naturalness on a 5-point scale, with higher scores indicating more naturalness as speech. The labels indicated were 5-very natural, 4- natural, 3-neutral, 2-unnatural, and 1-very unnatural.

(2) Likeliness: The same was done for the evaluation of likeliness as laughter, which was performed on 36 samples, including the read laughter speech. In this case, the higher scores indicate higher likeliness as laughter, and lower scores indicate its likeliness as normal (neutral) speech. The labels indicated were 5-definitel laughter, 4−likely laughter, 3-neutral, 2-likely read speech, and 1-definietely read speech.

3. RESULTS

3.1 Result of conversion

Figure 2 shows a spectrogram of the sound generated by voice conversion together with the voice before conversion (read laughter).

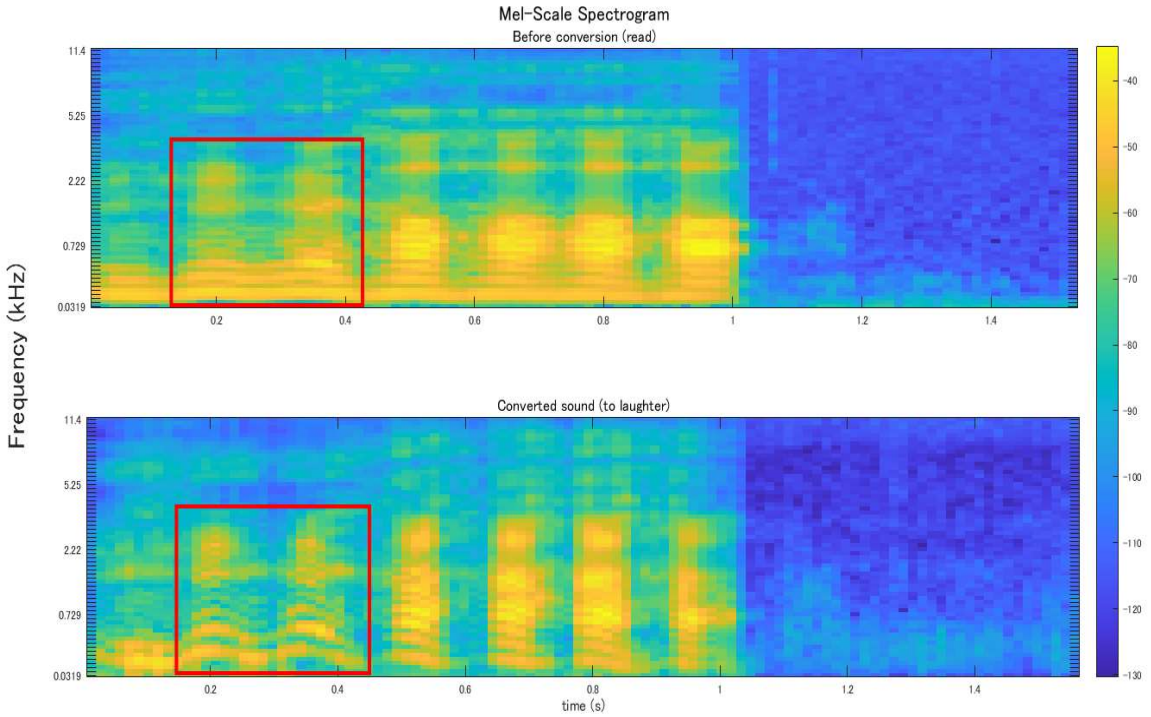


Figure 2 – Mel-scale spectrogram of laughter
(Read speech before conversion and converted laughter)

Note that inside the red box, you can see clearer formant bars, or clearer distinctions between peaks and valleys, in the converted voice. This may mean that the converted speech sound more "voiced" than the original read laughter. There is also a distinctive deviation in the formant frequencies.

3.2 Result of the subjective evaluation

Figure 3 shows the resulting MOS scores for each category and average scores (orange dots).

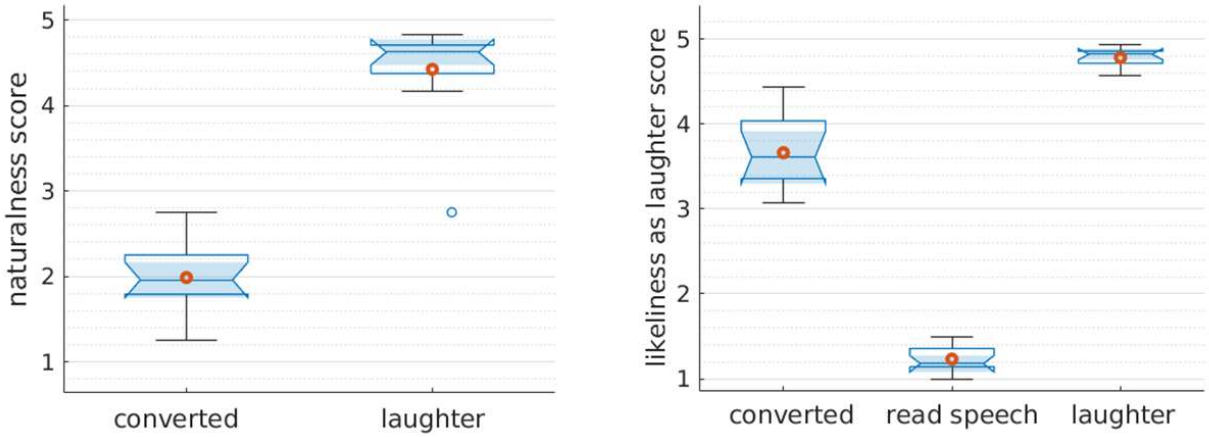


Figure 3 – The results of the subjective evaluation

The results in the left figure show that the converted speech was rated on average at about 1.9, i.e., unnatural. On the other hand, the average score for likeliness as laughter was 3.66, which is not as high as the original laughter, but much closer to natural laughter than the read-sentence (neutral) speech.

The reason why the naturalness evaluation of the converted voice was low (average of about 1.9) seems to be that the listeners could easily distinguish between a laughing voice and a converted sound. The converted sound contains noise compared to natural laughter and read-speech. The listeners seem to have noticed this noise and rated them as unnatural because of this, not necessarily from the quality of the synthesized laughter, which we were attempting to evaluate.

Considering these results together, it was surprising that even though the sound quality of the voice was evaluated as unnatural, it still was evaluated well in " laughter likeness". Overall, it may be said that the fluctuation of the frequency of the formants and the increase of the power difference seen in Figure 2 contribute to the feeling of "laughter likeness ".

4. CONCLUSIONS

In this study, we examined the use of deep learning to generate laughter from read-sentence speech through voice conversion. The result of the subjective evaluation was that even though the converted laughter was not perceived as natural (1.9 on a 5-point scale), it was still perceived as likely as laughter. This “laughter likeness” can be reproduced even with a small amount of data.

In the future, it will be necessary to generate more complex laughter, for example, “speech-laugh”. We also would like to compile a much larger dataset of laughter speech samples and categorize the types of laughter included in this dataset. A comprehensive evaluation method to measure the converted laughter quality is also needed.

ACKNOWLEDGEMENTS

Thank you very much for being selected as an ICA2022 Young Scientist Conference Attendance Grant and for helping me to attend the conference.

REFERENCES

1. H. Tanaka, N. Campbell, “Acoustic features of four types of laughter in natural conversational speech”, ICPhS 17, pp. 1958-1961, 2011.
2. C. Wang, Y. Yu, “CycleGAN-VC-GP: Improved CycleGAN-based Non-parallel Voice Conversion”, ICCT 20, pp. 1281-1284, 2020.
3. Morise M., Yokomori F., Ozawa K., “WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications,” IEICE Trans. Inf. Syst., E99-D (7), pp. 1877-1884, 2016.

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0309

Unsupervised singing voice conversion with star generative adversarial networks and autoencoder based speaker embedding

Gwantae Kim [(1)], Donghyeon Kim[(1)], Bokyeung Lee[(1)], Hanseok Ko[(1)]

[(1)]Department of Electrical Engineering, Korea University, South Korea

ABSTRACT

In this work, we propose a star generative adversarial network-based singing voice conversion with unparallel data. First, to embed the timbre and other important information of the target speaker to the embedding vector, we train the autoencoder and embedder networks with supervised training. Second, the generator and discriminators are trained by the unsupervised training method to achieve singing voice conversion. Especially, we propose a frequency-aware adaptive instance normalization to change formant. In the inference phase, the given source speech is converted to the target speech, which consists of the linguistic and melody of the source speaker and the timbre of the target speaker. With the several experiments using a public dataset, the proposed method successfully converts male to male, female to female, male to female, and female to male voice conversion with preserving melody and linguistic information.

Keywords: singing voice conversion, generative adversarial networks, unsupervised learning, speaker embedding, autoencoder

1 INTRODUCTION

Voice conversion and synthesis are already widely used in the speech research and industry field. Since the voice has characteristics of a human, it can make a more realistic artificial human or it helps to mimic real people. Moreover, speech-based human-computer interaction system must have own voice generation module. Current voice conversion system successfully control identity of the speakers. However, the voice also contains timbre, pitch and melody and the current voice conversion system does not focus on these parts. This is fine for normal speech, but voice conversion may not work properly if the pitch changes significantly. The Singing Voice Conversion(SVC) task aim to control not only identity but also other features of the voice. The goal of SVC is converting the timbre of a source singer to a target singer without changing the melody and linguistic information. The SVC is more challenging than conventional voice conversion because pitch and melody consistency must be guaranteed.

Since collecting parallel singing corpora is difficult, unsupervised learning-based approaches are proposed. Unsupervised Singing Voice Conversion(USVC)(16) constructs encoder-decoder networks based on WaveNet(17) and music translation(14) structure using waveform. Although USVC successfully converts the timbre of the singer, the melody and linguistic information, which do not want to change, are also converted. To overcome the restrictions of the USVC, Unsupervised cross-domain SVC(20) and PitchNet(2) are presented as a follow-up studies. Unsupervised cross-domain SVC proposed non-causal WaveNet-style model that trained with Generative Adversarial Network(GAN) loss and perceptual losses. The model used several perceptual features, such as wav2letter(22), CREPE(10), and loudness(13) feature. PitchNet has similar structure to USVC, but the pitch information extracted by Kaldi toolkit(21) is additionally used as ground-truth pitch. PitchNet found that the pitch information can preserve pitch and melody of the converted waveform.

Despite the success of the WaveNet-based models, they contain some problems. The models are hard to converge and outputs may contain noise. Moreover, since the WaveNet-based models are autoregressive models, they need a long inference time and are sensitive to outliers of the output samples.

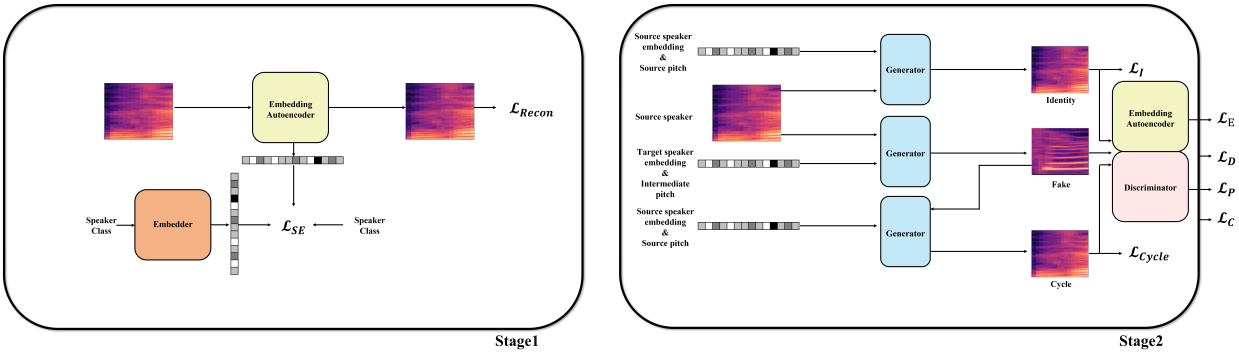


Figure 1. Proposed training methods. In the first stage, Embedder and Embedding Autoencoder are trained by supervised learning. In the second stage, Generator is trained by unsupervised learning with Discriminator and pre-trained Embedding Autoencoder.

Meanwhile, cycleGAN(9, 18, 26) and starGAN(1) successfully convert from source image style to target image style. Since audio signal can convert into image-like features by Fourier transform, similar approaches can be applied to audio signals. CycleGAN(7), starGAN(6, 8), and melGAN(12, 19) based voice conversion are trained with GAN loss. Especially, CycleGAN and starGAN can trained with unpaired training dataset. However, the results of the cycleGAN based unsupervised voice conversion methods contain some noise, and speaker information is hardly considered.

In this paper, we design a many-to-many SVC model based on the GAN, speaker autoencoder, feature attention, and AdaIN. We propose a convolution neural networks model that contains frequency axis AdaIN to inject speaker information and feature attention modules for refining features about speaker details and pitch information. We trained model with StarGAN with WasserteinGAN-Gradient Penalty(WGAN-GP) loss, embedding loss. We also propose the autoencoder-style speaker embedding model trained with deep metric learning. It can extract speaker information from both audio signal and one-hot speaker vector with joint mapping. The proposed model can convert the signal fast and converted speech has good sound quality.

To deliver the detailed description, the remainder of the paper is organized as follows. The proposed model are described in Section 2 The experimental process and results are presented in Section 3. Conclusions are drawn in Section 4.

2 PROPOSED METHOD

As shown in Fig. 1, the proposed method has two training stage and the model consists of embedder, embedding autoencoder, generator, and discriminator. In the first stage, embedder and embedding autoencoder are trained by supervised learning to find speaker embedding vector. In the second stage, generator and discriminator are trained with cycle generative adversarial networks. In the inference phase, input waveform and embedding vector of the target speaker pass through the generator and generate converted waveform.

Input features We used mel-spectrogram as input feature of the embedding autoencoder, generator, and discriminator. window length, hop length, nfft, the number of mel coefficients are 1024, 256, 1024, 80, respectively. This settings is same as Tacotron2(25) and WaveGlow(23). The pitch vector, which is used in transfer block, is extracted by CREPE(10) model. Since the pre-trained CREPE model cannot extract pitch from silence, we set the zero-vector as pitch when the input waveform is silent.

Vocoder WaveGlow(23) and WORLD(15) vocoder are used in the proposed model. The mel-spectrogram is converted into waveform using WaveGlow vocoder. We used pre-trained WaveGlow model that trained with LJSpeech dataset. The WORLD vocoder is used in post-processing stage.

Data augmentation We use phase-inversion and time-inversion data augmentation for generating audio clips. When a signal is played backward, the energy spectrum does not change(16). Therefore, training with time-inversion data augmentation does not hurt the sound quality and improve generalization. Moreover, the human

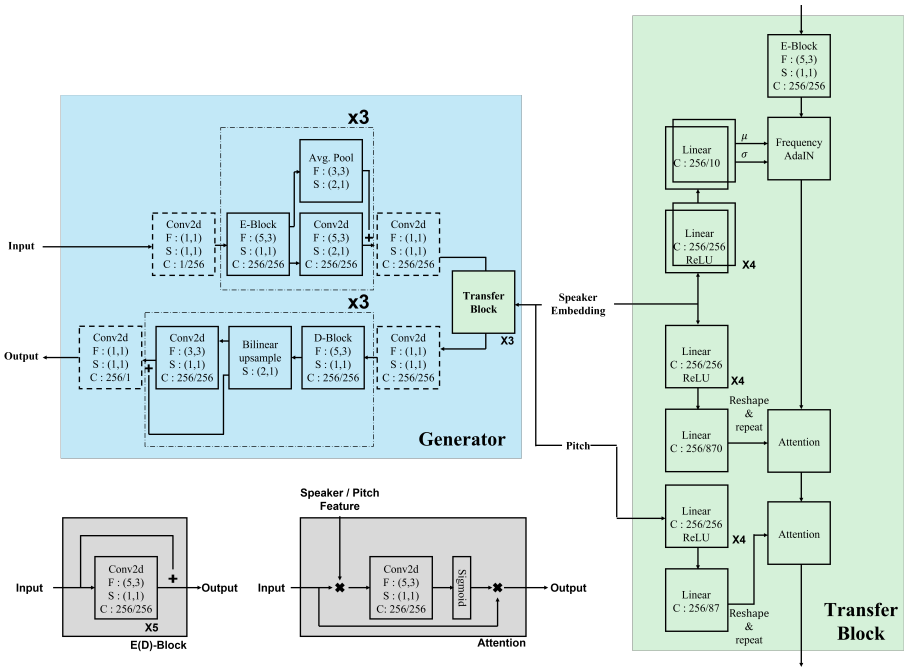


Figure 2. Proposed generator model architecture. F, S, C denote filter, stride, channel, respectively. The layers with dotted line do not have IN and activation function.

perception for monaural audio is not affected by the phase of the signal, hence one can shift the phase by 180 degrees without any effect on auditory perception(16).

Stage 1 In the stage 1, the embedding autoencoder and embedder are trained with mel-spectrogram and corresponding speaker label. The detailed model structures are shown in Fig. 3. Embedding autoencoder consists of encoder, decoder, and pooling-embedding layers. The encoder has 1x1 channel-expanding Convolution Neural Networks(CNN) layer, 3 residual CNN layers, and 1x1 CNN output layer. The statistical channel pooling and frequency pooling are applied to the output features of the encoder. Let the output features of the encoder X_{enc} has [B, C, F, T] dimension, then statistical channel/frequency pooling calculates mean and standard deviation of the X_{enc} along channel/frequency axis. The dimension of the channel and frequency pooling results are [B, 2C] and [B, 2F], respectively. These vectors are concatenated and pass through 3 fully-collected layers to get speaker embedding vector. The speaker embedding vectors are trained with triplet margin loss(24), which makes distance between same class goes minimum and distance between different class goes maximum. The loss function is defined as

$$L_{tri} = max(d(E_a, E_p) - d(E_a, E_n) + margin, 0) \tag{1}$$

where E_a, E_p, E_n denote anchor, positive, negative samples. d(·) is L1 distance and margin is 2.0, in this paper. The anchor, positive and negative samples are selected as all triplets in the mini-batch. The triplet margin loss is hard to converge, then one fully-collected layer is added to speaker embedding to calculate cross-entropy loss L_{cls}.

$$L_{cls} = -\sum_{c=1}^{N} y_c log(\hat{y}_c) \tag{2}$$

where c denotes class, N is the number of class, $\hat{y}$ is predicted label, and y is ground-truth label.

The decoder has 1x1 CNN layer, 3 residual upsample layers, and 1x1 output CNN layer. The reconstruction loss L_{rec} is calculated by L1 loss between input mel-spectrogram X_{in} and output mel-spectrogram X_{out}.

$$L_{rec} = ||X_{in} - X_{out}||_1 \tag{3}$$

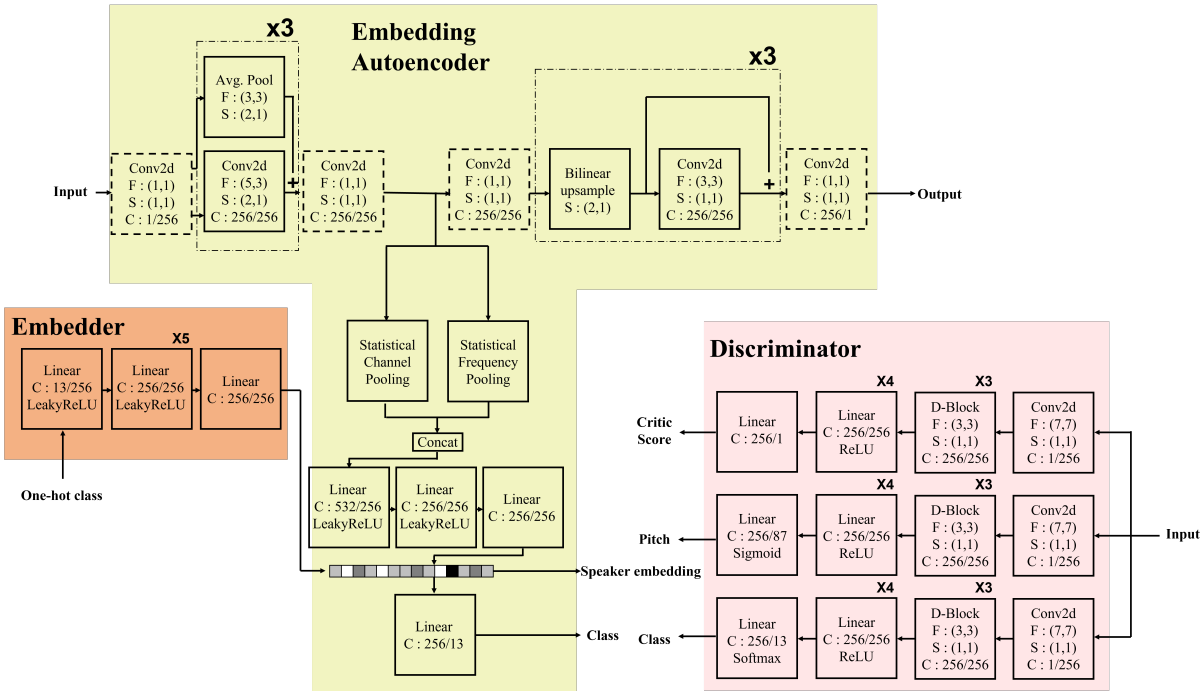


Figure 3. The discriminator and embedder model architecture. F, S, C denote filter, stride, channel, respectively. The layers with dotted line do not have IN and activation function.

The weights of embedding autoencoder are optimized by $L_{tri} + L_{cls} + L_{rec}$ with Adam optimizer(11), batch size 12 and learning rate 5e-5.

Embedder has 7 fully-connected layer with leaky-ReLU activation function. The output dimension of the embedding vector is same as speaker embedding vector, and calculate the L1 loss between embedding vectors from embedder and embedding autoencoder.

$$L_{emb} = ||E_{EA} - E_E||_1 \tag{4}$$

The weights of the embedder are optimized by L_{emb} with Adam optimizer, batch size 12 and learning rate 5e-5. If a speech sample of the speaker is given, the embedding autoencoder is used to acquire speaker embedding vector. With the embedding autoencoder, the embedding vector for the seen and unseen speaker can be generated. Moreover, the speaker embedding vector can be generated by the embedder with one-hot vector input for seen speaker. Therefore, we can generate a speaker embedding vector for seen and unseen speakers, whether and speech sample is given or not.

Stage 2 In the stage 2, the generator and discriminator are trained with input mel-spectrogram, pitch, and speaker label. The generator structure consists of encoder, transfer, and decoder block, as shown in Fig. 2 and Fig. 3. In the encoder block, one 1x1 CNN input block, three E-blocks with residual downsampling blocks, and last one 1x1 CNN block. Note that downsampling blocks only reduce frequency domain. In the transfer blocks, input features that pass through one E-block are transformed by frequency-domain Adaptive Instance Normalization(AdaIN)[**?**] and Attention blocks. The speaker embedding vector is transformed by two 5-stacked fully-connected layers to find style mean S_μ and style standard deviation S_σ. Next, the frequency-domain AdaIN is performed with content and style.

$$C_{out} = (C_{in} - C_\mu)/C_\sigma * S_\sigma + S_\mu \tag{5}$$

where C is style feature. Note that the mean and standard deviation is calculated along frequency axis, not channel axis. The attention layers, which is described in Fig. 2, also refine features from source domain to

target domain with speaker embedding vector and pitch vector. The decoder block makes the features from downsampled size to original size with residual CNN layers and upsampling blocks, as shown in Fig. 2. The discriminator consists of three individual discriminators, named critic discriminator, pitch discriminator, and speaker discriminator. The detailed structures are described in Fig. 3.

The losses of the stage 2 are based on cycleGAN loss and WGAN-GP loss(4). Let source mel-spectrogram is X_{src} and target mel-spectrogram is X_{tgt}, the identity loss and cycle consistency loss are calculated by

$$L_I = ||X_{src} - G(X_{src}, p_{src}, E_{src}||_1 \tag{6}$$

and

$$L_{cyc} = ||X_{src} - G(G(X_{src}, p_{tgt}, E_{tgt}), p_{src}, E_{src})||_1 \tag{7}$$

where p is pitch vector and E is speaker embedding vector. Next, critic discriminator loss, pitch discriminator loss, and speaker discriminator loss are computed by

$$L_{critic} = -D_{critic}(G(X_{src}, p_{tgt}, E_{tgt})) \tag{8}$$

$$\begin{aligned} L_{pitch} = &||p_{tgt} - D_{pitch}(G(X_{src}, p_{tgt}, E_{tgt}))||_1 \\ &+ ||p_{src} - D_{pitch}(G(X_{src}, p_{src}, E_{src}))||_1 \\ &+ ||p_{tgt} - D_{pitch}(G(G(X_{src}, p_{tgt}, E_{tgt})), p_{src}, E_{src})||_1 \end{aligned} \tag{9}$$

$$L_{speaker} = CE(D_{speaker}(G(X_{src}, p_{tgt}, E_{tgt})), y_{tgt}) \tag{10}$$

where y is label, CE(·) denotes cross entropy loss. To emphasize the reconstruction and speaker characteristics, the embed loss is added.

$$L_{embed} = ||E_{tgt} - EA(G(X_{src}, p_{tgt}, E_{tgt})||_1 \tag{11}$$

where EA(·) is embedding autoencoder. The critic discriminator assesses whether the sample is real or fake with the adversarial training between generator and discriminator. The pitch and class discriminator predict pitch vector and speaker class with the collaborative training between generator and discriminator, respectively.

The final generator loss function is

$$L_G = L_I + L_{cyc} + 0.5 * L_{critic} + L_{pitch} + L_{speaker} + 5.0 * L_{embed} \tag{12}$$

and the final discriminator loss is

$$L_D = L_{realD} + L_{fakeD} + L_{pitchD} + L_{speakerD} + GP \tag{13}$$

where

$$L_{realD} = -D(X_{tgt}) \tag{14}$$

$$L_{fakeD} = D(G(X_{src}, p_{tgt}, E_{tgt})) \tag{15}$$

$$L_{pitch} = ||p_{src} - D_{pitch}(X_{src})||_1 + ||p_{tgt} - D_{pitch}(X_{tgt})||_1 \tag{16}$$

$$L_{speaker} = CE(D_{speaker}(G(X_{src}, p_{tgt}, E_{tgt})), y_{tgt}) + CE(D_{speaker}(X_{tgt}), y_{tgt}) \tag{17}$$

and GP denotes gradient panelty with $\lambda = 10$.

The weights of the generator and discriminator are optimized by L_G and L_D with Adam optimizer, batch size 6 and learning rate 1e-5.

Post processing Since the output of the generator is mel-spectrogram, it must be changed into waveform. We manually used the WaveGlow Vocoder, which is trained with LJSpeech dataset.

Despite the generator model receives pitch information, pitch conversion in the female-to-male or male-to-female cases is not enough to change voice tone. To solve the problem, we extracted log-f0 using WORLD[**?**] analyzer, and calculating average mean and the average standard deviation for each speakers in the training set. After that, the f0 is converted by

$$f0_{tgt} = (f0_{src} - \mu_{src})/\sigma_{src} * \sigma_{avg_tgt} + \mu_{avg_tgt}. \tag{18}$$

Note that μ_{src} and σ_{src} is mean and standard deviation of the $f0_{src}$, but μ_{avg_tgt} and σ_{avg_tgt} are average of the training samples about target speaker. Finally the waveform is reconstructed by WORLD vocoder with spectrum features, aperiodic features, and converted f0.

3 EXPERIMENTS

We perform some experiments to evaluate the proposed method with singing voice dataset, named NUS-48E[**?**]. The dataset includes six male singers and six female singers. Each singer has 4 songs, and each song is almost 4 minutes. We select 3 songs per singers as training fold, and 1 song per singers as test fold. We train our models for 73000 iterations, with a batch size of 6 one second long audio segments. All audio was down-samapled to 22050Hz with a single channel.

Evaluation metrics We used Mean Opinion Score(MOS) to measure human perceptual performance. We selected three subjects: sound quality, human likeness, and speaker similarity. The range of the score is [1, 5] and higher is better. We also used two objective measures: NORESQA(27) and speaker embedding similarity. Since our dataset does not have source-target paired data samples, we cannot use speech quality measurements that need a paired clean reference, such as PESQ and SSNR. Instead, we used NORESQA, which is a framework for speech quality assessment using non-matching references, to measure speech quality. To calculate NORESQA score, we set the clean reference as the source speech. Next, the speaker embeddings are extracted by the trained speaker embedding model. In this paper, we used our embedding autoencoder as a speaker embedding model. The source speech and converted speech are split into short lengths and transformed into speaker embedding vectors. Then the average cosine similarity is calculated between the embedding vectors of the source speech and converted speech.

Results Subjective evaluation results on NUS-48E dataset are presented in Table 1 and Fig. 4. When the tasks are F→F and M→M, our approach outperforms the comparison methods in sound quality, human likeness, and speaker similarity. However, when the tasks are F→M and M→F, sound quality and human likeness go worse. Therefore, the proposed model has good reconstruction performance and conversion performance of the same gender. However, conversion performance of the different gender is worse than the state-of-the-art models. We explored the reason and found that f0 conversion with WORLD vocoder may hurts the sound quality.

Figure 4 describes the examples of the ground truth and converted spectrogram. (c), (d), (e), (f) are converted from the ground truth of the source speaker (a) to the target speaker (b). The conversion with USVC is clean, but almost same as source speaker, and the conversion with PitchNet is too noisy. The conversion with proposed method is clean enough and harmonics of the voice are changed. The reconstruction with proposed method is clean, and harmonics of the voice are almost same as ground truth.

Objective evaluation results on NUS-48E dataset are presented in Table 2.

4 CONCLUSION

We presented a novel deep learning-based method for unsupervised singing voice conversion using cycleGAN and speaker embedding autoencoder. We proposed speaker embedding autoencoder and embedder to make wealthy speaker information. We also proposed cycleGAN, WGAN-gp, and multi-domain discriminator style losses. Through a series of experiments, the proposed method successfully showed that it converted speaker in same-gender domain and reconstructed source domain. The proposed method can be extended to generate the fake human voice for the avatar-based interactive agents and service robots. In our future work, we plan to study a model that can converted speech in different-gender domain.

ACKNOWLEDGEMENTS

This work was supported by Korea Environment Industry & Technology Institute(KEITI) through Exotic Invasive Species Management Program, funded by Korea Ministry of Environment(MOE)(2021002280004).

REFERENCES

[1] Y. Choi, J.Choo, Stargan:Unified generative adversarial networks for multi-domain image-to-image translation, In Proceedings of the IEEE conference on computer vision and pattern recognition, page 8789-8797, 2018.

[2] C. Deng, D. Yu, Pitchnet: Unsupervised singing voice conversion with pitch adversarial network. In

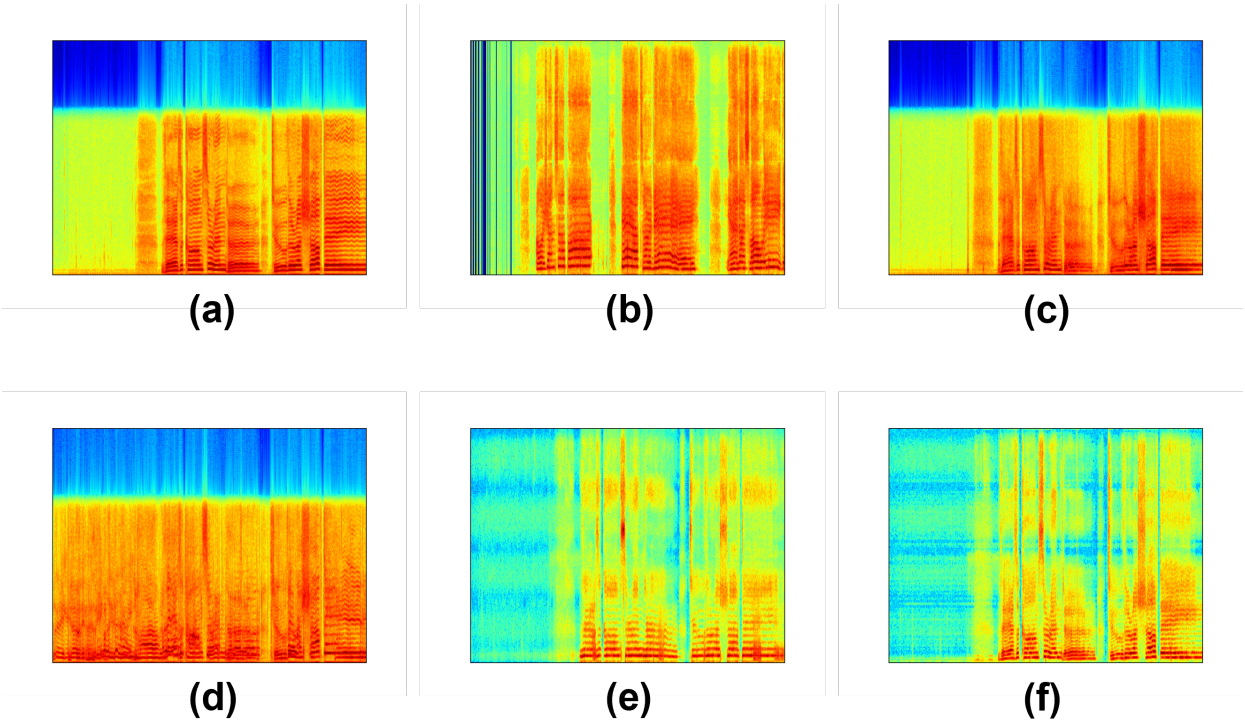


Figure 4. Spectrograms of the ground truth and converted waveforms. (a) ground truth of the source speaker. (b) ground truth of the target speaker. (c) USVC. (d) PitchNet (e) Proposed(convert). (f) proposed(recon).

Table 1. Mean opinion scores of the baselines and proposed method

methods	task	Sound quality	Human likeness	Speaker similarity
ground truth	F→ M	4.85	4.71	-
USVC	F→ M	1.71	**2.14**	1.00
PitchNet	F→ M	1.14	1.00	1.71
proposed(convert)	F→ M	**2.14**	1.71	**1.85**
proposed(recon)	F→ M	3.57	4.14	4.00
ground truth	M→ F	4.43	4.57	-
USVC	M→ F	**2.29**	**2.85**	1.43
PitchNet	M→ F	1.43	1.14	1.42
proposed(convert)	M→ F	1.86	1.85	**1.71**
proposed(recon)	M→ F	3.86	4.14	4.43
ground truth	F→ F	3.71	4.14	-
USVC	F→ F	1.86	2.29	2.86
PitchNet	F→ F	1.57	1.43	2.29
proposed(convert)	F→ F	**3.14**	**3.43**	**3.57**
proposed(recon)	F→ F	3.86	3.85	4.71
ground truth	M→ M	4.14	4.43	-
USVC	M→ M	2.00	2.57	2.43
PitchNet	M→ M	1.29	1.29	1.43
proposed(convert)	M→ M	**2.86**	**3.00**	**2.57**
proposed(recon)	M→ M	4.00	4.29	4.43

ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7749–7753. IEEE, 2020.

[3] Z. Duan,Y. Wang. The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pages 1–9. IEEE, 2013.

[4] I. Gulrajani, A. C. Courville. Improved training of wasserstein gans. Advances in neural information processing systems, 30, 2017.

[5] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE international conference on computer vision, pages 1501–1510, 2017.

[6] H. Kameoka, N. Hojo. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In 2018 IEEE Spoken Language Technology Workshop (SLT), pages 266–273. IEEE, 2018.

Table 2. Objective evaluation of the baselines and proposed method

methods	task	NORESQA	Cosine similarity(source)	Cosine similarity(target)
ground truth	F→ M	5.47	1.0	-0.29
USVC	F→ M	8,49	0.18	0.17
PitchNet	F→ M	11.15	-0.09	0.14
proposed(convert)	F→ M	14.13	-0.15	0.45
proposed(recon)	F→ M	11.01	-0.27	0.18
ground truth	M→ F	6.31	1.0	0.45
USVC	M→ F	8.65	0.45	0.11
PitchNet	M→ F	9.65	0.68	0.38
proposed(convert)	M→ F	13.54	0.23	0.09
proposed(recon)	M→ F	9.94	0.76	0.60
ground truth	F→ F	5.86	1.0	0.76
USVC	F→ F	9.64	0.51	0.42
PitchNet	F→ F	10.98	0.56	0.53
proposed(convert)	F→ F	11.53	0.68	0.65
proposed(recon)	F→ F	7.09	0.90	0.81
ground truth	M→ M	6.12	1.0	0.24
USVC	M→ M	11.35	-0.23	-0.15
PitchNet	M→ M	11.12	-0.41	-0.30
proposed(convert)	M→ M	10.34	0.24	0.50
proposed(recon)	M→ M	8.22	0.96	0.28

[7] T. Kaneko, N. Hojo. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6820–6824. IEEE, 2019.

[8] T. Kaneko, N. Hojo. Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion. arXiv preprint arXiv:1907.12279, 2019.

[9] G. Kim, H. Ko. Unsupervised real-world super resolution with cycle generative adversarial network and domain discriminator. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 456–457, 2020.

[10] J. W. Kim, J. P. Bello. Crepe: A convolutional representation for pitch estimation. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 161–165. IEEE, 2018.

[11] D. P. Kingma, J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[12] K. Kumar, A. C. Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. Advances in neural information processing systems, 32, 2019.

[13] B. C. Moore, T. Baer. A model for the prediction of thresholds, loudness, and partial loudness. Journal of the Audio Engineering Society, 45(4):224–240, 1997.

[14] N. Mor, Y. Taigman. A universal music translation network. arXiv preprint arXiv:1805.07848, 2018.

[15] M. Morise, K. Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE TRANSACTIONS on Information and Systems, 99(7):1877–1884, 2016.

[16] E. Nachmani, L. Wolf. Unsupervised singing voice conversion. arXiv preprint arXiv:1904.06590, 2019.

[17] A. v. d. Oord, K. Kavukcuoglu. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.

[18] J. Park, H. Ko. Adaptive weighted multi-discriminator cyclegan for underwater image enhancement. Journal of Marine Science and Engineering, 7(7):200, 2019.

[19] M. Pasini. Melgan-vc: Voice conversion and audio style transfer on arbitrarily long samples using spectrograms. arXiv preprint arXiv:1910.03713, 2019.

[20] A. Polyak, Y. Taigman. Unsupervised cross-domain singing voice conversion. arXiv preprint arXiv:2008.02830, 2020.

[21] D. Povey, P. Schwarz, et al. The kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding, number CONF. IEEE Signal Processing Society, 2011.

[22] V. Pratap, R. Collobert. Wav2letter++: A fast open-source speech recognition system. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6460–6464. IEEE, 2019.

[23] R. Prenger, B. Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3617–3621. IEEE, 2019.

[24] F. Schroff, J. Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 815–823, 2015.

[25] J. Shen, R. Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 4779–4783. IEEE, 2018.

[26] J.-Y. Zhu, A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 2223–2232, 2017.

[27] Manocha, P., Kumar, A. NORESQA: A framework for speech quality assessment using non-matching references. Advances in Neural Information Processing Systems, 34, 22363-22378, 2021.

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0801

Noisy-to-Noisy Voice Conversion with Pre-training Strategy

Chao Xie[(1)], Tomoki Toda[(2)]

[(1)]Graduate School of Information Science, Nagoya University, Japan, xie.chao@g.sp.m.is.nagoya-u.ac.jp

[(2)]Information Technology Center, Nagoya University, Japan, tomoki@icts.nagoya-u.ac.jp

ABSTRACT

Conventional voice conversion (VC), which transforms the speaker's identity without changing the linguistic content, has experienced significant progress with the advent of deep learning. However, extending VC to real-world scenarios faces several interferences, such as background sounds, causing adverse effects on VC performance. On the other hand, in some scenarios, such as VC in movie/video and VC in music, the background sounds are informative and should be retained. In our previous work, we have proposed a noisy-to-noisy (N2N) VC framework to do the conversion while retaining the background sounds. The framework consists of a denoising module to separate a noisy speech signal into speech and noise signals and a VC module to process the conversion. To alleviate adverse effects caused by the distortion introduced by the denoising module, we use the separated noise signal as the condition in the VC module to model the non-distortion noisy speech directly. We further improve our N2N VC framework by implementing a pre-training strategy using existing noise clips and clean speech data. The experimental results show that the pre-trained framework yields significant improvements against noisy environments.

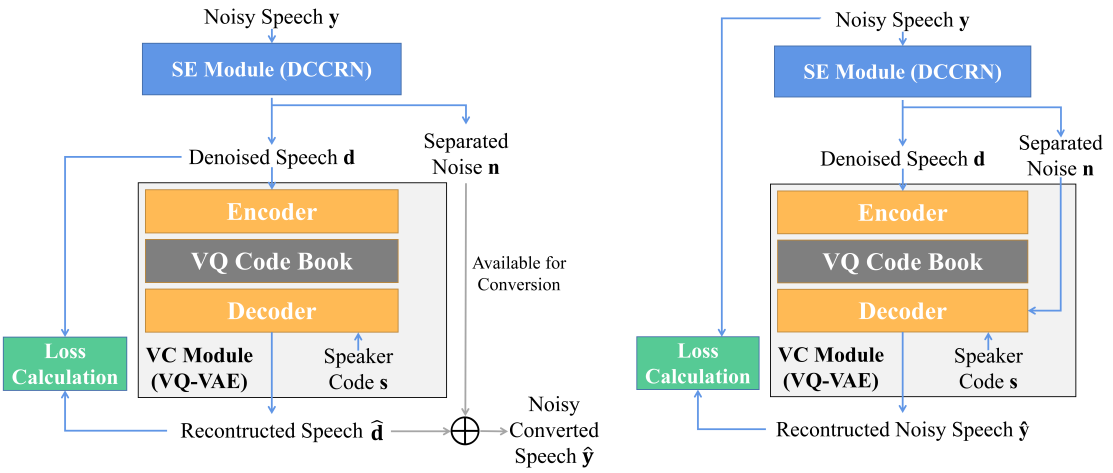
Keywords: Voice conversion (VC), Noisy-to-noisy VC, Pre-training

1 INTRODUCTION

Voice conversion (VC) is a technique for converting the vocal timbre of a speech from the source speaker to the target speaker without changing its linguistic contents. With the fast development of deep learning in recent years, neural network-based VC methods achieve significant improvements in terms of speech naturalness and speaker similarity, as demonstrated in the latest Voice Conversion Challenge (VCC) 2020 [12].

Recent trends have also emerged to introduce VC to real-world environments. Distinct from the experimental environments where high-quality data have been prepared beforehand for training and testing, the source/target speeches often have certain interference, the most common of which is background noise entangling with linguistic content and speaker identity. On the other hand, the background sounds are informative to be retained in some scenarios, such as dubbing [1] and audio data augmentation [7]. However, most related studies focus on noise-robust VC, where the background noise is suppressed, and only the clean converted speech is generated.

We have proposed a noisy-to-noisy (N2N) VC framework in [9]. The first "noisy" means all the source/target training data are noisy. The second "noisy" means the background sound in the converted sample is controllable. Then in [10] we further improved the framework by using noise as the condition for VC module to model the noisy speech. In this paper, the pre-training strategy, which has benefited many natural language processing (NLP) and VC studies, is introduced into the noise-conditioned VC module to improve the robustness in the noisy environment. Since the noise-conditioned VC method is novel, we first evaluate whether it is affected by the introduced noise condition and competitive with the original VC module when generating clean converted samples.



(a) Baseline framework (b) Noise-conditioned framework

Figure 1. Overview of N2N VC frameworks.

2 N2N VC FRAMEWORK

Figure 1(a) illustrates the baseline of the N2N VC framework. It is comprised of off-the-shelf denoising and VC modules, which are implemented as DCCRN [2] and VQ-VAE-based non-parallel VC method [8] as a case study. The denoising module is pre-trained on a publicly available dataset and separates the speech signal and noise signal in the time domain. The VC module is trained with the denoised speech. During inference, only the denoised speech is converted, and the separated noise is optional to be superimposed based on specific scenarios.

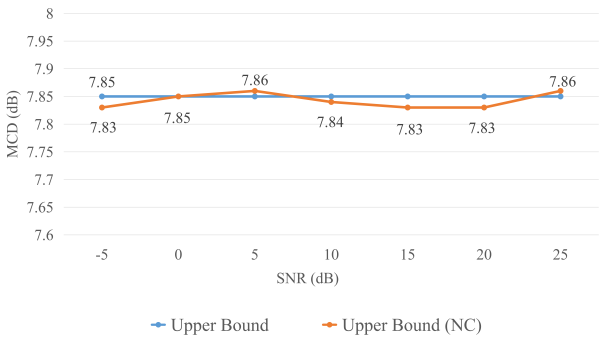
However, even the state-of-art SE method would introduce inescapable distortion when suppressing the background noise. The additional distortion is inconspicuous to perceptual listening, whereas it causes fatal deterioration to the quality of the converted speech in terms of naturalness and similarity. Unfortunately, the VC module is trained to reconstruct the distorted speech data, which further aggravates the degradation of the VC performance. Another drawback is that generating the noisy converted speech is redundant: the converted speech is generated first, then the separated background noise is superimposed.

Therefore, the framework is improved in [10] by leveraging the noise signal as the condition in the VC module to model the noisy speech with non-distortion, as shown in figure 1 (b). In the training stage, the VC module receives denoised speech as input and separated noise as a condition for the decoder to reconstruct the noisy speech in an auto-regressive manner, the loss of which is calculated with the original noisy speech. During the conversion stage, using noise signal to the condition results in noisy converted speech generation, while replacing the noise signal with zero sequences leads to clean converted speech. The decoder of the VC module can be described as the joint probability distribution:

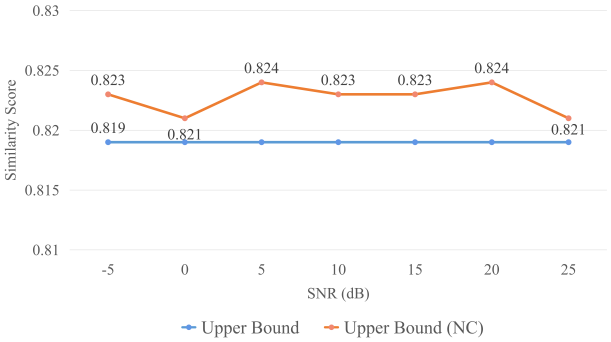
$$p(\mathbf{y} \mid \mathbf{n}, \mathbf{s}, \mathbf{z}) = \prod_{t=1}^{T} p(y_t \mid y_1, \ldots, y_{t-1}, n_1, \ldots, n_t, \mathbf{s}, \mathbf{z}), \tag{1}$$

where $\mathbf{y}$, $\mathbf{n}$, $\mathbf{s}$, $\mathbf{z}$ denote noisy speech, separated noise, speaker code, and quantized discrete vectors from the denoised speech, respectively.

To further improve the framework's robustness in noisy environments, pre-training is conducted on the noise-conditioned VC module. Specifically, the VC module receives the clean speech as input and original noise clips as the noise condition to reconstruct the noisy speech mixed by the two formers. Corpus from each speaker is corrupted with various noise categories and SNR levels so that the noisy conditions are less discriminative about speakers. When fine-tuning, the VC module inherits the pre-trained weights except for the speaker embedding layer, which is trained from scratch.



(a) MCD, lower is better



(b) Similarity score, higher is better

Figure 2. Evaluation results of Upper Bound and Upper Bound (NC).

3 EXPERIMENTAL EVALUATION

3.1 Experimental conditions

3.1.1 Datasets

The SE module was pre-trained on the dataset of Deep Noise Suppression (DNS) Challenge 2020 [6] The clean speeches and noise clips were mixed at uniformly sampled SNR levels between 0 to 20 dB. For pre-training the VC module, VCTK [11] was used as the clean corpus, and the evaluation set of DNS was used as the noise dataset for the noise-conditioned one. Clean corpus and the noise clips were mixed at randomly selected SNR levels within 0, 5, 10, and 15 dB.

VCC2018 dataset [4] was employed as the clean corpus for VC training and testing. ESC-50 [5] was used for conducting noisy conditions, where speech from one speaker was mixed with 41 subclasses at uniformly sampled SNR levels from 0, 5, 10, 15, and 20 dB. The rest 9 subclasses were for the testing set. Several noisy testing sets were parallelly constructed with a single SNR level within -5, 0, 5, 10, 15, 20, and 25 dB.

3.1.2 Models to be evaluated

Two kinds of the upper bound of the proposed framework are involved: the VQ-VAE trained on clean corpus and the noise-conditioned one using clean corpus and original noise to estimate noisy speech. They are denoted as **Upper Bound** and **Upper Bound (NC)**, respectively. The baseline of the framework constructing denoised speech is denoted as **Baseline**, and the noise-conditioned method using denoised speech and separated noise is named **N2N-VC**.

3.1.3 Evaluation methods

Mel cepstral distortion (MCD) [3] was employed to measure the speech quality. Besides, the word error rate (WER) was used to measure the quality of the linguistic content and calculated by a publicly available ASR model[1]. As for similarity, an open-source speaker verification method[2] was utilized to compute the score by comparing the converted sample with its target reference. Since these methods prefer clean speech data, all the methods generate the clean converted speech.

3.2 Evaluation Results

Figure 2 shows the results of MCD and Similarity score between two upper bounds. **Upper Bound** is noise-irrelevant, therefore it gets stable MCD of 7.85 and Similarity score of 0.819, while **Upper Bound(NC)** is noise-relevant. It is obvious that **Upper Bound(NC)** achieves comparable performance with **Upper Bound** under all SNR levels, showing that introducing noise into the VC module does not affect the performance of the converted speech generation in terms of quality and similarity.

[1]https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self.
[2]https://github.com/resemble-ai/Resemblyzer.

Table 1. Evaluation results of models w/ and w/o pre-training strategy at SNR levels 5 and 15 dB.

Models	Status	MCD (dB) ↓		WER (%) ↓		Similarity ↑	
		5 dB	15 dB	5 dB	15 dB	5 dB	15 dB
Noisy Testing Set	Denoised	-	-	6.41	3.81	-	-
Upper Bound (NC)	**w/ pre-trained**	7.86	**7.83**	**9.55**	**10.56**	**0.824**	0.823
	w/o pre-trained	7.84	7.84	14.57	14.93	0.821	0.823
Baseline	**w/ pre-trained**	**8.76**	**8.39**	**32.92**	**16.02**	**0.772**	**0.798**
	w/o pre-trained	8.89	8.64	56.62	40.59	0.757	0.766
N2N-VC	**w/ pre-trained**	**8.58**	**8.33**	**29.22**	**17.01**	**0.786**	**0.798**
	w/o pre-trained	8.62	8.38	39.27	27.62	0.777	0.786

Table 1 shows the evaluation results of the model with/without pre-training strategies. In general, the pre-trained methods show better performance, especially in terms of WER, which prove the effectiveness of pre-training strategies in improving the robustness of the N2N VC framework against noisy environments. **Upper Bound (NC)** ranks first place, the WER of which is improved by an average of 32.08% with the pre-trained model. **N2N-VC** outperforms **Baseline** in all metrics significantly without pre-training strategies. Still, **N2N-VC** with pre-trained model outperforms **Baseline** in terms of MCD, WER at 5 dB and similarity score at 5 dB. While using pre-training shortens their gaps, especially at SNR 15 dB, where **N2N-VC** and **Baseline** achieve the same similarity score of 0.798 and comparable MCD scores of 8.33 and 8.39, respectively, and **Baseline** even achieved lower WER of 16.02%, compared to **N2N-VC** of 17.01%.

4 CONCLUSION

In this paper, we utilize the pre-training model to improve its performance and robustness in noisy-to-noisy environments. The noise-conditioned VC model is evaluated first, and the results prove that introducing noise as a condition to the VC model does not affect conversion performance. Then our N2N VC framework with the pre-trained noise-conditioned VC model is evaluated. Results show that using pre-training strategies will achieve better performance, which is significant in WER. In future work, we will further perfect our N2N VC framework.

ACKNOWLEDGEMENTS

This work was partly supported by JST CREST Grant Number JPMJCR19A3, Japan.

REFERENCES

[1] W. Gan, B. Wen, Y. Yan, H. Chen, Z. Wang, H. Du, L. Xie, K. Guo, and H. Li. Iqdubbing: Prosody modeling based on discrete self-supervised speech representation for expressive voice conversion. *arXiv preprint arXiv:2201.00269*, 2022.

[2] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie. DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement. In *Proc. Interspeech 2020*, pages 2472–2476, 2020.

[3] R. Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128. IEEE, 1993.

[4] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. In *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, pages 195–202, 2018.

[5] K. J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.

[6] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke. The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results. In *Proc. Interspeech 2020*, pages 2492–2496, 2020.

[7] S. Shahnawazuddin, N. Adiga, K. Kumar, A. Poddar, and W. Ahmad. Voice conversion based data augmentation to improve children's speech recognition in limited data scenario. In *Interspeech*, pages 4382–4386, 2020.

[8] B. van Niekerk, L. Nortje, and H. Kamper. Vector-Quantized Neural Networks for Acoustic Unit Discovery in the ZeroSpeech 2020 Challenge. In *Proc. Interspeech 2020*, pages 4836–4840, 2020.

[9] C. Xie, Y.-C. Wu, P. L. Tobing, W.-C. Huang, and T. Toda. Noisy-to-noisy voice conversion framework with denoising model. In *Proc. 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Preprint: https://arxiv.org/pdf/2109.10608.pdf, 2021.

[10] C. Xie, Y.-C. Wu, P. L. Tobing, W.-C. Huang, and T. Toda. Direct noisy speech modeling for noisy-to-noisy voice conversion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6787–6791. IEEE, 2022.

[11] J. Yamagishi, C. Veaux, K. MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). 2019.

[12] Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z.-H. Ling, and T. Toda. Voice Conversion Challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pages 80–98, 2020.

PROCEEDINGS of the
24th International Congress on Acoustics
October 24 to 28, 2022 in Gyeongju, Korea

ABS-0921

Voiced/Unvoiced Classification on Expressive Speech Synthesis using Instantaneous Frequency Amplitude Spectrum to Improve Fundamental Frequency Estimation

Elok ANGGRAYNI[1]; Aprianto DWI PRASETYO[2]; Dhany ARIFIANTO[3]; Dipta NUSBANGGA HAIKAL AHMADA[4]

Department of Engineering Physics, Institut Teknologi Sepuluh Nopember, Indonesia

ABSTRACT

This paper presents the applicated frequency fundamental (F_0) estimation method on expressive speech synthesis based on instantaneous frequency amplitude spectrum (IFAS). F_0 is the quantity that is being estimated by virtually all pitch trackers and defined as harmonicity measure of the speech signal. Harmonicity measure is used as selector that selects the optimal frequency band on which the IF amplitude spectrum is evaluated. We describe a new technique for voice and unvoiced classification using IFAS-based F_0 with Gaussian window that can reduce discontinuity on the representation of clearly the harmonic structure. The performance of the proposed method is also compared against robust algorithm for pitch tracking (RAPT) to demonstrate the effectiveness of the proposed method. We do some variation in the number of sentences and the type of sentences which used in the training part. We use expressive speech corpus based on the phonetically balanced which contains 4 male and 4 female speakers. The result shown that the synthetic speech based on IFAS is higher than RAPT in term of naturalness. It is seen that the proposed technique can obtain better performance than the conventional method.

Keywords: Expressive Speech Synthesis, Instantaneous Frequency, Voiced/Unvoiced Classification

1. INTRODUCTION

Speech is the most effective communication for human beings to interact with each other. In parallel to speech processing development, people want to change the ways of interaction between humans and computers. A speech synthesis technique has recently been developed. The unit-selection synthesis is a speech synthesis technique that uses the database. This technique will automatically select the sub-word unit from the database given [1]. This technique can produce synthesized speech which similar to the original speech from the database. However, this technique requires a lot of databases to obtain comprehensive data coverage to build the models. So, it makes this technique requires a huge computing load and lacks the flexibility to be modified. In 1999, Yoshimura, et al., explain the method to model the spectral parameter, excitation parameter, and duration at once [2]. Then they sparked a speech synthesis technique-based statistical process known as statistical parametric speech synthesis that then began to grow today [1][3][4]. This technique uses Hidden Markov Model (HMM) to model the probability distribution of speech and linguistic features, it is called the HMM-based speech synthesis system (HTS). The formation of statistical models makes HTS has an advantage in flexibility to modify the acoustic models. Some of the advantages that can transform character voices, speaking styles, speaking adaptation, and supports for multilingual speech synthesis. In recent years, Deep Neural Networks (DNN) have emerged as an effective way to improve the performance of statistical parametric speech synthesis systems [5]. H. Zen et al. used DNN instead of the decision tree to model the relationship between contextual and acoustic features [6]. The DNN-based acoustic model can replace the decision-tree clustering model. Experiments show that the DNN-

[1] anggrayni.elok@gmail.com
[2] prasetyo.18023@mhs.its.ac.id
[3] dhany@ep.its.ac.id
[4] nusbanggadipta@gmail.com

based system performs better than the HMM-based speech synthesis system with the same number of experiments.

Generally, acoustic feature extraction is the main phase in a speech synthesis system. Every speech embedded the acoustic information and different individual characteristics in utterance. Extraction of the essential features from speech signals leads to achieving high performance in the synthesis process. Thus, research on extracting the most important and accurate set of features is the ongoing challenge to be figured out in the speech synthesis area. Fundamental frequency (F_0), as well as voiced/unvoiced information, is one of the most important features. Therefore, accurate and reliable fundamental frequency estimation and voiced/unvoiced classification are required, for example, to maintain the quality of synthesized speech. One of the oldest and most well-known techniques for F_0 estimation is based on the cepstrum [7]. The purpose of voiced/unvoiced classification is to classify the speech signal into voiced and unvoiced segments. F_0 is the quantity that is being estimated by virtually all pitch trackers and is defined as the harmonicity measure of the speech signal. Harmonicity measure is used as a selector that selects the optimal frequency band on which the IF amplitude spectrum is evaluated.

In this paper, we proposed the applicated frequency fundamental (F_0) estimation method on expressive speech synthesis based on instantaneous frequency amplitude spectrum (IFAS). Recently, time-frequency analysis has received considerable attention as a method to reveal time-varying speech parameters instantaneously. The instantaneous frequency is derived from the windowed Fourier transform of a signal as a function of time and frequency. The instantaneous frequency amplitude spectrum (IFAS) provides a better harmonic structure representation of speech signals than the Short-Time Fourier Transform (STFT) amplitude spectrum. We describe a new technique for voice and unvoiced classification using IFAS-based F_0 with a Gaussian window that can reduce discontinuity in the representation of clearly the harmonic structure. The performance of the proposed method is also compared against a robust algorithm for pitch tracking (RAPT) to demonstrate the effectiveness of the proposed method. We do some variations in the number of sentences and the type of sentences which used in the training part. We use expressive speech corpus based on the phonetically balanced which contains 4 male and 4 female speakers [8]. Besides that, we also try to compare the method to build the speech synthesis system, which is using the HMM-based speech synthesis system and HMM-DNN. Then we compare the speech quality of the synthesized speech using subjective and objective measurements.

2. EXPRESSIVE SPEECH SYNTHESIS SYSTEM

2.1 Development of Expressive Indonesian Speech Corpus

Language is an expression of the human mind and feeling which uses sound as its tool [9]. Every country has a different language with its own characteristic. Indonesian is the national language of Indonesia and is rooted in the Malay language. Besides Indonesian as the main language, most Indonesians are fluent in their own ethnic language according to the location of their tribe. Linguistic studies of Indonesian are divided into some levels, i.e., phonology, morphology, syntax, and lexicon [9]. The phoneme is an important role in natural language processing. Indonesian has 32 phonemes and contains six vocal phonemes, three diphthong phonemes, and 23 consonant phonemes. The Indonesian expressive speech database is the dataset of Indonesian language characteristics in accordance with Indonesian phonology and based on phonetically balanced. It consists of phoneme, speech, and transcription. The database contains 655 sentences with three expressive styles [8]. Three expressive styles are happy, sad, and angry. The sentence sequence is formed from some literature such as novels, books, films, and the internet which use the Indonesian language. The expressive Indonesian speech database was recorded by a total of four speakers with two male speakers (MDPA, MBAZ) and two female speakers (FYAT, FCIM). The recording process spent approximately 4-5 hours for each speaker. The recorded speech duration is 2-5 seconds for short sentences and 6-9 seconds for long sentences. The total duration of all recorded expressive Indonesian speech databases is 2.54 hours with the male voice for around 1.3 hours and for the female voice for around 1.24 hours.

2.2 Fundamental Frequency Estimation Using Instantaneous Frequency Amplitude System (IFAS)

The instantaneous frequency is the rate of change of the instantaneous phase angle with respect to time. First, the derivation will depart from the windowed Fourier transform point of view to estimate

the instantaneous frequency. By means of the notation of instantaneous frequency amplitude spectrum, then the harmonic structure of the signal can be revealed to determine the fundamental frequency. It is well-known that F_0 can be estimated using periodicity in the time domain or from the harmonics sequence of the speech signal in the frequency domain examined over a short-term window. Spectral-domain methods, to name several of them, cepstrum, maximum likelihood, and autocorrelation methods, estimate the fundamental period of a signal directly using windowed segments of speech. In this work, we consider an instantaneous frequency framework to estimate F_0 by introducing a harmonicity measure, a function to quantitatively ascertain the degree of regularity of the harmonic structure of the analyzed speech signal. We show that the estimation accuracy of the proposed method without any post-processing is better than that of the conventional method described in [12], WU, SWIPE, STRAIGHT, RAPT, and YIN. The instantaneous frequency amplitude spectrum (IFAS) at instantaneous frequency λ_0 is defined by the following equation (1).

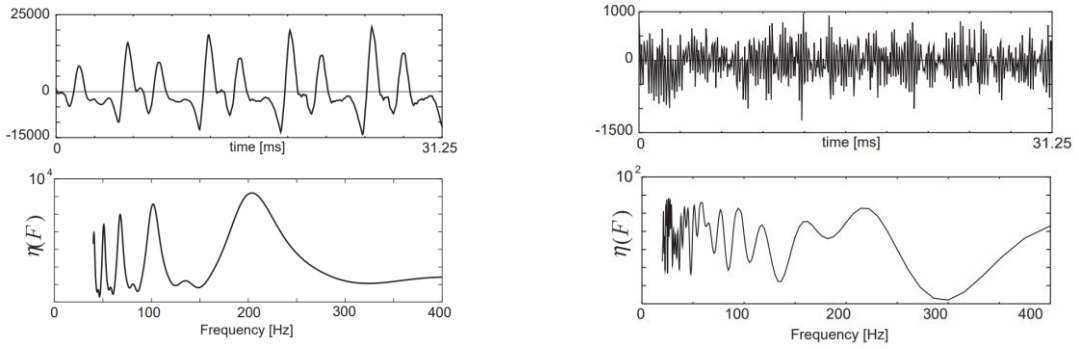
$$S(\lambda_0, t) = \lim_{\Delta\lambda \to 0} \frac{1}{\Delta\lambda} \int_{\Omega_0} |G(\omega, t)| \, d\omega, \tag{1}$$

Where $\Omega_0 = \{\omega | \lambda_0 \leq \lambda(\omega, t) \leq \lambda_0 + \Delta\lambda\}$. w(t) is an analysis window function. Without loss of generality, w(t) is real and of finite duration. If the Fourier transform of w(t) is a lowpass function, then G(ω, t) will be the output of a bandpass filter whose impulse response is $w(-t)e^{j\omega t}$.

2.3 Voiced/Unvoiced Classification

About a decade ago, voiced/unvoiced classification was reported by using temporal analysis [10]. The voiced/unvoiced decision was based on the pitch evidence and on the continuity of the pitch estimates. A frame is unvoiced unless the evidence was greater than a fixed voicing threshold. Until recently, the thresholding method is widely used for voicing decisions because of its simplicity and its considerable performance which depends on the underlying parameter to determine the threshold value. In the following, several thresholding methods will be described with the Instantaneous Frequency Amplitude Spectrum (IFAS)-based F_0 evaluation function. Since F_0 exists only in the voiced part of speech, applying a Fourier-type analysis for the determination of an F_0 found in the analysis window is widely used in voiced/unvoiced classification research. The first strategy for thresholding is by determining the value of the IFAS-based F_0 evaluation function of each frame, η(F) in equation (2), in one speech file. The threshold value is selected by examining the overall η(F) to single out the highest possible value for unvoiced speech while otherwise, the value is classified into voiced. For example, the IFAS-based F_0 evaluation function, η(F) is shown in Figure 1 (a) for voiced speech and Figure 1 (b) for unvoiced speech.

$$P_{\lambda l, \lambda u} = \max_{F} \xi_{\lambda l, \lambda u}(F) \tag{2}$$



(a) (b)

Figure 1– (a) Example of η(F) for voiced speech (b) Example of η(F) for unvoiced speech [11]

2.4 Speech Synthesis System Based on Hidden Markov Model-based Text to Speech System (HTS)

The Statistical parametric synthesis expresses the handicraft of experts from the rule-based model to the statistical model. HMM-based Text to Speech System (HTS) is one of the statistical parametric

synthesis techniques which widely known. In the HMM-based speech synthesis, the speech parameters of a speech unit such as fundamental frequency, phoneme duration, and spectrum are statistically modeled and generated by using HMMs based on the maximum likelihood criterion [13]. The HMM-based speech synthesis system consists of two main processes, which are the training and synthesis part. In the training part, the HMM model represents the excitation source, i.e., F_0, the spectrum, and the state duration of the context-dependent speech units. Each HMM model has a left-to-right state transition with no skip. The acoustic model in HTS is built from the application of maximum likelihood probabilistic equations in the training process (1) and in the synthesis process (2). The optimal model parameter can obtain by maximizing the likelihood of the training data which are given in the following equation (3) and equation (4).

$$\hat{\lambda} = \arg\max_{\lambda} P(O|T,\lambda) \tag{3}$$

Where $\hat{\lambda}$ is the model parameter estimation, O is the training data, T is a word derived from the label (transcription) and λ is a model parameter.

$$\hat{O} = \arg\max_{o} P(O|t,\hat{\lambda}) \tag{4}$$

where $\hat{O}$ is an estimation model speech, o is the speech parameter, t is the word to be synthesized which derived from the phrase labels, and $\hat{\lambda}$ is the estimation model [14].

The synthesis part has the inverse operation of the speech recognition system. The input system is contextual label sequence of the text possessing the same format but different text from the training part. From the context-dependent label of the given text, then an utterance HMM is constructed by concatenating the context-dependent HMMs according to the label. After that, the sequence of spectral and excitation parameter is generated by the speech parameter generation algorithm that maximizes their output probabilities. Finally, a speech waveform is synthesized directly from the generated spectral and excitation parameters using the mel log-spectrum approximation (MLSA) filter.

2.5 Speech Synthesis System Based on Hidden Markov Model-Deep Neural Network (HMM-DNN)

In the HMM-based speech synthesis system, we use more complicated speech units considering prosodic and linguistic contexts such as mora, accentual phrase, part of speech, breath group, and sentence information to model suprasegmental features in prosodic features appropriately. However, it is impossible to prepare training data that covers all possible context-dependent units, and there is great variation in the frequency of appearance of each context-dependent unit. In the traditional HMM speech synthesis system, decision tree clustering is used to cluster many context-dependent models to avoid overfitting caused by the lack of training data for each context model. To alleviate these problems, a few techniques are proposed to cluster HMM states and share model parameters among states in each cluster. The decision-tree-based context clustering technique [15][16] is applied separately to the spectral and log F_0 parts of the context-dependent phoneme HMM. This algorithm is often referred to as a decision-tree-based context clustering algorithm.

5

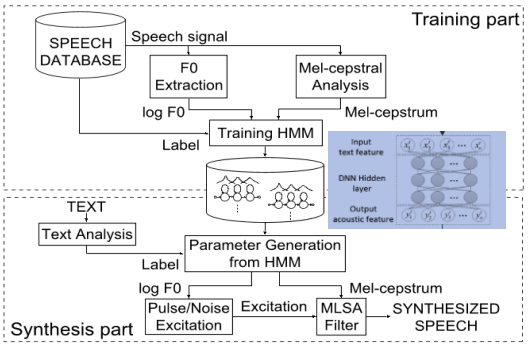


Figure 2 – Speech synthesis system based on HMM-DNN

Figure 2 shows the speech synthesis framework based on HMM-DNN. Firstly, the STRAIGHT synthesizer extracts the acoustic parameters, including fundamental frequency features, spectral parameters, and duration parameters. In this work, we also use the instantaneous frequency amplitude

spectrum (IFAS) framework to estimate F_0 by introducing a harmonicity measure, a function to quantitatively ascertain the degree of regularity of the harmonic structure of the analyzed speech signal. In the model training part, the inputs are context features, the outputs are the acoustic features, and the inputs and outputs are forcefully aligned by the trained HMM model. The DNN-based acoustic model can replace the decision-tree clustering model without restricting greedy search. It expresses complex contextual features through high-dimensional data inputs, establishes a matching model of contextual features and acoustic parameters, and effectively improves modeling accuracy through its powerful nonlinear modeling capabilities, but requires big training data.

3. MEASUREMENT TEST

In this paper, we are using the objective test to measure the quality of synthesized speech. The objective test uses the mel-cepstral distortion (MCD) method. The objective test is intended to assess the speech quality of the synthesized speech by analyzing mel-cepstrum distortion value from the original speech. The smaller MCD value indicates the closer synthesized speech to produce the natural speech. The mel-cepstral distortion (MCD) method is given in the following equation (5).

$$MCD = 10/ln10\sqrt{2\sum_{i=1}^{24}\left(mc_i^{(t)} - mc_i^{(c)}\right)^2} \quad (5)$$

with $mc_i^{(t)}$ is MFCC value which used as reference and $mc_i^{(c)}$ is MFCC value of predicted speech [16].

4. EXPERIMENT AND DISCUSSION

4.1 Experiment Set-up

In this section, we will describe our experiment to build synthesized speech of expressive Indonesian speech corpus using HMM-based speech synthesis system (HTS) and HMM-DNN. These experiments consist of some variations, first is a variation in the number of training sentences, second is a variation in the type of speech synthesis technique used for the training process, end third is some comparison of STRAIGHT and IFAS methods to estimate frequency fundamental.

The first experiment is making a variation in the number of expressive Indonesian speech corpus which used in the training part. We are using minimum, maximum, and combination numbers of speech corpus. Such variations are made according to the number of sentences. While for the minimum training, we construct sentences using the least number of phonemes according to the phonetically balanced of maximum training. The database contains 655 sentences with three expressive styles [8]. Three expressive styles are happy, sad, and angry. Variations training of expressive Indonesian speech corpus are given in the kind of sentences and in the training data amount. In the kind of sentences, variation is done with happiness sentences, anger sentences, sadness sentences, and the combination of all style of expressive. The variation applies to four speakers. This arrangement can be seen in Table 1.

Table 1 – Variation of training data number

Kind of training sentences	Minimum training amount	Maximum training amount	Synthesis sentences amount
Happiness	82	227	50
Anger	80	213	50
Sadness	81	215	50

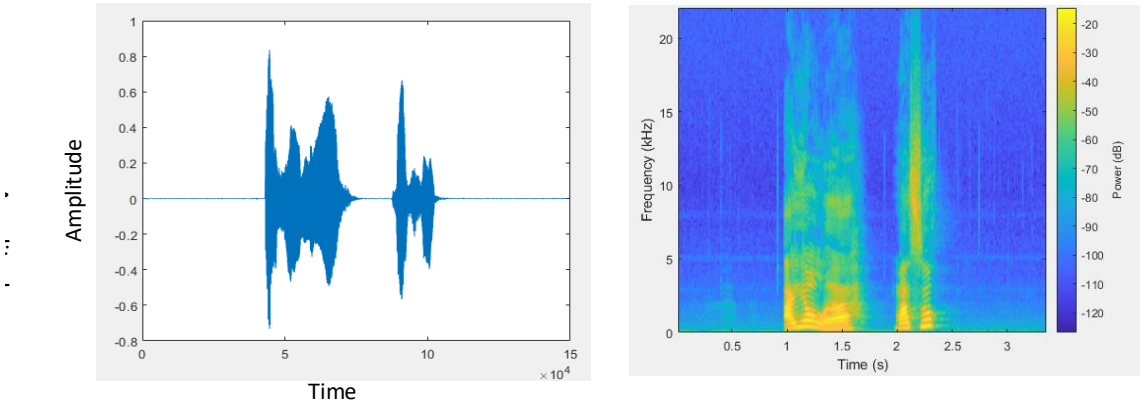
In this paper, we conducted some experiments to build a speech synthesis system by using HMM-based speech synthesis system demo for speaker dependent (HTS-demo-CMU-ARCTIC-SLT). The demo program work with some following software, i.e., SPTK, HMM Toolkit (HTK), HDcode, HTS-2.2, hts-engine API-1.05, festival, ActiveTcl and speech tools. All of them is an open source program on Linux. The HTS demo is available in English. Then we adapted into expressive Indonesian speech corpus with some modification, that are in the speech corpus which contain of speech unit and its context-label, and the question file to build the decision tree according to the phoneme rule of Indonesian [8]. All of them will be used for training part to build the parameter generation of HMM

model, then it will be used in synthesis part to generate the speech waveform by Mel log spectral approximation (MLSA) filter.

4.2 Result of Synthesized Speech

The different training processes will give different results in the training models. While variations in the training data number aimed to determine the lower limit of training data to keep producing the natural synthesized speech. The more speech corpus uses in the training process, the better acoustic models will be produced. It is because the distribution of phonemes in the speech corpus will affect the probability of acoustic model formation. However, with the more speech corpus used in the training part, it will take much computation load and time. In Table 2, shown the computation time while running the HTS demo for Indonesian speech corpus. The computation time varies from the minimum, maximum and combination sentences for both declarative and question sentences. It shows that the increasing number of training sentences, make the computation time longer.

The synthesis process is proceeded after the formation of model training is completed. The step is to combine the acoustics and linguistic features that have been formed in the training part to be desired synthesized speech. From several variations given, it has different synthesized speech quality which located in the level of naturalness. The combination training sentences HMM-DNN IFAS produced better synthesized speech than maximum and minimum training sentences. It can be seen from the comparison of fundamental frequency plot (excitation parameter) and mel-cepstral plot (spectral parameter) of speech signal. The fundamental frequency track show how the pitch of speech signal that show an intonation aspect in a sentence change in time. In Figure 3 show F_0 plot of the happiness sentence “ayolah, masuk!” in Indonesian, if translated in English become “Come on!". From that figure can be seen the waveform of speech signal followed by comparison among fundamental frequency (F_0) contour of the synthesized speech and original speech. From the F_0 contour can be identified voiced, unvoiced, and silent regions. Through the dotted line, can be seen the difference of F_0 from each synthesized speech. Aside from the fundamental frequency plot for the extraction parameter, we can see the spectral parameter by using mel-cepstral plot. It can be obtained by converting the speech signal from the time domain to the frequency domain in logarithmic scale (log FFT).



(a) (b)

Figure 3 – (a) Plot waveform and fundamental frequency (b) spectrogram of the female voice (FYAT) with happiness sentence “ayolah, masuk!”

MCEP plot for the synthesized speech of FYAT with the happiness sentence ″ayolah, masuk! can be seen in Figure 4. Plot MCEP obtained by us sing Speech Signal Processing Toolkit (SPTK) tools *mcep* with the condition of sampling frequency 16000 Hz, frame length 400 points (25 ms), frame period 80 points (5 ms), analysis of order 20, frequency warping parameter FFT size of 0.42 and 512 points, then stored in the file *.mcep*. Afterward plotted in MCEP graph using tools *glogsp* and *mgc2sp*. Mel-cepstrum plot has two main information, that are cepstral and cepstral envelope. Both of which will provide information of location, magnitude, and the characteristics of speech signal including duration, formant frequency (F1-F5), delta (the speed of speech, derived from the difference between the cepstrum peaks), delta-delta (the speech acceleration, derived from derivative of delta cepstrum). For voiced speech at cepstrum plot will have more energy at a lower frequency and have lower energy at high frequency (cepstral tilt). Whereas for the unvoiced speech will have energy that is almost evenly on each frequency. When compared based on MCEP plot of each synthesized speech both for FYAT, and for sadness, happiness, and anger sentences, it appears that

original speech and synthesized speech with maximum training data and using HMM-DNN IFAS have less distortion than synthesized speech with HMM-IFAS. This is because of the increasing number of training data that be used, the more acoustics model will be generated. So, the probability of the system generating synthesized speech will be even greater by maximizing the acoustics model of speech which will be synthesized with the acoustics model generated in the training process.

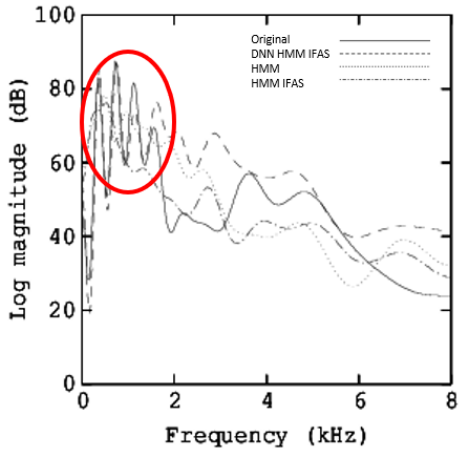


Figure 4 – Mel-cepstral plot of the female voice (FYAT) with happiness sentence "ayolah, masuk!"

In this paper, we are using the objective test to measure the quality of synthesized speech. First is using objective test, which using mel-cepstral distortion (MCD) method. MCD value indicate the closer synthesized speech to produce the natural speech. Figure 5 is the objective test result of synthesized speech for HMM IFAS and HMM-DNN IFAS, respectively.

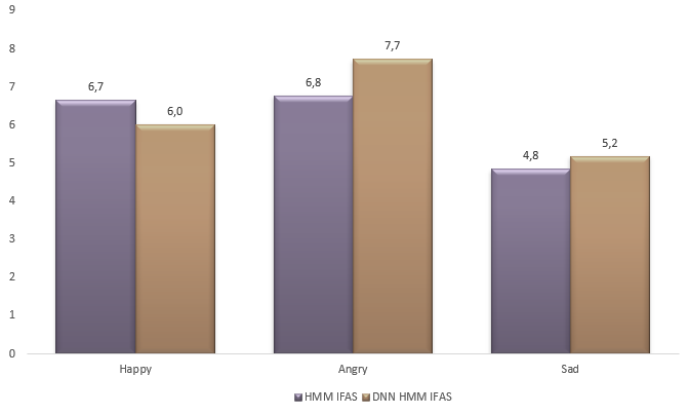


Figure 5 – Objective test of synthesized speech of female voice

Based on the results indicate that the speech quality of synthesized speech is still not enough. The smallest distortion value on female voice for sadness sentence is on HMM IFAS with a score 4.8 and for DNN-HMM IFAS has 5.2 scores with maximum training data which mean *degradation speech is slightly annoying*. Based on these data, can be concluded that the distortion of mel-cepstral will be smaller when the speech synthesis system uses DNN-HMM IFAS. That is because of the more probabilities of the appearance phonemes when using the DNN to replace the decision tree phase in HMM. Based on the explanation of this research above, can be concluded that a speech synthesis system for expressive Indonesian has been built. The system is built by a statistical parametric speech synthesis system that uses a statistical model to run the mapping between the speech and linguistic information. Some variation has been applied to the system and achieved the speech quality which measured by objective test.

5. CONCLUSIONS

Based on the explanation about this research above, can be concluded that expressive speech synthesis system for Indonesian has been built. Beside that also has been comparing the HMM-based text to speech (HTS) method and DNN-HMM method with STRAIGHT and IFAS to estimate the fundamental frequency. The system built by statistical parametric speech synthesis system which

using statistical model to run the mapping between the speech and linguistic information. Some variation has been applied to the system and achieved the speech quality which measured by objective and subjective test. From the evaluation acquired that the speech quality result of the synthesized speech by using HMM IFAS and DNN-HMM IFAS is not having a big difference. The objective test has shown that the synthesized speech still produces big distortion. The speech quality by objective test result is acquired the best value for sadness sentences is 4.8 using HMM IFAS and for DNN-HMM IFAS is 5.2 respectively.

In the future work, should be improved to increase the synthesized speech quality by modifying the system drawback, especially in HTS, by using better vocoder like STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) and IFAS make better acoustic model and reduce post filtering. The others future work is to build the speaker adaptation of Indonesian TTS with only using small adaptation data.

ACKNOWLEDGEMENTS

Authors would like to thank for Institut Teknologi Sepuluh Nopember (ITS) and who support this research under grant No. 1022/PKS/ITS/2019, and also thank for the speakers, Wins recording studio, and Institut Teknologi Bandung who take part in recording process of Indonesian speech corpus.

REFERENCES

1. A. J. Hunt and A. W. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In Proceedings of the Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings., 1996 IEEE International Conference - Volume 01 (ICASSP '96), Vol. 1. IEEE Computer Society, Washington, DC, USA, 373-376. DOI=http://dx.doi.org/10.1109/ICASSP.1996.541110
2. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, Proc. of Eurospeech, pp.2347-2350, Sept. 1999.
3. A. Black. "CLUSTERGEN: A Statistical Parametric Synthesizer Using Trajectory Modeling", in Interspeech 2006, Pittsburgh,PA., 2006.
4. H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda, The HMM-based speech synthesis system version 2.0, Proc. of ISCA SSW6, Bonn, Germany, Aug. 2007.
5. J. Wang and Y.Y. Zhang, "Title Research on Deep Neural Network Based Chinese Speech Synthesis," Computer Science, vol. 42, no. S1, pp. 75–78, 2015.
6. H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 7962–7966.
7. Noll, A. M., "Cepstrum pitch determination," Journ. Acoust. Soc. Am., vol. 41, pp. 293–309, February 1967.
8. E. Anggrayni and D. Arifianto. "HMM-based Speech Synthesis System with Expressive Indonesian Speech Corpus." 2019 Conference of the 23rd International Congress on Acoustics (ICA), 2019.
9. A. Black and J. Konimek "Optimizing Segment Label Boundaries for Statistical Speech Synthesis" ICASSP 2009, Taipei, Taiwan. 2009.
10. Immerseel, L. V. and Martens, J. P., "Pitch and voiced/unvoiced determination with auditory model," Journ. Acoust. Soc. Am., vol. 91, pp. 3511–3526, June 1992.
11. D. Arifianto, T. Kobayashi: "Voiced/Unvoiced determination of speech signal based on instantaneous frequency," Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on · May 2007 IEICE Journal, Trans. Information and System.
12. Abe, T., Kobayashi, T., and Imai, S., "Robust pitch estimation with harmonic enhancement in noisy environment based on instantaneous frequency," in Proc. 4th ICSLP, (Philadelphia, USA), pp. 1277–1280, October 1996.
13. K. Tokuda, H.Zen and A. BLack "An HMM-based Speech Synthesis System Applied to English", Proc. of 2002 IEEE SSW, Sept. 2002.
14. K. Shinoda and T. Watanabe. MDL-based context-dependent subword modeling for speech recognition. J. Acoust. Soc. Japan (E), 21:79–86, March 2000.
15. S.J. Young, J.J. Odell, and P.C. Woodland. Tree-based state tying for high accuracy acoustic modeling. In Proc. ARPA Human Language Technology Workshop, pages 307–312, March 1994.
16. Kondo, K. "Subjective Quality Measurement of Speech, its Evaluation, Estimation and Applications". 2012, XIV, p. 154, ISBN : 978-3-642-27505- 0

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0845

Immersive sound for social interactions in extended reality (XR) - insights and current challenges

Gunilla BERNDTSSON[1]; Janto SKOWRONEK[2]

[1]Ericsson Research, Sweden

[2]Hochschule für Technik - University of Applied Sciences, Germany

ABSTRACT

In the past, spatial audio has been investigated for contemporary telecommunication systems, often with a focus on multiparty, audio-only telemeetings. Considering new communication systems that use virtual, mixed and augmented reality technologies, also referred to as eXtended Reality (XR), activities on immersive sound for communication purposes have increased once more.

With the authors being active in standardization (ITU-T Study Group 12: Performance, QoS, and QoE), this conference contribution explores the field with the intention of bringing together knowledge from academia, industry and standardization bodies.

The focus lies on understanding current challenges regarding aspects of perception, immersion and communication. Topics addressed are, among others, ongoing work on assessment methods for spatial audio codecs and renderings for XR. With the new extended reality technologies, more complex use cases arise that need to be assessed in a new way, such as in the quality assessment of 6 degrees-of-freedom settings in the recent MPEG-I immersive audio test.

The contribution will provide an overview of relevant literature, complemented with insights that the authors obtained during expert interviews with active persons in the field.

Keywords: Extended reality, Spatial audio evaluation, Social interactions

1. INTRODUCTION

There has been an unprecedented increase in remote meetings due to the COVID-19 pandemic and advances in communication technology has made it easier for us to connect with others and work remotely. Now that higher bandwidths and lower latencies are possible with 5G networks, more advanced extended reality (XR) applications can be implemented. This is an active research field in both industry and academia. Standardization organizations are also active in order to make these technologies well-functioning and interoperable in the so-called metaverse, where the physical and digital worlds meet (1).

New tools for collaborations and meetings are emerging based on XR technologies such as augmented reality (AR) and virtual reality (VR). These technologies add a spatial dimension that can make the experience come closer to that of a physical meeting. To make a good experience possible for all participants in a virtual meeting, there is a need to assess the Quality of Experience (QoE) of the users. This is especially important now that we foresee more variants of hybrid meetings including XR components.

The authors are working on methods for conferencing and telemeeting assessment in ITU-T Study group 12/Question 10 (2). This group has recently developed a recommendation on aspects of importance for QoE assessment of telemeetings with extended reality elements (3). The intent is to work further on assessment methods for immersive audio suitable for XR meetings.

The content in this paper is based on work in ITU-T SG12 and other standardization organizations as well as research in industry and academia. In addition, we have interviewed audio experts and conducted a literature survey.

[1] gunilla.berndtsson@ericsson.com
[2] janto.skowronek@hft-stuttgart.de

2. BENEFIT OF SPATIAL AUDIO FOR SOCIAL INTERACTIONS

The benefit of spatial audio rendering in telemeetings lies predominantly in positive effects on the way participants can process shared information to follow and contribute to the conversation. On a perceptual level, spatial audio rendering enables participants to use binaural hearing processes to easier listen to individual speakers in a group of simultaneously talking persons, widely known as the cocktail party effect (4). Further effects are for instance an increased focal assurance (5,6), i.e., an increased ability to identify the different speakers and remember who said what, a reduced cognitive load or listening effort experienced during the telemeeting (7,8), and a higher effectiveness in fulfilling a certain task (9). In that context, (10) discusses how spatial audio can, among many other factors, reduce video-conferencing fatigue.

Next to such effects, spatial audio can enhance the Quality of Experience by having a positive impact on the communication processes (11). Spatial audio leads to the perception of spatial attributes of sounds, which in turn contribute to a spatial quality component (12) of QoE. In addition, spatial audio can increase the feeling of immersion and presence (13,14), which again positively impacts QoE; and in the context of audiovisual telemeetings, an adequate spatial alignment of auditory and visual information can also increase QoE (15,16).

In an XR meeting participants can have the possibility to meet and move around in the same virtual room, which leads to new possibilities to interact (17). New design requirements and opportunities for spatial audio arise (18). One research question is about the exact placement of sound sources, such as other meeting participants, in the virtual auditory space in order to ensure a coherent spatial experience between the visual and auditory information (19).

3. TECHNOLOGY CONSIDERATIONS FOR IMMERSIVE SOUND

A lot of information about technologies for immersive sound can be found in the literature. For example, (20,21,22,23) provide overviews on sound field representations such as ambisonics and wave field synthesis (WFS), which may be rendered to different loudspeaker configurations or binaural playback. Further pointers to related literature are (24) on coding technologies for spatial audio and (25,26) on the use of spatial audio for AR and VR scenarios.

Focusing on social interaction scenarios, there are a number of technical and also non-technical factors that should be considered in order to find an optimal mapping of the technology to the use case. To structure such factors, we developed a Telemeeting Profile Template, a tool that is designed for systematically characterizing telemeetings from a QoE perspective (27). For this purpose, the tool utilizes Quality Influence Factors (QIFs), which comprise System, Human, Context and Mixed Influence Factor; see Figure 1 and (11,28).

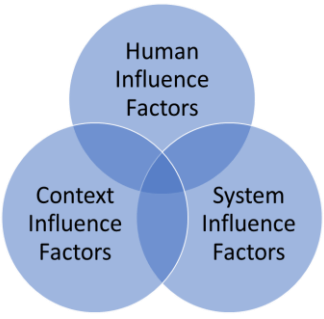


Figure 1 – Visualization of the three types of Quality Influence Factors (28), i.e., technical and non-technical factors that influence the Quality of Experience. The overlap illustrates that there are also Mixed Factors (11,27).

When designing a communication system or an assessment method, the Profile Template can be used to identify factors that are relevant for the considered use cases and extract from this information the technical requirements for the system. The following considerations illustrate how this can be achieved for spatial audio telemeeting systems.

With respect to sound capture and signal processing, a number of System and Mixed Influence Factors need to be considered that are similar to non-spatial setups, e.g., type and positioning of the microphone(s), as well as signal processing algorithms for noise reduction, echo cancellation, and so forth.

Concerning the sound reproduction, the requirements for the spatial audio technology are similar to non-interactive use cases. For loudspeaker-based systems typical challenges are the reduction of artifacts for higher-order ambisonics (HOA) and WFS systems when using different loudspeaker configurations (29,30). For headphone-based systems typical considerations concern the performance of the binaural synthesis, especially the processing of the used head-related transfer functions or binaural room impulse responses (31,32) and the spatial and temporal precision of tracking the participants (23,33).

In use cases in which a strong feeling of co-presence (34) or social presence (35) is desired, requirements for a virtual environment concern System Influence Factors such as the auditory and visual representation of the participants, the degree of realism and the degree of freedom. This can be challenging due to potential perceptual effects such as the uncanny valley effect (36) and simulator sickness (37).

When the *purpose* of a telemeeting, which is a Context Influence Factor, is to achieve a good collaboration of remote participants using VR or AR settings, further requirements arise with respect to the auditory and visual representations of the meeting participants and especially the environments. Here, the capturing, reproduction, and creation of the (virtual) communication environment, and for AR in particular the mixture of real and virtual environments, is a topic of ongoing research. For the audio domain, respective applications could build on work on plausible representations of real acoustic rooms in virtual and augmented environments (25,38).

4. EVALUATION METHODS FOR SPATIAL AUDIO

4.1 Standardized methods relevant for spatial audio telemeetings

There are different approaches to evaluate the benefits of spatial audio in telemeetings, ranging from collecting ratings from test participants over measuring task performance and physiological measurements to analyzing participant behavior. For instance, ITU-T Recommendations P.1301 on ”Subjective quality evaluation of audio and audiovisual multiparty telemeetings” (39) and P.1310 on ”Spatial audio meetings quality evaluation” (12) provide guidance on choosing an appropriate test method according to the test scenario at hand. Here, the ITU-T also provides details on the collection of quality ratings from test participants as well as information on the collection of task performance indicators for the intelligibility of concurrent speakers (12,40), communication effectiveness (9), and cognitive load (12).

4.2 Further approaches for assessment of spatial audio

Next to these standardized assessment methods, additional approaches that may also be used for assessment of XR meetings are reported in the literature. To start with, there is a lot of work on assessing individual attributes of spatial audio (41,42,43,44), which can also form the basis for assessing individual aspects of spatial audio in XR meetings. Going beyond the use of verbal descriptions for spatial audio assessment, works such as (45) address the potential to use non-verbal descriptions. As an example, (46) describes a sophisticated approach to not only assess direction but also distance, dispersion and trajectory.

Behavioral analysis is another assessment approach. Here, a method based on behavior tracking is proposed in (47) which is particularly designed to assess binaural renderers while being immersed in multimodal virtual scenes.

Physiological measurements form another assessment approach. A survey of QoE assessment using physiological measurements for different video, audiovisual and speech-based services can be found in (48). With respect to XR scenarios, Keighrey et al. (49) for example, report on how to combine physiological measurements with user ratings in order to compare VR, AR and tablet applications in a speech memory test.

Also crowdsourced tests are used more and more for audio tests (50) and have been used for spatial audio as well (51).

4.3 Current test method developments in standardization

There is a need for measuring the performance of audio systems, and subjective test methods published by the ITU are frequently used for this purpose.

In 3GPP SA4 there is ongoing work to decide on suitable audio test methods for the ”Immersive Voice and Audio Services (IVAS)” codec (52). Traditionally, test methods specified in ITU-T Rec. P.800 (53) have been used for codec evaluation by non-expert test participants. However, as IVAS offers immersive audio, which may go beyond what listeners experienced in communication services have experienced so far, the spatial dimension may need to be brought to their attention.

A subjective test method for evaluating speech oriented stereo communication systems over headphones was developed in ITU-T Rec. P.811 (54). It is based on the degradation category rating (DCR) methodology described in ITU-T Rec. P.800, where test participants rate the quality degradation of the processed test signal relative to the unprocessed signal (the reference). Similar to the ITU-T Rec. P.835 (55), there are three sub-trials in P.811 (54), for signal, spatial, and overall quality. When judging the spatial quality, test participants are instructed to attend only to the spatial localization accuracy degradation of the test sample compared to the reference sample. Spatial anchors have been developed to ensure that there are known degradations in the spatial dimension.

As it is time consuming with three sub-trials in an experiment, it was investigated if similar results for the overall quality could be achieved in a more time-efficient way. Tests showed that this was possible using only the scale for the overall quality if modified instructions and the same type of spatial anchors as in P.811 were added in the test material, see Appendix II in (54). In ITU-T SG12 there is now an effort to collect examples and recommendations for how methods described in P.800 can be used for assessment of spatial audio.

The target for the MPEG Audio Coding group (ISO/IEC JTC 1/SC 29/WG 6) is to deliver a spatial audio renderer for virtual and augmented reality applications in 6 degrees-of-freedom (6-DoF). 6-DoF allows translation along the X, Y and Z axis and rotational movements along those axes. This renderer will support position and spatial extent of audio objects, occlusions, diffractions, reverberation, early reflections, movement through connected scenes, etc. It will be described in the MPEG-I Immersive Audio standard ISO/IEC 23090-4.

In order to assess the performance of different technologies an audio evaluation platform was developed, where users equipped with VR and AR head mounted displays and headphones can explore virtual and augmented reality environments and evaluate audio rendering technologies in real-time. A Call for Proposals (CfP) for 6DoF immersive audio technology was issued in April 2021. The MPEG-I immersive audio listening tests were performed using the A-B methods described in ITU-R BS.1284-2 (56) comparing rendering technologies from different proponents. The main focus was on binaural rendering and encoding of audio scene metadata. All aspects of the CfP process are described in (57).

5. CHALLENGES

After the overview of approaches concerning the technology and assessment of spatial audio, this section discusses several challenges that we have identified based on the literature survey and the expert interviews.

One important aspect is how to prepare test participants for subjective tests. They need to familiarize themselves with immersive audio content to become aware of spatial aspects, and suitable listener instructions are needed. Persons that are not used to assessing immersive audio tend to focus on signal quality and might disregard the spatial dimension in their judgments.

The difficulty and complexity of subjective tests conducted in VR and AR (as in the MPEG-I immersive audio tests mentioned above) is much greater than in traditional audio coding evaluations. The test participants need to get used to move around in the XR environment and experience the content and acoustic effects that are apparent in the different test environments. Also, the test method needs to be tried out.

In several audio test methods such as the ITU recommendations P.800, P.811, BS.1284-2 and BS.1534-3 (53,54,56,58), a reference signal is presented to the test subject. The reference provides a target for the comparison. When comparing different renderings of AR and VR applications, there is usually no such reference available. The test subjects must instead rely on what they think sounds more pleasant or perceive as realistic.

Another challenge in AR and VR tests is the freedom to move around in the scene. If test participants are allowed to move along different paths in scenes, they will have different experiences to judge. For instance, effects like occlusion, reverb, dampening, and distance attenuation will depend on their positions in the scene. To get more consistent experiences among the different test participants, they might need instructions on which positions they should take, and how they should move around in the test scenes. This aspect becomes even more complex in an AR communication setup, where local, remote, and virtual acoustic environments may need to be combined.

It is also often a challenge to understand why test participants vote as they do, and it can be helpful to ask them to comment on the background to their votes during the voting procedure and after the test session. This can lead to a more complete understanding of the perceived Quality of Experience.

In listening tests, where there is a comparison of two or more test samples, it is important that the loudness of these samples is equal, so that differences in loudness does not affect the test results. This can be especially challenging regarding tests in AR and VR, as there can be many different sounding objects. As an example, a realistic rendering of a distant audio source may sound muffled, and only be heard at a low level. How should that be compared to an unrealistic rendering, that presents the audio source loudly and clearly?

Also, audio-visual synchronization impacts the perceived QoE, and needs to be considered. How to evaluate media synchronicity and the effects of delay in real-time communications is addressed in (59), but new assessment methods need to be developed when it comes to social interactions in XR.

6. CONCLUSIONS

There are currently many development activities in both academia and industry, in which XR technology is used for different use cases, among them social interaction scenarios. As discussed in the previous section, several challenges and open questions need to be addressed, both in technology and evaluation. Here, we conclude that there is further need for a good understanding of the interplay between the technology and important auditory and cognitive aspects. This is especially true when the purpose of using XR technology is to create a strong feeling that a person is in an environment (immersion, presence) and that she or he is connected with others (co-presence, social presence).

Moreover, in order to achieve such a goal, new methods for assessing the auditory and audiovisual experience of XR meetings are required. While designing such test methods, there is a need to keep different technical and non-technical factors (i.e., Quality Influence Factors) in mind to provide reliable and comparable results. One important aspect in this context is the observation that more variants of telemeetings with different combinations of XR elements will occur, which will lead to an increased importance of understanding the resulting QoE for all meeting participants.

For that reason, the active parties in Question 10 of ITU-T Study Group 12 will continue to develop such assessment methods, both looking at the complex mixture of factors and detailed aspects concerning the resulting test protocols.

ACKNOWLEDGEMENTS

The authors would like to thank all interview partners from industry, academia and standardization bodies for sharing their insights.

REFERENCES

1. The Metaverse Standards Forum, Online resource, accessed on July 28, 2022, https://metaverse-standards.org/
2. ITU-T Study Group 12/ Question 10. Online resource accessed on July 28, 2022. https://www.itu.int/net4/ITU-T/lists/q-text.aspx?Group=12&Period=17&QNo=10&Lang=en
3. ITU-T Rec. P.1320. QoE assessment of extended reality (XR) meetings. International Standard, International Telecommunication Union, Geneva, Switzerland, 2022.
4. Bronkhorst AW. The cocktail-party problem revisited: Early processing and selection of multi-talker speech. Attention, Perception, Psychophys. 2015; vol. 77, no. 5, p. 1465-1487.
5. Baldis JJ. Effects of spatial audio on memory, comprehension, and preference during desktop conferences. Proc. ACM CHI Hum. Factors Comput. Syst. Conf., Beaudouin-Lafon M, Jacob RJK, editors. Seattle, WA, USA, 2001, vol. 3, no. 1, pp. 166173, doi: 10.1145/365024.365092.
6. Kilgore R, Chignell M, Smith P. Spatialized audioconferencing: What are the benefits? Proc. Conf. Centre Adv. Stud. Collaborative Res. (CASCON); 2003, p. 135-144.

7. Skowronek J, Raake A. Assessment of cognitive load, speech communication quality and quality of experience for spatial and non-spatial audio conferencing calls. Proc. Speech Commun., 2015, p. 154-175, doi: 10.1016/j.specom.2014.10.003.
8. Fintor E, Aspöck L, Fels J, Schlittmeier SJ. The role of spatial separation of two talkers' auditory stimuli in the listener's memory of running speech: listening effort in a non-noisy conversational setting. International Journal of Audiology; 2021. p. 1–9.
9. ITU-T Rec. P.1312. Method for the measurement of the communication effectiveness of multiparty telemeetings using task performance. International Standard, International Telecommunication Union, Geneva, Switzerland, 2016.
10. Raake A, Fiedler M, Schoenenberg K, De Moor K, Döring N. Technological factors influencing videoconferencing and zoom fatigue: 2022, arXiv:2202.01740.
11. Qualinet white paper on definitions of Quality of Experience, version 1.2. European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003). Le Callet P, Möller S, Perkis A, editors. Lausanne, Switzerland, March 2013. Available online: http://www.qualinet.eu/resources/qualinet-white-paper/
12. ITU-T Rec. P.1310. Spatial audio meetings quality evaluation. International Standard, International Telecommunication Union, Geneva, Switzerland, 2017.
13. Västfjäll D. The subjective sense of presence, emotion recognition, and experienced emotions in auditory virtual environments. CyberPsychology & Behavior 6.2 (2003): p.181-188.
14. Agrawal S, Bech S, Bærentsen K, De Moor K, Forchhammer S. Method for subjective assessment of immersion in audiovisual experiences. J. Audio Eng. Soc. 2021; vol. 69, no. 9, p. 656-671.
15. De Bruijn WP. Application of wave field synthesis in videoconferencing. Ph.D. dissertation, Lab. Acoust. Imag. Sound Control, Fac. Appl. Sci., Delft Univ. Tech., Delft, The Netherlands, 2004.
16. You J, Reiter U, Hannuksela MM, Gabbouj M, Perkis A. Perceptual-based quality assessment for audio–visual services: A survey. Signal Processing: Image Communication, vol 25, Issue 7, 2010, p. 482-501, ISSN 0923-5965, https://doi.org/10.1016/j.image.2010.02.002.
17. Schäfer A, Reis G, Stricker D. A Survey on Synchronous Augmented, Virtual and Mixed Reality Remote Collaboration Systems. ACM Comput. Surv; 2022. https://doi.org/10.1145/3533376
18. Kailas G, Tiwari N. Design for Immersive Experience: Role of Spatial Audio in Extended Reality Applications. In: Chakrabarti A, Poovaiah R, Bokil P, Kant V, editors. Design for Tomorrow—Volume 2. Smart Innovation, Systems and Technologies; 2021 vol 222. Springer, Singapore. https://doi.org/10.1007/978-981-16-0119-4_69
19. Kim H, Remaggi L, Jackson PJB, Hilton A. Immersive Spatial Audio Reproduction for VR/AR Using Room Acoustic Modelling from 360° Images. IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 2019, pp. 120-126, doi: 10.1109/VR.2019.8798247
20. Rumsey F. Spatial Audio (1st ed.). Routledge; 2001. https://doi.org/10.4324/9780080498195
21. Spors S, Wierstorf H, Raake A, Melchior F, Frank M, Zotter F. Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State. In: Proceedings of the IEEE 101.9 (2013), p. 1920–1938.
22. Frank M, Zotter F, Sontacchi A. Producing 3D Audio in Ambisonics. AES 57th International Conference, USA 2015
23. Hacihabiboglu H, De Sena E, Cvetkovic Z, Johnston J, Smith JO. Perceptual Spatial Audio Recording, Simulation, and Rendering: An overview of spatial-audio techniques based on psychoacoustics. In IEEE Signal Processing Magazine, vol. 34, no. 3, p. 36-54, May 2017, doi: 10.1109/MSP.2017.2666081
24. Herre J, Dick S. Psychoacoustic Models for Perceptual Audio Coding—A Tutorial Review. Applied Sciences. 2019; 9(14):2854. https://doi.org/10.3390/app9142854
25. Brandenburg K, Cano E, Klein F, Köllmer T, Lukashevich H, Neidhardt A, Sloma U, Werner S. Plausible Augmentation of Auditory Scenes Using Dynamic Binaural Synthesis for Personalized Auditory Realities," Paper P8-3, (2018 August.).
26. Gupta R. et al., Augmented/Mixed Reality Audio for Hearables: Sensing, control, and rendering. In IEEE Signal Processing Magazine, vol. 39, no. 3, pp. 63-89, May 2022, doi: 10.1109/MSP.2021.3110108
27. Skowronek J, Raake A, Berndtsson G, Rummukainen O, Usai P, Gunkel SNB, Johanson M, Habets EAP, Malfait L, Lindero D, Toet A. Quality of Experience in telemeetings and videoconferencing: A comprehensive survey. In IEEE Access, vol. 10, p. 63885-63931, 2022, doi: 10.1109/ACCESS.2022.3176369
28. Reiter U, Brunnström K, de Moor K, Larabi M, Pereira M, Pinheiro A, You J, Zgank A. Factors influencing quality of experience. In Quality of Experience: Advanced Concepts, Applications, Methods, Möller S, Raake A, editors., Switzerland: Springer, 2014, p. 5572.

29. Spors S, Ahrens J. Comparison of higher-order ambisonics and wave field synthesis with respect to spatial aliasing artifacts. In 19th International Congress on Acoustics 2007.
30. Vilkaitis A. WFS and HOA: Simulations and evaluations of planar higher order ambisonic, wave field synthesis and surround hybrid algorithms for lateral spatial reproduction in theatre.' Proceedings of the 4th International Conference on Spatial Audio, Graz, 7th-10th September 2017.
31. Andreopoulou A, Katz BFG. Comparing the effect of HRTF processing techniques on perceptual quality ratings. Audio Engineering Society Convention 144. Audio Engineering Society, 2018.
32. Arend JM, Garí SV, Schissler C, Klein F, Robinson PW. Six-degrees-of-freedom parametric spatial audio based on one monaural room impulse response. Journal of the Audio Engineering Society. 2021 Jul 2;69(7/8):557-75.
33. Algazi VR, Duda RO. Headphone-based spatial sound. IEEE Signal Processing Magazine, 2010;28(1):33-42.
34. Zhao S. Toward a taxonomy of copresence. Presence: Teleoperators Virtual Environ., vol. 12, no. 5, p. 445-455, Oct. 2003, doi: 10.1162/105474603322761261.
35. Biocca F, Harms C, Burgoon JK. Toward a more robust theory and measure of social presence: Review and suggested criteria. Presence, Teleoperators Virtual Environments, vol. 12, no. 5, p. 456-480, Oct. 2003.
36. Mori M, MacDorman KF, Kageki N. The uncanny valley," IEEE Robot. Autom. Mag., vol. 19, no. 2, p. 98-100, Jun. 2012, doi: 10.1109/MRA.2012.2192811.
37. Duzmanska N, Strojny P, Strojny A. Can simulator sickness be avoided? A review on temporal aspects of simulator sickness. Frontiers Psychol., vol. 9, p. 21-32, Nov. 2018, doi: 10.3389/fpsyg.2018.02132
38. Neidhardt, A, Schneiderwind, C, Klein F. Perceptual matching of room acoustics for auditory augmented reality in small rooms - Literature Review and Theoretical Framework. Trends in Hearing 2022. doi: 10.1177/23312165221092919
39. ITU-T Rec. P.1301. Subjective Quality Evaluation of Audio and Audiovisual Telemeetings. International Standard, International Telecommunication Union, Geneva, Switzerland, 2017.
40. ITU-T Rec. P.1311. Method for Determining the Intelligibility of Multiple Concurrent Talkers. International Standard, International Telecommunication Union, Geneva, Switzerland, 2014.
41. Berg J. (2006). Evaluation of perceived spatial audio quality. Journal of Systemics, Cybernetics and Informatics, 4(2):10-14.
42. Bech S, Zacharov N. Perceptual audio evaluation-Theory, method and application. John Wiley & Sons, 2007.
43. Zacharov N, Pedersen T, Pike C. A common lexicon for spatial sound quality assessment - latest developments. Eighth International Conference on Quality of Multimedia Experience (QoMEX), 2016, p. 1-6, doi: 10.1109/QoMEX.2016.7498967.
44. Berndtsson G, Krokstad A. A room acoustic experiment with an artificial reverberation system using wooden loudspeakers. Acta Acustica 1994; Vol 2(1) p. 37-48.
45. Mason R, Ford N, Rumsey F, de Bruyn B. Verbal and non-verbal elicitation techniques in the subjective assessment of spatial sound reproduction. 109th AES Convention 2000; Paper 5225.
46. Darcy D, Terry K, Davidson G, Graff R, Brandmeyer A, Crum P. Methodologies for High-Dimensional Objective Assessment of Spatial Audio Quality. In Audio Engineering Society Convention 140. Audio Engineering Society 2016.
47. Rummukainen O, Robotham T, Schlecht S, Plinge A, Herre J, Habets, EA. Audio quality evaluation in virtual reality: Multiple stimulus ranking with behavior tracking. In Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality. Audio Engineering Society, August 2018.
48. Engelke U, Darcy DP, Mulliken GH, Bosse S, Martini MG, Arndt S, Antons J, Chan KY, Raman N, Brunnström K. Psychophysiology-based QoE assessment: A survey. IEEE Journal of Selected Topics in Signal Processing 2016; 11(1):6-21.
49. Keighrey C, Flynn R, Murray S, Murray N. A physiology-based QoE comparison of interactive augmented reality, virtual reality and tablet-based applications. IEEE Transactions on Multimedia. 2020;23, 333-341.
50. ITU-T Rec. P.808 Subjective evaluation of speech quality with a crowdsourcing approach. International Standard, International Telecommunication Union, Geneva, Switzerland, 2018.
51. Volk T, Keimel C, Moosmeier M, Diepold K. Crowdsourcing vs. laboratory experiments–QoE evaluation of binaural playback in a teleconference scenario. Computer Networks. 2015;90, 99-109.
52. IVAS work item description: Online resource, accessed on July 28, 2022, https://www.3gpp.org/ftp/tsg_sa/TSG_SA/TSGS_96_Budapest_2022_06/Docs/SP-220608.zip).

53. ITU-T Rec. P.800. Methods for subjective determination of transmission quality. International Standard, International Telecommunication Union, Geneva, Switzerland, 1996.
54. ITU-T Rec. P.811. Subjective test methodology for evaluating Speech oriented stereo communication systems over headphones. International Standard, International Telecommunication Union, Geneva, Switzerland, 2019.
55. ITU-T Rec. P.835. Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. International Standard, International Telecommunication Union, Geneva, Switzerland, 2003.
56. ITU-R Rec. BS.1284-2. General methods for the subjective assessment of sound quality. International Telecommunication Union, Geneva, Switzerland, 2019.
57. MPEG-I Immersive Audio CfP Document Set - Final. Online resource, retrieved on July 28, 2022, https://www.mpeg.org/wp-content/uploads/mpeg_meetings/136_OnLine/w20922.zip
58. ITU-R Rec. BS.1534-3. Method for the subjective assessment of intermediate quality level of audio systems. International Telecommunication Union, Geneva, Switzerland, 2015.
59. Berndtsson G, Schmitt M, Hughes P, Skowronek J, Schoenenberg K, Raake A. Methods for human-centered evaluation of mediasync in real-time communication. In Montagud M, Cesar P, Boronat F, Jansen J, editors, MediaSync: Handbook on Multimedia Synchronization Springer International Publishing; 2018. p. 229–270.

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0048

Data Augmentation for Korean Far-field Conversational Speech Recognition

Sung Joo LEE; Hoon CHUNG; Byung Ok KANG

Electronics and Telecommunication Research Institute, Republic of Korea

ABSTRACT

In this paper, we introduce useful data augmentation methods for Korean distant conversational speech recognition. The main issues with speech recognition are background noise, reverberation, speaking style, and so on. Our speech recognition results show how the end-to-end speech recognition system is affected by each issue. And then, we adopt data augmentation methods to alleviate the problems. The adopted augmentation approaches are as follows speed perturbation, noise addition, room impulse response filtering, volume perturbation, speech enhancement, hyper-and hypo-articulated speech synthesis, and voice conversion based on CycleGAN. In general, the reason why data augmentation in deep neural networks is effective can be explained by two aspects. One is that it provides multi-conditioned data and the other is that it is a simple method to obtain a generalized model. Therefore, we evaluate the performance of each data augmentation method in the two aspects. Finally, we suggest an effective combination of data augmentation approaches in the conclusion.

Keywords: End-to-end, Speech Recognition, Data Augmentation, Voice Conversion, Speech Synthesis

1. INTRODUCTION

Automatic speech recognition (ASR) as known as speech-to-text (STT) has been attracting research attention for many decades since it is considered a core technique in man-machine interface. After a long period of performance stagnation, deep neural network (DNN) technology brings remarkable breakthroughs. Nowadays, we are easily experiencing speech recognition applications such as Apple Siri, Samsung Bixby, Google Assistant, and so on. Unlike traditional STT composed of an explicit acoustic model, lexicon, and language model, end-to-end (E2E) speech recognition is an integrated neural network that directly outputs text token sequences from an input signal. The E2E ASR becomes popular because its recognition accuracy is higher when abundant training data are available [1][2].

A large amount of data is essential since the DNN algorithm learns a set of factorized functions including non-linearity in a data-driven way. However, in many cases, it is difficult to get enough data for target tasks. In particular, the characteristics of conversational talk have too many style variations in nature. So, it is hard to cover in-domain data for training. Data augmentation is to increase the amount of data by modifying existing data into possible variants likely to be. Therefore, artificial data augmentation has been investigated from the early stage of the artificial intelligence era and it is well known to be effective. In the ASR research field, various approaches have been proposed. Speed perturbation variates speech rate by warping speech signals along the time domain using interpolation and decimation techniques [3]. Vocal tract length normalization (VTLP) converts timbre by controlling the speech spectrum in the frequency domain. Volume perturbation is to change the amplitude of speech [3]. Unlike these perturbation approaches, SpecAugment achieves significant improvements in various ASR tasks by applying an on-the-fly time/frequency masking technique [4]. Text-to-speech (TTS) is also a good approach for increasing training data because it synthesizes various speech styles. Adding ambient noise to speech signals is useful to imitate possible noisy environments and room impulse response (RIR) filtering is good for synthesizing reverberant speech signals by applying various RIRs.

Although a state-of-the-art ASR system achieves substantial improvements, it is still challenging to accurately recognize distant spontaneous speech in noisy environments. In the case of a single-channel microphone, the problem becomes worse. Therefore, most far-field speech recognition

systems are based on speaker localization using microphone arrays. Ambient noise and reverberation are mainly considered the cause of performance degradation.

In this study, we analyze the reason why Korean distant spontaneous voices are hard to recognize in noisy environments. To do so, we try to establish a reliable speech recognition test bed. 1,100 hours of speech data are prepared for training an E2E ASR model. The training set is composed of isolated words, navigation commands, digits, and read sentences except for conversational talks. All the test data are spontaneous voices with different styles from verious situations. This is a so-called open test that is good for estimating the reliability of an ASR system. In this work, considered data augmentation approaches are as follows: speed perturbation, volume perturbation, noise addition, RIR filtering, speech enhancement, hyper-and hypo-articulated voice synthesis, husky voice synthesis, voice conversion, and SpecAugment. All the approaches provide explicit data sets except for SpecAugment. In order to test the generality of a speech recognition model, adversarial samples are provided as follows: wiener filtering and adding weak white noise. Emotional speech data from broadcast drama, entertainment programs, and call-center are also prepared to test the influence of conversational variants. On the basis of the series of evaluations, we propose one possible data augmentation combination. We hope that our result tables are useful to predict performance degradation in many tasks caused by ambient noise, reverberation, speaking style, and so on.

2. DATA AUGMENTATION

Data augmentation can significantly improve the performance of an ASR system, especially in mismatched conditions, or when the training data is insufficient. In this work, we try to estimate its efficiency in the different talking styles in addition to the aforementioned factors such as background noise, acoustic environment, speaking styles, and so on.

2.1 Speed Perturbation

Speed perturbation is to make the speech rate up or down by resampling techniques such as decimation, and interpolation. Therefore, it modifies both the pitch and tempo of the original speech. It is proven to be useful for performance improvement [3]. We use SoX's "speed" option [0.9, 1.0, 1.1] for speed perturbation.

2.2 Volume Perturbation

In general, the intensity of sound waves such as audio transmission in the air decreases with distance from the source. According to the inverse-square law, for every doubling of distance away from the source, the sound is going to be four times less intense. This sound attenuation phenomenon makes distant speech recognition difficult. Volume perturbation is a simple technique to control the amplitude of speech. It makes a speech recognition model robust to audio volume changes, especially in the case of far-field talks. We use SoX's "vol" option for volume perturbation. The volume perturbation factors are between 0.1 and 1.2.

2.3 Noise Addition

Ambient noise is known as a major source that degrades the performance of an ASR system. The purpose of adding noise to clean speech signals is to improve the robustness of ASR in noisy environments. Noise addition is also able to increase the generality of an ASR model by masking random components of the input spectrum. Our noise signals are recorded in subway stations, automobiles, restaurants, bus stations, and everyday home environments including various home appliances. We apply signal-to-noise (SNR) factors in decibels between 5 and 15.

2.4 Reverberation

Reverberation is a complex acoustic phenomenon caused by diverse time delays while sound wave propagates indoors. A soundwave travels in numerous spatial paths sometimes it is reflected or absorbed by the surfaces of objects in the space. Reverberation is noticeable when the sound source stops but the reflections continue. Reverberation causes acoustic distortion that makes speech recognition difficult. Therefore, it needs to be overcome for real-world speech recognition applications. In this case, ASR significantly benefits from an increased amount of training dataset with room impulse responses (RIR). RIRs are a set of filters that represent a given acoustic situation

when a sound wave propagates from a source to a microphone. However, capturing real-life impulse responses is expensive and time-consuming work. Therefore, researchers have been attracted to computational approaches that simulate RIRs. Traditionally, RIRs have been estimated based on the numerical solutions of the wave equation which are computationally very expensive. Recently, low computational approaches that allow online generation during training have been proposed for audio data augmentation in machine learning applications [5,6]. Among the online RIR generators, we adopt the stochastic RIR generation method [7].

2.5 Hyper-and Hypo- Articulated Speech Synthesis

Among various speech data augmentation approaches, hyper-and hypo-articulated speech synthesis is a method focused on the variability of speech depending on the different degrees of articulation [8]. The main idea is related to acoustic-phonetic studies on articulation variants and this is based on traditional signal processing techniques. From a neutral style speech dataset, this approach modifies the characteristics of speech such as amplitude, pitch, harmonics, formant, duration, spectral tilt, speech rate, and articulatory movement. Therefore, this approach can control the parameters related to the aforementioned speech attributes and the synthesized signal will likely include the effects of the speed/volume perturbation at the same time. In particular, this method is proven to be helpful in recognizing a conversational speech [8].

2.6 Speech Enhancement

The aim of speech enhancement is to maximize the perceptual quality of speech signals in ambient noise environments. In most enhancement approaches, it is done by removing background noise components. Speech enhancement has been attracting much research interest since there are many applications such as audio and video calls, hearing aids, and communication devices. Like many other research fields, in speech enhancement, there also is a significant improvement in the aspects of PESQ, STOI, SNR, MOS, and so on by adopting deep learning algorithms.

As mentioned before, noisy signals are enhanced by suppressing noise components from the input. Throughout this process, some speech components can be removed when they are overlapped with noise. This undesired algorithm concept often results in the distortion of original speech signals. This distortion becomes severe when signal-to-noise ratios are low. In this study, by augmenting enhanced data, we want to show that applying a traditional speech enhancement technique to a speech recognition system is similar to time/frequency masking effects. We apply the denoiser in [9] for speech data augmentation.

2.7 Husky Voice Synthesis

In this study, we want to directly manipulate excitation signals while maintaining original formants. To do so, white Gaussian noise is added to excitation signals after linear predictive coding (LPC) of speech signals. And then we resynthesize a speech signal. As a result, we can obtain a husky voice. The purpose of this process is to get diverse speech signals by scattering harmonic amplitudes of speech.

2.8 Voice Conversion based on Cycle-GAN

Voice conversion (VC) is a technique that transforms the non/para-linguistic information of a given speech to another while preserving the linguistic information. This technique can be found in entertainment and speaking-aid systems, and ASR systems. Recently, data augmentation using a cycle-consistent generative adversarial network (Cycle-GAN) has been applied to map two different audio characteristics [10, 11]. The main advantage of Cycle-GAN is to make many-to-many VC possible in an unsupervised manner. So, it doesn't require parallel speech data sets. In this study, we adopt Cycle-GAN-VC2 to convert a given close talk to a distant speech [12]. In this study, we convert only 20 % of the training data.

3. EXPERIMENTS

3.1 Baseline System

We exploit ESPnet for end-to-end speech recognition based on transformer architecture [13]. 1,100 hours of speech data are prepared for training and 11 hours for validation. Both the training and the validation don't include spontaneous speech data and our test sets are composed of speech data in

different domains that represent diverse talking styles such as far-field speech, broadcast drama, entertainment, call-center consultation, and so on. Our test sets are follows.

1. bcast: broadcast news interview speech, 891 files
2. debate: Korea university students' debate speech, 1,308 files
3. present: academic presentation by Korea university students, 1,184 files
4. array: array microphone, simulated conversation, 3,617 files
5. ai: multi-channel acoustic beam-former, simulated conversation, 1,951 files
6. pin: pin microphone, simulated conversation, 2,572 files
7. stand: stand microphone, simulated conversation, 2,246 files
8. call-center: 16kHz up-sampled telephony speech, 133 files
9. drama: broadcast drama program recordings with background sound effect, 260 files
10. ent: broadcast entertainment program recordings with background sound effect, 260 files

It is difficult to distinguish speaking styles in 1-10 but all the test sets are spontaneous talks whether their situations are simulated or not. Speakers in 1-8 try to talk their intention or make a claim to strangers in a natural way and most of the emotions are neutral. Unlike the others, the channel environment in 8 is telephony. Distance from a receiver to a speaker is in order of array > ai > pin > stand and the reverberation gets worse as distance increases. Although all the categories of speech are conversational, the emotional states in 9 and 10 are very different from the others. Therefore, the variations of speaking styles in 9 and 10 are not easy to be recognized by ASR systems. Due to the nature of the broadcast program, background sound effects are also included in 9-10.

Table 1 – ASR results on relatively clean speech data (WER, %)

	bcast	debate	present	array	ai	pin	stand
No-Aug	11.9	14.6	10.8	32.5	23.4	15.8	12.1
Speed(3xfold)	11.0	14.0	10.7	27.7	21.1	14.3	11.3
Volume(2xfold)	11.5	13.6	10.2	23.3	21.9	14.1	11.5
Add noise(2xfold)	10.9	12.6	**9.2**	20.8	20.1	**11.5**	**10.6**
Reverb(2xfold)	11.3	**12.1**	9.9	**17.3**	**19.4**	13.2	11.3
Articulation(3xfold)	11.1	12.3	9.7	21.0	20.7	12.6	11.1
Enhance(2xfold)	**10.8**	12.5	9.5	20.4	19.9	12.6	11.1
Husky(2xfold)	11.2	14.0	10.4	26.0	23.3	14.6	12.1
VC(1.2xfold)	11.5	14.3	10.7	27.0	21.0	15.5	12.0

(Nxfold) indicates that training data are increased by N times after data augmentation. As shown in Table 1, noise addition, reverberation, and speech enhancement approaches are effective in performance improvement in the case of relatively clean speech data sets. In the case of distant talks, data augmentation by RIR filters is also helpful to the robustness of the ASR model. As shown in Table 1, data augmentation approaches such as noise addition, reverberation, and speech enhancement are relatively effective in recognizing clean speech signals. Voice conversion based on Cycle-GAN doesn't perform well in far-field speech recognition because the conventional Cycle-GAN is not enough to convert a source signal's detailed styles and acoustic characteristics to the corresponding target.

Table 2 – ASR results on different style and channel (WER, %)

	Call-center	Drama	Ent
No-Aug	43.5	48.7	62.1
Speed(3xfold)	40.0	51.5	65.1

Volume(2xfold)	40.2	50.2	63.5
Add noise(2xfold)	38.3	44.6	57.3
Reverb(2xfold)	42.4	45.9	62.7
Articulation(3xfold)	39.2	48.7	65.4
Enhance(2xfold)	**38.2**	**42.7**	**57.1**
Husky(2xfold)	40.0	49.6	65.6
VC(1.2xfold)	41.8	49.3	63.5

Table 2 shows the experimental result in mismatched conditions. All the data are recorded from telephony call-center, broadcast drama, and entertainment programs. And the data recorded from broadcast programs include background sound effects. As shown in Table 2, the speech enhancement-based segmentation approach is slightly better than the noise addition approach. Sometimes data augmentation doesn't work for improving ASR performance when the speaking style in the test condition (i.e. broadcast entertainment programs) is much different from the training. As shown in Table 2, we can see that the speaking styles in broadcast entertainment programs are more diverse and exaggerated than the others. We think that this is related to the emotional state of speakers.

Table 3 – ASR results on Wiener filtered speech data (WER, %)

	bcast	debate	present	array	ai	pin	stand
No-Aug	12.4	15.9	11.3	36.2	23.5	17.1	12.4
Speed(3xfold)	11.6	16.1	11.0	34.2	22.5	16.5	11.8
Volume(2xfold)	12.1	14.0	11.0	25.1	22.5	14.0	11.7
Add noise(2xfold)	11.2	12.8	**9.6**	22.3	19.6	**11.8**	10.8
Reverb(2xfold)	11.6	12.7	10.4	**19.2**	20.1	13.4	11.5
Articulation(3xfold)	11.6	12.7	10.1	23.7	21.1	13.2	11.5
Enhance(2xfold)	**11.0**	**12.5**	9.7	21.8	**19.6**	12.0	**10.6**
Husky(2xfold)	11.6	12.7	10.1	23.7	21.1	14.4	12.4
VC(1.2xfold)	12.4	16.4	11.2	32.0	21.8	18.3	12.6

We apply a conventional Wiener filter to perturb speech signals while maintaining linguistic information. Table 3 shows that the performance of a E2E ASR system is affected by these tiny changes in speech signals.

Table 4 – ASR results on adversarial speech data (WER, %)

	bcast	debate	present	array	ai	pin	stand
No-Aug	12.1	20.9	11.7	50.4	26.5	28.8	17.0
Speed(3xfold)	11.2	19.3	10.8	45.3	25.3	26.3	15.1
Volume(2xfold)	11.6	18.8	11.4	37.7	24.4	22.6	14.4
Add noise(2xfold)	10.9	15.1	**10.1**	**26.4**	21.2	**15.3**	12.0
Reverb(2xfold)	11.1	15.4	10.8	27.7	21.2	21.1	14.2
Articulation(3xfold)	11.0	16.2	10.5	34.3	22.4	19.7	13.7
Enhance(2xfold)	**10.8**	**14.5**	**10.1**	26.5	**21.0**	15.5	**11.9**
Husky(2xfold)	11.4	18.2	11.1	36.3	25.5	20.4	14.2

VC(1.2xfold)	12.2	21.0	10.9	44.4	25.5	26.8	15.6

Table 4 shows that recognition performance changes after adding a small amount of white noise to speech signals. Although the linguistic information in the test sets is slightly affected by the tiny noise, the changes in performance are not negligible. As shown in Tables 3 and 4, the ASR model trained by enhanced speech data augmentation is more robust to adversarial samples of speech.

Table 5 – ASR results on simulated noisy speech data (WER, %)

	bcast	debate	present	array	ai	pin	stand
No-Aug	41.3	62.7	48.2	68.1	57.7	48.5	48.3
Speed(3xfold)	40.1	62.3	47.0	64.6	53.8	45.9	46.5
Volume(2xfold)	39.1	59.5	44.7	63.3	54.2	45.4	45.6
Add noise(2xfold)	**20.4**	**36.4**	**23.4**	**42.0**	**35.3**	**24.0**	**23.9**
Reverb(2xfold)	36.1	54.5	42.5	54.3	48.1	42.4	43.1
Articulation(3xfold)	37.3	56.7	42.1	61.8	53.2	43.2	43.2
Enhance(2xfold)	20.7	36.8	23.9	42.6	35.9	24.2	24.2
Husky(2xfold)	37.8	58.4	43.8	61.5	53.7	44.2	44.7
VC(1.2xfold)	43.1	63.5	49.0	64.5	55.5	48.4	48.5

Table 5 shows the ASR results in noisy conditions. All the test sets are obtained by artificially adding noise sources to original speech signals. However, the noise sources (i.e. sound effects in broadcast drama and entertainment programs) are not included in the training phase for correct evaluations.

Table 6 – ASR results on simulated reverberant speech data (WER, %)

	bcast	debate	present	array	ai	pin	stand
No-Aug	27.3	34.4	23.7	54.2	47.1	31.7	27.8
Speed(3xfold)	26.5	32.1	21.1	45.7	40.8	26.9	24.8
Volume(2xfold)	22.3	26.8	19.1	40.7	38.9	25.3	21.2
Add noise(2xfold)	20.7	29.3	19.6	37.8	37.5	24.0	19.7
Reverb(2xfold)	**13.6**	**15.3**	**12.5**	**22.4**	**25.2**	**16.2**	**13.8**
Articulation(3xfold)	20.6	24.0	18.0	39.1	36.8	22.6	19.2
Enhance(2xfold)	20.5	30.6	19.5	37.4	37.8	23.5	19.7
Husky(2xfold)	21.3	26.1	19.3	40.3	39.0	24.7	21.5
VC(1.2xfold)	27.8	30.9	20.9	44.0	39.1	27.6	24.6

Table 6 indicates the ASR results in simulated reverberant conditions. In this case, the data augmentation approach based on RIR filters shows better recognition performance.

Table 7 – ASR results on relatively clean speech data (WER, %)

	bcast	debate	present	array	ai	pin	stand
No-Aug	11.9	14.6	10.8	32.5	23.4	15.8	12.1
Proposed(6xfold)	10.4	11.2	9.7	15.5	18.2	11.5	10.3

Proposed+SpecAug(6xfold)	**9.8**	**9.7**	**8.9**	**13.7**	**15.4**	**10.0**	**9.3**

We combine efficient data augmentation approaches (i.e. additive noise, RIR filtering, hyper- and hypo-articulated speech synthesis, and husky voice synthesis) for an ASR model. As a result, the proposed training data are 6 times larger than the original. Table 7 shows that a more generalized ASR model can be trained by the proposed method.

Table 8 – ASR results on different style and channel (WER, %)

	Call-center	Drama	Ent
No-Aug	43.5	48.7	62.1
Proposed(6xfold)	39.3	42.8	58.4
Proposed+SpecAug(6xfold)	**31.8**	**40.0**	**54.1**

People make a phone call to the call center usually when they are in trouble. Broadcasters and drama actors also express their emotions when they speak. So, their speaking styles are far from neutral. However, it is also seen in Table 8 that the proposed method is also helpful for speech recognition with different speaking styles and acoustic channels.

Table 9 – ASR results on Wiener filtered speech data (WER, %)

	bcast	debate	present	array	ai	pin	stand
No-Aug	12.4	15.9	11.3	36.2	23.5	17.1	12.4
Proposed(6xfold)	10.7	11.6	10.0	16.6	17.4	11.6	10.8
Proposed+SpecAug(6xfold)	**9.8**	**10.0**	**9.3**	**14.7**	**15.5**	**10.4**	**10.1**

Table 10 – ASR results on adversarial speech data (WER, %)

	bcast	debate	present	array	ai	pin	stand
No-Aug	12.1	20.9	11.7	50.4	26.5	28.8	17.0
Proposed(6xfold)	10.4	12.9	9.7	22.4	19.6	14.6	11.7
Proposed+SpecAug(6xfold)	**9.8**	**11.1**	**9.1**	**19.3**	**16.7**	**12.8**	**10.5**

Wiener filtering on a relatively clean signal makes an artifact but ordinary people are hard to recognize it. Adversarial speech samples are obtained by adding a small amount of Gaussian white noise. Therefore, there is no problem with speech recognition by a human. However, these changes in speech signals affect the performance of an E2E ASR system. As shown in Tables 9 and 10, the proposed data augmentation is also robust to tiny changes in speech signals.

Table 11 – ASR results on simulated noisy speech data (WER, %)

	bcast	debate	present	array	ai	pin	stand
No-Aug	41.3	62.7	48.2	68.1	57.7	48.5	48.3
Proposed(6xfold)	20.1	35.6	23.9	39.8	34.7	24.5	24.0
Proposed+SpecAug(6xfold)	**17.5**	**32.5**	**20.9**	**35.5**	**30.2**	**21.2**	**21.1**

As shown in Table 11, the performance of an ASR system is drastically corrupted in the case of simulated noisy conditions without data augmentation. It is obvious that the proposed data augmentation+SpecAug is helpful. But, the recognition performance is not enough to satisfy commercial requirements. So, automatic speech recognition in the ambient noise condition still remains in challenging areas.

Table 12 – ASR results on simulated reverberant speech data (WER, %)

	bcast	debate	present	array	ai	pin	stand
No-Aug	27.3	34.4	23.7	54.2	47.1	31.7	27.8
Proposed(6xfold)	12.9	14.7	11.7	21.3	24.3	15.3	12.8
Proposed+SpecAug(6xfold)	**11.6**	**12.9**	**10.8**	**18.1**	**20.2**	**13.0**	**11.3**

Table 12 shows that the reverberation issue can be handled by the proposed data augmentation method.

4. CONCLUSIONS

In this paper, we evaluate various data augmentation approaches for ASR and a useful augmentation combination is suggested. We hope that our test results are helpful to predict performance gain in various ASR tasks when each data augmentation approach is applied. In the future, we are going to focus on voice conversion for realistic conversion including more detailed acoustic characteristics of target speech since conventional voice conversion is not enough to help the performance of an ASR.

REFERENCES

1. Karita S, Chen N, Hayashi T, Hori T, Inaguma H, Jiang Z, Someki M, Soplin N. E. Y, Yamamoto R, Wang X et al. A comparative study on transformer vs RNN in speech applications. Proc ASRU Workshop 2019; 14-18 December 2019; Sentosa, Singapore 2019. p. 449–456
2. Gulati A, Qin J, Chiu C -C, Parmar N, Zhang Y, Yu J, Han W, Wang S, Zhang S, Wu Y et al. Conformer: Convolution-augmented transformer for speech recognition. Proc Interspeech2020; 25-29 October 2020; Shanghai, China 2020. p. 5036–5040.
3. Ko T, Peddinti V, Povey D, and Khudanpur S. Audio augmentation for speech recognition. Proc Interspeech2015; 6-10 September 2015; Dresden, Germany 2015. p. 3586-3589.
4. Park D. S, Chan W, Zhang Y, Chiu C.-C, Zoph B, Cubuk E. D, and Le Q. V. SpecAugment: A simple data augmentation method for automatic speech recognition. Proc Interspeech2019; 15-19 September 2019; Graz, Austria 2019. p. 2613-2617
5. Habets E. RIR-Generator. [Online]. Available: https://github. com/ehabets/RIR-Generator
6. Scheibler R, Bezzam R, and Dokmanic R. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. Proc ICASSP 2018, 15-20 April 2018; Calgary, Alberta, Canada 2018. p. 351-355
7. Masztalski P, Matuszewski M, Piaskowski K, Romaniuk M. StoRIR: Stochastic Room Impulse Response Generation for Audio Data Augmentation. Proc Interspeech 2020, 25-29 October 2020; Shanghai, China 2020. p.2857-2861
8. Lee S J, Kang B, Chung H, Park J G, Lee Y K, Hypo and Hyperarticulated Speech Data Augmentation for Spontaneous Speech Recognition. Proc EUSIPCO 2018, 3-7 September 2018; Rome, Italy 2018. p. 2080-2084
9. Defossez A, Synnaeve G, Adi Y. Real Time Speech Enhancement in the Waveform Domain. Proc Interspeech 2020, 25-29 October 2020; Shanghai, China 2020. p. 3291-3295
10. Tsunoo E, Shibata K, Narisetty C, Kashiwagi Y, Watanabe S. Data Augmentation Methods for End-to-End Speech Recognition on Distant-Talk Scenarios. Proc Interspeech 2021, 30 August -3 September 2021; Brno, Czechia 2021. p.301-305
11. Singh D, Amin P, Sailor H, Patil H. Data Augmentation Using CycleGAN for End-to-End Children ASR. Proc EUSIP 2021, 23-27 August 2021; Dublin, Ireland, 2021. p. 511-515
12. Kaneko T, Kameoka H, Tanaka K, Hojo N. CYCLEGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion. Proc ICASSP 2019, 12-17 May 2019; Brighton, UK 2019. p. 6820-6824
13. Watanabe S, Hori S, Karita S, Hayashi T, Nishitoba J, Unno Y, Enrique Yalta Soplin N, Heymann J, Wiesner M, Chen N, Renduchintala A, Ochiai T. ESPnet: End-to-End SPeech Processing Toolkit. Proc Interspeech 2018, 2-6 September 2018, Hyderabad, India 2018. p.2207-2211

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0059

A Data-Augmented Transfer Learning Method for the Speech Recognition in Domains with Sparse Speech Data

Byung Ok KANG[1]; Hyeong Bae JEON[2]

[1,2] Electronics and Telecommunications Research Institute, Korea

ABSTRACT

In this paper, we propose a data-augmented transfer learning method for the purpose of improving the performance of speech recognition in the domains with sparse training data where it is difficult to collect large amounts of labeled/unlabeled speech data. As the first step, the proposed method augments speech corpus of acoustic characteristics such as speaker and channel/noise environment similar to that of the target domain using speech corpus of other domains that is relatively easy to collect speech data for training. Next, it performs the proposed transfer learning in the form of combination of self-training and teacher/student learning using source/augmented speech corpus with input. For evaluation, we performed experiments on the AMI corpus task and the call-center speech-to-text (STT) task, and the proposed approach outperformed the existing teacher/student-based transfer learning method.

Keywords: Speech recognition, Transfer learning, Data augmentation

1. Introduction

The speech recognition system can be expected to achieve optimal performance when it is trained on large-capacity data that reflects the matched acoustic characteristics of speakers using the service and the noise and channel characteristics that match the service environment.

However, there are speech recognition services that are difficult to acquire large amounts of learning data. For example, there are speech recognition services in which it is difficult to collect speech data matching the acoustic characteristics of a speaker's utterance in a large amount, such as speech recognition of a specific language targeting non-native speakers who are limited to a small number compared to native speakers. As another example, there is a case where the collection of a large amount of speech data is limited due to a problem of personal information security, such as speech recognition for speech to text in a call-center domain.

Various approaches have been studied to deal with these difficulties. The most representative studies are approaches such as domain adaptation and multi-task learning. Domain adaptation approach is based on a source model trained on speech data collected from various environments, adapting to a target domain with target task matching the domain [1-4]. Multi-task learning approach based on the semi-supervised learning linearly combines cost functions of a deep classifier and a deep auto-encoder, and then minimizes the combined cost combination [5,6]. Domain adaptation methods are widely applied and have the ability to improve stability and performance. However, to obtain satisfactory performance improvement, most domain adaptation approaches require significant amounts of domain speech data along with transcription. Multi-task learning methods based on semi-supervised learning have no additional cost for transcription of target domain speech data, but have the disadvantage of requiring a significant amount of un-transcribed speech data in target domain.

To handle these problems, this work proposes a transfer learning method for speech recognition in domains with sparse speech data. The transfer learning method proposed in this paper is a method for speech recognition in a domain where it is difficult to collect large volumes of speech data, such as non-native speaker speech recognition and call-center speech recognition. For speech recognition in domains where it is difficult to collect large-capacity speech data, a native speaker's speech data and broadcast speech data, which are relatively easy to acquire large-capacity speech data, are used. In

[1] bokang@etri.re.kr
[2] hbjeon@etri.re.kr

the proposed method, training speech data having acoustic characteristics close to the speaker, noise, and channel environment similar to the target sparse data domain is augmented from native speaker's speech and broadcast speech data, and the obtained target domain augmented speech data and large-capacity speech corpus are used as the input training data to perform teacher/student-based transfer learning.

This paper is organized as follows. Section 2 describes our proposed method in detail. Section 3 describes the experimental setting and results to verify the proposed method. Finally, Section 4 draws our conclusions of this study.

2. Proposed Method

Figure 1 is the graphical illustration explaining the data augmentation step for the sparse data area in transfer learning proposed in this paper [7,8].

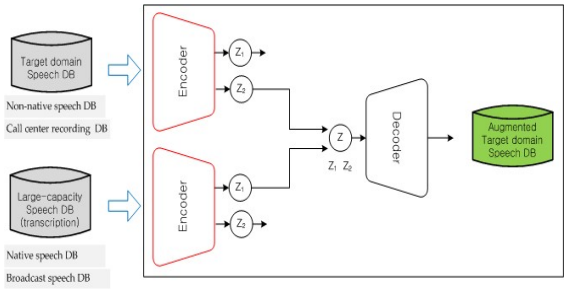


Figure 1 – Graphical illustration of the step of data augmentation

The basic structure is identical to that of a variable autoencoder, which consists of an encoder that infers latent variables and a decoder that generates the same output speech as the input based on the latent variables. The variable autoencoder applied in this paper has a graph structure composed of properties that vary in short units such as phonetic/linguistic content information and properties that vary in units of entire utterances such as channel/noise environment and speaker information. It differs from a typical auto-encoder in that it is possible to infer each attribute separately [9]. While maintaining the necessary attributes such as phonetic/linguistic content information of the transcribed speech corpus that can be collected in a large capacity, by replacing other attributes with the channel/noise environment and speaker information attributes inferred from the speech data from the target domain with sparse speech data, the target sparse data region of voice data is augmented to a large capacity.

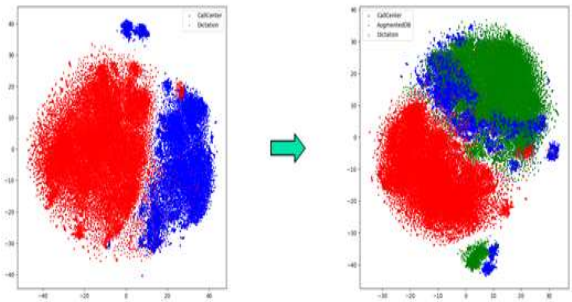


Figure 2 – Examples of data augmentation into the domain with sparse speech data

Figure 2 shows the results of augmenting speech data through attribute separation and substitution with broadcast and call-center speech data as inputs [7,8]. The distribution in the acoustic space composed of channel/noise environment and speaker information properties is visually shown through t-distributed stochastic neighbor embedding (t-SNE). In Figure 2, red is broadcast speech data, blue is call-center speech data, and green is augmented speech data. It can be confirmed that the augmented speech data is distributed in the same acoustic space of the call-center speech data composed of the same channel/noise environment and speaker information properties while having contents information of the broadcast speech data.

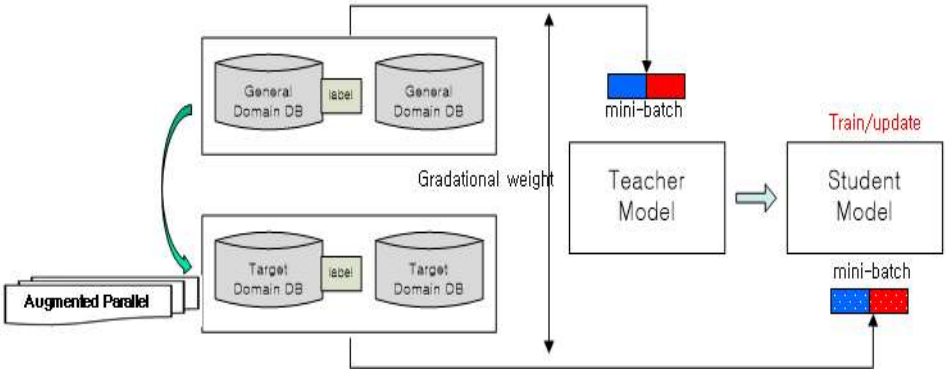


Figure 3 – The block diagram of the proposed transfer learning method

Figure 3 shows the block diagram of the proposed transfer learning method for the speech recognition in domains with sparse speech data. First, pre-training is performed by applying a small amount of training data of the target domain. Next, by the method described above, the target domain speech data having similar acoustic characteristics as that of a non-native speakers and call-center speech recognition is augmented from the transcribed native speaker's speech and broadcast speech data. In the proposed transfer learning, a sophisticated longitudinal speech recognition system trained with a large-capacity transcribed speech corpus in the universal domain serves as the teacher, and the speech recognition system in the target sparse data domain serves as a student. Learning is performed by inputting transcribed and un-transcribed speech data from the universal domain and speech data from the target sparse data domain augmented therefrom, respectively. The knowledge transfer method in teacher/student learning is a knowledge distillation learning method that learns by transferring the posterior distribution of the end of each system [10].

3. Experimental Results

3.1 AMI meeting corpus task

In order to verify the proposed method, we evaluated the performance on the ASR task on the AMI meeting corpus composed of the speech data with near/far channels for each speaker's utterance. The variable auto-encoder was trained using the AMI meeting corpus, and through this, the speech data with the channel characteristics of the long-distance channel speech data (Single Distance Microphone 1, SDM1) was augmented, while having the contents information of the short-distance channel speech data (Independent Headset Microphones, IHM) which is capable of being collected with large amounts. The model trained by the proposed transfer learning method with data augmentation was compared with a model trained only with far-channel voice data (SDM1), and a model trained with a conventional teacher/student-based transfer learning method using short-channel and far-channel speech data as inputs. We used ESPnet, an open source end-to-end ASR platform [11] to train the end-to-end ASR in all of the experiments. Table 1 shows the recognition performance results for the evaluation set (Eval set) and the development set (Dev set) of the far-channel speech data (SDM1). The performance indicated in Table 1 corresponds to the character error rate.

Table 1 – Comparison of character error rate (%) on the AMI meeting corpus task

Model	Eval set	Dev set
SDM1 only	47.6	44.8
Conventional T/S	45.0	42.1
Proposed T/S	41.3	38.2
Error Reduction Rate	13.2	14.7

As confirmed in Table 1, the model trained by the teacher/student-based transfer learning method with the short-distance channel speech data (IHM) and the long-distance channel speech data (SDM1)

shows improved performance compared to the model trained only with the far channel voice data (SDM1) [8]. When the proposed transfer learning method with data augmentation is applied, it can be seen that there is an additional performance improvement of 13.2% and 14.7% in error reduction rates in the evaluation set and development set, respectively, compared to the conventional teacher/student-based transfer learning.

3.2 Call-center STT task

In order to verify the proposed method to the ASR on the domain sparse speech data, we evaluated the performance on the ASR task on the call-center STT task. First, 500 hours of call-center data were applied for pre-learning. Thereafter, paired data for teacher-student learning was generated by augmenting target domain speech data through attribute separation and substitution using 3,000 hours of broadcast speech data. The student model obtained through pre-learning is file-tuned by the sophisticated teacher model trained with a large-capacity transcribed speech data using broadcast speech data and sparse speech domain data augmented based on attribute separation and substitution as input training data. From Table 2, it can be seen that there is a performance improvement of 14.6% in the error reduction rate compared to the performance obtainable through pre-learning.

Table 2 – Comparison of character error rate (%) on the call-center STT task

Model	Eval set
Pre-trained	31.5
Proposed T/S	26.9
Error Reduction Rate	14.6

4. Conclusions

In this paper, we propose a transfer learning method to improve speech recognition performance in the domain where training data is sparse, where it is difficult to collect a large amount of labeled/unlabeled speech data. For evaluation, experiments were performed on the AMI corpus task and the call-center speech-to-text (STT) task, and the proposed approach improved performance by 13.2% and 14.6% in the evaluation set compared to the conventional teacher/student-based transfer learning method and pre-trained model.

ACKNOWLEDGEMENTS

This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (2019-0-00004, Development of semi-supervised learning language intelligence technology and Korean tutoring service for foreigners)

REFERENCES

1. Sun S, Zhang B, Xie L., An unsupervised deep domain adaptation approach for robust speech recognition. Neurocomputing 2017, 257, 79–87.
2. Asami T, Masumura R., Yamaguchi Y, Masataki H, Aono Y., Domain adaptation of dnn acoustic models using knowledge distillation. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5185–5189.
3. Meng, Z.; Li, J.; Gong, Y.; Juang, B.-H. Adversarial teacher-student learning for unsupervised domain adaptation. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5949–5953.
4. Meng Z, Li J, Gaur Y, Gong Y., Domain adaptation via teacher-student learning for end-to-end speech recognition. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Sentosa, Singapore, 14–18 December 2019; pp. 268–275.
5. Ranzato M, Szummer M., Semi-supervised learning of compact document representa tions with deepnetworks. In Proceedings of the 25th International Conference on Machine learning. ACM, Helsinki, Finland, 5–9 July 2008; pp. 792–799.
6. Dhaka A.K, Salvi G., Sparse autoencoder based semi-supervised learning for phone classification with

limited annotations. In Proceedings of the GLU 2017 International Workshop on Grounding Language Understanding, Stockholm, Sweden, 25 August 2017; pp. 22–26.
7. B. O. Kang, H. B. Jeon, J. G. Park, Speech recognition for task domains with sparse matched training data. Applied Sciences 10.18 (2020): 6155.
8. B. O. Kang, H. B. Jeon, J. G. Park, A Study on Transfer Learning Method for Speech Recognition in Domains with Sparse Speech Data. In Proceedings of the Winter Annual Conference of KICS, Kangwon, Korea, 3–5 February 2021.
9. Hsu, Wei-Ning, Yu Zhang, and James Glass., Unsupervised learning of disentangled and interpretable representations from sequential data. Advances in neural information processing systems. 2017.
10. Li. Jinyu, et al., Large-scale domain adaptation via teacher-student learning. arXiv preprint arXiv:1708.05466 (2017).
11. S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, E. Soplin, J. Heyman, M. Wiesner, T. Ochiai, et al. ESPnet: End-to-End Speech Processing Toolkit. arXiv 2018, arXiv:1804.00015. Available online: https://arxiv.org/abs/1804.00015 (accessed on 8 August 2020).

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0154

End-to-End ASR semi-supervised training using adversarial self-training

Hoon CHUNG[(1)], Yoonhyung KIM[(2)], Byung-Ok KANG[(3)]

[(1)]Electronics and Telecommunications Research Institute, Daejeon, Korea, hchung@etri.re.kr

[(2)]Electronics and Telecommunications Research Institute, Daejeon, Korea, yhkim1127@etri.re.kr

[(3)]Electronics and Telecommunications Research Institute, Daejeon, Korea, bokang@etri.re.kr

ABSTRACT

This paper proposes an interleaved self-training using adversarial augmentation for semi-supervised end-to-end automatic speech recognition (ASR) model training. As a method to use unlabelled speech corpora for ASR model training, a consistency regularized self-training is a common approach, which uses the model's highly confident predictions of weakly augmented data as target labels for strongly augmented versions of the same data. It means that augmentation and training strategy are important, and it is desirable to sample the augmented data around the decision boundary in classification problems. Even though there are various speech augmentations techniques such as speed perturbation, noise addition, channel distortion and so on, these methods are not directly concerned with generating augmented data around decision boundaries. Therefore, to handle the issue, we propose to use adversarial augmentation which generates examples misclassified by a model, and we also investigate batch-wise interleaved training strategy to prevent ASR model overfitted to unlabelled data. The proposed approach was evaluated on the Wall Street Journal task domain. The experimental results show that the proposed method is effective by reducing the character error rate from 10.4% to 6.8%.

Keywords: Automatic speech recognition, Semi-supervised learning, Consistency regularization, Adversarial example

1 INTRODUCTION

Automatic speech recognition (ASR) systems using end-to-end models have recently become popular because of their simplicity and state-of-the-art performance. They can integrate separate acoustic, pronunciation, and language models into a single neural network [2, 7, 6], and outperform conventional ASRs in certain general tasks [7]. Despite these models' popularity, there are some problems for practical use. One of these problems is a shortage of labeled training corpora. A large amount of labeled data is necessary for end-to-end ASR systems to achieve high performance [1, 22, 12, 3, 13]. However, it is expensive and time consuming job to collect a large amount of labeled speech corpus, whereas it is cheap and easy to collect unlabeled speech corpus in public. Therefore, to handle the shortage of labeled corpora, semi-supervised training approaches have been actively conducted as a way to exploit unlabeled corpora.

Although various semi-supervised end-to-end ASR training methods have been proposed, the following approaches are related to our proposed method. The first approach is pseudo-labeling or self-training, which generates machine transcriptions for unlabeled speech data using a pre-trained ASR systems [14, 31, 16, 32, 15]. The second approach consists of methods that use multi loss minimization. For example, shared encoder loss is composed of cross-entropy loss for labeled data and reconstruction loss for unpaired speech/text data, and divergence loss for embedding space between speech and text [13], and cycling loss is composed of cross-entropy loss and reconstruction loss by integrating ASR and Text-to-Speech (TTS) or Text-to-Encoder (TTE) [12, 27]. The third approach is based on reinforcement learning (RL). The RL-based method focuses on rewarding an end-to-end ASR to generate more sensible sentences for both labeled and unlabeled speech data [11]. The fourth approach is consistency regularization which uses highly confident pseudo-labels of weakly augmented data as target for strongly augmented version of the same data [5, 4, 25].

This work was based on consistency regularization and our contributions can be summarized as follows:

- We use adversarial augmentation instead of domain-dependent strong augmentation
- We generate k-best pseudo-labels for original unlabeled data instead of using weakly augmented data
- We use interleaved training strategy.

The rest of this paper is organized as follows. Section 2 briefly describes semi-supervised end-to-end ASR. Section 3 reviews the consistency regularization. Section 4 details our proposed approach. Section 5 presents the experimental setting and results. Section 6 concludes the paper and discusses future works.

2 SEMI-SUPERVISED END-TO-END ASR

In this section, we briefly review end-to-end ASR systems based on an encoder-decoder architecture and general semi-supervised training problems.

2.1 end-to-end ASR

In this work, an end-to-end ASR system is composed of an encoder-decoder architecture. The system estimates the posterior probability $P_\theta(\mathbf{y}|\mathbf{x})$, where $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ is a sequence of input feature vectors, $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$ is a sequence of output characters, and θ denotes the model parameters. The posterior probability $P_\theta(\mathbf{y}|\mathbf{x})$ is factorized as follows:

$$P_\theta(\mathbf{y}|\mathbf{x}) = \prod_{n=1}^{N} P_\theta(y_n|y_{1:n-1}, \mathbf{x}) \tag{1}$$

where $y_{1:n-1}$ is the sub-sequence $\{y_1, y_2, \ldots, y_{n-1}\}$, and $P_\theta(y_n|y_{1:n-1}, \mathbf{x})$ is calculated by the encoder-decoder network as follows [12, 19]:

$$\mathbf{H} = \text{enc}(\mathbf{x}) \tag{2}$$

$$\mathbf{c}_n = \text{Attention}(\mathbf{a}_{n-1}, \mathbf{H}) \tag{3}$$

$$y_n = \text{dec}(\mathbf{c}_n, y_{1:n-1}) \tag{4}$$

where $\mathbf{H}$ is a sequence of state vectors of the encoder's top layer for a given $\mathbf{x}$, $\mathbf{a}_n$ is an attention weight vector, and $\mathbf{c}_n$ is a context vector integrated using all encoder outputs $\mathbf{H}$ by the attention mechanism. In recognition, inference is conducted through beam search using an external language model, $p_{LM}(\mathbf{y})$, as follows [28, 34]:

$$\hat{\mathbf{y}} = \text{argmax}_{\mathbf{y}} \log P_\theta(\mathbf{y}|\mathbf{x}) + \beta \log P_{LM}(\mathbf{y}) \tag{5}$$

where β is the language model scale.

2.2 Semi-supervised ASR training problem

The semi-supervised end-to-end ASR training problem is tackled with a general optimization problem to find model parameters $\hat{\theta}$, that minimizes the loss function, $\mathscr{L}(\theta)$, for given labeled data, $(\mathbf{X}_l, \mathbf{Y}_l)$, and unlabeled speech data, $(\mathbf{X}_u)$, and text data $(\mathbf{Y}_u)$, as follows:

$$\hat{\theta} = \text{argmin}_\theta \mathscr{L}_s(\mathbf{X}_l, \mathbf{Y}_l; \theta) + \gamma \mathscr{L}_u(\mathbf{X}_u, \mathbf{Y}_u; \theta) \tag{6}$$

where $\mathscr{L}_s(\mathbf{X}_l, \mathbf{Y}_l; \theta)$ is the labeled data loss, $\mathscr{L}_u(\mathbf{X}_u, \mathbf{Y}_u; \theta)$ is the unlabeled data loss, and γ is a scalar hyper-parameter denoting the relative weight of the unlabeled data loss. In general, cross entropy loss is used for labeled data loss. So, the semi-supervised ASR training problem can be approached by how to define unlabeled data loss $\mathscr{L}_u(\mathbf{X}_u, \mathbf{Y}_u; \theta)$.

3 CONSISTENCY REGULARIZATION

For unlabeled data loss, consistency regularization can be defined as follows:

$$\mathscr{L}_u = \frac{1}{U}\sum_{u=1}^{U} \mathbf{1}(max(q_u \geq \tau))H(\hat{y}_u, P_\theta(y|\mathscr{A}(\mathbf{x}_u)) \tag{7}$$

where $\mathscr{A}()$ is statistically strong augmentation, and $\hat{y}_u$ is pseudo-label for unlabeled data $\mathbf{x}_u$ as follows:

$$\hat{y}_u = \text{argmax}_y(q_u) \tag{8}$$

where q_u is

$$q_u = P_\theta(y|\alpha(\mathbf{x}_u)) \tag{9}$$

In other words, consistency regularization is characterized by how to define weak augmentation $\alpha()$, strong augmentation $\mathscr{A}()$, and the threshold τ.

4 Proposed semi-supervised learning

In this section, we describe the proposed approach in detail. For simplicity, we use $f_\theta(\mathbf{x}) = \text{argmax}_\mathbf{y} \text{log} P_\theta(\mathbf{y}|\mathbf{x})$ in the following.

4.1 Adversarial examples

An adversarial example is an example misclassified by a model, but it is only slightly skewed from the original correctly-classified one [9]. Such examples can be generated by adding some well-designed small perturbations to the original examples as follows:

$$\mathbf{x}_i^{adv} = \mathbf{x}_i + \delta_i \tag{10}$$

$$y_i^* \neq f_\theta(\mathbf{x}_i^{adv}) \tag{11}$$

$$\|\delta_i\| \ll \|\mathbf{x}_i\| \tag{12}$$

In this work, we used a popular method, the fast gradient sign method (FGSM), which is a linear perturbation that adds an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input proposed by [9, 26, 29].

$$\delta_i^{FGSM} = \varepsilon \, \text{sign} \nabla \mathbf{x}_i H(y_i^*, f_\theta(\mathbf{x}_i))$$

Figure 1 shows an adversarial example generated by setting $\varepsilon = 0.5$ for original features whose $\varepsilon = 0.0$.

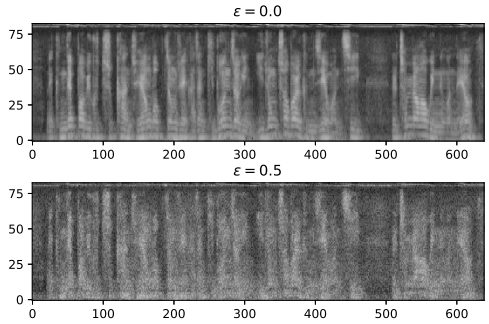


Figure 1. An FGSM example of an 83-dimensional filter bank with pitch features

4.2 Proposed semi-supervised training procedure

Algorithm 1 describes the proposed training procedure. The algorithm consists of pre-training, and interleaved training. Pre-training was performed to build a robust seed model using the labeled data and their adversarial examples scaled by ε_1. Semi-supervised training was then performed in an interleaved way for labeled and unlabeled data [20, 18]. The batch size of the labeled data and unlabeled data was the same. Since the number of unlabeled data batches, N_u, was much larger than that of the labeled ones N_l, we cycled the labeled data. The unlabeled data was processed first to avoid additional fine-tuning for the validation test at the end of each epoch. Standard cross entropy was used as a loss function for both the labeled and unlabeled data. For the labeled data, cross entropy loss was used between ground truth labels and model predictions from the original labeled data. For the unlabeled data, cross entropy loss was also used between highly confident top-k pseudo-labels of the original unlabeled data and adversarial examples from the same unlabeled data.

The proposed algorithm is controlled by five hyper-parameters: ε_1 controls the intensity of adversarial examples for pre-training, τ controls the pseudo-label pruning threshold, γ controls the contribution of unlabeled data loss, ε_2 controls the intensity of adversarial examples for interleaved training, and k controls the number of pseudo labels.

Algorithm 1 Proposed semi-supervised training procedure

Require: A training set $(\mathbf{X}_l, \mathbf{Y}_l), (\mathbf{X}_u)$, θ_0

1. Adversarial pre-training

while `not converged` **do**
  for $i = 1$ to N_l **do**
    Select labeled data $(\mathbf{x}_i, \mathbf{y}_i^*) \in (\mathbf{X}_l, \mathbf{Y}_l)$
    $\theta_{t+1} \leftarrow \theta_t - \alpha_t \nabla_\theta H(\mathbf{y}_i^*, f_\theta(\mathbf{x}_i))$
    $\mathbf{x}_i^{adv} \leftarrow \mathbf{x}_i + \varepsilon_1 \nabla_{\mathbf{x}_i} H(\mathbf{y}_i^*, f_\theta(\mathbf{x}^i))$
    $\theta_{t+1} \leftarrow \theta_t - \alpha_t \nabla_\theta H(\mathbf{y}_i^*, f_\theta(\mathbf{x}_i^{adv}))$
  end for
end while

2. Interleaved semi-supervised training

while `not converged` **do**
  for $i = 1$ to N_u **do**
    Select unlabeled data $\mathbf{x}_i \in (\mathbf{X}_u)$
    $\hat{\mathbf{y}} = \text{argmax}_k P_\theta(\mathbf{y}|\mathbf{x}_i)$
    for $j = 1$ to k **do**
      if $log(P_\theta(\hat{\mathbf{y}}_j|\mathbf{x}_i))/T \geq log(\tau)$ **then**
        $\mathbf{x}_i^{adv} \leftarrow \mathbf{x}_i + \varepsilon_2 \nabla_{\mathbf{x}_i} H(\hat{\mathbf{y}}_j, f_\theta(\mathbf{x}_i))$
        $\theta_{t+1} \leftarrow \theta_t - \alpha_t \gamma \nabla_\theta H(\hat{\mathbf{y}}_j, f_\theta(\mathbf{x}_i^{adv}))$
      end if
    end for
    $j = i \pmod{N_l}$
    Select labeled data $(\mathbf{x}_j, \mathbf{y}_j^*) \in (\mathbf{X}_l, \mathbf{Y}_l)$
    $\theta_{t+1} \leftarrow \theta_t - \alpha_t \nabla_\theta H(\mathbf{y}_j^*, f_\theta(\mathbf{x}_j))$
  end for
end while

5 Experiments

5.1 Settings

We used the Wall Street Journal (WSJ) dataset LDC93S6B and LDC94S13B [21] to evaluate the proposed training approach. The dataset is composed of a small 15-hour (7138 utterances) dataset called si84, and a large 81-hour (37416 utterances) dataset called si284. We used si84 as a labeled dataset and si284 as an unlabeled dataset. We employed the official test dataset dev93 for a hyper-parameter and decoding parameter search and eval92 for performance evaluation. An 83-dimensional filter-bank with pitch features were used as the input feature. The encoder-decoder network utilized location-aware attention [2, 8]. The encoder comprises six bi-directional Long Short Term Memory (LSTM) layers [6, 23, 24] each with 320 units and the decoder comprises one (uni-directional) LSTM layer with 300 units. The cross entropy and Connectionist Temporal Classification (CTC) [1, 10, 17] objective was optimized using AdaDelta [33] with an initial learning rate set to 1.0. The

training batch size was 10 and the number of training epochs was 15. ESPnet [30] is used to implement and execute the experiments.

We pre-trained a seed model with the si84 dataset in an adversarial training manner and then retrain the model with the labeled si84 and unlabeled si284 datasets in a semi-supervised manner. The performance was measured in terms of character error rates (CER) without a language model, and the performance is compared between three conventional methods: shared encoder [13], cycle consistency loss [3], and RL-based method [11].

5.2 Baseline performance

Table 1 shows the performance of the baseline systems or seed models trained only on the si84 corpus. The shared encoder [13] and cycle consistency [3] reported 15.8% and 10.2% CERs, respectively. Our re-implementation of the end-to-end model of the shared encoder and supervised training achieved a CER of 10.4%. The difference in CERs were due to the different numbers of encoder layers, decoder units, batch sizes and numbers of epochs. When using adversarial training, 8.7% CER was achieved. Figure 2 shows CERs on

Table 1. CERs(%) of baseline systems trained using WSJ-SI84 dataset

System	**dev93**	**eval92**
Shared encoder [13]	25.4	15.8
Cycle consistency [3]	-	10.2
Reinforcement learning [11]	15.2	10.4
This work	15.2	10.4
+ Adversarial training	14.3	8.7

different ε_1. According to the baseline results, we performed semi-supervised training by fixing $\varepsilon_1 = 0.05$.

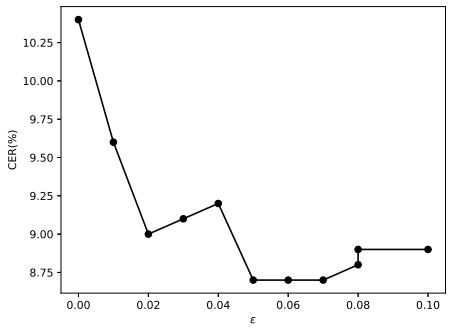


Figure 2. CERs(%) of adversarial training for different ε

Table 2. CERs(%) by hyper-parameters

τ	ε_2	γ	k	**dev93**	**eval92**	τ	ε_2	γ	k	**dev93**	**eval92**
0.875	0.04	0.04	1	10.7	6.8	0.9	0.04	0.04	2	10.9	7.0
0.875	0.04	0.06	1	10.7	6.9	0.85	0.04	0.04	1	11.0	7.3
0.875	0.04	0.05	2	10.8	7.0	0.85	0.04	0.04	2	11.0	7.3

5.3 Semi-supervised training performance

In this experiment, we measured the performance of the proposed approach by varying the hyper-parameters in the range $\tau = \{0.825, 0.85, 0.875, 0.9\}$, $\varepsilon_2 = \{0.03, 0.04, 0.05\}$, $\gamma = \{0.03, 0.04, 0.05, 0.06\}$, $k = \{1, 2, 3, 4\}$, and Table 2 shows the top performances for eval92 and dev93 sorted by CERs. Table 3 shows the best performance of the reported and proposed methods. The proposed method achieved 6.8% CER with the $\varepsilon_1 = 0.05, k = 1, \tau = 0.875, \varepsilon_2 = 0.04, \gamma = 0.04$ setting.

Table 3 also shows the importance of seed model accuracy. The CER increases from 6.8% to 8.7% when using a less accurate seed model. It can only achieve 7.8% even when re-tuning the hyper-parameters in semi-supervised training.

Table 3. Semi-supervised training performance using WSJ-SI84 and WSJ-SI284

System	dev93	eval92
Shared encoder [13]	24.8	14.4
Cycle consistency [3]		9.1
Reinforcement learning [11]	13.0	8.7
The work	**10.8**	**6.8**
Without adversarial pretraining	12.6	8.7
+ higher threshold ($\tau = 0.9$)	12.1	7.8

As shown in Table 2, the best performance for the eval92 test set was achieved by using a 1-best pseudo-label, but it seems reasonable to use 2-best pseudo labels in practice due to best and stable performance for the dev93 set. ε_2 showed the best contribution at 0.04, but the γ parameter showed little correlation with performance in this experiment.

6 Conclusions

This work proposes a semi-supervised end-to-end ASR training approach based on the consistency regularization and adversarial augmentation. Although we share the idea of consistency regularization between highly confident pseudo-labels of weakly augmented data and predictions of strongly augmented data, our implementation details are different. We used n-best pseudo labels and adversarial augmentation instead of domain specific weak and strong augmentation. We also used an adversarially pre-trained seed model to improve robustness.

We evaluated the proposed approach in the WSJ domain. The experimental results showed a significant character error rate reduction from 10.4% to 6.8%. The results demonstrate that confidence threshold τ is the most important parameter, followed by the k number of pseudo labels.

Although the proposed method is effective, it has some issues. One is that it requires high computational costs because it must evaluate whole end-to-end ASR systems at least twice to generate pseudo-labels and adversarial examples. The other is related to which augmentation is more suitable for weak and strong augmentation. In future studies we intend to investigate these two issues.

ACKNOWLEDGEMENTS

This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (2019-0-00004, Development of semi-supervised learning language intelligence technology and Korean tutoring service for foreigners).

REFERENCES

[1] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182, 2016.

[2] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4945–4949. IEEE, 2016.

[3] M. K. Baskar, S. Watanabe, R. Astudillo, T. Hori, L. Burget, and J. Černockỳ. Self-supervised sequence-to-sequence asr using unpaired speech and text. *arXiv preprint arXiv:1905.01152*, 2019.

[4] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019.

[5] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5050–5060, 2019.

[6] W. Chan, N. Jaitly, Q. Le, and O. Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE, 2016.

[7] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, et al. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. IEEE, 2018.

[8] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585, 2015.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[10] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1764–1772, 2014.

[11] H. B. J. Hoon Chung and J. G. Park. Semi-supervised training for sequence-to-sequence speech recognition using reinforcement learning. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 2020. to appear.

[12] T. Hori, R. Astudillo, T. Hayashi, Y. Zhang, S. Watanabe, and J. Le Roux. Cycle-consistency training for end-to-end speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6271–6275. IEEE, 2019.

[13] S. Karita, S. Watanabe, T. Iwata, A. Ogawa, and M. Delcroix. Semi-supervised end-to-end speech recognition. In *Proc. Interspeech 2018*, pages 2–6, 2018.

[14] L. Lamel, J.-L. Gauvain, and G. Adda. Lightly supervised and unsupervised acoustic model training. *Computer Speech & Language*, 16(1):115–129, 2002.

[15] B. Li, T. N. Sainath, R. Pang, and Z. Wu. Semi-supervised training for end-to-end models via weak distillation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2837–2841. IEEE, 2019.

[16] J. Ma and R. Schwartz. Unsupervised versus supervised training of acoustic models. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[17] Y. Miao, M. Gowayyed, and F. Metze. Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 167–174. IEEE, 2015.

[18] L. Mošner, M. Wu, A. Raju, S. H. K. Parthasarathi, K. Kumatani, S. Sundaram, R. Maas, and B. Hoffmeister. Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6475–6479. IEEE, 2019.

[19] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey. Multichannel end-to-end speech recognition. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2632–2641. JMLR. org, 2017.

[20] S. H. K. Parthasarathi and N. Strom. Lessons from building acoustic models with a million hours of speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6670–6674. IEEE, 2019.

[21] D. B. Paul and J. M. Baker. The design for the wall street journal-based csr corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 357–362. Association for Computational Linguistics, 1992.

[22] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly. A comparison of sequence-to-sequence models for speech recognition. In *Interspeech*, pages 939–943, 2017.

[23] J. Schmidhuber and S. Hochreiter. Long short-term memory. *Neural Comput*, 9(8):1735–1780, 1997.

[24] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[25] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[26] S. Sun, C.-F. Yeh, M. Ostendorf, M.-Y. Hwang, and L. Xie. Training augmentation with adversarial examples for robust speech recognition. *arXiv preprint arXiv:1806.02782*, 2018.

[27] A. Tjandra, S. Sakti, and S. Nakamura. End-to-end feedback loss in speech chain framework via straight-through estimator. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6281–6285. IEEE, 2019.

[28] S. Toshniwal, A. Kannan, C.-C. Chiu, Y. Wu, T. N. Sainath, and K. Livescu. A comparison of techniques for language model integration in encoder-decoder speech recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 369–375. IEEE, 2018.

[29] X. Wang, S. Sun, C. Shan, J. Hou, L. Xie, S. Li, and X. Lei. Adversarial examples for improving end-to-end attention-based small-footprint keyword spotting. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6366–6370. IEEE, 2019.

[30] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, et al. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*, 2018.

[31] F. Wessel and H. Ney. Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(1):23–31, 2005.

[32] K. Yu, M. Gales, L. Wang, and P. C. Woodland. Unsupervised training and directed manual transcription for lvcsr. *Speech Communication*, 52(7-8):652–663, 2010.

[33] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[34] A. Zeyer, K. Irie, R. Schlüter, and H. Ney. Improved training of end-to-end attention models for speech recognition. *arXiv preprint arXiv:1805.03294*, 2018.

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0291

An empirical study on semi-supervised transfer learning schemes for out-of-domain application of wav2vec 2.0

Yoonhyung KIM[(1)], Hyeong Bae JEON[(2)], Byung Ok KANG[(3)], Hoon CHUNG[(4)]

[(1)]Electronics and Telecommunications Research Institute (ETRI), Republic of Korea, E-mail yhkim1127@etri.re.kr

[(2)]Electronics and Telecommunications Research Institute (ETRI), Republic of Korea, E-mail hbjeon@etri.re.kr

[(3)]Electronics and Telecommunications Research Institute (ETRI), Republic of Korea, E-mail bokang@etri.re.kr

[(4)]Electronics and Telecommunications Research Institute (ETRI), Republic of Korea, E-mail hchung@etri.re.kr

ABSTRACT

Pre-trained speech models such as wav2vec 2.0 show decent performance for in-domain transfer learning scenarios. However, those models are not robust to domain discrepancy between pre-training and fine-tuning corpora. Preparing a new in-domain pre-trained model could be a naive solution, but it requires huge amounts of speech corpus and computational burdens. Thus, how to conduct transfer learning with off-the-shelf pre-trained models with restricted amounts of out-of-domain data becomes a crucial issue to enhance the usability of pre-trained models. Based on this motivation, in this paper, we present an extensive comparative study on out-of-domain and resource-scarce (i.e., semi-supervised) fine-tuning setups using the wav2vec 2.0 model. In addition, we present self-training results of small in-domain corpus and large out-of-domain corpus. We consider three native-to-nonnative (i.e., pre-train-to-fine-tune) corpora for automatic speech recognition (ASR) task, which are English spoken by Korean, Japanese, and Indian. Comparative evaluation results show that the ASR accuracy with 100 hours (10 hours labeled) in-domain data is better than that of 60k hours (960 hours labeled) out-of-domain data. Our experimental results would be a useful benchmark for researchers who are interested in utilizing pre-trained speech models in practice.

Keywords: Automatic speech recognition, Unsupervised learning, Semi-supervised learning, Domain shift, Transfer learning

1 INTRODUCTION

1.1 Backgrounds

With the rise of deep neural networks (DNNs), various recognition tasks have been widely addressed by means of data-driven manners. Automatic speech recognition (ASR) is one of the representative tasks, and the early works are focused on substituting traditional acoustic models with deep networks. In DNN-HMM models [1], the Gaussian mixture model (GMM) of traditional GMM-HMM models is replaced by the DNN module. Subsequently, end-to-end (E2E) models [2–4] are proposed as the fully DNN-based ASR framework. Those DNN-based frameworks demonstrate dramatic performance improvement of ASR, but large-scale annotated corpus is required for training. Preparing large-scale paired data (i.e., collecting audio files with corresponding transcriptions) is not only time-consuming and but also costly. To alleviate annotation costs, several recent studies are focused on representation learning for ASR model, which aims to learn feature representation by means of unsupervised learning. Several representative ASR pre-training models are contrastive predictive coding [7], HuBERT [8], wav2vec [9], wav2vec 2.0 [10], and data2vec [12].

Among those pre-trained models, wav2vec 2.0 [10] is the first framework which demonstrates that pre-training with large-scale raw audios followed by fine-tuning with small-scale transcribed audios can lead to state-of-the-art ASR. The encoder of wav2vec 2.0 consists of a multi-layer convolutional network and a Transformer network with multiple self-attention blocks [5]. For fine-tuning and inference stages, the network is attached

with a linear projector which is trained via the connectionist and temporal classification (CTC) loss [6]. The training process of wav2vec 2.0 is composed of pre-training and fine-tuning stages. In the pre-training stage, the model is induced to predict the correspondence of masked contextualized representation (i.e., Transformer output) and convolutional representation. In the fine-tuning stage, a linear projector is added on the top of the encoder and trained with the CTC loss. The output dimension of the linear projector is equal to the number of class tokens. According to the literature [10], the wav2vec 2.0 model requires around 100 times less amount of transcribed data (1 hour versus 100 hours) for fine-tuning to obtain a comparable ASR accuracy with the previous state-of-the-art ASR model.

1.2 Motivation

Though the wav2vec 2.0 model demonstrates its robustness to small-scale transcribed data, its performance improvement is limited to in-domain scenarios. In other words, the wav2vec 2.0 model assumes that the domain characteristic of speech data for pre-training should be similar to that of fine-tuning speech data. For example, in [10], the Librivox corpus (LV-60k [14]) is adopted for pre-training and the Librispeech (LS-960h [13]) corpus is used for fine-tuning. Both datasets share similar domain characteristics, i.e., English dictation spoken by native speakers. Apart from this example, every experiment in [10] considers in-domain setups, posing a question on the robustness of the wav2vec 2.0 model to domain shift.

In [11], a comparative study on the out-of-domain (OOD) fine-tuning scenarios of wav2vec 2.0 is addressed. The study considers scenarios that pre-training with dictational speaking corpus and fine-tuning with conversational speaking corpus. The results show that the wav2vec 2.0 model is fragile to domain shift (i.e., different speaking styles), and several empirical approaches are addressed to relieve the problem. However, the empirical observations in [11] are focused on the pre-training stages of wav2vec 2.0 models, overlooking the costs for preparing large amount of computational resources and pre-training data. In practice, there may be needs for adopting off-the-shelf pre-trained wav2vec 2.0 models and fine-tuning on their own transcribed data.

1.3 Contributions of this work

To address the above-mentioned issue, in this paper, we consider various fine-tuning schemes for cost-effective usage of the wav2vec 2.0 model on OOD scenarios. Based on a given pre-trained wav2vec 2.0 model and resource-scarce (i.e., semi-supervised) setups, we explore the optimal fine-tuning approach for ASR. To this end, we consider three native-to-nonnative datasets for ASR task, which are English spoken by Korean, Japanese, and Indian. In addition, we report various analysis on pre-training schemes such as incremental pre-training on small-scale splits and pre-training on a dataset of different language yet same accent. The following sections are as follows. In Sec. 2, we explain experimental setups. In Sec. 3, we report experimental results along with analysis and observations. In Sec. 4, concluding remarks and future research directions are given.

2 EXPERIMENTAL SETUPS

2.1 Datasets

For comparative study on various OOD scenarios, we select three datasets, which contain English spoken by Korean, Japanese, and Indian, respectively. Nonnative speakers have a different speech style leading to domain shift. For Korean English, we use Korean speaker files in the AESOP corpus [15] which we denote as K-AESOP. For Japanese and Indian English, King-ASR-050 [16] and King-ASR-383 [17] are used, respectively. Special characters (e.g., <non>, <spk>, <nps>) in the transcriptions are removed. For King-ASR-050, only the first channel is adopted for the experiment. The statistics of the corpora are summarized in Table 1. To investigate results on smaller transcribed subsets, we generate 1-hour and 10-hours splits by randomly selecting audio files.

For further investigation on pre-training corpus, which is addressed in Sec. 3.2, we use our Korean dictational speech corpus called R6. The Korean R6 corpus contains around 12k hours of speech data, and we adopt them for unsupervised pre-training of the wav2vec 2.0 large model. Following [10], we selected speech files whose length is longer than 2 seconds and shorter than 20 seconds. The statistic of the corpus is summarized in Table 2. The total duration of the audio files after the selection becomes around 11k hours.

Table 1. Statistics of speech corpora for transfer learning

Nationality (Accent)	Dataset	Number of utterances (train/dev/test)
Korean	K-AESOP [15]	26,599 (53h) / 1,000 / 1,000
Japanese	King-ASR-050 [16]	60,074 (89h) / 2,000 / 2,000
Indian	King-ASR-383 [17]	79,892 (98h) / 2,000 / 2,000

Table 2. Statistic of Korean R6 corpus for wav2vec 2.0 pre-training

	2∼3 sec	3∼4 sec	4∼5 sec	5∼10 sec	10∼15 sec	15∼20 sec	Total
Num. files	2,614,896	1,873,127	1,349,538	2,504,122	279,622	35,249	8,656,554
Duration (hours)	1788.4	1806.7	1675.6	4661.5	903.3	164.6	11000.1

2.2 Pre-trained models

For fine-tuning on the nonnative speech datasets, we adopt the publicly released LV-60k large model which is pre-trained on the Libri-Light corpus [14] (53k hours without transcription). The LV-60k model is used for supervised and semi-supervised fine-tuning experiments by adding a linear projection layer whose length is the number of class tokens. Following [10], the number of class tokens is set to 28, including uppercase alphabets along with space and apostrophe characters.

For comparative study on pre-training corpus in Sec. 3.2, the same model (i.e., wav2vec 2.0 large model) is pre-trained from scratch with the R6 corpus. The training process took around 6 weeks using 16 A6000 GPUs. Since the R6 model is pre-trained on Korean, its feature representation conceives Korean accents, which is closely related with the accents of Korean English.

2.3 Fine-tuned models

For empirical study, two kinds of fine-tuned models are compared, i.e., native and nonnative models. A 'native model' indicates a fine-tuned model which is trained on native speaker corpus. Since the target domain is nonnative speakers, the native speaker corpus involves domain discrepancy (i.e., OOD). In this work, we adopt four native models, which are fine-tuned on Librispeech 100h/960h (denoted as LS-100 and LS-960) and further self-trained on LV-60k (denoted as LS-100+ST, LS-960+ST).

A 'nonnative model' is a fine-tuned model which is trained on nonnative speaker corpus. For each of the three nationalities, we make three transcription splits, i.e., 1h, 10h, and overall splits. The nonnative models are trained by in-domain transcribed speech data, but the amount of training data is much smaller than that of the native models.

2.4 Semi-supervised fine-tuning

Due to the cost of transcribing speech files, a training corpus is often given as partially transcribed. To address this practical situation, we conduct semi-supervised transfer learning by means of pseudo labeling [18]. To this end, we propose a simple transfer learning curriculum as follows. First, the LV-60k pre-trained model is fine-tuned on transcribed speech data, e.g., Korean English 1h, 10h. Second, raw audio files are transcribed with the fine-tuned model (i.e., pseudo-labeling). Finally, the pre-trained model is trained on both of ground-truth transcribed and pseudo-labeled speech data.

2.5 Implementation details

For all experiments, we used the Fairseq toolkit [19]. For fine-tuning and semi-supervised transfer learning, the maximum iteration is 50K and learning rate is 3e-5. The rest of the setups are identical to [10]. For pre-training, we used the same parameters as [10] such as masking frequency and quantization. For decoding, we adopt viterbi decoder with beam size 50.

3 EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Supervised and semi-supervised fine-tuning results

The ASR results of Korean, Japanese, and Indian English corpora are reported in Table. 3, 4, and 5, respectively. In the tables, the 'out-of-domain' results are word error rates of author-released models, which are fine-tuned and self-trained (ST) by the native speech corpus. In the out-of-domain results, '+ST' indicates self-training on the Libri-Light corpus. The 'in-domain' results are obtained by fine-tuning and self-training on each nonnative speech corpus. For every case, the ASR accuracy of small-scale in-domain training data is better than that of large-scale out-of-domain training data. Specifically, 10h of in-domain transcribed data leads to better ASR accuracy than LS-960h+ST. The ASR accuracy on 1h of in-domain transcribed speech data is competitive to that of LS-960h+ST. For in-domain setups, self-training with pseudo labeled in-domain data leads to the better ASR accuracy.

Based on the above results, we can find the following empirical observations. First, despite of the huge scale of pre-training (LV-60k) and fine-tuning data (LS-960h), the wav2vec 2.0 model is not robust to domain shift, showing abrupt performance degradation on OOD test corpus. Second, additional transcribed data are much more effective than raw data (for pseudo labeling) for performance enhancement. For example, for Indian English in Table 5, collecting 9h of transcribed data lead to much greater WERR than 97h of pseudo labeled data. Finally, the error rates are reduced as the amount of transcribed data is increased. This indicates that the magnitude of transcribed data is important for OOD application of the wav2vec 2.0 model.

Table 3. Comparision of ASR error rates (%) - Korean English

out-of-domain				in-domain			
FT split	CER	WER	WERR via ST	FT split	CER	WER	WERR via ST
LS-100h	11.5	21.2	-	1h	8.2	17.8	-
LS-100h+ST	9.9	17.8	16.0	1h+ST	8.0	16.4	7.9
LS-960h	9.8	17.5	-	10h	1.6	3.9	-
LS-960h+ST	9.3	16.3	6.9	10h+ST	1.5	3.7	5.1
				All (53h)	1.0	2.4	-

Table 4. Comparision of ASR error rates (%) - Japanese English

out-of-domain				in-domain			
FT split	CER	WER	WERR via ST	FT split	CER	WER	WERR via ST
LS-100h	14.9	28.9	-	1h	9.3	20.5	-
LS-100h+ST	11.3	21.6	25.3	1h+ST	7.4	17.0	17.1
LS-960h	11.3	21.7	-	10h	4.8	9.5	-
LS-960h+ST	10.1	19.3	11.1	10h+ST	4.1	8.3	12.6
				All (89h)	3.6	6.7	-

3.2 Comparative study on pre-training corpus

Since the target domain in our setup is nonnative English, a speech signal of the target training corpus may conceive accents of its native language. For example, a Korean English speech signal covers an English sentence, yet it has Korean accents. To investigate which of the two components is more significant, a comparative study

Table 5. Comparision of ASR error rates (%) - Indian English

out-of-domain				in-domain			
FT split	CER	WER	WERR via ST	FT split	CER	WER	WERR via ST
LS-100h	9.4	22.1	-	1h	8.5	22.2	-
LS-100h+ST	7.1	16.8	24.0	1h+ST	7.0	18.9	14.9
LS-960h	7.2	17.2	-	10h	4.9	12.2	-
LS-960h+ST	6.6	15.6	9.3	10h+ST	4.2	10.9	10.7
				All (98h)	4.0	10.1	-

on pre-training corpus is conducted. In Table 6, comparative evaluation results on the two pre-trained models are reported. As we can see, the higher accuracy is obtained by the English pre-trained model (LV-60k). This implies that language correspondence is more significant than accent similarity for ASR accuracy.

Table 6. Comparision of ASR error rates (%) on the two pre-trained models

	LV-60k		Korean-R6-11k	
FT split	CER	WER	CER	WER
1h	8.2	17.8	16.4	33.6
10h	1.6	3.9	4.8	12.2
53h	1.0	2.4	2.5	6.4

3.3 Incremental pre-training with small corpus

To investigate the impact of pre-training on in-domain data, we conducted incremental pre-training on each corpus. To be specific, incremental pre-training is conducted on the LV-60k large model, and the model is subsequently fine-tuned on each transcribed data. However, this approach causes catastrophic forgetting after pre-training, leading to performance degradation after fine-tuning. We expect that the performance degradation is due to the lack of pre-training data. Thus, vanilla pre-training on a small corpus is not encouraged for OOD applications.

4 CONCLUSION AND FUTURE WORK

In this paper, we presented an extensive comparative study on out-of-domain and resource-scarce (i.e., semi-supervised) fine-tuning setups using the wav2vec 2.0 model. We considered three native-to-nonnative (i.e., pretrain-to-finetune) corpora for automatic speech recognition (ASR) task, which are English spoken by Korean, Japanese, and Indian. Comparative evaluation results show that the ASR accuracy with a small amount of in-domain data is much more effective than large amount of out-of-domain data. We believe that our results would be a useful benchmark for researchers who are interested in applying large-scale pre-trained speech models such as wav2vec 2.0 to with a restricted amount of transcribed speech data.

ACKNOWLEDGEMENTS

This work was supported by Institute of Information Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (2019-0-00004, Development of semi-supervised learning language intelligence technology and Korean tutoring service for foreigners)

REFERENCES

[1] Hinton, Geoffrey, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal processing magazine, Vol. 29, No. 6, pp. 82-97, 2012.

[2] Dzmitry Bahdanau, et al., "End-to-end attention based large vocabulary speech recognition," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4945-4949, 2016.

[3] Chung-Cheng Chiu, et al., "State-of-the-art speech recognition with sequence-to-sequence models," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4774-4778, 2018.

[4] William Chan, et al., "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4960-4964, 2016.

[5] Vaswani, Ashish, et al. "Attention is all you need," Advances in neural information processing systems 30 (2017).

[6] Graves, Alex, et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," Proceedings of the 23rd international conference on Machine learning. 2006.

[7] Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding," arXiv preprint arXiv:1807.03748 (2018).

[8] Hsu, Wei-Ning, et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021): 3451-3460.

[9] Schneider, Steffen, et al. "wav2vec: Unsupervised pre-training for speech recognition," arXiv preprint arXiv:1904.05862 (2019).

[10] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations," Advances in Neural Information Processing Systems 33 (2020): 12449-12460.

[11] Hsu, Wei-Ning, et al. "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," arXiv preprint arXiv:2104.01027 (2021).

[12] Baevski, Alexei, et al. "Data2vec: A general framework for self-supervised learning in speech, vision and language," arXiv preprint arXiv:2202.03555 (2022).

[13] Panayotov, Vassil, et al., "Librispeech: an asr corpus based on public domain audio books," 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5206-5210, 2015.

[14] J. Kahn et al. "Libri-light: A benchmark for asr with limited or no supervision," In Proc. of ICASSP, 2020.

[15] Visceglia, Tanya, et al. "Phonetic aspects of content design in AESOP (Asian English Speech cOrpus Project)," 2009 Oriental COCOSDA International Conference on Speech Database and Assessments, IEEE, 2009.

[16] "King-ASR-050: Japanese English Speech Recognition Corpus (Desktop)," Speechocean, 2014.06. (link: https://en.speechocean.com/datacenter/details/1419.html)

[17] "King-ASR-383: Indian English Speech Recognition Corpus (Desktop)," Speechocean, 2015.10. (link: https://en.speechocean.com/datacenter/details/1671.html)

[18] Lee, Dong-Hyun. "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," Workshop on challenges in representation learning, ICML. Vol. 3. No. 2. 2013.

[19] Ott, Myle, et al. "fairseq: A fast, extensible toolkit for sequence modeling." arXiv preprint arXiv:1904.01038 (2019).

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0472

Knowledge distillation learning for the lightweight dynamic convolution in keyword spotting

Donghyeon Kim[(1)], Kyungdeuk Ko[(1)], Gwantae Kim[(1)], Hanseok Ko[(1)]

[(1)]Korea University, South Korea

ABSTRACT

This paper presents a Knowledge Distillation (KD) learning for a dynamic convolution in the front end. In our previous work, we confirmed that applying a dynamic filtering to the front end of a classifier would improve the performance of the classification in noisy environments. The main goal of this study is to develop a Relational Knowledge Distillation (RKD) framework for dynamic convolution. The teacher model consists of six layers of dynamic convolution. The student model is constructed with a single layer of dynamic convolution, trained by the RKD loss and classifier loss. For performance evaluation, the experiments are carried out by a classification task using Keyword Spotting (KWS). The experimental results show that the proposed KD method improves the KWS performance in noisy environments over the baseline student model, which is trained only by the classifier loss.

Keywords: Sound classification, Knowledge distillation, Light weight, Dynamic convolution, Key word spotting

1 INTRODUCTION

Recent speech-based Artificial Intelligence (AI) and sound classification made advances by deep learning architecture, neural networks with high computational power (many parameters and FLOPS.) [13, 12, 4]. Keyword Spotting (KWS) to wake up AI system thus is applied to real stream speech applications [16]. The KWS applied to the real streaming under an on-device environment should be minimally designed for model efficiency, and to this goal, various small footprint models made advances. For memory-efficient Convolutional Neural Networks (CNN), a Depth-Separable Convolution (DSConv) [31] and a temporal 1D CNN models [1, 14] show reasonably acceptable performance with relatively low computational power over conventional CNN model. In addition, raw audio based deep learning models [11, 17, 29], Neural Architecture Search (NAS) [18, 30] and far-field data augmentation [3] have been developed.

In our earlier work [7], we proposed a small footprint dynamic convolution in the front-end as a feature extractor. By utilizing a dynamic filter framework, we generate weights of the CNN via another neural network. Unlike the general dynamic convolution process that would produce a CNN kernel for each patch unit, we split the filter generator process into two branches (pixel-level and instance-level). This two-branch method would reduce all computational costs over the conventional method. Although the proposed model would show robust results in unseen noise environments, the performance might be further improved by using a Multi-layer operation [9]. However, applying Multi-layer Dynamic Convolution (MDC) requires high computational power over a Single layer Dynamic Convolution (SDC). This paper thus explores a Knowledge Distillation framework for the lightweight dynamic convolution model [7]. We assume that the MDC model as teacher model and leading the SDC model produce similar output to the MDC model. To this end, we first train the MDC (teacher model) and classifier by using the classifier loss (Cross-Entropy (CE) loss between a prediction and its label). Then, the SDC model (student model) is trained with fine-tuned classifier by using the classifier loss and distillation loss of the classifier embedding between the teacher model and student model. We utilize Relational Knowledge Distillation (RKD) metric for the distillation loss.

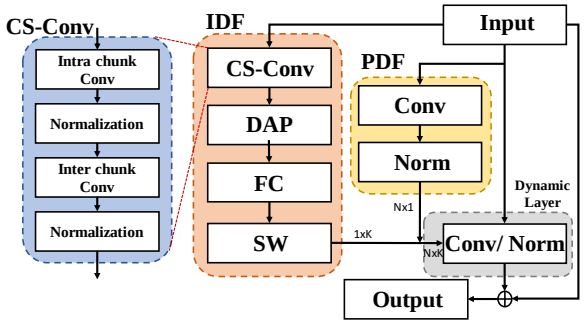


Figure 1. Model structure of our proposed lightweight model.

The KWS experiments are carried out on Speech Command datasets v1 and v2 [26]. In unseen noise situations, we compare our method to tiny footprint KWS models and other KD methods [21, 20]. In the WHAM [27] 0dB scenario, our proposed method shows improved performance over the SDC model, which is trained only by CE loss.

2 RELATED WORKS

2.1 Dynamic filter model

Dynamic Filter Network (DFN) is an adaptive deep neural network architecture [6] that generates weights of a neural network by the learning process. In the part of a filter generator, weights are produced from input data, and the subsequently generated weights are employed to compute the deep learning operations with the input data, which is utilized to produce the weights. As the weights are driven from input data, the weights of the network get updated by the input. As a result, it would make the deep learning model more robust and flexible to various domains of applications. Kim *et al.* [9] proposed a convolution-based dynamic filter network to enhance salient features from the unseen noisy audio stream, and their experimental results showed that their approach outperforms conventional feature enhancement methods. Fujita *et al.* [2] also utilized a dynamic filter-based method for a lightweight ASR model. They confirmed that applying dynamic convolution to the decoder part in the encoder-decoder model improves accuracy and reduces computational load over transformer[25] based models. Although dynamic filter-based approaches have shown progressive results, the implementation typically requires high computational cost as the weights are produced depending on each basis. In our earlier work [8, 7], we employed the concept of Decoupled Dynamic Filter (DDF) [32] from computer vision to address this issue. Instead of a single filter to take up the channel-spatial features directly, DDF divides the filter into two parts: channel and spatial. The channel-wise filter applied 1-D convolution to produce the channel weights, while the spatial part used a simple global average pooling and Fully Connected (FC) layers to produce Kernel weights. These two filters are trained at separate branches of the network and then combined to compute dynamic convolution. By conducting two branches of models with simple computations, this method achieved a significant reduction in memory and floating-point operations (FLOPs). In our application, we apply dynamic convolution to a single channel of the T-F feature as described in Figure 1. We split the dynamic filter process into Pixel Dynamic Filter (PDF) and Instance-level Dynamic Filter (IDF). In the PDF, pixel weights get produced by a single layer of convolution. In the IDF, the kernel weights are produced by processing through a simple feature averaging with FC layers. Then these two weights are combined to compute a single layer of dynamic convolution. We apply this method to the front-end of the KWS model, and the results show robust performance in various noisy environments.

2.2 Relational Knowledge Distillation

In general, conventional KD methods [5, 21] minimize the distance metric between the teacher and the student model to produce similar distribution. As the methods compute the distance of the teacher and student pair

individually, the student model could not train the structural distribution among the learning instance. To combat this issue, a Relational Knowledge Distillation (RKD) [20] proposed KD metric which considers structural knowledge using mutual relations. The objective for RKD is as follows:

$$L_{RKD} = \sum_{(x_1,x_2,..,x_n)\in\chi^N} D(\psi(t_1,t_2,..,t_n),\psi(s_1,s_2,..,s_n)), \tag{1}$$

where $(x_1,x_2,..,x_n)$ is a n-tuple drawn from χ^N; ψ is a relational function that represents relational distribution. D denotes a distance loss function. As the RKD trains the class-wise distributions between the teacher model and student model, the student model would produce tight class clusters with the teacher model over Individual Knowledge Distillation (IKD). The proposed RKD method has two different ways: distance-wise and angle-wise distillation losses. In the distance-wise distillation loss, ψ_D measures the Euclidean distance between the two instances in the batch unit. The related loss function as follows:

$$\psi_D(t_i,t_j) = \frac{1}{\mu}||t_i - t_j||_2, \tag{2}$$

$$\mu = \frac{1}{\chi^2}\sum_{(x_i,x_j)\in\chi^2}||t_i - t_j||_2, \tag{3}$$

where μ is a scale weight of normalization. To focus on relative distances among other pairs, we set μ to as a averaging distance between all possible pairs. Using the distance-wise distribution of both the teacher and student model, the distance-wise distillation loss is as follows:

$$L_{RKD_D} = \sum_{(x_i,x_j)\in\chi^N} D_\delta(\psi_D(t_i,t_j),\psi_D(s_i,s_j)), \tag{4}$$

where D_δ denotes a Huber loss function [20], which conditionally utilises L_1 and L_2 losses. In angle-wise distillation loss, cosine based distance metric is utilized to build structural relationship from triplet of examples. The KD loss is as follows:

$$\psi_A(t_i,t_j,t_j) = cos\angle t_i t_j t_j = <\mathbf{e}^{i,j},\mathbf{e}^{k,j}>, where\ \mathbf{e}^{i,j} = \frac{t_i - t_j}{||t_i - t_j||_2}, \mathbf{e}^{k,j} = \frac{t_k - t_j}{||t_k - t_j||_2}, \tag{5}$$

$$L_{RKD_A} = \sum_{(x_i,x_j,x_k)\in\chi^3} D_\delta(\psi_A(t_i,t_j,t_k),\psi_A(s_i,s_j,s_k)), \tag{6}$$

where D_δ denotes Huber loss function. These L_{RKD_D} and L_{RKD_A} encourage that the student model would follow the structural relationship of the teacher model by penalizing distance differences between their output distributions.

3 PROPOSED MODEL

3.1 Learning pipeline

Our KD learning pipeline is depicted in Figure 2. We utilize the same classifier model, while the number of dynamic convolution layers is changed in both the teacher and student models. In the teacher model training, we employ six layers of dynamic convolution. They are hierarchically connected with a dropout [24] at the final layer. The output layer of the MDC model feeds to the KWS classifier, and subsequently, they are trained jointly by the classifier loss. In the student model training, the output of the SDC model feeds to the KWS classifier, where the weights are initialized from the teacher model classifier. Both the SDC and the fine-tuned classifier weights are trained jointly by the classifier loss and the RKD loss.

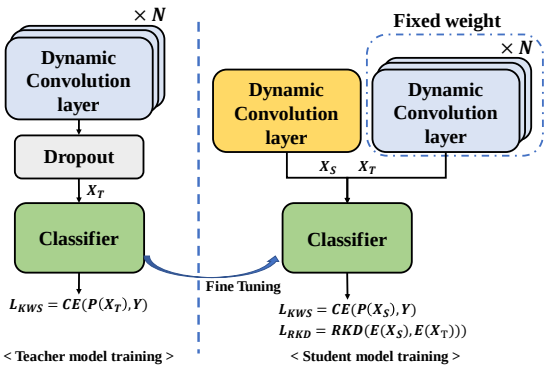


Figure 2. Pipeline of the KD training. X_T and X_S denote the output of the teacher and student network respectively. $P(\cdot)$ and Y denote prediction of the classifier and its label respectively. For the student model training, weighs of the teacher model classifier is utilized as initial weighs of the student classifier.

3.2 Knowledge Distillation learning

To compute the KD metric, we utilize the classifier embedding vector. The embedding vector is the output of the final convolution layer in the TENet model. The output of the pre-trained MDC feeds to the classifier. Here, the embedding vector acts as the teacher embedding. Then, the RKD loss gets computed between the teacher embedding and student embedding. The loss function for the student model training is as follows:

$$L_{total} = L_{classifier} + \lambda L_{RKD}, \tag{7}$$

where $L_{classifier}$ and L_{RKD} denote the classifier loss and RKD loss, respectively. λ denotes the coefficient of the L_{RKD}. From this process, it is ancipated that the dynamic convolution layer gets trained to produce similar classifier embedding with the teacher model. In short, this process should better train the entire model over the student model, which is trained only by the classifier loss function.

4 EXPERIMENT

4.1 Experimental Setup

In this section, we discuss our experimental setup and learning details for the KWS experiment.

Dataset. we used Speech Command dataset v1 and v2 [26] to train and evaluate our models. We follow the same data protocol of the DB guideline. We utilized 10 keywords with two extra classes (unknown and silence). We split 80% of the dataset for the model training, 10% for validation, and the remaining 10% for the test. We also injected background noise (mike noise) and performed random time-shifting. For evaluating the KWS performance in unseen noise environments, we use DCASE [15], Urbansound8K [23] and WHAM [27] datasets which contain various urban data. To set up the test environments, randomly selected noise segments and the original test data are mixed with five different Signal-to-Noise Ratios (SNR) levels.

Training detail. For the model training, we use a batch size of 100, a learning step of 30K, and ADAM optimizer [10] with a 0.001 initial learning rate. At every 10K-th step, the learning rate is cut down to 0.1. We use 30ms of windows with 10ms overlap and 64 Mel filters to produce 40 dimensions of Mel Frequency Cepstral Coefficients. For the total loss computation, we used $\lambda = 0.1$ as the coefficient of the loss function.

4.2 Baselines

We implement our KD training on the EDy-TENet12 [7] and compared KWS performance with the following baseline models.

TCNet. TCNet [1] contains two layers of temporal convolutions with a skip-connection as a bundle model and TCNet14 is composed of 6 bundles and 1 FC layer.

Table 1. Comparison with lightweight models on Speech Command v1 and v2. For an accurate experiment, 8 times averaging accuracy and the best performance are presented.

Model	**(Par.,FLOPS.)**	V1		V2	
		Acc	**Best**	**Acc**	**Best**
TCNet14[1]	(305K,8.26M)	-	96.6	96.53	96.8
TENet12[14]	(100K,6.42M)	-	96.6	97.10	97.3
TENet12†[14]	(100K,6.42M)	97.19	97.3	97.43	97.6
NAS2[18]	(886K,-)	-	97.2	-	-
Random[30]	(196K,8.8M)	96.58	96.8	-	-
DARTS[30]	(93K,4.9M)	96.63	96.9	96.92	97.1
F-DARTS[30]	(188K,10.6M)	96.70	96.9	97.11	97.4
N-DARTS[30]	(109K,6.3M)	96.79	97.2	97.18	97.4
LightConv[28]	(105K,7.40M)	96.88	97.0	97.24	97.3
DyConv[28]	(107K,7.69M)	96.89	97.1	97.26	97.4
MHA-RNN†[22]	(743K,87.2M)	97.50	-	98.36	-
LDy-TENet12[8]	(102K,6.64M)	96.95	97.1	97.35	97.6
LDy-TENet12†[8]	(102K,6.64M)	97.42	97.6	97.66	97.7
EDy-TENet12[7] (Student baseline)	(102K,6.68M)	**97.07**	**97.4**	**97.42**	**97.8**
Teacher model[7]	(109K,7.96M)	**97.06**	**97.2**	**97.42**	**97.7**
RKD-A	(102K,6.68M)	**97.17**	**97.4**	**97.51**	**97.7**
RKD-D	(102K,6.68M)	**97.01**	**97.4**	**97.29**	**97.5**

TENet. TENet [14] uses depth-separable temporal convolutions. TENet12 contains 12 convolution blocks with 1 FC layer. Every convolution block has 32 output channels.

Neural Architecture Search. NAS is a learning-based model designing method that uses computational cost-based loss functions (FLOPS., memory, accuracy, etc.). We compare our method with various NAS methods including Differentiable Architecture Search (DARTS). Please see details of the model in [18, 30].

Self attention. MHA-RNN[22] employed Convolutional Recurrent Neural Network (CRNN) and self-attention mechanism. The output of CRNN feeds to the dot product-based self-attention model, and two layers of MLP produce finial output probability.

LDy-TENet. LDy-TENet12 [8] uses a lightweight dynamic convolution on the TENet12 [14]. They split the dynamic filter model into two parts (Pixel and Instance) to perform a low computational Dynamic (LDy) filter. This filter extracts the T-F feature at the front-end of the TENet12 model.

EDy-TENet. EDy-TENet12 [7] is a variation of the LDy model to mitigate an overfitting problem caused by feature averaging. Instead of utilizing a simple feature mean in the LDy filter, a learnable feature pooling is applied.

Table 2. Comparison of KWS performance depending on the KD methods.

Noise	SNR (dB)	Method				
		RKD_D[20]	RKD_A[20]	*Fit*[21]	*Teacher*[7]	*Student*[7]
		v1, v2	**v1, v2**	**v1, v2**	**v1, v2**	**v1, v2**
DCASE [15]	**20**	96.89, 97.10	96.81, 97.07	96.65, 96.85	96.77, 96.97	96.72, 97.04
	15	96.79, 96.63	96.51, 96.49	96.46, 96.25	96.56,96.52	96.5, 96.45
	10	95.59, 95.89	95.41, 95.81	95.42, 95.67	95.64, 95.97	95.32, 95.64
	5	94.12, 94.01	93.76, 93.87	93.52, 93.30	94.08, 94.05	93.54,93.42
	0	89.87, 89.75	89.67, 89.35	89.10, 88.66	89.90, 90.00	89.3, 88.73
Urban [23]	**20**	96.36, 96.35	96.07, 96.42	96.01, 96.07	96.20, 96.39	96.08, 96.35
	15	95.30, 95.35	95.01, 94.99	94.85, 94.80	95.29, 95.26	95.2, 94.95
	10	93.45, 93.39	92.91, 93.12	92.87, 92.71	93.56, 93.46	93.10, 93.04
	5	90.36, 88.47	89.65, 88.05	89.39, 87.32	90.24, 88.92	89.63, 87.51
	0	80.65, 81.16	79.89, 79.84	78.91, 79.26	81.27, 81.94	80.50, 79.63
WHAM [27]	**20**	96.33, 96.36	96.23, 96.30	96.07, 96.13	96.25, 96.23	96.17, 96.30
	15	95.83, 95.51	95.70, 95.49	95.64, 95.30	95.77, 95.45	95.50, 95.40
	10	93.40, 93.65	93.44, 93.47	93.07, 93.22	93.24, 93.39	93.06, 93.39
	5	89.95, 88.90	89.54, 88.56	88.66, 88.10	88.88, 88.8	89.10, 88.10
	0	78.83, 78.83	78.61, 78.25	77.35, 76.89	77.9, 78.42	77.79, 76.97
Noisy-AVG.		92.25, 92.09	91.95, 91.81	91.60, 91.37	92.10, 92.12	91.87, 91.58

4.3 Result discussion

For an accurate evaluation, eight repeated experiments are carried out. The best result and model parameters are used for performance evaluation. The values in Tables 1 and 2 show the KWS results in the original test data and noise augmented test data, respectively. For a fair comparison, we indicate the results which used spec-augmentation [19] by †. It is noted that spec-augmentation is not applied to our models. From the results, we observe that the teacher model does not increase the performance in the original test data while it takes robust performance in a noisy environment. In particular, the teacher model takes 1.5% of performance improvement in the WHAM 0 dB condition. The RKD loss-based KD model, on the other hand, outperforms the student model and shows similar performance to the teacher model. Especially, RKD_D achieves 1.1% (v1) and 1.9% (v2) of performance improvement in the WHAM 0dB with the same computational power (FLOPS. and parameters). In some cases of the RKD results, KD trained model shows better performance over the teacher model. As the teacher model provides guidelines for student training, some values might show better results than the teacher model.

5 CONCLUSIONS

The main goal of this study was to develop a knowledge distillation learning framework to reduce the performance gap between a single layer of dynamic convolution and multi-layers of dynamic convolution. Based on the prescribed classifier embedding procedure, the relational knowledge distillation loss was obtained and

applied to compute the feature distance between the student and teacher model. The experimental results verify that applying the knowledge distillation learning to the dynamic convolution would improve the KWS performance in noisy environments without incurring additional costs.

ACKNOWLEDGEMENTS

This work was supported by Korea Environment Industry & Technology Institute(KEITI) through Exotic Invasive Species Management Program, funded by Korea Ministry of Environment(MOE) (2021002280004)

REFERENCES

[1] S. Choi, S. Seo, B. Shin, H. Byun, M. Kersner, B. Kim, D. Kim, and S. Ha. Temporal convolution for real-time keyword spotting on mobile devices. *arXiv preprint arXiv:1904.03814*, 2019.

[2] Y. Fujita, A. S. Subramanian, M. Omachi, and S. Watanabe. Attention-based asr with lightweight and dynamic convolutions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7034–7038. IEEE, 2020.

[3] Y. Gao, Y. Mishchenko, A. Shah, S. Matsoukas, and S. Vitaladevuni. Towards data-efficient modeling for wake word spotting. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7479–7483. IEEE, 2020.

[4] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[5] G. Hinton, O. Vinyals, J. Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[6] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool. Dynamic filter networks. In *Advances in neural information processing systems*, pages 667–675, 2016.

[7] D. Kim, G. Kim, B. Lee, J. Kwak, D. K. Han, and H. Ko. Efficient dynamic filter for robust and low computational feature extraction. *arXiv preprint arXiv:2205.01304*, 2022.

[8] D. Kim, K. Ko, J. Kwak, D. K. Han, and H. Ko. Lightweight dynamic filter for keyword spotting. *arXiv preprint arXiv:2109.11165*, 2021.

[9] D. Kim, J. Park, D. K. Han, and H. Ko. Dual stage learning based dynamic time-frequency mask generation for audio event classification. *Proc. Interspeech 2020*, pages 836–840, 2020.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] K. Kumatani, S. Panchapagesan, M. Wu, M. Kim, N. Strom, G. Tiwari, and A. Mandai. Direct modeling of raw audio with dnns for wake word detection. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 252–257. IEEE, 2017.

[12] S. Lee, D. K. Han, and H. Ko. Multimodal emotion recognition fusion analysis adapting bert with heterogeneous feature unification. *IEEE Access*, 9:94557–94572, 2021.

[13] Y. Lee, J. Min, D. K. Han, and H. Ko. Spectro-temporal attention-based voice activity detection. *IEEE Signal Processing Letters*, 27:131–135, 2019.

[14] X. Li, X. Wei, and X. Qin. Small-footprint keyword spotting with multi-scale temporal convolution. *arXiv preprint arXiv:2010.09960*, 2020.

[15] A. Mesaros, T. Heittola, and T. Virtanen. Acoustic scene classification in dcase 2019 challenge: Closed and open set classification and data mismatch setups. 2019.

[16] A. H. Michaely, X. Zhang, G. Simko, C. Parada, and P. Aleksic. Keyword spotting for google assistant using contextual speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 272–278. IEEE, 2017.

[17] S. Mittermaier, L. Kürzinger, B. Waschneck, and G. Rigoll. Small-footprint keyword spotting on raw audio data with sinc-convolutions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7454–7458. IEEE, 2020.

[18] T. Mo, Y. Yu, M. Salameh, D. Niu, and S. Jui. Neural architecture search for keyword spotting. *arXiv preprint arXiv:2009.00165*, 2020.

[19] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[20] W. Park, D. Kim, Y. Lu, and M. Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.

[21] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[22] O. Rybakov, N. Kononenko, N. Subrahmanya, M. Visontai, and S. Laurenzo. Streaming keyword spotting on mobile devices. *arXiv preprint arXiv:2005.06720*, 2020.

[23] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[26] P. Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

[27] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux. Wham!: Extending speech separation to noisy environments. In *Proc. Interspeech*, Sept. 2019.

[28] F. Wu, A. Fan, A. Baevski, Y. N. Dauphin, and M. Auli. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*, 2019.

[29] N. Zeghidour, O. Teboul, F. d. C. Quitry, and M. Tagliasacchi. Leaf: A learnable frontend for audio classification. *arXiv preprint arXiv:2101.08596*, 2021.

[30] B. Zhang, W. Li, Q. Li, W. Zhuang, X. Chu, and Y. Wang. Autokws: Keyword spotting with differentiable architecture search. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2830–2834. IEEE, 2021.

[31] Y. Zhang, N. Suda, L. Lai, and V. Chandra. Hello edge: Keyword spotting on microcontrollers. *arXiv preprint arXiv:1711.07128*, 2017.

[32] J. Zhou, V. Jampani, Z. Pi, Q. Liu, and M.-H. Yang. Decoupled dynamic filter networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6647–6656, 2021.

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea
Extended abstract

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0493

Accelerating the Centerline Processing of Vocal Tract Shapes for Articulatory Synthesis

Romain Karpinski, Vinicius Ribeiro, Yves Laprie

Université de Lorraine, CNRS, Inria, LORIA, Nancy, F-54000, France

ABSTRACT

Acoustic simulations used in the articulatory synthesis of speech take a series of vocal tract shapes as an input. Acoustic simulations assume a plane wave propagation, simplifying and limiting the calculation time. It is, therefore, necessary to split 2D vocal tract shapes into small tubes perpendicular to the centerline simulating the plane wave propagation. The algorithm developed previously used a time-consuming regularization step whose computation time was close to that of acoustic simulations. Therefore, we explored the possibility of using deep learning to perform this step and accelerate the whole synthesis process. We used a database with a large number of rt-MRI images (150 000) and our regularizing algorithm for training. Two architectures were tested, one using a regression strategy applied to the two curves defining the vocal tract and one exploiting the classification of pixels in 2D images of the vocal tract. The first turned out to be much faster, even if it requires checking that the center line is correct and, in some sporadic cases using the initial algorithm as a fallback solution.

Keywords: Speech, Vocal tract shape, Centerline, Articulatory synthesis

1 INTRODUCTION

The process of articulatory synthesis comprises generating a series of 2D or 3D vocal tract shapes corresponding to the target utterance and synthesizing the audio signal using numerical aero-acoustical simulations [1]. An important intermediate step is generating the acoustic parameters used as input for the simulations. In order to limit the computation time of the simulations, we used an approach that assumes the propagation of a plane wave in the vocal tract. The intermediate step consists of splitting the vocal tract into small tubes, which requires the determination of the centerline assumed to represent the propagation of the wave inside the vocal tract.

The determination of the centerline has therefore received sustained attention, which led to several algorithms [2], [3]. In [4], we presented a heuristic algorithm that relies on dynamic programming to generate a first guess of the centerline and then a regularization step inspired by active curves [5] to obtain a smooth and relevant curve. The first step was optimized to reduce the space explored by dynamic programming, but the second step requires more computation. The work reported here is intended to accelerate the determination of the centerline by using neural networks, which are increasingly used to solve problems and surrogate optimization algorithms in many engineering domains ([6] for instance).

2 ACCELERATING THE CENTERLINE DETERMINATION

The time required to determine the centerline appears comparable to that of the acoustic simulation in order of magnitude. The time required for this step is about 0.250 s on a recent 2021 laptop (11th Gen Intel(R) Core(TM) i7-11850H @ 2.50GHz), and it seems reasonable to reduce the time required to the determination of the centerline to accelerate the overall synthesis. Moreover, even if the shapes of the vocal tract present variability, they always have invariant characteristics (e.g., the location of the extremities, the fixed or quasi-fixed walls of the pharynx, the presence of one or more constrictions), which means that a machine learning

approach can be considered. We thus have experimented with two types of neural networks:

- The classification approach which converts the initial problem into a semantic segmentation task and aims to find the points (pixels) in the vocal tract corresponding to the center line.
- The regression approach which uses the vocal tract contours directly as an input and generates the desired centerline.

In both cases, the training uses the centerline determined by the heuristic algorithm [4].

2.1 Dataset

The centerline algorithm is intended to be used on synthetic vocal tract shapes. In [7] we developed a deep learning approach for generating vocal tract shapes for a given series of phonemes (corresponding to a target sentence). The training of this approach exploited a database composed of rt-MRI (real-time Magnetic Resonance Imaging) sequences recorded by one male French native speaker [8] at Max Plank Institute, Göttingen, Germany. The recordings have a frame rate of 55 `fps`, pixel spacing of 1.412 `mm`, and an image resolution of 136×136 pixels for the 2D images of the vocal tract in the mid-sagittal plane. The corpus contains 38 acquisitions, with a median acquisition time of 81.8 seconds, a minimum of 36.3 seconds, and a maximum of 90.1 seconds. The sentences were selected to provide a phonetically balanced coverage of French, and the whole dataset comprises 161 570 images. We developed a deep learning automatic tracking of the tongue [9] which was extended to all the speech articulators [10]. By concatenating those contours, the two edge contours of the vocal tract C_{int} and C_{ext} (for inner and outer) are obtained and give the complete 2D geometrical shape of the vocal tract from the glottis to the lips (see Figure 1).

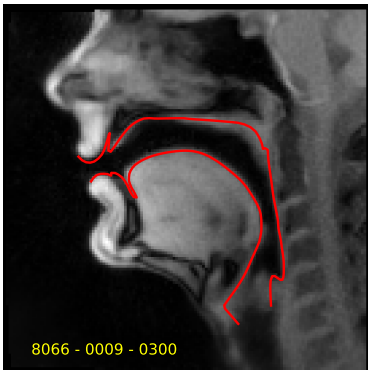


Figure 1. C_{int} and C_{ext} contours of the vocal tract.

Since the shapes used for articulatory synthesis derive from the contours of this dataset through training, we used them to train and test the acceleration of the centerline determination. Indeed, they cover a substantial variability of vocal tract shapes. The database is divided into three parts: 104 808 images for the training set, 19 307 images for the validation set, and 37 455 images for the test set. Some images do not correspond to speech since several images correspond to pauses, breaths, and swallowings.

2.2 Classification approach

This approach converts the centerline determination problem into a semantic segmentation task. The model aims to determine whether a pixel in a 2D image belongs to the centerline or the background. Once the segmentation is achieved, a second post-processing step transforms the network's output into an actual curve.

The network's inputs are binary images describing the two vocal tract walls (see Figure 1), while the target is a binary image describing the centerline calculated by the heuristics algorithm. To perform the semantic segmentation, we chose MobileNetV3 [11] as a backbone CNN to extract features and took advantage of pre-trained weights provided by the `PyTorch` [12] framework. However, the architecture of MobileNetV3 does

not enable semantic segmentation directly, and a decoder is thus needed to retrieve the original image size required to classify each pixel. We stripped the classification layer of MobileNetV3 and added upsampling blocks composed of an upsampling layer followed by a 2D convolution layer with `ReLU` activation. A total of three upsampling blocks are required to obtain the original image size. Finally, we added two convolutional layers, the last one performing classification. Figure 2 summarizes the neural network architecture used for this experiment.

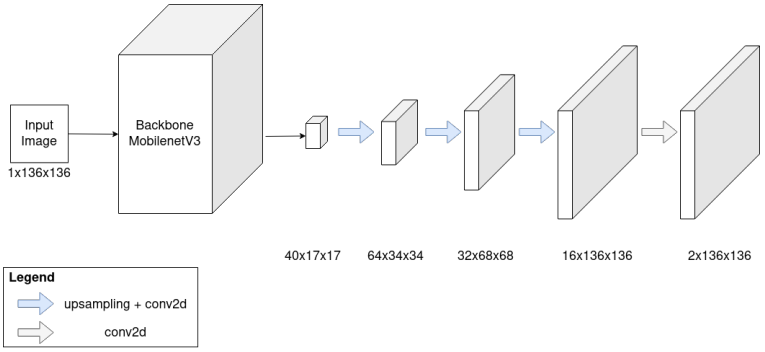


Figure 2. Architecture of the neural network used for the semantic segmentation.

The resulting probability map is then post-processed to find the centerline, which amounts to finding the shortest path between the extremities of the vocal tract (the middle points at the glottis and lips). Those two extremities are found from C_{int} and C_{ext} contours. The weight of each point equals $1 - p(j,i)$ where $p(j,i)$ is the probability that pixel j,i belongs to the centerline. Since the region corresponding to the vocal tract is given by the two edge contours, the graph can be drastically reduced, guaranteeing that the shortest path, i.e., the centerline, is inside the vocal tract.

2.3 Regression approach

The regression approach aims at estimating a function through a neural network. This function uses the contours of the vocal tract to extract the centerline. The goal is to obtain a function f' such that $f' \approx f$ with f being the existing function to replace. The method can be trained by minimizing the mean squared error (MSE) between f' and f.

The advantages of this solution are:

- Fast: the dimensionality being very low, the number of calculations is limited.
- Simple: the problem consists in approximating a function which is the expected result.
- Direct: there is no need for post-processing since the centerline is directly obtained.

The input corresponds to the two curves C_{int} and C_{ext} with their normalized coordinates in the interval $[0;1]$. These two curves are concatenated in a matrix $X \in \mathbb{R}^{4\times N}$ with N the number of points of both curves. It is important to note that the points of the two contours (inner and outer) are not synchronized, i.e., the *i-th* point of the inner contour is independent of the *i-th* point of the outer contour. The output curve corresponds to the centerline C_{center} in a form similar to inputs, i.e., a vector of N points.

Figure 3 shows the regression network's architecture. It uses convolutional layers that allow features to be extracted and fully connected layers that allow features to be related to each other. After each convolutional layer, and `Dense 1` and `Dense 2` layers, `ReLU` activation function is applied. This architecture enables the centerline to be obtained at a low computational cost.

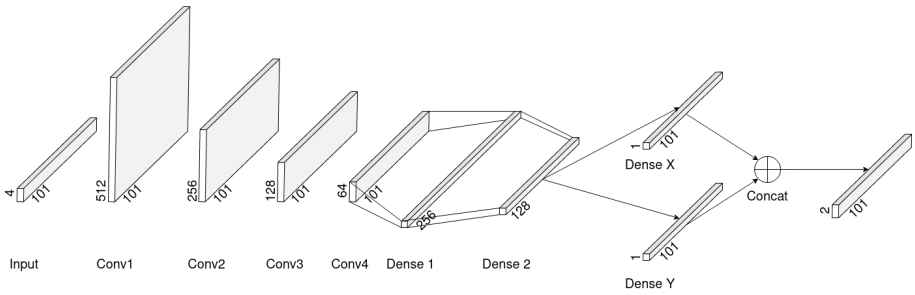


Figure 3. Architecture of the neural network used to perform regression.

2.3.1 Loss function

Let the predicted matrix $Y' \in \mathbb{R}^{2\times N}$. The is to minimize the MSE described by Equation 1 and Equation 2.

$$L(Y,Y') = \frac{1}{N} \times \sum_{i=0}^{N-1} L(Y_i, Y'_i) \tag{1}$$

$$L(Y_i, Y'_i) = \frac{1}{2} \times \sum_{j=0}^{1} (Y_{j,i} - Y'_{j,i})^2 \tag{2}$$

The objective described above does not guarantee the conformity of the predictions to the physical constraint, i.e., the predicted coordinates must lie between the vocal tract walls. An additional cost was introduced to penalize the predictions outside the vocal tract for overcoming this. If a point is inside the vocal tract, the sum of its distance with both contours should be equal to the distance of the two contours at the same position (illustrated by Figure 4) as written in Equation 3 and Equation 4.

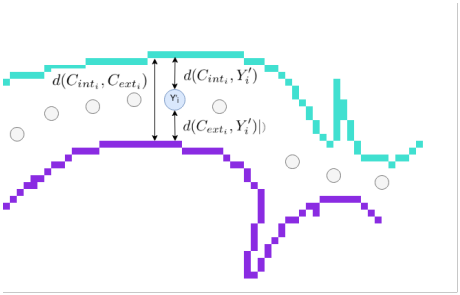


Figure 4. Illustration of the distance property between a centerline point and the nearest vocal tract contours points.

$$L_{out}(Y', C_{int}, C_{ext}) = \frac{1}{N} \times \sum_{i=0}^{N-1} L_{out}(Y'_i, C_{int_i}, C_{ext_i}) \tag{3}$$

$$L_{out}(Y'_i, C_{int_i}, C_{ext_i}) = |d(C_{int_i}, C_{ext_i}) - d(C_{int_i}, Y'_i) - d(C_{ext_i}, Y'_i)| \tag{4}$$

In addition, all centerline points are not equally important. For example, shifting a centerline point near a constriction gives rise to a more significant acoustic error than when the vocal tract is wider. Thus, the narrower the vocal tract, the more accurate the centerline at that point should be. The points are thus weighted according to the inverse of their distance with the vocal tract contours as shown by Equation 5 where ε was set to 10^{-4}.

$$Weight(C_{int_i}, C_{ext_i}) = \frac{1}{d(C_{int_i}, C_{ext_i}) + \varepsilon} \tag{5}$$

The complete updated loss is given by Equation 6.

$$Loss(C_{int}, C_{ext}, Y, Y') = \frac{1}{N} \times \sum_{i=0}^{N-1} Loss(C_{int_i}, C_{ext_i}, Y_i, Y_i') \tag{6}$$

$$Loss(C_{int_i}, C_{ext_i}, Y_i, Y_i') = (L(Y_i, Y_i') + L_{out}(Y_i', C_{int_i}, C_{ext_i})) \times Weight(C_{int_i}, C_{ext_i}) \tag{7}$$

2.3.2 Training parameters

The `PyTorch` library was used to create and train the neural network. Examples are also shifted during training by a small value to make the network more robust against translation. The Adam optimizer [13] was used with an initial learning rate of 10^{-3} which is decayed by a factor of 0.9 after five epochs without validation loss improvements. We used 500 epochs to train the network.

3 RESULTS

3.1 First results

Figure 5 illustrates some examples of the determination of the centerline by the Regression-B method and Table 1 shows assessment of both approaches. Regression-A experiment is the version of the regression without improving the loss function, and Regression-B with the improved loss function. A centerline is rejected as soon as it lies outside the vocal tract defined by the two contours C_{int} and C_{ext}. The Classification experiment has no rejection rate since it decodes the centerline within the vocal tract.

Table 1. Centerline distance and rejection rates for the three experiments.

	Regression-A	Regression-B	Classification
Centerline distance (in mm)	0.56	**0.50**	0.88
Rejection rate (in %)	36.30	40.75	0
Rejection rate for speech frames without extremities (in %)	10.25	8.86	0

It turns out that Regression-B gives the lowest distance between the reference and computed centerline but with a rejection rate of 23%. When looking at the results, it turns out that the high rejection rate is mainly due to the extremities slightly outside the vocal tract without changing the splitting of the vocal tract into tubelets. When the extremities are not considered, the rejection rate is only 8.86% for vocal tracts corresponding to speech and 23.07% for non-speech vocal tracts (silences, pauses, breaths, and swallowings). It should be noted that most rejections corresponding to speech are due to a small error. Concerning the Classification experiment, the lower accuracy is due to the low resolution of the rt-MRI images (136×136). The precision could be increased artificially by re-scaling images to a resolution of 1024×1024 but this would largely increase the time to perform inference and post-processing. We, therefore, decided to abandon this avenue.

4 Speed evaluation and improvements

Experiments were done using the following hardware with `PyTorch` framework:

- CPU: 11th Gen Intel(R) Core(TM) i7-11850H @ 2.50GHz
- GPU: NVIDIA T600

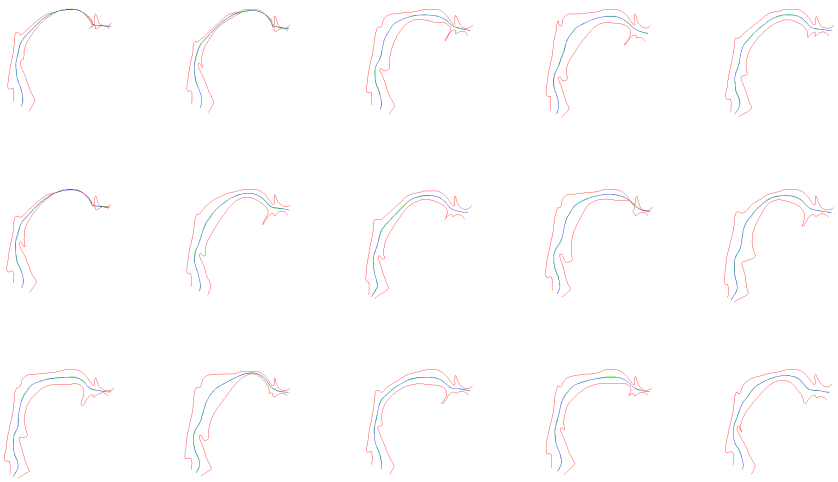


Figure 5. Results of the Regression-B on the test set. In red the vocal tract shape contours. In green the reference centerline and in blue the predicted centerline.

The Regression method is a lot faster (0.1044 ms with a batch size of 32, and the time used by the rejection test is negligible) than the Classification (38.57 ms on average, including post-processing). However, the time required by the Regression method has to consider the rejection cases. Indeed, the initial heuristic algorithm is used as a fallback solution in case of rejection. The average time required by the Regression method is thus 12.94 ms, approximately 20 times faster than the original solution.

Some of the hyperparameters used in the neural network architecture were set arbitrarily. This means that we may not have the best combination of feature maps and/or the number of layers. We thus searched for the best hyperparameters by providing the range of values for the convolution and dense networks in the Tree-Structured Parzen Estimator (TPE) [14] implemented in the Optuna framework [15].

We found that the best results on the validation test set were obtained with a network using 200 points for the vocal tract contours and 100 points for the centerline. Also, the optimal convolution sub-network uses 512 initial feature maps and a depth of 5. The optimal dense sub-network uses 256 neurons and a depth of 3. These values are not far from the original network since the new architecture only adds one depth to each sub-network.

With this network configuration, the rejection rate decreases to 6.18% for vocal tracts corresponding to speech with an average centerline error of 0.56 mm.

5 CONCLUSIONS

This work thus enables the determination of the centerline of the vocal tract to be drastically accelerated, which was the objective. Indeed, the results presented above show that the accelerated version is 20 times faster than the original algorithm while keeping the same level of precision. However, the training database relies on a heuristic algorithm using "common sense" acoustic criteria to express a cost function. There is no proof that this heuristic corresponds to the acoustical ground truth. The strength of this approach presented in this paper is that it can be applied to other centerline determination algorithms or even real data, provided that wave propagation in the vocal tract can be observed easily.

ACKNOWLEDGEMENTS

Authors acknowledge the CNRS for funding the engineer involved in this project and the ANR for funding the Full3DTalkingHead project in which this work takes place.

REFERENCES

[1] S. Stone, Y. Gao, and P. Birkholz, "Articulatory synthesis of vocalized /r/ allophones in german," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 879–889, 2022. DOI: 10.1109/TASLP.2021.3130969.

[2] A. Poznyakovskiy, A. Mainka, I. Platzek, and D. Mürbe, "Fast semiautomatic algorithm for centerline-based vocal tract segmentation," *Biomed Res Int.*, vol. Epub 2015 Oct 18. 2015. DOI: 10.1155/2015/906356.

[3] Z. I. Skordilis, A. Toutios, J. Töger, and S. Narayanan, "Estimation of vocal tract area function from volumetric magnetic resonance imaging," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 924–928. DOI: 10.1109/ICASSP.2017.7952291.

[4] Y. Laprie, M. Loosvelt, S. Maeda, E. Sock, and F. Hirsch, "Articulatory copy synthesis from cine x-ray films," in *Interspeech 2013 (14th Annual Conference of the International Speech Communication Association)*, Lyon, France, Aug. 2013.

[5] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1987.

[6] K. Singh and R. K. Kapania, "Accelerated optimization of curvilinearly stiffened panels using deep learning," *Thin-Walled Structures*, vol. 161, p. 107 418, 2021, ISSN: 0263-8231. DOI: https://doi.org/10.1016/j.tws.2020.107418. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263823120312817.

[7] V. Ribeiro, K. Isaieva, J. Leclere, P.-A. Vuissoz, and y. Laprie, "Automatic generation of the complete vocal tract shape from the sequence of phonemes to be articulated," *Speech Communication*, vol. 141, pp. 1–13, Apr. 2022. DOI: 10.1016/j.specom.2022.04.004. [Online]. Available: https://hal.univ-lorraine.fr/hal-03650212.

[8] I. Douros, J. Felblinger, J. Frahm, K. Isaieva, A. Joseph, Y. Laprie, F. Odille, A. Tsukanova, D. Voit, and P. Vuissoz, "A multimodal real-time mri articulatory corpus of french for speech research," in *INTERSPEECH 2019-20th Annual Conference of the International Speech Communication Association*, 2019.

[9] K. Isaieva, Y. Laprie, N. Turpault, A. Houssard, J. Felblinger, and P. Vuissoz, "Automatic tongue delineation from mri images with a convolutional neural network approach," *Applied Artificial Intelligence*, vol. 34, no. 14, pp. 1115–1123, 2020.

[10] V. Ribeiro, K. Isaieva, J. Leclere, P. Vuissoz, and Y. Laprie, "Towards the Prediction of the Vocal Tract Shape from the Sequence of Phonemes to be Articulated," in *Proc. Interspeech 2021*, 2021, pp. 3325–3329. DOI: 10.21437/Interspeech.2021-184.

[11] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, *Searching for mobilenetv3*, 2019. DOI: 10.48550/ARXIV.1905.02244. [Online]. Available: https://arxiv.org/abs/1905.02244.

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[14] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *International conference on machine learning*, PMLR, 2013, pp. 115–123.

[15] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

PROCEEDINGS of the
24th International Congress on Acoustics
October 24 to 28, 2022 in Gyeongju, Korea

ABS-0723

Estimation Of Japanese Word Intelligibility Using DNN-Based Speech Recognition Systems

Masaki HATTORI[1]; Kazuhiro KONDO[2]

[1] Graduate School of Science and Engineering, Yamagata University, Japan

[2] Graduate School of Science and Engineering, Yamagata University, Japan

ABSTRACT

Noise degrades speech, leading to the degradation of speech intelligibility and speech quality. We have been studying methods for estimating the speech intelligibility of degraded speech. There are two classifications of speech intelligibility estimation methods. One is the complete full reference method, which uses both degraded speech and the original speech before degradation. The other is the non-reference method, which uses only degraded speech. Estimation by speech recognition systems is one of the non-reference methods, and in our previous work, we have used GMM-HMM. We simulated JDRT with the system and achieved a certain degree of accuracy. However, a model that adapts to each speaker, noise, and noise level is needed. For practical use, a robust method that can support unknown speakers and noises is desired. In this research, we consider a method using DNN-HMM, which is known to perform significantly better than GMM-HMM provided enough data is available for training. The results for the speaker-independent model show correct response rates improved by 17% with the original speech and by 20% with 10 dB, relative to the untrained GMM-HMM. Additionally, we attempted feature adaptation using VTLN, SS, and CMVN as a countermeasure against speaker variation and noise.

Keywords: Speech intelligibility, Speech recognition, Deep neural network (DNN)

1. INTRODUCTION

In today's advanced wireless communication systems, people communicate by mobile phone and the Internet in noisy environments. It is known that ambient noise and reverberation can significantly degrade speech intelligibility. In the design and operation of voice communication systems, it is necessary to evaluate speech intelligibility for quality. Speech intelligibility is one of the speech quality scales and measures how accurately words and sentences are conveyed to listeners. The subjective evaluation method by human subjects is highly reliable, but it needs time and cost, and there are still issues regarding its stability. On the other hand, the objective evaluation method estimates the subjective evaluation based on the physical features of the observed speech signal. If a highly accurate objective evaluation method is developed, the problems with subjective evaluation can be solved. Objective evaluation methods fall into two categories. One is a full reference method that uses both the degraded speech to be evaluated and the original speech before degradation. The other is a non-reference method that estimates speech intelligibility only from the degraded speech. Although the full reference method is generally considered superior, the non-reference method is preferable because clean reference sounds are rarely available in practice.

Estimation using speech recognition is one method of objective evaluation of non-reference methods. Since speech recognition systems convert speech signals into text, they can simulate subjective evaluation tests. Recently, it has also attracted attention as a speech understanding process that reflects human auditory characteristics well. In our previous paper, Gaussian Mixture Model – Hidden Markov Model (GMM-HMM) was used [1]. Although a certain degree of accuracy was achieved by speaker and noise adaptation, some issues remain. For example, it is not possible to completely reproduce the trends in subjective evaluation, and a model for each speaker, noise, and

[1] t222987m@ st.yamagata-u.ac.jp
[2] kkondo@yz.yamagata-u.ac.jp

noise level is needed. In the future, robust models for unknown speakers and noise would be ideal.

In this paper, we estimate speech intelligibility using a Deep Neural Network – Hidden Markov Model (DNN-HMM) speech recognition system. DNN-HMM has been shown to outperform GMM-HMM in a general speech recognition task. The main reasons for this are the minimization of the probability of incorrect answers by the discriminative model and the natural capture of speech by the use of high-dimensional features. DNNs are also known to generalize well to unseen environments. Thus, we will attempt to improve the estimation accuracy using this higher accuracy speech recognition.

2. SPEECH INTELLIGIBILITY ESTIMATION

2.1 Japanese Diagnostic Rhyme Test (JDRT)

The JDRT is one of the subjective evaluation methods to measure Japanese word intelligibility. DRT uses word pairs that differ only in the first phoneme. In the test, subjects are given the word pair and made to listen to one of the words. The subjects choose the word they hear, and their intelligibility is calculated based on the percentage of correct responses. Most Japanese syllables are either a single vowel or a vowel-consonant combination. In JDRT, only words with the consonant-vowel combinations as the first syllable is used. Word pairs whose first phoneme differs by a single feature are selected. For example, the word pair "Zai" and "Sai" differ in the initial consonant by the voiced /z/ and unvoiced /s/. There are no other differences in the phonemes following these. Japanese consonants are categorized into the following six phonetic feature attributes:

- Voicing: Voiced and voiceless.
- Nasality: Nasal and oral.
- Sustention: Continuant and interrupted.
- Sibilation: Strident and mellow.
- Graveness: Grave and acute.
- Compactness: Compact and diffuse.

For each attribute, ten pairs are defined. The DRT word list consists of 60-word pairs, for a total of 120 words. Refer to [2] for a complete word list.

2.2 Automatic Speech Recognition (ASR)

ASR consists of two main models: an acoustic model and a language model. The acoustic model estimates phoneme sequences from speech features, and calculates the posterior probability $P(x|w)$ that a speech feature x is observed when a text sequence w is spoken. Furthermore, by dividing the acoustic model into a phoneme model and an utterance dictionary, words that do not exist in the speech data can be recognized. Define p as the phoneme sequence, and then calculate $P(x|p)$ for the phoneme model and $P(p|w)$ for the utterance dictionary. The language model defines the probability of the occurrence of a word $P(w)$. When simulating JDRT in a speech recognition system, words are chosen from two options, so the output probabilities of words are equal. This corresponds to isolated word recognition in the speech recognition task. Therefore, $P(w)$ is a uniformly distributed probability and can be ignored. From the above, the speech recognition system in this paper can be formulated as in Equation (1). The performance difference between recognition and estimation is determined by the acoustic model.

$$\hat{w} = arg\max_{w}\left\{\max_{p} P(x|p)P(p|w)\right\} \tag{1}$$

2.3 Feature conversion

An important factor in improving the accuracy of ASR is to reduce the mismatch between the input data and the training data of the model. Therefore, it is necessary to adapt the model itself to fit the features or to convert the features to fit the model. In DNN-HMM, adaptive training of the model itself, such as Maximum-Likelihood Linear Regression (MLLR), which was the mainstream method in GMM-HMM, cannot be easily applied. Hence, with DNN-HMM, the feature conversion is often employed to match the characteristics of the input data to the training data. In this paper, Vocal Tract Length Normalization (VTLN) and Spectral Subtraction (SS) are used as feature conversion methods. VTLN is one of the speaker adaptation methods. It transforms the features to compensate for the speaker's vocal tract length and the changes in vocal tract shape, which is considered to be equivalent

to a warping operation on the frequency axis. SS is the simplest noise suppression method. It subtracts the power spectrum of noise estimated from the voiceless segment of the input speech from the noisy speech.

3. EXPERIMENTS

We will attempt to use feature conversion with the DNN-HMM to estimate the speech intelligibility of the JDRT speech, and evaluate its estimation accuracy.

3.1 Sound Source

The clean speech of the JDRT word list to be evaluated was recorded by four male speakers and four female speakers. The recording conditions were monaural, 16-bit quantization, and 16 kHz sampling frequency. Added noises are Gaussian white noise, babble (multi-talker) noise, and pseudo-noise (white noise filtered with frequency characteristics matching the average multi-talker noise spectrum). These three types of noise were added at S/N ratios of +10dB, 0dB, -10dB, and -15dB with respect to the original voice.

3.2 ASR setup

We used a trained model from the Julius dictation kit ver. 4.5 [3] as the acoustic model. This model is a gender-independent (GID) and speaker-independent model. The features are mean and variance normalized (Cepstrum Mean and Variance Normalization; CMVN). Each layer of the DNN is constructed by initialization using the Restricted Boltzmann Machine (RBM), fine-tuned using the cross-entropy criterion, and series training using the state-level Minimum Bayes Risk (sMBR) criterion. The decoder is the Julius ver. 4.6 [4]. In addition, several settings were made to simulate JDRT. The language model was defined for 60 pairs, limiting the network of grammars to a two-way choice. During recognition, the language model corresponding to a word pair is selected. The definition of the utterance dictionary includes 120 additional words and their phoneme sequences. For feature transformation, Julius can apply SS and VTLN by applying the options -sscalc and -vtln. The parameters were set at static default values in Julius.

3.3 Evaluation methods

We simulated the JDRT using the speech recognition system and speech. Recognition rate is determined from the output text and the percentage of correct responses is calculated using the standard formula (2) used in the DRT.

$$S = (R - W)/T \times 100[\%] \quad (2)$$

Here, S is the correct response rate (intelligibility), R is the number of correct answers, W is the number of incorrect answers, and T is the total number of attempts. The correct response rate is 100% for all correct answers and -100% for all incorrect answers. Random responses ideally result in 0%.

4. RESULTS AND DISCUSSIONS

Table 1 shows the percentage of correct responses when each feature conversion method is applied. In general, speaker adaptation is expected to improve the recognition rate, especially for the original speech. Noise adaptation is expected to improve the recognition rate for noisy speech, and decrease the recognition rate for the original speech.

The application of VTLN showed a 1.49% increase in the percentage correct versus the default for the original speech, but contrary to expectations decreased for all noisy speech. The application of SS showed a 1.92% decrease in correct responses relative to the default for the original speech, but an increase when noise was added except for SNR -10 dB. The results of the applying SS showed a 1.92% decrease in the percentage correct versus the default for the original speech, and an increase except at SNR -10 dB for noisy speech. CMVN+VTLN+SS increased the correct rate by an average of 1.81 versus the default, but the correct rate for the original speech did not recover, decreasing by 1.5%. It may be necessary to optimize the VTLN application for each speaker. SS ignored not only noise, but also important speech features.

Fig. 1 shows the trend of correct responses by noise type, while Fig. 2 shows the trend of correct responses by consonant attribute. The correct response rate for Sibilation was very high for the original and SNR 10 dB but decreased rapidly as the noise increased. Since the characteristics of phonemes in the sibilation category are similar to those of white noise, these phonemes most likely

became indistinguishable from noise. The other consonant attributes showed more than 60% of correct responses for the original speech. However, as the noise increases, the degradation is severe, and at SNR -10B and -15dB, words were not recognized at all.

Fig. 3 shows a comparison of DNN-HMM with subjective and GMM-HMM. As can be seen, the GMM-HMM with adaptation is very close to subjective intelligibility. DNN-HMM, which does not employ adapted models, shows a much lower rate at all conditions but is better at high SNR conditions compared to unadapted GMM-HMM. We believe that with model adaptation, DNN-HMM will show an accuracy much closer to subjective intelligibility.

Table 1 – Percentage of correct responses in applying each feature conversion method

Condition	SNR (dB)				Original	Average
	-15	-10	0	10		
CMVN (Default)	1.97	6.07	24.12	54.54	78.23	32.99
VTLN	2.50	5.03	19.06	50.69	79.72	31.40
SS	3.32	5.20	28.71	58.80	76.31	34.47
CMVN+VTLN+SS	3.41	5.75	29.49	58.59	76.73	34.80

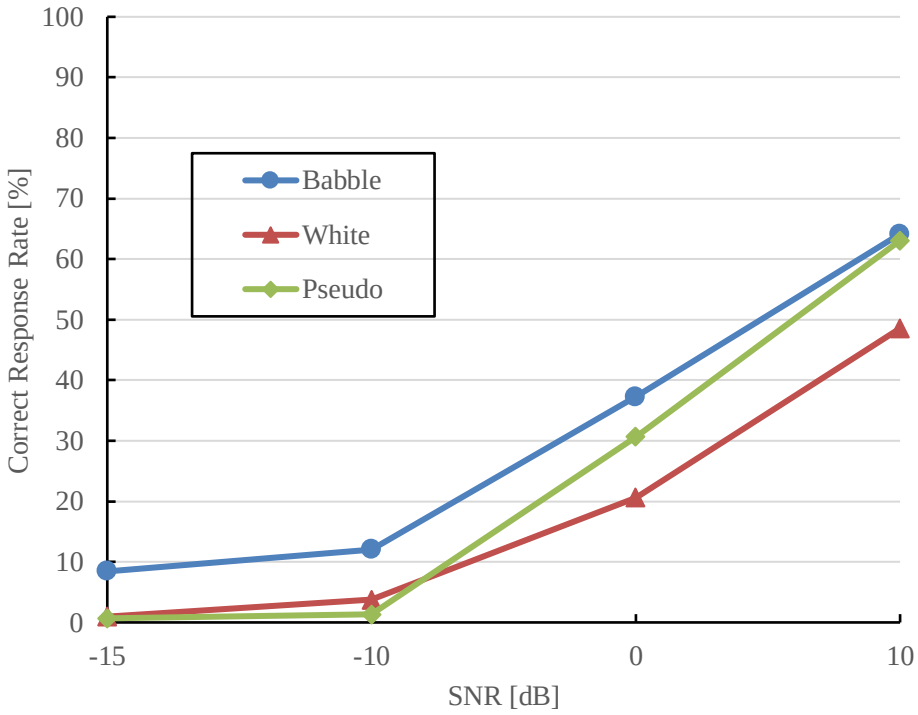


Fig. 1 – Correct response rate with DNN-HMM for each noise vs. SNR

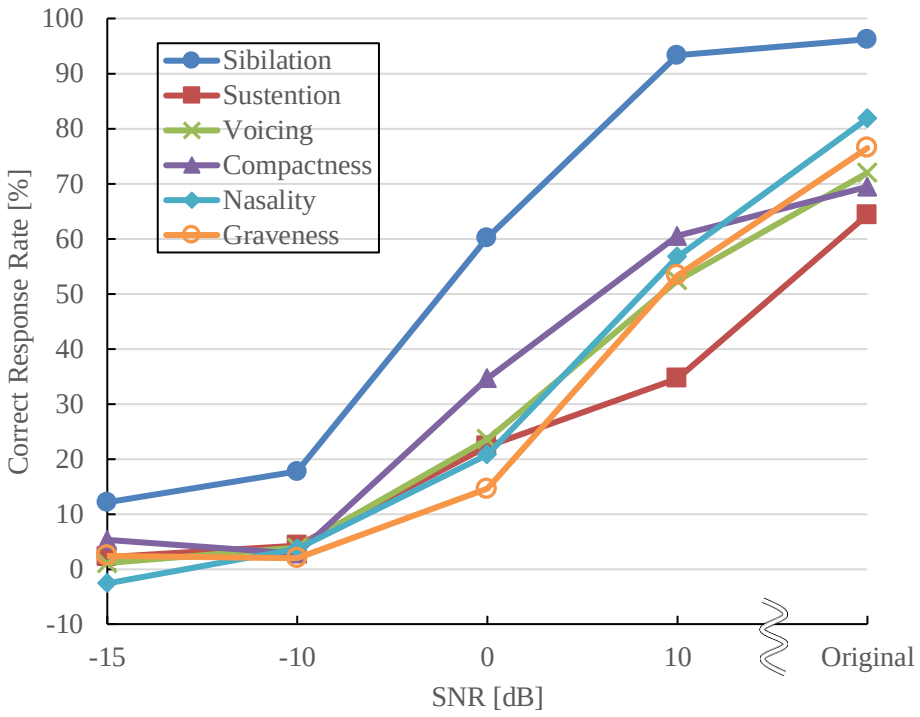


Fig. 2 – Correct response rate with DNN-HMM for each attribute vs. SNR

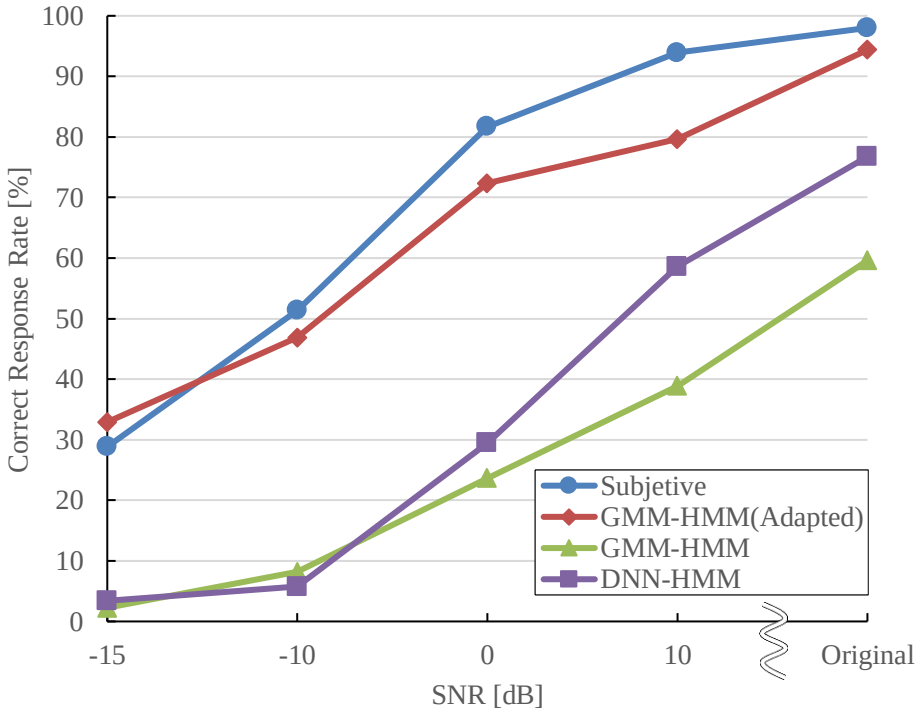


Fig. 3 – Correct response rate with DNN-HMM, GMM-HMM and Subjective vs. SNR

5. CONCLUSIONS

In this paper, we examined the estimation of Japanese word intelligibility using a speech recognition system. To improve the estimation performance compared to previous results using GMM-HMM, we attempted to use DNN-HMM. First, we simulated the JDRT using the speaker-independent model DNN-HMM and obtained estimation results. The results showed a significant improvement in accuracy for the original sound and low noise compared to the unadapted GMM-HMM. Next, VTLN and SS, a simple feature conversion method, were used to further improve the original speech recognition rate and as a solution against noise. The results showed only a slight improvement overall, and the negative effects of each method were also observed. However, full-scale adaptive learning on DNN-HMM is also expected to be better than the noise- and speaker-adapted GMM-HMM.

In the future, we plan to use Kaldi to build a speech recognition system with speaker and noise adaptation. We are also interested in recent speech recognition systems, such as the end-to-end (E2E)-ASR without HMMs and cloud-based ASR. We plan to develop intelligibility estimation using these technologies.

REFERENCES

1. Y. Takano, K. Kondo, "Estimation of speech intelligibility using speech recognition systems", IEICE Trans. Inf. & Syst., Vol. E93-D, No. 12, Dec. 2010
2. K. Kondo, et al. "Two-to-one selection-based Japanese speech intelligibility test", J. of Jap. Acoust. Soc., vol. 63, no.4, p. 196-205, 2007.4
3. A. Lee, et al. 2019, Julius-dictation-kit, https://github.com/julius-speech/dictation-kit, (2022).
4. A. Lee and T. Kawahara: Julius v4.5 (2019). https://doi.org/10.5281/zenodo.2530395

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0758

Development of a vocal-coaching system for disaster prevention broadcasts based on the advices of a professional vocal-coach and the environmental sound feedback

Sayoko TAKANO[1]; Yoshio TSUCHIDA[2]

[1] Kanazawa Institute of Technology, Japan

ABSTRACT

We develop a vocal-coaching system of disaster prevention broadcasts for people to sense its emergency and evacuate immediately, because the people who may announce at the crises in the local area are not professional speakers. We have been investigated the instruction words and advices of an professional vocal-coach. In this research, we developed a vocal-coaching system that implements 1) the order of the instructions were considered based on the professional vocal coach and 2) the recorded sound was played after the subject's speech with simulated environmental sound. The instructions were louder, impatiently, clearly, with pause and all with "eager to escape". The simulated environmental sounds have five echo of reflections and frequency characteristics of loud speakers.

Fourteen adults joined the experiment. We expected that instructions and the played sounds might realize the subjects to speak louder and articulate precisely. The subjects speak louder and higher fundamental frequency, however, the results were not better than the professional vocal-coach's instructions. We need further investigations, i.e., the vocal-coach's empirical knowledge and skills of the vocal-coaching.

Keywords: Disaster prevention broadcasts, Vocal-coaching system, Feedback of environmental sound simulation

1. INTRODUCTION

Japan is a disaster-prone country, and evacuation announcements from loudspeakers have saved many lives. At the same time, it has been also pointed out that some people were not evacuated due to the normalcy bias even the hardware worked fine to call their evacuation.

One of this co-author had built an sound simulation feedback system for announcers to notify that the sound in reality with loudspeakers is degraded with noise and echoes, however this system was not much used so far (1).

Many studies have been done on professional speakers' analysis (2) and synthesized speech for evacuation broadcasting (3). We have trained announcers and general person to know the best instruct words. We found that effective instruction were; "to be hasty", "in a commanding tone", and "strongly" (4),

On the other hand, the general people's speech improvement is relatively weak, so we decided to employed a professional vocal coach. We asked him to train people as his way within ten times without any limitations (5). He gives advice for the speaker from the basics in the beginning, and more specific one in the latter. His advice by ten times was effective, rather exploring specific words in a random order. Therefore we decided to built a simple vocal coaching system based on his advice (5).

In this research, we investigate the improvements of the instructions mimicking the vocal coach, and the effects of the sound simulation of the environment for feedback which was developed previously.

[1] tsayoko@neptune.kanazawa-it.ac.jp
[2] tsuchida@ neptune.kanazawa-it.ac.jp

2. DEVELOPMENT OF THE VOCAL COACHING SYSTEM

2.1 Training System and the sound analysis

Our future goal is to build an adaptive vocal coaching system for immediate evacuation, however, this system only records the speech sound and gives the determined instructions to the speaker. These instructions should be described in an Excel file in advance. They are based on the previous coaching experiment from an professional vocal coach, and the effects of these instructions are already examined (5).

In this experiment, five selected instructions are shown in Table 1.

Table 1 – The five instructions in vocal-coaching experiment

	Instructions in Japanese	Translation in English
1	もっと大きな声で発話してください。	Please speak louder.
2	いいですね！ では、もう少し焦って発話してください。	Very good! Now, please be in a hurry.
3	次は言葉が明瞭になるように注意しましょう。	Next, please speak clearly.
4	その調子です！ さらに、言葉の 間 をとることを意識しましょう。	Right on! Pay attention to the pause between the words.
5	最後です。では、今までのことをふまえて 逃げる気になるようにアナウンスしてください。	This is the last recording. Please announce as if people to feel “eager to escape”.

The fundamental frequency is calculated, and also the Japanese speech sound are segmented by each phonemes. The sound analysis are realized with WORLD by (6. 7) and the values of fundamental frequency is exported as CSV format in this system. Also, the automatic segmentation is realized by Julius (8).

The total system arrows the feedback of the sound simulation of environment, such as echo, rainy sound, and degradation. The “play” button plays the recorded sound with the 5 delayed sounds, recorded rainy sound and frequency response imitating the speaker’s frequency response.

Inside of the training system is independent from the sound simulation software. The training system copies the sound file to the target folder, and the sound simulation software observes the target folder every one second. The sound simulation software activates and play with the simulation when a new specified file is found.

2.2 Sound simulation of the environment

This sound simulation software is built to imitate the disaster prevention loudspeaker and its environment echo with rainy sound (1). The interface of this software is named “TRANSIS (previously, AEGIS)”, made by Mr. Koyama based on Max 8 (cycling ’74) (1). This software applies the echoes, pre-set rainy sound, and a pre-set frequency degradation of the loud speaker. The echoes (delay) can be set automatically (random) or manually.

In this experiment, five independent delayed sound are added; 150 m, 210 m, 390 m, 715 m, and 1000 m. The speed of the sound is calculated as 340 m/s.

3. VOCAL COACHING EXPERIMENT

3.1 Agenda

The vocal coaching with instructions for five times and the feedback of the sound environment simulation will be tested in this experiment. The improvement of the subject s’ speech is evaluated by the change of the fundamental frequency. It is known that the relatively higher frequency in the same person is judged to be more effective for the evacuation broadcasts based on the production experiments and the speech re-synthesis experiments (3). It is probably because the higher tension

naturally make the higher expiratory pressure and higher fundamental frequency.

The fundamental frequency is different from person to person, so the first speech without sound simulation of the environment is used for the basis of the each person's original fundamental frequency.

3.2 Methods

The vocal coaching system which were described in this paper in the section 2 is used. The settings are described above. Each subjects are asked to say "Chikaku no kawaga hanran shimashita. Tadachini shiteino syougakkouni hinan shitekudasai," meaning "A river in your neighbor is flooded. Please evacuate an assigned elementally school immediately."

The recording were performed in a quiet office. The condenser microphone (NT2-A, RODE) is placed about 30 cm from the speaker's lip. The recordings were 44.1 kHz sampling with 16 bit with the system.

All subjects are asked two conditions, with feedback of the sound environment simulation: 'Y'(yes), and without any: 'N' (no). The feedback sound were played by a set of loud speakers (A80, AIRPULSE) with comfortable volume. The order of 'Y' and 'N' were changed for each subject. Fourteen male subjects joined the experiment,

The conversions from [Hz] to [cent]

Each person has different fundamental frequency, also logarithmic unit for the sound perception should be used to compare the change of the fundamental frequency. In this paper, the first speech of each person's fundamental frequency without sound simulation feedback is used to for this basis.

The definition of the musical intervals 'n' in cent is described as equation (1).

$$n = 1200 \cdot \log_2\left(\frac{b}{a}\right) \approx 3986 \cdot \log_{10}\left(\frac{b}{a}\right) \quad (1)$$

Where 'a' is the each person's fundamental frequency of the basis, and 'b' is the target fundamental frequency compared with the value 'a'. Therefore the first speech without sound simulation feedback is value 0 in cent.

3.3 Results

3.3.1 Improvement from the instructions

Ordinarily subjects change their speech reacting to the instructions. One of the easiest value to examine the change is fundamental frequency, although the goal is not to produce higher fundamental frequency. Unfortunately, one subject did not show any change for his speech sound, probably because he did not understand enough the instructions.

Figure 1 shows the change of the fundamental frequency according to the instructions. The sample "0" means the first take without any instructions, and the sample "1" indicates the recording after the first instruction and so on. The 'silB' and 'sp' stands for 'silent beginning' and 'space'.

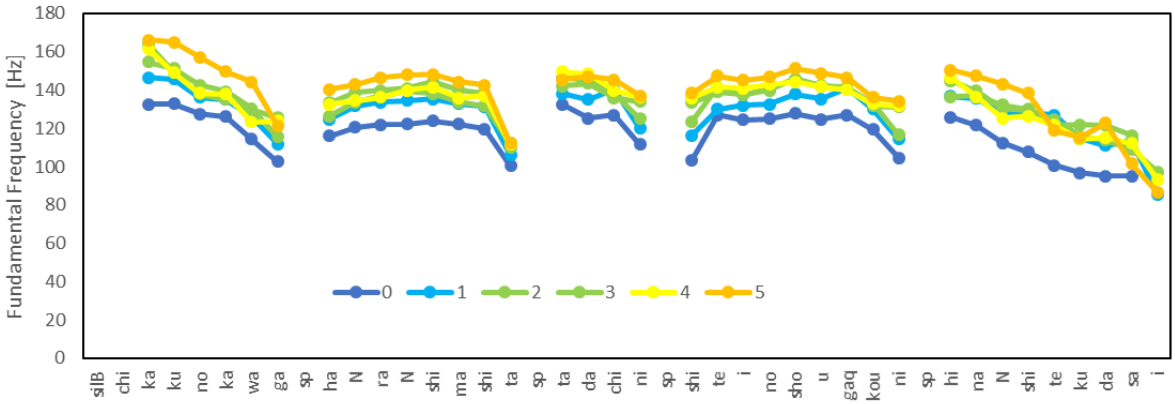


Figure 1 – One example of the fundamental frequency from a subject 8.

This subject's speech segmentation was successful, however, the segmentation or extraction of the fundamental frequency could fail. We observed and analyzed the least fail words "hanran (flood)" for the comparison of the standardization for the next section.

3.3.2 Effects of the sound simulation of the environment

The fundamental frequency of each instructions and the feedback of the sound environment simulation are compared in a logarithmic unit. The conversion from [Hz] to [cent] is described in the method section. All the subjects' average (except the one subject) is shown in Figure 2.

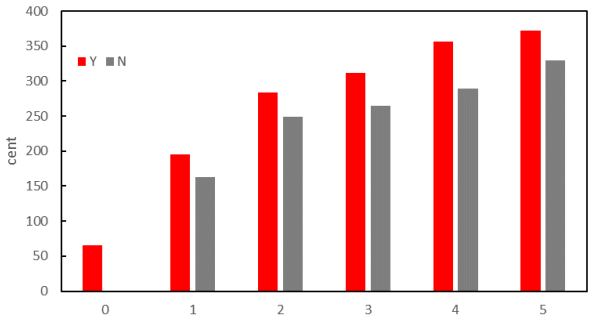


Figure 2 – The change of the fundamental frequency in cent (average of 13 subjects)

Figure 2 shows the average change of the fundamental frequency in cent. The fundamental frequency becomes higher from 1st to 5th as described in the previous section. The differences in fundamental frequency from the beginning to 5th were about 300 in cent; i.e., about 1/4 octave.

'Y' and 'N' stand for with or without the sound simulation feedback, respectively. 'Y' is lightly higher than 'N', and these indicate the feedback of the sound simulation slightly effects their speech production.

3.4 Discussions

The vocal coaching with instructions and the feedback of the sound environment simulation were examined. We have been reported vocal coaching with right instructions with right order may effective for improvement of the speech sound for evacuation broadcasts. This effect agreed to elicit the better speech.

In this experiment, also the feedback of the sound environment simulation was examined. The feedback of the own voice adding the sound environment simulation is slightly higher than without feedback. These mean that the instructions and feedback with sound environment simulation can change the speech. Probably people pay more attention for their own voice and they notice that they need more effort for their announcement.

In reality vocal coaching migh rarely give recorded sound, specially for the broadcasts in the air with echoes and rainy sound in the air. These environmental sound simulations with own speech sound can also help to improve their speech.

4. CONCLUSIONS

In this study, we examined the improvements for disaster prevention broadcasting by five selected instructions and the effects of environmental sound simulation. The five instructions could have changed the fundamental frequency about 1/4 octave higher. Also, the environmental sound simulation effected positively.

In the future, it may be possible to provide appropriate instructions automatically by collecting voices in the system,

ACKNOWLEDGEMENTS

We appreciate people who have joined this work, specially the volunteers and the experimenters. The original vocal-coaching system was mainly programmed by Mr. Ito who graduated Kanazawa Institute of Technology. The interface of the sound simulation software was built by Mr. Koyama. The vocal-coaching experiments were advised by a professional voice coach, Mr. Nagatsuka.

This work is supported by JSPS Kakenhi 18H03490, JST JST-Mirai Program Grant Number JPMJMI18D1, and SCOPE 201605002, Japan.

REFERENCES

1. Tsuchida, Y. Basic consideration on training system for announcement of public address in disasters (in Japanese), Summaries of technical papers of annual meeting, Architectural Institute of Japan, p. 435-436, 2013.
2. Kobayashi,M., Hamada,Y.,and Akagi.M. Acoustic features in speech for emergency perception. The Journal of the Acoustical Society of America. 2018. 144. 3. p. 1835-1835.
3. Takano, S. and Tsuchida, Y. Effective announcements during disasters for immediate evacuation: Emphasis on each sentence, Acoustical Science and Technology, 2021. 42. 4, p. 200-201.
4. Takano, S. and Tsuchida, Y. Studies on suggesting words to speakers in order to evoke the evacuation behavior by announcements for disaster prevention, Part 2: Characteristics of the acoustical impressions (in Japanese), ASJ Reports of autumn meeting the Acoustical Society of Japan. 2018. p413-414.
5. Nagatsuka Z., Takano. S. and Tsuchida, Y. Toward building a voice training system for motivating people to evacuate －Instructions to elicit emotional voice－ (in Japanese), ASJ-SP. 2022. 2. 1. p. 51-54..
6. Morise,M., Yokomori,F.,and Ozawa, K. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, IEICE transactions on information and systems, 2016. vol. E99-D, no. 7, p.1877-1884.
7. Morise,M.“D4C, a band-aperiodicity estimator for high-quality speech synthesis, Speech Communication. 2016. 84. p. 57-65.
8. Lee, A., and Kawahara, T. Julius v4.5. https://doi.org/10.5281/zenodo.2530395 2019..

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022, in Gyeongju, Korea

Extended abstract

ABS-0819

Deep Learning to Estimate Vocal Tract Geometry from Acoustic Impedance Spectrum

Balamurali B T; Jer-Ming Chen

balamurali_bt@sutd.edu.sg, jerming_chen@sutd.edu.sg

Science, Mathematics and Technology, Singapore University of Technology and Design, Singapore

ABSTRACT

Determining accurate and acoustically meaningful vocal tract geometry directly from speech sounds is a classic inverse problem. This problem is further exacerbated by the paucity of data - because the voice, with typical frequency >100 Hz, samples the frequency domain poorly, and direct methods of measuring vocal tract geometry are mostly invasive, uncomfortable, or slow. In this study, a data-driven approach using artificial neural networks (ANNs) is proposed to address the inverse area function problem, i.e., to determine the vocal tract geometry from input acoustic impedance spectrum. The vocal tract geometry was modelled as a tube of nonuniform cylindrical cross-sections, and four different combinations of cylinders were considered. The predicted radii for various cylindrical pipe combinations by ANNs when compared against the actual radii were found to have high correlation for all the selected radii in the three- and four-cylinder model (Pearson correlation coefficient and Lin concordance coefficient exceeded 95%); however, for the subsequent five- and six-cylinder model, the correlations were expectedly lower (Pearson correlation around 75% and Lin concordance around 69%). Upon standardizing the impedance value and retraining however, the correlation between predicted and actual radii improved significantly for all the cases (Pearson correlation and Lin concordance exceeded 90%).

Keywords: Vocal tract, Area function, Inverse problem, Artificial neural networks.

1. INTRODUCTION

Determining accurate and acoustically meaningful vocal tract geometry information directly from speech sounds is a classic inverse area function problem in speech science. Further, understanding how the shape of the vocal tract (its articulation) influences the output speech sound (from the lips) generated by an acoustic current (made at the glottis) is crucial to vocal tract acoustics and voice research. X-ray radiography, computed tomography, and magnetic resonance imaging (MRI) are some of the direct methods for determining the geometry of the vocal tract [1], [2] . The techniques are unsatisfactory because they are either highly invasive (endoscopic study: anaesthetics are required, and the probe must pass through the velum), expensive and potentially harmful (x-ray fluoroscopy), poorly resolved/difficult to interpret (ultrasound imaging), or unnatural (MRI: subject lies supine, often long measurements, and very noisy). The audio output (voice) has been used in studies employing indirect measurement methods to elucidate the vocal tract geometry [3], [4]. Another indirect method is to use the acoustic impedance of the vocal tract [5], [6], which could be measured directly at the lips and provides highly resolved frequency information in a non-invasive manner.

In this investigation, a data-driven approach using artificial neural networks (ANNs) was employed to sidestep the inverse area function problem while determining the non-linear relationship between vocal tract impedance (output) when compared with its corresponding vocal tract geometry (input). The vocal tract was approximated using concatenated cylindrical pipes (See Figure 1) and the resulting impedance at the glottis is determined analytically [7]. Altogether four different combinations of cylindrical pipes were considered in modelling the vocal tract [8]. The distance from lips for all the four combinations are also shown in Figure 1.

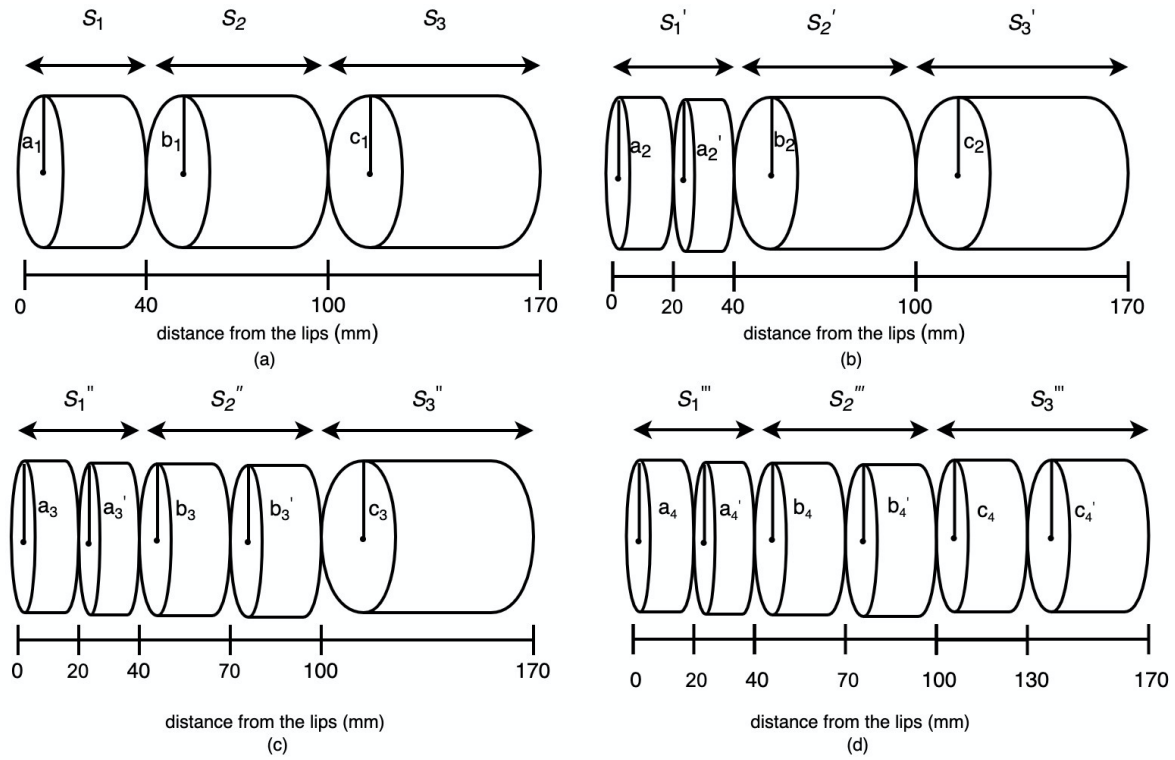


Figure 1Vocal tract combinations modelled (a) three-cylinder approximation (b) four-cylinder approximation (c) five-cylinder approximation (d) six-cylinder approximation ([8]).

In three-cylinder approximation, the radii of first two sections (S_1, S_2) closer to the lips (i.e., a_1 and b_1) vary from 2 mm to 25 mm and 2 mm to 20 mm, respectively, in steps of 1 mm. The radius of the final section S_3 (i.e., c_1) varies from 5 mm to 12 mm, in steps size of 1 mm (See Figure 1(a)). The radii of four-cylinder approximation (i.e., a_2, a_2', b_2 and c_2) are varied similar to a_1, b_1 and c_1. Finally, for the five- and six-cylinder approximation, the radii of sections are varied across the same range as that of the earlier three- and four-cylinder approximations, however, with a slightly lower resolution (i.e., with a slightly larger step size of 1.5 mm compared to 1 mm in five-cylinder and 2 mm in six-cylinder) [8]. The radii combinations for the various cylindrical pipe sections were selected in accordance with [9], [10].

A deep neural network (DNN) was developed using the keras machine learning package to estimate the radii of voice tract section approximated as a combination of cylindrical pipes. DNN with six layers was deployed in this investigation [11], [12]. Except for the final hidden layer, which has 16 neurons, each hidden layer has 100 neurons. With different cylindrical combinations, the number of neurons in the output layer varies (from 3 to 6), and this final layer predicted the regressed radii. Each hidden layer in this model has a rectified linear unit (ReLU) as an activation function, followed by a linear activation unit in the output layer. To stochastically optimize the weights of neurons in hidden layers, the 'Adam' optimization method was applied.

With different cylindrical models, the number of combinations for radii length and the resulting acoustic impedance vary. For instance, in a three-cylinder model, 3648 input-output vector pairs were generated and for the four-cylinder model, there would be a total of more than 80 000 combinations; similarly, increasing larger datasets for five-/six-cylinder models were created. As this would lead to an impossibly large number of combinations to examine for latter two, a subset of 'only' 100 000 randomly selected combinations and their impedances was chosen; 70% of the combinations of all cylindrical model was used for training the deep neural network and then tested on the remaining 30%. The maximum-normalized impedance vectors along with the radii values are used to train the neural network to solve this vocal tract area-function inverse problem.

Pearson's correlation coefficient ρ and Lin's concordance coefficient ρ_c are estimated for deriving the model prediction accuracy. ρ and ρ_c can take values between [-1, +1]; the coefficients are measures of closeness of the prediction (the higher the coefficient, the better is the prediction accuracy) to the actual value [13][14].

2. Result and Conclusions

The predicted and actual radii for various cylinder combinations were compared, and the resulting ρ and ρ_c are shown in Table 1. The effect of input standardization on prediction accuracy was also explored, which was accomplished by changing the distribution of max-normalized impedance values to have a zero mean and unit standard deviation. As can be seen, ρ_c was always less than ρ, which is to be expected. The ρ and ρ_c were found to be high for all selected radii combinations in the three- and four-cylinder approximations; however, the correlation declined in the five- and six-cylinder approximations. But with the standardization procedure, the correlation in the prediction accuracy improved for all radii and especially for a_4 in which improvement is by almost 25% suggesting the procedure has a positive impact on the prediction accuracy. In conclusion, our predictions using a data-driven approach with or without standardization in place had resulted in similar or better predictions for most of the simulated cases when compared against predicted vocal tract shape reported in [9] where their fitting predicted vocal tract shape to a resolution of about a centimeter.

Table 1Resulting prediction accuracy with out and with standardization in place.

Simulation	Radius Symbol	Without Standardization		With Standardization	
		ρ (%)	ρ_c (%)	ρ (%)	ρ_c (%)
Three-cylinder Approximation	a_1	98.2	96.6	99.8	99.8
	b_1	99.2	99.2	99.9	99.9
	c_1	97.9	97.4	99.5	99.4
Four-cylinder Approximation	a_2	95.1	94.9	98.6	98.5
	$a_2{}'$	99.5	99.4	99.9	99.9
	b_2	99.8	98.6	99.9	99.9
	c_2	99.3	99.3	99.9	99.9
Five-cylinder Approximation	a_3	87.7	86.6	96.9	96.7
	$a_3{}'$	92.1	91.7	99.0	98.9
	b_3	98.7	98.6	99.9	99.9
	$b_3{}'$	99.8	99.8	99.9	99.9
	c_3	99.2	97.6	99.9	99.9
Six-cylinder Approximation	a_4	73.1	69.4	92.8	92.7
	$a_4{}'$	87.0	86.2	97.3	97.1
	b_4	98.1	98.1	99.7	99.7
	$b_4{}'$	99.9	99.9	99.9	99.9
	c_4	99.5	99.5	99.9	99.9
	$c_4{}'$	98.6	97.6	99.9	99.9

Although not strictly physiological, this proof-of-concept approach strategically complemented recent studies applying similar deep neural network techniques to address other intractable inverse problems in voice science such as vocal fold mechanics, subglottal pressure, and other physiological control parameters [15]–[17]. This unified insight could pave the way to meaningfully resolve fundamental questions of vocal tract geometry, articulatory conditions, and voice mechanics for both speech and singing, and thus offers the potential of a diagnostic voice tool for applications in vocal disorders, speech pathology, voice therapy, and language pronunciation training in a natural, ecological, and non-invasive context.

REFERENCES

[1] T. Baer, J. C. Gore, L. C. Gracco, and P. W. Nye, “Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels,” *J. Acoust. Soc. Am.*, vol. 90, no. 2, pp. 799–828, Aug. 1991, doi: 10.1121/1.401949.

[2] A. Soquet, V. Lecuit, T. Metens, and D. Demolin, “Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI,” *Speech Commun.*, vol. 36, no. 3–4, pp. 169–180, Mar. 2002, doi: 10.1016/S0167-6393(00)00084-4.
[3] H. Wakita, “Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms,” *IEEE Trans. Audio Electroacoustics*, vol. 21, no. 5, pp. 417–427, Oct. 1973, doi: 10.1109/TAU.1973.1162506.
[4] S. Dusan and L. Deng, “Recovering vocal tract shapes from MFCC parameters,” in *5th International Conference on Spoken Language Processing (ICSLP 1998)*, Nov. 1998, p. paper 0367-0. doi: 10.21437/ICSLP.1998-795.
[5] N. Hanna, J. Smith, and J. Wolfe, “Frequencies, bandwidths and magnitudes of vocal tract and surrounding tissue resonances, measured through the lips during phonation,” *J. Acoust. Soc. Am.*, vol. 139, no. 5, p. 2924, May 2016, doi: 10.1121/1.4948754.
[6] M. Garnier, N. Henrich, J. Smith, and J. Wolfe, “Vocal tract adjustments in the high soprano range,” *J. Acoust. Soc. Am.*, vol. 127, no. 6, pp. 3771–3780, Jun. 2010, doi: 10.1121/1.3419907.
[7] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*. New York, NY: Springer New York, 1998. doi: 10.1007/978-0-387-21603-4.
[8] B. B T, S. Kapoor, and J.-M. Chen, “Estimating vocal tract geometry from acoustic impedance using deep neural network,” *JASA Express Lett.*, vol. 2, no. 3, p. 034801, Mar. 2022, doi: 10.1121/10.0009599.
[9] Rodriguez, A, Hanna, N, Goios Borges De Almeida, A, Smith, J, and Wolfe, J, “Estimation of vocal tract and trachea area functions from impedance spectra measured through the lips,” 2018, doi: 10.26190/ASF1-SM30.
[10] B. H. Story, I. R. Titze, and E. A. Hoffman, “Vocal tract area functions from magnetic resonance imaging,” *J. Acoust. Soc. Am.*, vol. 100, no. 1, pp. 537–554, Jul. 1996, doi: 10.1121/1.415960.
[11] F. Chollet, *“Keras*. [Online]. Available: ” https://keras.io Last viewed 23/10/2021
[12] S. Lakkam, B. T. Balamurali, and R. Bouffanais, “Hydrodynamic object identification with artificial neural models,” *Sci. Rep.*, vol. 9, no. 1, p. 11242, Dec. 2019, doi: 10.1038/s41598-019-47747-8.
[13] J. Liu, W. Tang, G. Chen, Y. Lu, C. Feng, and X. M. Tu, “Correlation and agreement: overview and clarification of competing concepts and measures,” *Shanghai Arch. Psychiatry*, vol. 28, no. 2, pp. 115–120, Apr. 2016, doi: 10.11919/j.issn.1002-0829.216045.
[14] L. I. Lin, “A concordance correlation coefficient to evaluate reproducibility,” *Biometrics*, vol. 45, no. 1, pp. 255–268, Mar. 1989.
[15] P. Gomez, A. Schutzenberger, M. Semmler, and M. Dollinger, “Laryngeal Pressure Estimation With a Recurrent Neural Network,” *IEEE J. Transl. Eng. Health Med.*, vol. 7, p. 2000111, 2019, doi: 10.1109/JTEHM.2018.2886021.
[16] E. J. Ibarra *et al.*, “Estimation of Subglottal Pressure, Vocal Fold Collision Pressure, and Intrinsic Laryngeal Muscle Activation From Neck-Surface Vibration Using a Neural Network Framework and a Voice Production Model,” *Front. Physiol.*, vol. 12, p. 732244, Sep. 2021, doi: 10.3389/fphys.2021.732244.
[17] Z. Zhang, “Estimation of vocal fold physiology from voice acoustics using machine learning,” *J. Acoust. Soc. Am.*, vol. 147, no. 3, p. EL264, Mar. 2020, doi: 10.1121/10.0000927.

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022, in Gyeongju, Korea

Extended abstract

ABS-0857

Predicting vocal tract geometry trained from acoustic impedance of elliptical cylinders

Prachee PRIYADARSHINEE; Balamurali BT; Christopher Johanne CLARKE; Jer-Ming CHEN

Audio Research Group, Singapore University of Technology and Design, Singapore

ABSTRACT

The finite element method (FEM) was used to derive the acoustic impedance spectra of the vocal tract where the vocal tract is modelled as a combination of elliptical cylinders with different ellipticity factors (EF). A data-driven approach is then developed to get around the inverse problem of estimating the geometry of the vocal tract. In this study, a deep neural network is trained with the impedance data from the FEM generated models for three-cylinder approximation of the vocal tract. Earlier efforts to model the impedance of vocal tract was limited simply to circular cross-sections, but FEM provides the flexibility to model complex shapes of the vocal tract, making it closer and more realistic to the actual shape. Accordingly, a comparison of elliptical and circular cross-sections of cylinders is performed to examine the differences in the corresponding impedances of the vocal tract. The effect of the number of cylinders and the different EF are explored to better approximate and understand the benefits and difficulties of such a data-driven approach, opening new avenues for approaching other similar inverse problems in acoustics.

Keywords: Vocal Tract, Impedance, Elliptical-cylinders

1. INTRODUCTION

Elliptical cylinders have been shown to be closer in approximation to the realistic vocal tract geometry for speech and voice production models [2-4]. A data-driven approach on the classic inverse problem of predicting the geometry of vocal tract from acoustic response spectra based on the area function of a circular-cylindrical geometry for vowel ‘ə’ has been addressed earlier by Balamurali et al. [2]. In this work, we use elliptical-cylinders to model the vocal tract geometry. We refer to the term ellipticity factor (EF) as the aspect ratio of major and minor axes of the ellipse.

The data generated for the numerical model was using a three-dimensional finite element method. The vocal tract is modelled as a series of three straight elliptic cylinders. The walls of the cylinder are modelled as sound-hard (or sound reflecting). The aim of this work is to be able to predict the geometry from the input impedance, measured at the glottis. The radiation load at the lips is modelled by a hemisphere.

The vocal tract is modelled as a simplified elliptical tube constituting of three sections: the oral region (40 mm long), second section (60 mm long), and a pharyngeal region (70 mm long) [2]. The area of all three elliptic cylinders are constant. Keeping the lengths of the three sections fixed, and the cross-sectional area of the ellipses equal, the parameter for which we train the machine learning model is ellipticity factor. The neural network is trained on the input acoustic impedance spectra (at the glottis) of the vocal tract models.

Amongst the different machine learning models tried in this investigation, Convolutional Neural Networks (CNN) was found to outperform others. CNN are a class of deep neural networks which are very helpful at identifying spatial features and non-linear relationships.

2. RESULTS

In the three-cylinder model, EF1, EF2, and EF3 are the ellipticity factors of the first, second and third elliptic-cylinder, respectively. The predicted EF when compared against the actual EF, exceeded 99% accuracy for both Pearson correlation and Lin concordance coefficients. These denote high accuracy. This observation was consistent for all three elliptic cylinders.

As shown in Figure 1, the Pearson Correlation for Predicted EF1 vs Actual EF1 is 0.9984, for

Predicted EF2 vs Actual EF2, is 0.9980, and for Predicted EF3 vs Actual EF3, is 0.9980, respectively.

As shown in Figure 2, the Lin Concordance for Predicted EF1 vs Actual EF1 is 0.9980, for Predicted EF2 vs Actual EF2, is 0.9972, and for Predicted EF3 vs Actual EF3, is 0.9972, respectively.

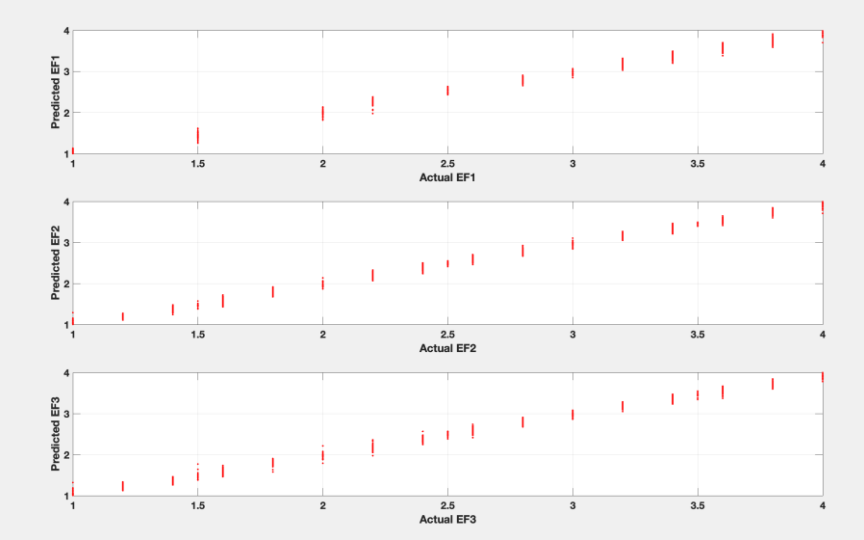


Figure 1 Pearson correlation coefficients

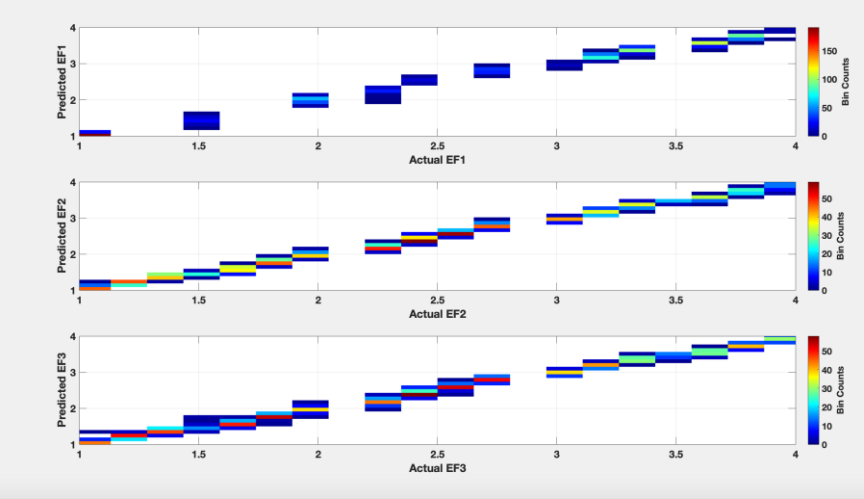


Figure 2 Lin concordance coefficients

2.1 Machine Learning model

As shown in the network architecture in Figure 3, we train embedding that returns higher dimensional input to the convolutional layers based on the dictionary size of 1000X801X300. The total number of cases in the dataset was 1114. The train-test split was 75/25, with the data shuffled at each epoch. The batch size was 835. A preliminary grid-search was conducted to find the optimal model size between wide and shallow, square, triangular, and diamond-shaped dense layered network. The narrow network returned a preliminary low mean square error (MSE) value of 0.6, compared to the other models which returned >1.5. Optimizer: we trained the model for 2000 epochs; the first 1000 epochs were trained at every 200 intervals, which was then progressively scaled for the learning rate to decrease from 1e-2 to 1e-7, in which the power term was scaled logarithmically. Our loss function was mean squared error.

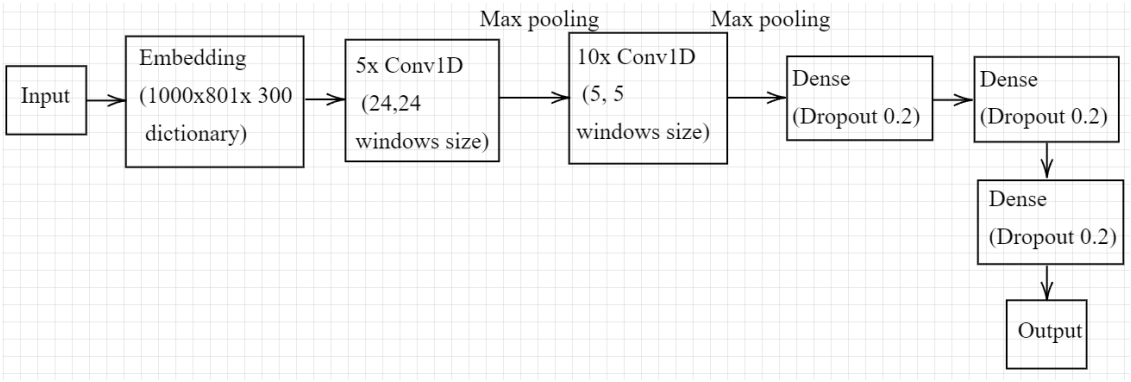


Figure 3 CNN model architecture

3. CONCLUSION AND FUTURE WORK

We proposed an efficient CNN algorithm to find the elliptical geometry from the acoustic response of a three-cylinder model. A data-driven approach using CNNs was used to solve the inverse area function problem to derive the non-linear relationship between the vocal tract impedance and the corresponding vocal tract geometry. A narrow but deep neural network was trained using acoustic impedance spectra, and the predicted radii, associated with the vocal tract geometry approximated using cylindrical tubes, were found to be highly correlated with the actual radii, showing reasonable agreement. This preliminary investigation while keeping area and lengths of cylinders fixed, while only varying the EF opens doors for further investigation of a combination of lengths and area functions.

REFERENCES

1. Fletcher, N. H., & Rossing, T. D. (2012). The physics of musical instruments. Springer Science & Business Media.
2. BT, B., Kapoor, S., & Chen, J. M. (2022). Estimating vocal tract geometry from acoustic impedance using deep neural network. JASA Express Letters, 2(3), 034801.
3. Arnela, M., & Guasch, O. (2013). Finite element computation of elliptical vocal tract impedances using the two-microphone transfer function method. The Journal of the Acoustical Society of America, 133(6), 4197-4209.
4. Matsuzaki, H., & Motoki, K. (2000). FEM analysis on acoustic characteristics of vocal tracts shape with different geometrical approximation. Proc. ICSLP2000, Beijing, 897-900.
5. Matsuzaki, H., Miki, N., & Ogawa, Y. (2000). 3D finite element analysis of Japanese vowels in elliptic sound tube model. Electronics and Communications in Japan (Part III: Fundamental Electronic Science), 83(4), 43-51.

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-1005

Respiratory sound classification using multi-resolution features

HyungJin SEO; Chang-Hyeon JEONG; DaHye KIM; JongMin WOO; In-Chul YOO; Dongsuk YOOK

Artificial Intelligence Laboratory, Department of Computer Science, Korea University, Korea

ABSTRACT

With the advance of deep learning technologies, medical image analysis using deep neural networks is achieving very high performance. On the other hand, while sounds contain medical information just as images, medical sound analysis such as automatic identification of adventitious respiratory sound has still been a challenging problem. Most of previous studies for automatic identification of adventitious respiratory sound mainly focused on applying various deep neural networks to the respiratory sound classification using the features developed for automatic speech recognition. In this paper, after the close analysis of adventitious respiratory sounds, we propose to consider a multi-resolution feature to capture the characteristics of widely varying adventitious respiratory sounds. The efficiency of the proposed method is evaluated using the International Conference on Biomedical and Health Informatics (ICBHI) 2017 dataset.

Keywords: Respiratory sound classification, Multi-resolution, Deep neural network

1. INTRODUCTION

Automatic detection of problematic respiratory sounds can be used for pre-filtering, greatly reducing the burden on experienced professionals at a very low cost. Previous work on automatically detecting respiratory sounds deals with feature extractions, classification models, and data augmentations. The features used for respiratory sound recognition includes mel-frequency cepstral coefficient (MFCC), inverted MFCC (IMFCC) (1), power spectrogram (2), and log mel-spectrogram (3, 4). Combining multiple features is also studied, as in (5, 6, 7). For classification models, random forest (8), convolutional neural network (CNN)-based model (9, 10, 11), VGG16-based model (3), and ResNet-based model (2) have been proposed. Combining several models as in (4, 12), using time delayed neural network (TDNN)-based models (1), BiGRU-attention-based models (13) have also been proposed. In (14), various models such as VGG16, ResNet50, and AlexNet, and various features such as spectrogram, scalogram, mel-spectrogram, and gammatonegram were compared. In (15), various fusion methods based on InceptionV3 were explored. The data augmentation aimed at increasing the size of the training datasets includes concatenation-based method (4), time domain and time-frequency domain methods (16), vocoder-based method (13), spectrogram cropping method (9). In (4), which is based on ResNet50, the performance was increased through data augmentation and fine tuning on the recording devices.

In this work, we focus on the frame length in the spectral feature analysis. That is, the conventional frame window size of about 25 ms has been tuned to process human speech composed of vowels and consonants. On the other hand, respiratory sounds have different duration and spectral characteristics from human speech, so that different feature extraction hyperparameters are required. Rather than finding the optimal window size for respiratory sound recognition, we propose using multi-resolution spectral analysis ranging from 20ms to 120ms and combining them to have achieve high accuracy.

2. METHOD

Figure 1 summarizes the extraction process of various features used to classify respiratory sounds in this work. A short-time Fourier transform (STFT) is performed on the raw wave to produce a spectrogram. Applying mel-filterbanks to a spectrogram produces a mel-spectrogram. Additionally, applying a discrete cosine transform (DCT) and logarithm to the mel-spectrogram produces a mel-frequency cepstrum (MFC). The spectrogram and mel-spectrogram can be converted to log spectrogram and log mel-spectrogram, respectively, after applying logarithms. As a baseline experiment, deep neural network (DNN) models are trained and evaluated over various window length

to find the best window length for each type of features. For each best case, performances of those with and without the first-order and second-order time derivatives of the features are also compared.

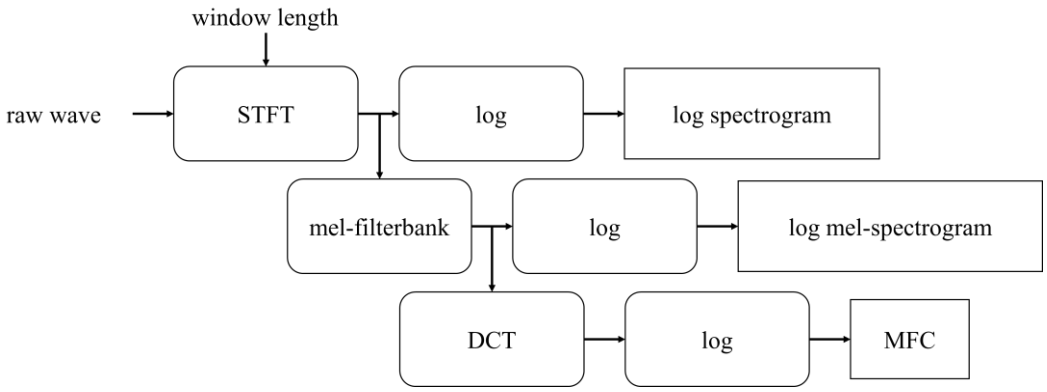


Figure 1 – Feature extraction process for respiratory sound classification

There is a trade-off between time and frequency resolutions for frame window sizes. Therefore, combining the various window sizes can help to increase the classification accuracies. Combining these different features can be accomplished in several methods. First, all features are combined into a single input feature which a single classification model uses. This is similar to the multi-resolution feature map approach in (14) developed for anti-spoofing. Second, each feature is trained separately with a corresponding classification model, and the outputs of these models are collected to produce final results. Combining the outputs of each model can be done by voting, applying a softmax function to the outputs, or utilizing a simple multi-layer perceptron (MLP) that accepts the classification models' outputs as inputs. Figure 2 summarizes the proposed method of utilizing the multi-resolution features.

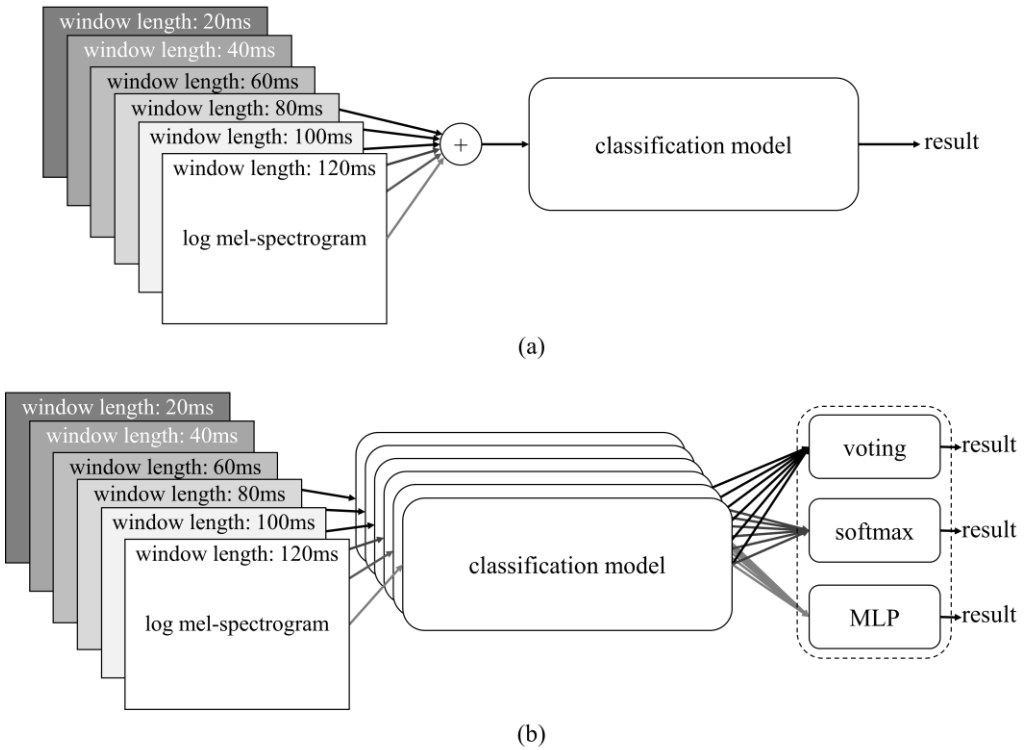


Figure 2 – Overview of the proposed multi-resolution classification. (a) a single classification model combines all features extracted with various window sizes and uses it as a single input feature (b) each classification model is trained for a given window length and the outputs of these models are collected by voting, softmax, or MLP.

3. EXPERIMENTS

3.1 Data

The International Conference on Biomedical and Health Informatics (ICBHI) 2017 dataset (17) was used in the experiments. It consists of 6898 labeled instances of four categories: "normal," "crackle," "wheeze," and "both" (which indicates that crackle and wheeze occur simultaneously in an instance). For training, we split the official training data in the ICBHI 2017 dataset into a training set and a development set in a ratio of 4:1. Table 1 summarizes the number of instances in the training, development, and test sets. We trained the models using the training set and selected the optimal hyperparameters based on the classification result of the development set.

Table 1 – Number of instances in the training, development, and test sets

	Training	Development	Test	Total
Normal	1651	412	1579	3642 (53%)
Crackle	972	243	649	1864 (27%)
Wheeze	401	100	385	886 (13%)
Both	291	72	143	506 (7%)
Total	3315 (48%)	827 (12%)	2756 (40%)	6898 (100%)

Table 2 summarizes the lengths of the shortest and longest instances in each class. As shown in Table 2, the length of each instance varies from 0.20 seconds to 16.16 seconds. We used 10 seconds of audio segments as the input to the classification models. For the instances of which length is less than 10 seconds, we repeated the audio segments to make a fixed size model input.

Table 2 – Length (in seconds) of the shortest and longest instances in each class

	Shortest	Longest
Normal	0.20	16.16
Crackle	0.37	8.74
Wheeze	0.23	9.22
Both	0.57	8.59

For the input feature, we used three most widely used features for respiratory sound classification: log spectrogram, log mel-spectrogram, and MFC. We fixed the hop size to 10 ms and used window sizes of 20, 40, 60, 80, 100, 120 ms. The first-order and second-order time derivatives of the original features, i.e., the delta and accel features, were also evaluated. For the number of input channels, we used a value of 1 for single-resolution tasks and a value of 6 for multi-resolution tasks. Resnet50 (18) was used as the classification DNN model.

3.2 Results

Table 3 summarizes the classification accuracy of three input features with various frame window lengths on the development data. Log spectrogram, log mel-spectrogram, and MFC achieved best accuracies at 20 ms, 60 ms, and 120 ms, respectively.

Table 3 – Classification accuracy (%) of three input features: log spectrogram (LS), log mel-spectrogram (LMS), and mel-frequency cepstra (MFC), with various frame window lengths

Window length	LS	LMS	MFC
20 ms	**77.4**	77.3	65.5
40 ms	75.6	77.1	67.5
60 ms	75.2	**80.3**	73.6
80 ms	70.7	77.4	71.6
100 ms	70.0	78.7	72.9
120 ms	69.6	76.2	**75.7**

We performed first-order time derivative (delta) and second-order time derivative (accel) feature experiments for each best case, which are summarizes in Table 4. The results confirm that dynamic features such as delta and accel do not improve the performance.

Table 4 – Effect of dynamic features: first-order time derivatives (delta) and second-order time derivatives (accel)

Feature	Accuracy (%)
LS (20 ms)	77.4
+Delta	**78.1**
+Delta+Accel	77.8
LMS (60 ms)	**80.3**
+Delta	78.1
+Delta+Accel	80.2
MFC (120 ms)	**75.7**
+Delta	74.1
+Delta+Accel	71.6

Since the log mel-spectrogram has generally shown better accuracies than log spectrogram and MFC, we used the log mel-spectrogram as the feature for the following multi-resolution experiments. Table 5 summarizes the results of the multi-resolution methods, described in the previous section, where the single-resolution method using log mel-spectrogram with 60 ms window size is also included for comparison. It can be seen that the output combining methods generally have higher accuracies for the development data, but input combining method has greatly improved the accuracy for the test data. In the experiments so far, the performance improvement for the development data and test data showed quite different trends. Though we selected the best hyperparameters based on the development data, those were not always best for the test data. We suspect that the characteristics of the official training data and test data of the ICBHI 2017 dataset are quite different.

Table 5 –Classification accuracies (%) of various multi-resolution methods, on the development and test data

Fusion level	Development data	Test data
Single-resolution	80.3	45.2
Input	77.8	**53.3**
Output: voting	81.1	47.3
Output: softmax	**81.6**	45.2
Output: MLP	81.1	45.9

Since the development data and test data showed quite different results in multi-resolution fusion methods, we analyzed each result using confusion matrices. Table 6 shows the confusion matrices of the single-resolution and multi-resolution methods, respectively, on the development data. By observing the diagonal components in Table 6, it can be seen that the improvement mostly comes from correctly classifying normal sounds.

Table 6 – Confusion matrices for the single-resolution method (LMS 60 ms) and the multi-resolution method (output softmax), respectively, on the development data.

	Normal	Crackle	Wheeze	Both
Normal	**351**	49	9	3
Crackle	42	**190**	1	10
Wheeze	9	4	**73**	14
Both	0	7	15	**50**

(a) Single-resolution method

	Normal	Crackle	Wheeze	Both
Normal	**359**	45	5	3
Crackle	46	**192**	1	4
Wheeze	12	3	**74**	11
Both	1	7	14	**50**

(b) Multi-resolution method

Table 7 shows the confusion matrices of the test data for the baseline log mel-spectrogram using a 60 ms window size and the multi-resolution method using input fusion, respectively. By observing the diagonal components in Table 7, it can be seen that the proposed input fusion method greatly improved the accuracy of normal and crackle sounds. The amount of mistaking crackle sounds as normal sounds has been greatly reduced, while misidentifying wheeze sounds as normal sounds has been slightly increased.

Table 7 – Confusion matrices for single-resolution method (LMS 60 ms) and multi-resolution method (input fusion), respectively, on the test data

	Normal	Crackle	Wheeze	Both
Normal	**824**	352	310	93
Crackle	**352**	232	35	30
Wheeze	101	74	**143**	67
Both	39	24	33	**47**

(a) Single-resolution method

	Normal	Crackle	Wheeze	Both
Normal	**999**	265	255	60
Crackle	**323**	293	21	12
Wheeze	**151**	47	147	40
Both	**49**	26	39	29

(b) Multi-resolution method

4. CONCLUSION

In this study, we proposed the multi-resolution methods for respiratory sound classification. While the conventional window size of about 25 ms has been adjusted to handle human speech, respiratory sounds require different feature extraction hyperparameters since their duration and spectral characteristics are different from those of human speech. Rather than finding the optimal window size for respiratory sound classification, we proposed to use multi-resolution spectral analysis using various frame window sizes ranging from 20 ms to 120 ms. We also analyzed how to combine the multi-resolution features. Since the proposed methods do not depend on the choice of classification model, DNN models other than ResNet50 can also be used. Also, various data augmentation methods can be combined with the proposed methods.

ACKNOWLEDGEMENTS

This work was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Science, ICT, and Future Planning (NRF-2017R1E1A1A01078157), and by the NRF of Korea under project BK21 Four.

REFERENCES

1. Liu L, Li L, Li S, Wu J, Guo D. An End-to-end System Based on TDNN for Lung Sound Classification. Proc IEEE International Conference on Anti-counterfeiting, Security, and Identification. 2020. p. 20–4.
2. Petmezas G, Cheimariotis GA, Stefanopoulos L, Rocha B, Paiva RP, Katsaggelos AK, Maglaveras N. Automated Lung Sound Classification Using a Hybrid CNN-LSTM Network and Focal Loss Function. Sensors. 2022;22(3):1232.
3. Kim Y, Hyon Y, Jung SS, Lee S, Yoo G, Chung C, Ha T. Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning. Sci Rep. 25 Aug 2021;11(1):17186.
4. Gairola S, Tom F, Kwatra N, Jain M. RespireNet: A Deep Neural Network for Accurately Detecting Abnormal Lung Sounds in Limited Data Setting. Proc Annual International Conference of the IEEE Engineering in Medicine & Biology Society. 2021. p. 527–30.
5. Xu L, Cheng J, Liu J, Kuang H, Wu F, Wang J. ARSC-Net: Adventitious Respiratory Sound Classification Network Using Parallel Paths with Channel-Spatial Attention. Proc IEEE International Conference on Bioinformatics and Biomedicine. 2021. p. 1125–30.
6. Jung SY, Liao CH, Wu YS, Yuan SM, CT Sun. Efficiently Classifying Lung Sounds through Depthwise Separable CNN Models with Fused STFT and MFCC Features. Diagnostics. 2021;11(4).
7. Uppin R, Ambesange S, Sangameshwar, Aralikatti S, Gowda V M. Respiratory Sound Abnormality Classification using Multipath Deep Learning Method. Proc Third International Conference on Inventive Research in Computing Applications. 2021. p. 903–15.
8. Elsetronning A, Rasheed A, Bekker J, San O. On the effectiveness of signal decomposition, feature extraction and selection on lung sound classification. [Internet] https://arxiv.org/abs/2012.11759.

9. Bardou D, Zhang K, Ahmad SM. Lung sounds classification using convolutional neural networks. Artificial Intelligence in Medicine. 1 Jun 2018;88:58–69.
10. Minami K, Lu H, Kim H, Mabu S, Hirano Y, Kido S. Automatic Classification of Large-Scale Respiratory Sound Dataset Based on Convolutional Neural Network. Proc International Conference on Control, Automation and Systems. 2019. p. 804–7.
11. Nguyen T, Pernkopf F. Lung Sound Classification Using Snapshot Ensemble of Convolutional Neural Networks. Proc Annual International Conference of the IEEE Engineering in Medicine & Biology Society. 2020. p. 760–3.
12. Naves R, Barbosa BHG, Ferreira DD. Classification of lung sounds using higher-order statistics: A divide-and-conquer approach. Computer Methods and Programs in Biomedicine. 1 Jun 2016;129:12–20.
13. Zhao X, Shao Y, Mai J, Yin A, Xu S. Respiratory Sound Classification Based on BiGRU-Attention Network with XGBoost. Proc IEEE International Conference on Bioinformatics and Biomedicine. 2020. p. 915–20.
14. Neili Z, Sundaraj K. A comparative study of the spectrogram, scalogram, melspectrogram and gammatonegram time-frequency representations for the classification of lung sounds using the ICBHI database based on CNNs. Biomedical Engineering 2022. https://doi.org/10.1515/bmt-2022-0180.
15. Pham L, Ngo D, Hoang T, Schindler A, McLoughlin I. An Ensemble of Deep Learning Frameworks Applied For Predicting Respiratory Anomalies. [Internet] https://arxiv.org/abs/2201.03054.
16. Nguyen T, Pernkopf F. Lung Sound Classification Using Co-tuning and Stochastic Normalization. IEEE Transactions on Biomedical Engineering. 2022. https://doi.org/10.1109/TBME.2022.3156293
17. Rocha BM, Filos D, Mendes L, Serbes G, Ulukaya S, Kahya YP, Jakovljevic N, Turukalo TL, Vogiatzis IM, Perantoni E. An open access database for the evaluation of respiratory sound classification algorithms. Physiol Meas. Mar 2019;40(3):035001.
18. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Proc IEEE Conference on Computer Vision and Pattern Recognition. 2016. p. 770-8.