INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

A Space-and-Speaker-Aware Iterative Mask Estimation Approach to Multi-channel Speech Recognition in the CHiME-6 Challenge

Yan-Hui Tu[1], *Jun Du*[1], *Lei Sun*[1], *Feng Ma*[1], *Jia Pan*[1], *Chin-Hui Lee*[2]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]Georgia Institute of Technology, Atlanta, Georgia, USA
{jundu, tuyanhui}@ustc.edu.cn

Abstract

We propose a space-and-speaker-aware iterative mask estimation (SSA-IME) approach to improving complex angular central Gaussian distributions (cACGMM) based beamforming in an iterative manner by leveraging upon the complementary information obtained from SSA-based regression. First, a mask calculated by beamformed speech features is proposed to enhance the estimation accuracy of the ideal ratio mask from noisy speech. Second, the outputs of cACGMM-beamformed speech with given time annotation as initial values are used to extract the log-power spectral and inter-phase difference features of different speakers serving as inputs to estimate the regression-based SSA model. Finally, in decoding, the mask estimated by the SSA model is also used to iteratively refine cACGMM-based masks, yielding enhanced multi-array speech. Tested on the recent CHiME-6 Challenge Track 1 tasks, the proposed SSA-IME framework significantly and consistently outperforms state-of-the-art approaches, and achieves the lowest word error rates for both Track 1 speech recognition tasks.

Index Terms: speech recognition, CHiME-6 Challenge, multi-channel speech enhancement, SSA-IME

1. Introduction

Automatic speech recognition (ASR) in distant-talking scenarios based on the use of microphone arrays has become an important part of everyday life with the emergence of speech-enabled applications on multi-microphone portable devices due to its convenience and flexibility [1]. Many limited tasks were first investigated, such as the TIdigits [2], the TIMIT [3], the Wall Street Journal (WSJ) [4] and LibriSpeech [5] corpora, which do not consider noisy or reverberant conditions. The CHiME (1-4) [6, 7, 8] series were also launched to investigate the effects of background noise in far-field cases, focusing on solving ASR problems in real-world applications. To improve ASR robustness, multi-channel speech enhancement was usually adopted as front-end system. The representative algorithms in this category include multi-channel Wiener filtering [9], blind source separation methods [10, 11, 12, 13], and beamforming methods [14, 15, 16]. Beamforming became a popular approach in the CHiME-3 Challenge [17]. In CHiME-4 Challenge, the best system proposed an approach combining the conventional multi-channel speech enhancement and deep learning methods [18] to improve multi-channel speech recognition.

Recently, the CHiME-5 Challenge provides the first large-scale corpus of real multi-talker conversational speech recorded via commercially available microphone arrays in multiple realistic homes [19]. It essentially congregates a large number of acoustic problems that may exist in real life, which poses a great challenge to existing ASR systems, especially for front-end processing with noisy, reverberant, and overlapping speech. In this challenge, the best system [20] proposed a speaker-dependent speech separation framework, exploiting advantages of both deep learning based and conventional preprocessing techniques. In the latest CHiME-6 Challenge [21], the data set for Track 1 is generated from the CHiME-5 data with array synchronization. The word error rates (WERs) of the worn microphone and multi-array data in the official baseline report are 41.21% and 51.76%, respectively, fully illustrating the difficulty of and issues confronted with the CHiME-6 ASR tasks.

In this paper, we propose a novel space-and-speaker-aware iterative mask estimation (SSA-IME) approach to multi-channel speech recognition in the CHiME-6 Challenge. It aims to improve the complex angular central Gaussian distributions (cACGMM)-based beamforming approach in an iterative manner by leveraging upon the complementary information obtained from space-and-speaker-aware (SSA)-based regression model. Although cACGMM has been recently demonstrated to be quite effective for multi-channel, ASR in operational scenarios, the corresponding mask estimation, however, is not always accurate in multi-talker environments due to the lack of prior or context information. To train this model, we construct a simulated dataset based on the official real multi-channel training data. First, to avoid the impact of noise on accuracy of the ideal ratio mask, the beamformed mask calculated by beamformed features is proposed. Second, The log-power spectral (LPS) and inter-phase difference (IPD) features of different speakers as the input of the proposed SSA model are extracted from the beamformed outputs of cACGMM-based beamforming with time annotation as initial values. These features contain rich space and speaker information which can make the regression model distinguish the different speakers by itself from multi-channel noisy data without any prior information. Finally, the mask estimated by SSA model is also used to refine cACGMM-based mask estimation, yielding an ASR performance improvement. Tested on the recently launched CHiME-6 Challenge Track1 tasks (multiple-array speech recognition), the proposed SSA-IME approach significantly and consistently outperforms the GSS approach [22]. Furthermore, the SSA-IME approach plays a key role in the ensemble system that achieves the best performance in the CHiME-6 Challenge Track 1 tasks.

2. The SSA-IME Framework

The overall SSA-IME framework is shown in Fig. 1. The SSA model is trained using the concatenated features which contain the space and speaker information. To reduce the impact of noise on the accuracy of ideal ratio mask [23], the learning target of the SSA model is calculated by beamformed signals.

The decoding process of SSA-IME is divided into four successive steps, namely, beamforming initialization, SSA-based signal statistics estimation, beamforming, and recogni-

http://dx.doi.org/10.21437/Interspeech.2020-2150

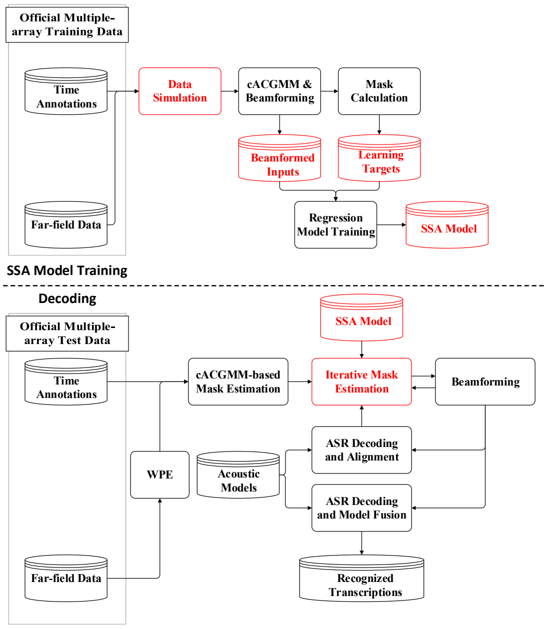

Official Multiple-array Training Data
Time Annotations
Far-field Data
Data Simulation
cACGMM & Beamforming
Mask Calculation
Beamformed Inputs
Learning Targets
Regression Model Training
SSA Model
SSA Model Training
Decoding
Official Multiple-array Test Data
Time Annotations
SSA Model
cACGMM-based Mask Estimation
Iterative Mask Estimation
Beamforming
WPE
Acoustic Models
ASR Decoding and Alignment
ASR Decoding and Model Fusion
Far-field Data
Recognized Transcriptions

Figure 1: *An illustration of SSA-IME framework.*

tion. First, beamformed speech is initialized and a T-F mask of test speech is obtained by cACGMM-based beamforming [24] using time annotation as initial prior values. Then, the mask estimated by our SSA model is used to improve the initial mask where the SSA model uses the features of the initial beamformed speech. And the ASR-based voice activity detection (VAD) information from the segmentation results of a recognizer with beamformed speech [18] also can be used to improve the initial mask. Next, the improved mask is used as the initial values of the cACGMM-based approach to generate the estimated mask which steers the beamforming, thereby obtaining the beamformed speech for ASR.

2.1. Multi-channel beamforming

At the beginning, we use a weighted prediction error (WPE) [25] algorithm on the multi-channel signals of the reference array, which is commonly used as a dereverberation preprocessor. We use generalized eigenvalue (GEV) beamformer which aims to maximize the signal-to-noise power ratio in the output [26]. Using the information provided by SSA model, a cACGMM is adopted to better estimate the cross-power density matrices in the GEV beamformer, while avoiding the speaker permutation problem.

The goal of a GEV beamformer is to find a linear vector of filter coefficients $\boldsymbol{W}_{\text{GEV}}(f) \in \mathbb{R}^{M\times 1}$ to maximize the signal-to-noise power ratio in each frequency bin [26]:

$$\boldsymbol{W}_{\text{GEV}}(f) = EV\{\boldsymbol{R}_{\boldsymbol{nn}}^{\mathbf{-1}}(f)\boldsymbol{R}_{\boldsymbol{ss}}(f)\} \tag{1}$$

where f is the frequency bin index and $EV\{\}$ denotes the eigenvector corresponding to the largest eigenvalue. $\boldsymbol{R}_{\boldsymbol{ss}}(f) \in \mathbb{R}^{M\times M}$ and $\boldsymbol{R}_{\boldsymbol{nn}}(f) \in \mathbb{R}^{M\times M}$ are the cross-power density matrices of the speech and noise terms, respectively. The above cost function has the same form as the Rayleigh coefficient.

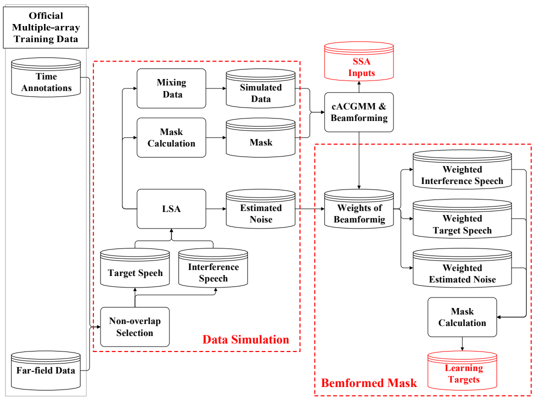


Figure 2: *Training data generation for building SSA models.*

The cross-power density matrices can be defined as:

$$\boldsymbol{R}_{\boldsymbol{vv}}(f) = \sum_{t=1}^{T} M_v(t,f)\boldsymbol{X}(t,f)\boldsymbol{X}^H(t,f) \tag{2}$$

where f is the frequency bin index and t is the frame index. $\boldsymbol{X}(t,f) \in \mathbb{C}^{M\times 1}$ is the observed signal from M microphones of the reference array. v can represent the speech of different speakers or noise class, and $M_v(t,f)$ denotes the probabilities of v in the time-frequency bin (t,f).

Finally, the estimate for the source signal is achieved as:

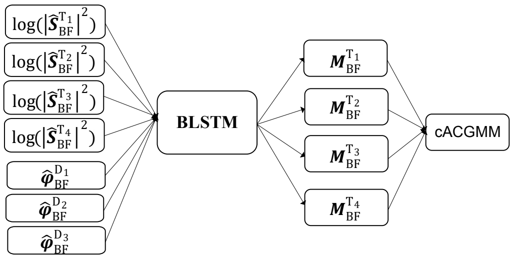
$$\hat{S}(t,f) = \boldsymbol{W}_{\text{GEV}}^H(f)\boldsymbol{X}(t,f). \tag{3}$$

Obviously, the key of the GEV beamformer is the estimation of time-frequency masks $M_v(t,f)$.

2.2. Training data generation for SSA model

In this section, we will give a detailed description on the training data generation of the SSA model as shown in Fig. 2. Based on the speech analysis, the most challenging part of CHiME-6 is about the dialogue style. Unlike reading speech, the complexity of conversational and spontaneous speech greatly increases the difficulty of a speech recognition system. For instance, casual pronunciation and frequent overlapping speech severely decrease the discriminating ability of acoustic models.

First, to investigate the speech overlapping problem, we excluded non-speech regions and aligned the time stamps of all speakers to locate the overlapped speech regions. The non-overlapped speech of each speaker is obtained by removing the overlapped speech regions from the aligned time stamps of each speaker. According to the introduction of the CHiME-6 dataset, there are a fixed number of four speakers in each session. Therefore, the non-overlapped speech of the four speakers in each session is used for generating training data. And STFT features of these mixed speech are denoted as $\boldsymbol{X}^{\mathrm{T}_1}(t,f), \boldsymbol{X}^{\mathrm{T}_2}(t,f), \boldsymbol{X}^{\mathrm{T}_3}$and $\boldsymbol{X}^{\mathrm{T}_4}(t,f) \in \mathbb{C}^{M\times 1}$, respectively. Note that the four speakers in one session are in turn considered as target speakers. Because the speech is directly selected from the far-field data, it contains much background noise. To reduce the noise influence on data simulation, we first perform single-channel noise estimation and suppression based on Log-Spectral Amplitude Estimator (LSA) [27]. We can obtain the estimated noise and enhanced speech of each speaker as $\hat{\boldsymbol{N}}_{\text{LSA}}^{\mathrm{T}_i}(t,f) \in \mathbb{C}^{M\times 1}$ and $\hat{\boldsymbol{S}}_{\text{LSA}}^{\mathrm{T}_i}(t,f) \in \mathbb{C}^{M\times 1}$, respectively, to calculate:



$\log(|\hat{\boldsymbol{S}}_{\text{BF}}^{\text{T}_1}|^2)$
$\log(|\hat{\boldsymbol{S}}_{\text{BF}}^{\text{T}_2}|^2)$
$\log(|\hat{\boldsymbol{S}}_{\text{BF}}^{\text{T}_3}|^2)$
$\log(|\hat{\boldsymbol{S}}_{\text{BF}}^{\text{T}_4}|^2)$
$\hat{\boldsymbol{\varphi}}_{\text{BF}}^{\text{D}_1}$
$\hat{\boldsymbol{\varphi}}_{\text{BF}}^{\text{D}_2}$
$\hat{\boldsymbol{\varphi}}_{\text{BF}}^{\text{D}_3}$
BLSTM
$\boldsymbol{M}_{\text{BF}}^{\text{T}_1}$
$\boldsymbol{M}_{\text{BF}}^{\text{T}_2}$
$\boldsymbol{M}_{\text{BF}}^{\text{T}_3}$
$\boldsymbol{M}_{\text{BF}}^{\text{T}_4}$
cACGMM

Figure 3: *An illustration of SSA model training.*

$$M_{\text{LSA}}^{\text{T}_i}(t,f)=\frac{\left\|\hat{\boldsymbol{S}}_{\text{LSA}}^{\text{T}_i}(t,f)\right\|_2^2}{\sum_{j=1}^4\left\|\hat{\boldsymbol{S}}_{\text{LSA}}^{\text{T}_j}(t,f)\right\|_2^2+\sum_{j=1}^4\left\|\hat{\boldsymbol{N}}_{\text{LSA}}^{\text{T}_j}(t,f)\right\|_2^2} \tag{4}$$

where $\|\cdot\|_2$ denotes the L2 norm of a vector. $M_{\text{LSA}}^{\text{T}_i}$ denotes the mask calculated by LSA-based speech. To generate the simulated data, the enhanced target speech and interference speech are linearly added.

Then, because the enhanced speech, $\hat{\boldsymbol{S}}_{\text{LSA}}^{\text{T}_i}$, is obtained by conventional single-channel speech enhancement, it also contains non-linear residue noises. Accordingly the mask, $M_{\text{LSA}}^{\text{T}_i}$, can not accurately present the speech presence probability, but it can provide more elaborate information at time-frequency bin level comparing to the time annotation at frame level. The mask is just used as the initial value for cACGMM and the outputs of cACGMM-based beamforming are used for SSA model training. And the noise estimated by LSA is directly adopted as the initial value. According to the Eqs. (1) and (2), different mask estimations, $M_{\text{LSA}}^{\text{T}_i}(t,f)$, will result in different beamforming weights, $\boldsymbol{W}_{\text{GEV}}^{\text{T}_i}(f)$, which not only suppress noises but also provide space and speaker information. The beamformed features of each target can be obtained as:

$$\begin{aligned}\hat{S}_{\text{BF}}^{\text{T}_i}(t,f)&=(\boldsymbol{W}_{\text{GEV}}^{\text{T}_i}(f))^H\hat{\boldsymbol{S}}_{\text{LSA}}^{\text{T}_i}(t,f)\\ \hat{S}_{\text{BF}}^{\text{T}_{ij}}(t,f)&=(\boldsymbol{W}_{\text{GEV}}^{\text{T}_i}(f))^H\hat{\boldsymbol{S}}_{\text{LSA}}^{\text{T}_j}(t,f)\\ \hat{N}_{\text{BF}}^{\text{T}_{ij}}(t,f)&=(\boldsymbol{W}_{\text{GEV}}^{\text{T}_i}(f))^H\hat{\boldsymbol{N}}_{\text{LSA}}^{\text{T}_j}(t,f)\end{aligned} \tag{5}$$

where $\hat{S}_{\text{BF}}^{\text{T}_i}(t,f)$, $\hat{S}_{\text{BF}}^{\text{T}_{ij}}(t,f)$ and $\hat{N}_{\text{BF}}^{\text{T}_{ij}}(t,f)$ are weighted target speech, weighted interference speech and weighted estimated noise. Finally the learning target of each speaker can be computed as:

$$M_{\text{BF}}^{\text{T}_i}(t,f)=\frac{\left|\hat{S}_{\text{BF}}^{\text{T}_i}(t,f)\right|^2}{\sum_{j=1}^4\left|\hat{S}_{\text{BF}}^{\text{T}_{ij}}(t,f)\right|^2+\sum_{j=1}^4\left|\hat{N}_{\text{BF}}^{\text{T}_{ij}}(t,f)\right|^2} \tag{6}$$

2.3. SSA model training

In this section, we will describe the training process of the SSA model in detail. To improve the mask estimation accuracy, a neural-network-based mask estimator learned from a multi-feature concatenation data set is proposed. The beamformed STFT features, $\hat{\boldsymbol{S}}_{\text{BF}}^{\text{T}_i}$, are composed of the elements in Eq. (6). Unlike conventional regression model for mask estimation, the beamformed features of four speakers are used together as the input of the BLSTM-based regression model as shown in Fig. 3. Specifically, $\log(|\hat{\boldsymbol{S}}_{\text{BF}}^{\text{T}_i}|^2)$ $(i=1,2,3,4)$ denotes the log-power spectral (LPS) features of four speakers on a whole utterance. And $\hat{\boldsymbol{\varphi}}_{\text{BF}}^{\text{D}_j}$ $(j=1,2,3)$ denotes the inter-phase difference (IPD) between a target speaker and three other interfering speakers on a whole utterance, which contains the spatial information between different speakers. Based on the above introduction, the BLSTM-based regression model can learn both space and speaker information at the same time. Therefore, we defined this regression model as space-and-speaker-aware (SSA) model which is also a speaker-independent speech separation model.

To train the BLSTM-based SSA model, the learning targets generated in Section 2.2 are used because they are calculated by beamformed features which are more reliable than the conventional masks. The optimization function of the BLSTM-based model is defined as:

$$E_{\text{SSA}}=\sum_{i=1}^4\sum_{t,f}\left(\hat{M}_{\text{SSA}}^{\text{T}_i}(t,f)-M_{\text{BF}}^{\text{T}_i}(t,f)\right)^2 \tag{7}$$

where $\hat{M}_{\text{SSA}}^{\text{T}_i}(t,f)$ and $M_{\text{BF}}^{\text{T}_i}(t,f)$ are the BLSTM estimated mask and the reference mask, respectively. By using E_{SSA}, the model can not only distinguish four speaker as much as possible by taking advantage of the space and speaker information but also yield robust and refined masks. After training, the one single SSA model of all four speakers can be generated.

3. Experiments

3.1. Data corpus

The latest CHiME-6 Challenge provides the first large-scale corpus of real multi-talker conversational speech recorded via commercially available microphone arrays in multiple realistic homes [28]. Speech material is elicited using a dinner party scenario with efforts taken to capture data that is representative of natural conversational speech. The parties have been made using multiple 4-channel microphone arrays and have been fully transcribed. This corpus essentially congregates a large number of acoustic problems that may exist in real life, which poses a great challenge to existing ASR systems, especially for the front-end processing in the case of noise, reverberation, overlapping speech. The CHiME-6 Challenge contains two tracks, namely Track 1 for multiple-array speech recognition and Track 2 for multiple-array diarization and recognition. Here, we focus on Track 1 where annotations can be used to recognize a given test utterance.

3.2. Implementation detail

For front-end configurations, speech waveform is sampled at 16 kHz, and the corresponding frame length is set to 1024 samples (or 64 msec) with a frame shift of 256 samples. The STFT analysis is used to compute the DFT of each overlapping windowed frame. To train the SSA model, the four reference masks were concatenated to the size of 513×4 as the learning targets. Four beamformed LPS features and three IPD features were concatenated to the size of 513×7 as the input. PyTorch was used for neural network training [29]. The learning rate for the first 3 epochs was initialized as 0.01 and then decreased by 90% after each epoch, and the number of epochs was 10. For beamforming, we stack all arrays into one big array according to [30]. The channel selection [31] and online beamforming [32] are also adopted. The CHiME-6 Challenge training set was used as our training data. BLSTM with 2 hidden layers and 1024 cells for each layer was employed as mask estimator.

For the back-end configurations, the baseline ASR recognition system is trained on the speech recognition toolkit Kaldi [33]. For factorized time delay neural network (TDNN-F) acoustic model training, backstitch optimization method is used. The decoding is based on 3-gram language models with explicit pronunciation and silence probability modeling.

3.3. Results and analysis

Table 1: *WER (%) comparison of official BeamformIt, GSS-based approach and our SSA-IME based approach for multi-channel speech enhancement using baseline ASR system on the development and evaluation set.*

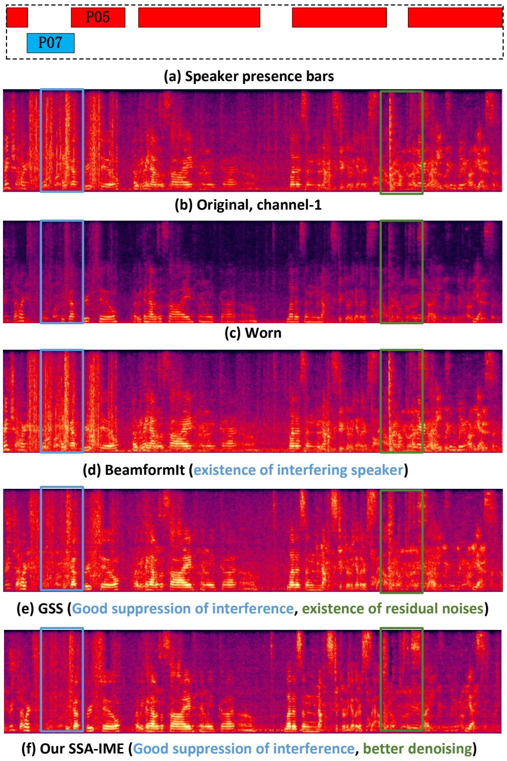
Enhancement	Dataset	DINING	KITCHEN	LIVING	AVG
BeamformIt	Dev Eval	68.54 53.69	74.11 67.55	65.74 64.15	69.48 61.19
GSS	Dev Eval	50.61 42.18	50.13 58.13	45.30 48.49	48.34 48.89
SSA-IME	Dev Eval	48.35 39.36	46.05 54.45	42.56 45.33	45.23 45.71

In Table 1 we show a WER (%) comparison of official BeamformIt, GSS-based and our SSA-IME based approach for multi-channel speech enhancement using the baseline ASR system on the development and evaluation sets.

First, "BeamformIt" [34] and "GSS" [22] are two baseline multi-channel speech enhancements, respectively. "GSS" used a spatial mixture model initiated with time annotations and the ASR-based VAD information from the segmentation results of a recognizer, while "BeamformIt" is a conventional multi-channel beamforming without any prior information. Comparing the two methods, we could find that the "GSS" significantly outperformed the "BeamformIt", e.g., the AVG WERs were significantly reduced from 69.48% to 48.34% and from 61.19% to 48.89% on development and evaluation sets, respectively. Based on the above results, it indicates that the speaker prior information is very important to improve the performance of multi-channel speech enhancement.

Second, "SSA-IME" denotes the proposed method which estimated the mask in an iterative manner from different pieces of complementary information sources, such as, the mask estimated by SSA model and the ASR-based VAD information from the segmentation results of a recognizer, yielding absolute WER reductions of 3.11% and 3.18% over GSS approach on development and evaluation sets, respectively. The proposed SSA-IME framework significantly and consistently outperforms the state-of-the-art GSS approach, and achieves the lowest ASR word error rates for both Track 1A and Track 1B.

To better understand the effectiveness of the proposed SSA-IME approach, an utterance of Speaker P05 selected from Session 02 was illustrated in Fig. 4. In the top panel, the boundaries from different speakers shown with the red areas indicating the target speaker P05 and the blue area denoting the interfering Speaker P07. The spectrograms of speech recorded with channel-1 and worn microphones are plotted in Figs. 4 (b) and (c) respectively. Compared with the spectrogram after BeamformIt shown in Fig. 4(d), speech processed by GSS shown in Fig. 4(e) removed most of the interferences. Though it also retains some residual noises, it shows that GSS greatly improves the speech intelligibility. In Fig. 4(f), the proposed SSA-IME method cannot only removes the interfering speaker well but also have better denoising effect than GSS, yielding a better recognition performance.



(a) Speaker presence bars

(b) Original, channel-1

(c) Worn

(d) BeamformIt (existence of interfering speaker)

(e) GSS (Good suppression of interference, existence of residual noises)

(f) Our SSA-IME (Good suppression of interference, better denoising)

Figure 4: *Spectrogram comparison of an utterance of speaker P05 from Session 02. (d) and (e) are the spectrograms from BeamformIt and GSS methods, respectively. The spectrogram after our final multi-channel beamforming is plotted in (f).*

4. Summary

In this paper, we have proposed an effective SSA-IME speech preprocessing framework to accurately estimate speech masks in an iterative manner from different pieces of complementary information sources with comprehensive and promising results on a state-of-the-art ASR challenge corpus. By using multi-feature concatenation, the SSA model not only makes a full use of the space and speaker information but also distinguishes different speakers from multi-channel noisy data. In the future, we can improve SSA-IME further by leveraging upon better spatial beamforming approaches, better deep learning architectures for mask estimation, and more informative feedback from the ASR systems. Finally, our back-end acoustic modeling effort, a key to our overall Track 1 ASR system, is described in another companion paper submitted to the same conference.

5. Acknowledgements

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grant Nos. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005. This work was also funded by Huawei Noah's Ark Lab.

6. References

[1] L. Deng, G. Hinton, and B. Kingsbury, “New types of deep neural network learning for speech recognition and related applications: an overview,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 8599–8603.

[2] R. Leonard, “A database for speaker-independent digit recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 1984 IEEE International Conference on*, vol. 9, pp. 328–331.

[3] J. Garofolo, “Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database,” 1988.

[4] D. Paul and J. Baker, “The design for the Wall Street Journal-based CSR corpus,” in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 5206–5210.

[6] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, “The PASCAL CHiME speech separation and recognition challenge,” *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013.

[7] E. Vincent, J. Barker, S. Watanabe, J. Le R., F. Nesta, and M. Matassoni, “The second ‘CHiME’ speech separation and recognition challenge: datasets, tasks and baselines,” in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*. IEEE, 2013, pp. 126–130.

[8] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: dataset, task and baselines,” in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pp. 504–511.

[9] B. Cornelis, M. Moonen, and J. Wouters, “Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1368–1381, 2011.

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

[11] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pp. 189–192.

[12] H. Buchner, R. Aichner, and W. Kellermann, “A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics,” *IEEE transactions on speech and audio processing*, vol. 13, no. 1, pp. 120–134, 2005.

[13] L. Wang, H. Ding, and F. Yin, “A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures,” *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 3, pp. 549–557, 2011.

[14] H. Cox, R. Zeskind, and M. Owen, “Robust adaptive beamforming,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[15] R. Talmon, I. Cohen, and S. Gannot, “Convolutive transfer function generalized sidelobe canceler,” *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 7, pp. 1420–1434, 2009.

[16] M. Souden, J. Benesty, and S. Affes, “A study of the LCMV and MVDR noise reduction filters,” *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4925–4935, 2010.

[17] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ¡®chime¡¯ speech separation and recognition challenge: Dataset, task and baselines,” in *Proc. IEEE Automat. Speech Recognition and Understanding Workshop.(ASRU)*, 2015.

[18] Y. Tu, J. Du, L. Sun, F. Ma, H. Wang, J. Chen, and C. Lee, “An iterative mask estimation approach to deep learning based multi-channel speech recognition,” *Speech Communication*, vol. 106, pp. 31–43, 2019.

[19] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, “The fifth ’CHiME’ speech separation and recognition challenge: dataset, task and baselines,” *arXiv preprint arXiv:1803.10609*, 2018.

[20] L. Sun, J. Du, T. Gao, Y. Fang, F. Ma, and C. Lee, “A speaker-dependent approach to separation of far-field multi-talker microphone array speech for front-end processing in the chime-5 challenge,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 827–840, 2019.

[21] S. Watanabe, M. Mandel, J. Barker, and E. Vincent, “Overview of the 6th chime challenge,” in *CHiME6 Workshop*, 2020.

[22] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, “Front-End Processing for the CHiME-5 Dinner Party Scenario,” in *CHiME5 Workshop*, 2018.

[23] C. Hummersone, T. Stokes, and T. Brookes, “On the ideal ratio mask as the goal of computational auditory scene analysis,” *Blind Source Separation*, pp. 349–368, 2014.

[24] N. Ito, S. Araki, and T. Nakatani, “Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing,” *european signal processing conference*, pp. 1153–1157, 2016.

[25] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, “Nara-wpe: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing,” in *ITG 2018, Oldenburg, Germany*, 2018.

[26] E. Warsitz and M. R. Haebumbach, “Blind acoustic beamforming based on generalized eigenvalue decomposition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[27] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[28] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, “The fifth ’chime’ speech separation and recognition challenge: Dataset, task and baselines,” *INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association*, pp. 1561–1565, 2018.

[29] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.

[30] N. Kanda, C. Boeddeker, J. Heitkaemper, Y. Fujita, S. Horiguchi, and R. Haebumbach, “Guided source separation meets a strong asr backend: Hitachi/paderborn university joint investigation for dinner party asr,” *conference of the international speech communication association*, 2019.

[31] K. Wojcicki and P. C. Loizou, “Channel selection in the modulation domain for improved speech intelligibility in noise,” *Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2904–2913, 2012.

[32] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, “Robust mvdr beamforming using time-frequency masks for online/offline asr in noise,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The Kaldi speech recognition toolkit,” in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society.

[34] X. Anguera, C. Wooters, and J. Hernando, “Acoustic beamforming for speaker diarization of meetings,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.