Speaker Separation Using Speaker Inventories and Estimated Speech

Peidong Wang, Zhuo Chen, DeLiang Wang, *Fellow, IEEE*, Jinyu Li, and Yifan Gong, *Fellow, IEEE*

***Abstract*—We propose speaker separation using speaker inventories and estimated speech (SSUSIES), a framework leveraging speaker profiles and estimated speech for speaker separation. SSUSIES contains two methods, speaker separation using speaker inventories (SSUSI) and speaker separation using estimated speech (SSUES). SSUSI performs speaker separation with the help of speaker inventory. By combining the advantages of permutation invariant training (PIT) and speech extraction, SSUSI significantly outperforms conventional approaches. SSUES is a widely applicable technique that can substantially improve speaker separation performance using the output of first-pass separation. We evaluate the models on both speaker separation and speech recognition metrics.**

***Index Terms*—Estimated speech, speaker inventory, speaker separation, speech recognition.**

I. Introduction

SPEECH overlaps occur commonly in daily conversations. They make automatic speech recognition (ASR) and speaker diarization in conversations difficult. The task of separating overlapped speech is referred to as speaker (or speech) separation and has long been an active research area.

A key challenge in speaker separation is the so-called permutation problem as defined in [8]. When multiple speakers are involved in a speech mixture, different orders of output signals may lead to conflicting gradients across training utterances. Two kinds of algorithms were proposed to handle the permutation problem, namely speaker separation and speech extraction. Speaker separation uses specifically designed training objectives that are invariant to the order of the outputs. Deep clustering [8], [10] and permutation invariant training (PIT) [11], [37] are two representative approaches. Many studies have been conducted to improve these two approaches, including new objective functions [2], [18], [29], end-to-end training [1], [14], [15], [17], [34], new model architectures [13], [16], [35], and different input features [7], [19], [24], [30], [32], [36]. Speech extraction avoids the permutation problem by extracting only one output signal using the bias information distinguishing the target speaker from others. The bias information can be in the form of visual signals [5], [32], [38], speaker locations [3], [22], [39], and speaker identities (SIDs) [4], [20], [25], [27], [33], [40]. Among these three types of bias information, SIDs are easier to acquire since they do not need extra hardware such as cameras and microphone arrays. Speaker identities are readily available in many scenarios such as meetings. For SID based speech extraction, Delcroix *et al.* proposed a method called SpeakerBeam to adapt sub-layers in a context-adaptive deep neural network to a target speaker [4], [40]. The VoiceFilter proposed by Wang *et al.* [27] concatenates spectral features with a d-vector generated by an SID model to extract the speech of the target speaker. Wang *et al.* performed speech extraction using a deep extractor network (DENet) [25] formed by stacking two deep attractor networks (DANets) [2]. The output of an "anchor" (i.e. speaker profile) based DANet is used as input features to another DANet. Xiao *et al.* proposed an attention based speech extraction model [33], which uses an attention mechanism to generate context-dependent biases for speech extraction. Recently, Ochiai *et al.* proposed ASENet, a unified framework for speaker separation and extraction [20]. They use an attention mechanism to combine the internal embedding vectors of overlapped speech and the embedding of the target speaker profile. Both speaker separation and speech extraction have limitations. Although speaker separation can be used in cases when speaker profiles are not available, they cannot obtain very high separation performance due to the lack of ability to leverage speaker information. Since speech extraction can only generate one output signal, its computation cost would be proportional to the total number of speakers in a meeting; even if a speaker does not say anything in the whole meeting, one would need to launch a speech extraction model for the speaker. Also, speech extraction is performed without the awareness of competing speakers, which may result in insufficient discrimination between some speaker pairs.

We thus propose SSUSI to deal with the issues in both speaker separation and speech extraction. SSUSI leverages bias information to improve separation performance and generates all separated signals in overlapped speech simultaneously. It works equally well or better than speaker separation when some speaker profiles are missing; in such cases, speech extraction is not able to function.

Manuscript received May 11, 2020; revised September 23, 2020 and October 20, 2020; accepted December 4, 2020. Date of publication December 18, 2020; date of current version January 6, 2021. The work of P. Wang was supported by internship at Microsoft, and finished at The Ohio State University. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wenwu Wang. *(Corresponding author: Peidong Wang.)*

Peidong Wang is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: wang.7642@osu.edu).

Zhuo Chen, Jinyu Li, and Yifan Gong are with the Microsoft, Redmond, WA 98052 USA (e-mail: zhuc@microsoft.com; jinyli@microsoft.com; ygong@microsoft.com).

DeLiang Wang is with the Department of Computer Science and Engineering and the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASLP.2020.3045556

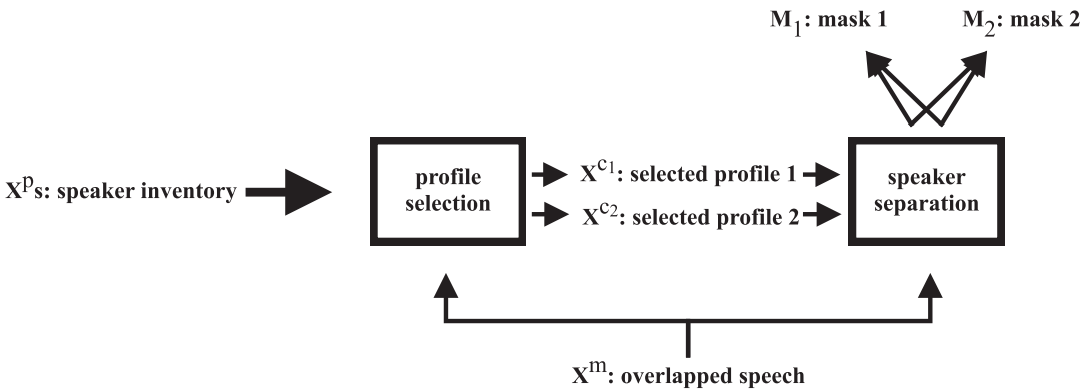


Fig. 1. Overview of SSUSI. The features of selected speaker profiles are denoted as $\mathbf{X}^{c_1}$ and $\mathbf{X}^{c_2}$. The arrows between the speaker separation stage and estimated speech denote PIT.

Although SSUSI improves speaker separation in cases when speaker profiles are available, it has two limitations. First, when the number of speaker profiles increases, it is more likely for SSUSI to select wrong speaker profiles and the performance of SSUSI degrades accordingly. Second, when speaker profiles are not available, SSUSI reduces to a normal PIT-based system and its separation performance is relatively low. We thus propose SSUES to deal with these limitations. SSUES takes the estimated speech of first-pass separation as the bias information for another iteration of speaker separation. Since estimated speech is guaranteed to be from a speaker in overlapped speech, the wrong profile selection problem is alleviated. SSUES shows substantial improvement when used with not only SSUSI but also PIT-based first-pass separation. This suggests the wide applicability of SSUES. There are few prior studies on using estimated speech to improve speaker separation performance. Hu and Wang proposed to alleviate the signal level mismatch problem between training and test by adapting a Gaussian mixture model based speaker separation model with estimated signal-to-noise ratios (SNRs) [9].

Preliminary results of this paper are presented in [26]. This paper expands [26] in the following ways. First, we propose SSUES, which makes SSUSI robust to increasing meeting participants. SSUES can also improve the performance of conventional speaker separation methods. Second, we integrate SSUSI and SSUES into SSUSIES, a new speaker separation approach different from existing speaker separation and speech extraction methods in its ability to leverage bias information in a multi-output separation framework.

The rest of this paper is organized as follows. We describe SSUSI and SSUES in Sections II and III, respectively. Experimental setup and evaluation results are presented in Section IV and V. Concluding remarks are given in Section VI.

II. Speaker Separation Using Speaker Inventories

A speaker inventory consists of a list of speaker profiles collected from the speakers that are possibly involved in overlapped speech. We denote the speakers in the speaker inventory as candidate speakers, and those that are actually involved in overlapped speech as relevant speakers at a certain time. In a scheduled business meeting scenario, for example, speaker profiles can be the prior voice recordings from all meeting invitees. In this paper, the number of relevant speakers is assumed to be two and the speaker inventory only contains voice recordings.

An overview of SSUSI is shown in Fig. 1. SSUSI performs speaker separation in two stages. First, it selects relevant speaker profiles from candidate profiles using an attention mechanism measuring the correlations between overlapped speech and speaker profiles. After that, two selected profiles are incorporated in the speaker separation stage by a different attention mechanism. The speaker separation stage is designed to exploit the speaker information for separation.

A. Profile Selection Stage

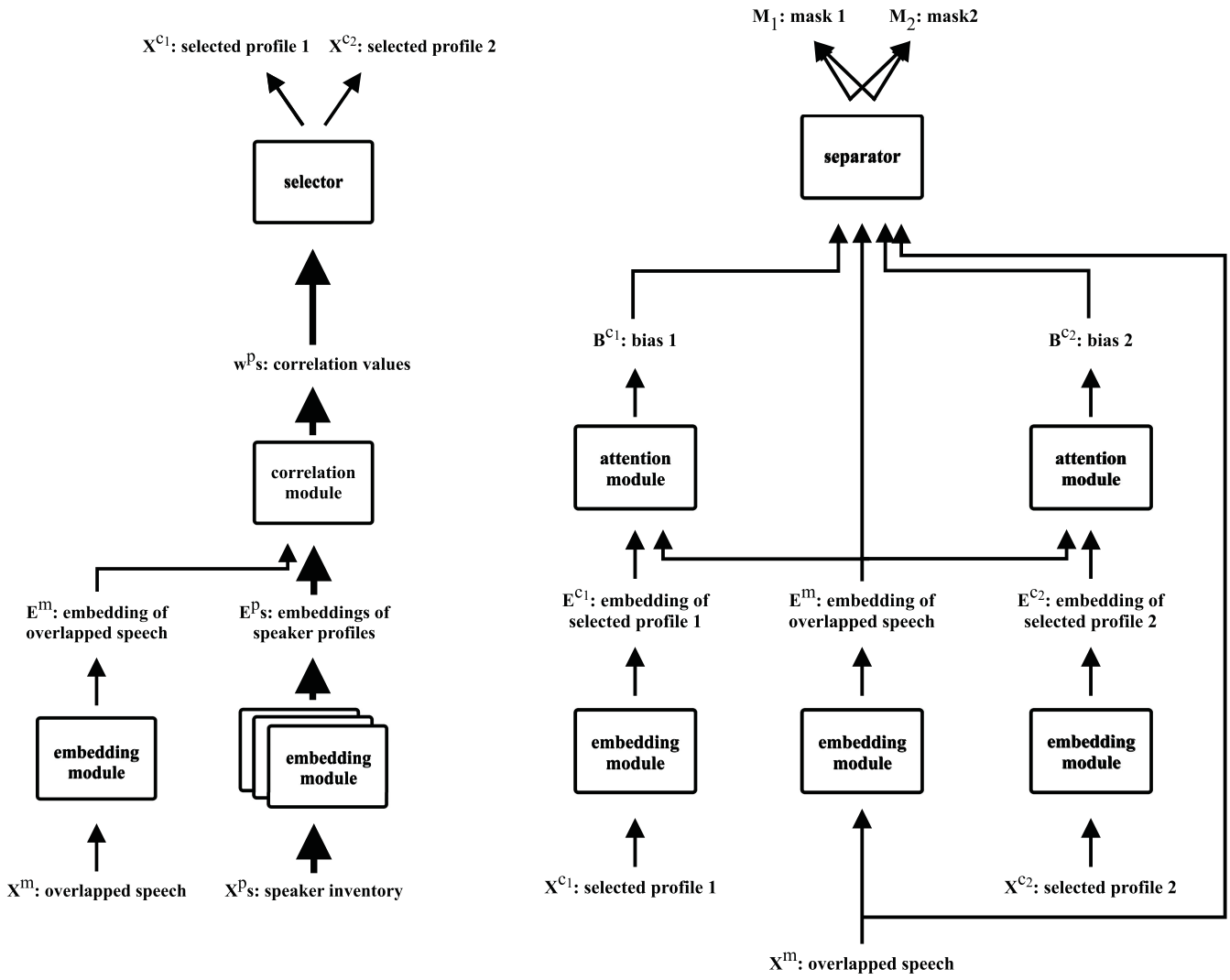
This stage consists of three components, an embedding module, a correlation module, and a profile selector. Fig. 2(a) depicts the profile selection stage in SSUSI.

We use a learnable embedding module to extract features for correlation. The embedding module maps input features $\mathbf{X} \in \mathbb{R}^{T\times F}$ to embeddings $\mathbf{E} \in \mathbb{R}^{T\times E}$, where T denotes the number of frames, F refers to input feature dimension, and E is the embedding dimension. For overlapped speech $\mathbf{X}^m \in \mathbb{R}^{T_m\times F}$, the embedding can be denoted as $\mathbf{E}^m \in \mathbb{R}^{T_m\times E}$. For a profile p in speaker inventory $\mathbf{P}$, the embedding can be written as $\mathbf{E}^p \in \mathbb{R}^{T_p\times E}$. Here T_m and T_p denote the numbers of frames in overlapped speech and speaker profile p, respectively.

The correlation module measures the correlation between the embedding of overlapped speech and that of each speaker profile. We use $\boldsymbol{e}_i^m$ to denote the vector in $\mathbf{E}^m$ at time frame i and $\boldsymbol{e}_j^p$ the vector in profile embedding $\mathbf{E}^p$ at frame j, with i ranging from 1 to T_m and j from 1 to T_p. We perform correlation in three steps. First, for each profile p, we calculate the dot product between each $\boldsymbol{e}_i^m$ and $\boldsymbol{e}_j^p$. Second, we normalize the dot products using the softmax function below. Finally, we average the correlation values over both i and j. These three steps are expressed as:

$$d_{i,j}^p = \boldsymbol{e}_i^m \cdot \boldsymbol{e}_j^p \tag{1}$$

$$w_{i,j}^p = \frac{\exp(d_{i,j}^p)}{\sum_{p\in\mathbf{P}}\sum_{j=1}^{T_p}\exp(d_{i,j}^p)} \tag{2}$$



(a) profile selection stage (b) speaker separation stage

Fig. 2. Illustrations of profile selection and speaker separation in SSUSI. The embeddings of selected profiles are denoted as $\mathbf{E}^{c_1}$ and $\mathbf{E}^{c_2}$, and the corresponding speaker biases are $\mathbf{B}^{c_1}$ and $\mathbf{B}^{c_2}$.

$$w^p = \frac{\sum_{i=1}^{T_m}\sum_{j=1}^{T_p} w_{i,j}^p}{T_m T_p} \tag{3}$$

where symbol $\cdot$ denotes the dot product operation and $d_{i,j}^p$ is the dot product of embedding vectors $\boldsymbol{e}_i^m$ and $\boldsymbol{e}_j^p$. Note that the denominator in equation (2) is a summation over both profile time steps j and profiles p. Symbol w^p is the mean correlation value for speaker profile p. The higher w^p is, the more likely that the speaker corresponding to p is involved in the overlapped speech.

The profile selector then selects the first and second largest w^p. We denote the selected two profiles as c_1 and c_2. The profile selection functions are:

$$c_1 = \arg\max_{p\in\mathbf{P}}\{w^p\} \tag{4}$$

$$c_2 = \arg\max_{p\in\mathbf{P}-\{c_1\}}\{w^p\} \tag{5}$$

B. Speaker Separation Stage

This stage has three components, an embedding module, an attention module, and a separator. Fig. 2(b) shows the speaker separation stage in the SSUSI framework.

Similar to the profile selection stage, the embedding module in the speaker separation stage maps input features to embeddings for subsequent attention calculation. For this module, we can re-use the one in the profile selection stage or train a new one specifically for speaker separation, as will be discussed in Section II-C.

The attention module in the speaker separation stage is slightly different from the correlation module in the profile selection stage. It is used to softly align speaker profiles so that they have

the same length as overlapped speech. We denote the aligned speaker profiles as speaker biases since they bias speaker separation towards selected speakers. Speaker bias $\boldsymbol{b}_i^{c_1}$ for selected profile c_1 at time i is calculated by the following equations:

$$d_{i,j}^{c_1} = \boldsymbol{e}_i^m \cdot \boldsymbol{e}_j^{c_1} \tag{6}$$

$$\alpha_{i,j}^{c_1} = \frac{\exp(d_{i,j}^{c_1})}{\sum_{j=1}^{T_{c_1}} \exp(d_{i,j}^{c_1})} \tag{7}$$

$$\boldsymbol{b}_i^{c_1} = \sum_{j=1}^{T_{c_1}} \alpha_{i,j}^{c_1} \boldsymbol{e}_j^{c_1} \tag{8}$$

where $\alpha_{i,j}^{c_1}$ denotes element (i, j) of the attention matrix. Speaker bias $\boldsymbol{b}_i^{c_2}$ is calculated similarly. Note that equation (6) is the same as equation (1) for profile selection. Attention matrix element $\alpha_{i,j}^{c_1}$ in equation (7) differs from correlation matrix element $w_{i,j}^p$ in equation (2) in that $\alpha_{i,j}^{c_1}$ is normalized over selected profile c_1, whereas $w_{i,j}^p$ is normalized over all the profiles in the speaker inventory. Because of this difference, $\alpha_{i,j}^{c_1}$ is able to softly align the embeddings of the selected profiles, whereas $w_{i,j}^p$ is used for comparisons between different profiles.

The separator takes as input the original input features of overlapped speech, the embedding of overlapped speech, and the speaker biases generated from the attention module. The output of the separator are time-frequency masks $\mathbf{M}_1$ and $\mathbf{M}_2$. The training objective is to minimize a signal restoration loss [6], [31] based on PIT. Let $\mathbf{Y}_1$ and $\mathbf{Y}_2$ be the target clean features. An utterance-wise PIT loss can be expressed as equations (9) and (10) below:

$$L(\theta) = \min\{l_{1,1} + l_{2,2}, l_{1,2} + l_{2,1}\} \tag{9}$$

where L denotes the loss of a training sample and θ refers to learnable parameters. $l_{u,v}$ means the loss between estimated and clean speech, which is defined as:

$$l_{u,v} = ||\mathbf{M}_u \otimes \mathbf{X}^m - \mathbf{Y}_v||_F^2, \tag{10}$$

where $||\cdot||_F$ denotes matrix Frobenius norm and $\otimes$ is the element-wise multiplication.

C. Ssusi-Sep, Ssusi-Pse and Ssusi-Jt

There are three modules in SSUSI that contain learnable parameters, i.e. the embedding module in the profile selection stage, the embedding module in the speaker separation stage, and the separator in the speaker separation stage. We can thus design three methods to train SSUSI, namely SSUSI that only trains the speaker separation stage (SSUSI-SEP), SSUSI with profile selection embedding (SSUSI-PSE), and SSUSI with joint training (SSUSI-JT).

SSUSI-SEP only trains the speaker separation stage and reuses the embedding module for profile selection. The rationale behind this method is that both embedding modules are used for the subsequent correlation calculation, as given in equations (1) and (6). To train the speaker separation stage, oracle relevant profiles are used as $\mathbf{X}^{c_1}$ and $\mathbf{X}^{c_2}$. By sharing the embedding module, the size of the whole model is also reduced.

SSUSI-PSE encourages the correct selection of relevant profiles. In addition to the speaker separation stage, we train the embedding module in the profile selection stage using a specifically designed training objective. The speaker separation stage and the profile selection stage are trained separately. The loss function to train the embedding module in the profile selection stage is as follows:

$$L(\theta) = (1 - w^{o_1} - w^{o_2})^2 + \sum_{o_k \in \mathbf{P} - \{o_1, o_2\}} (w^{o_k})^2 \tag{11}$$

where o_1 and o_2 are the oracle relevant profiles, and w^{o_1} and w^{o_2} are the corresponding correlation values calculated by equation (3). Speaker inventory $\mathbf{P}$ is divided into two subsets, oracle relevant profiles $\{o_1, o_2\}$ and irrelevant profiles $\mathbf{P} - \{o_1, o_2\}$. For relevant profiles, the training objective is to make their summation equal one, whereas for each irrelevant profile $o_k \in \mathbf{P} - \{o_1, o_2\}$, the objective is to set its weight to zero.

Different from SSUSI-SEP and SSUSI-PSE, which train the speaker separation stage and the profile selection stage separately, SSUSI-JT jointly optimizes the whole SSUSI framework using a single PIT objective. This way, the speaker separation stage may be more robust to wrong profile selections of the profile selection stage. Note that there is an argmax function in the profile selection stage, as shown in equations (4) and (5). During back-propagation, although the gradients with respect to the indices selected by argmax are hard to derive, we can still calculate the gradients with respect to the selected profiles. Fig. 3 shows a diagram illustrating SSUSI-JT.

III. Speaker Separation Using Estimated Speech

Although SSUSI can substantially improve separation performance and efficacy, it has two limitations. First, when the number of candidate speakers is large, the profile selection stage in SSUSI tends to select a wrong profile. Second, SSUSI may downgrade to a simple PIT-based separation stage when speaker inventories are not available. The separation performance in such cases would be relatively low.

SSUES solves SSUSI's problems by treating estimated speech (i.e. speaker separation output) $\mathbf{X}^{e_1}$ and $\mathbf{X}^{e_2}$ as speaker profiles. SSUES can be performed iteratively by feeding the estimated speech from a previous iteration to a subsequent separation iteration. Since estimated speech is part of overlapped speech, it is guaranteed to be from a relevant speaker. The negative influence of wrong profile selection can thus be alleviated. Moreover, SSUES provides a feedback loop for both SSUSI and speaker separation, which is able to improve separation performance after each iteration. An illustration of SSUES is presented in Fig. 4.

A. Ssues

SSUES requires first-pass separation to get initial estimated speech. As mentioned above, the first-pass separation can be SSUSI when a speaker inventory is available, or a speaker separation approach such as PIT when there is no speaker inventory.

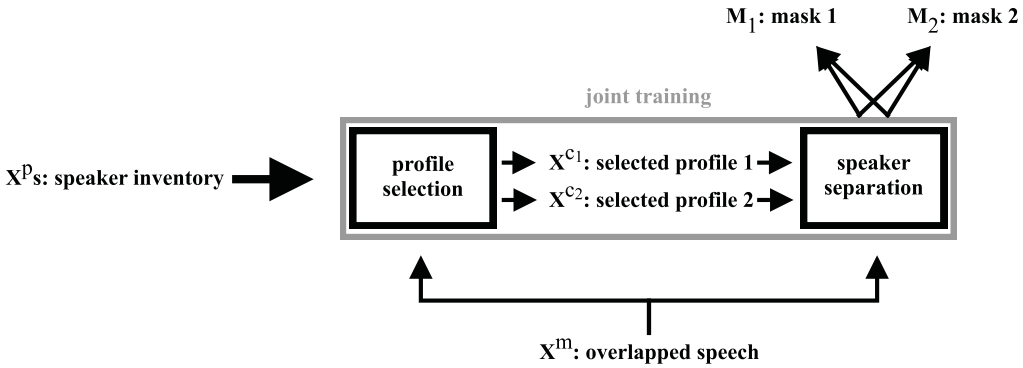


Fig. 3. Illustration of SSUSI-JT. The gray box and text indicate the joint training of the profile selection stage and the speaker separation stage.

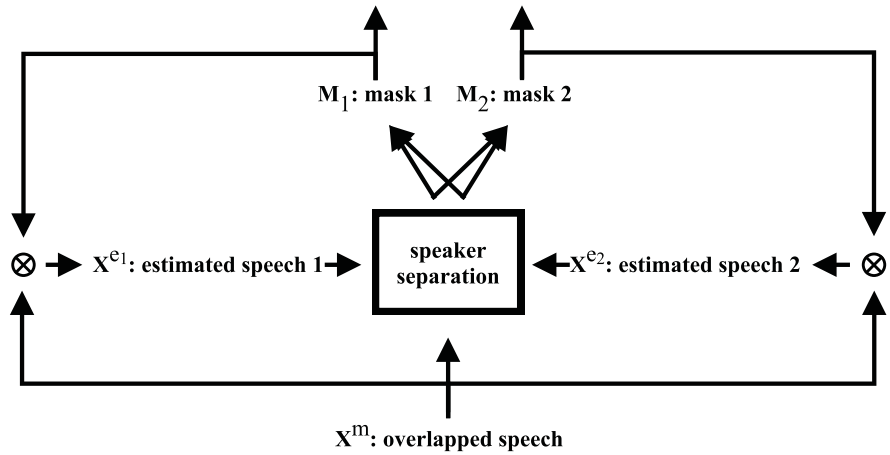


Fig. 4. Illustration of SSUES.

After obtaining initial estimated masks $\mathbf{M}_1$ and $\mathbf{M}_2$, we calculate estimated speech as:

$$\mathbf{X}^{e_k} = \mathbf{M}_k \otimes \mathbf{X}^m \tag{12}$$

where e_k indicates estimated speech and $k \in \{1, 2\}$ is the speaker index.

Note that for notational simplicity, we only present the spectral magnitude representation of estimated speech in equation (12). In implementation, the input to SSUES may be other types of feature and equation (12) may change accordingly.

B. Ssues-Nt and Ssues-Jt

SSUES can be performed by re-using the speaker separation stage in SSUSI. We denote this no-training method as SSUES-NT. This method, however, may cause an input data mismatch problem between training and test. During training, speaker profiles can be viewed as "clean" speech, whereas at test time, estimated speech may contain distortions. To handle this problem, we design SSUES with joint training (SSUES-JT), which jointly optimizes SSUSI and SSUES by a single PIT objective on the output of SSUES. Fig. 5 depicts SSUES-JT.

C. Speaker Separation Using Speaker Inventory and Estimated Speech

SSUSI and SSUES are closely related. The speaker separation stage in SSUSI makes it possible for SSUES to use estimated speech, and SSUES expands the application scenarios of SSUSI to cases when the number of candidate speakers is large or a speaker inventory is missing. We thus integrate SSUSI and SSUES into SSUSIES. The key component in SSUSIES is the speaker separation stage that leverages bias information such as speaker profiles and estimated speech. When a speaker inventory is available, SSUSIES performs SSUSI for first-pass separation and uses SSUES to leverage the information in estimated speech. In speaker separation tasks, conventional speaker separation is performed for first-pass separation and the SSUES method in SSUSIES can be used to further improve the separation result.

IV. Experimental Setup

A. Dataset

Our experiments are conducted on the LibriSpeech corpus [21] following the same recipe as [33]. The training set is generated using both train-clean-100 and train-clean-360. At test time, overlapped speech is generated using the test-clean set.

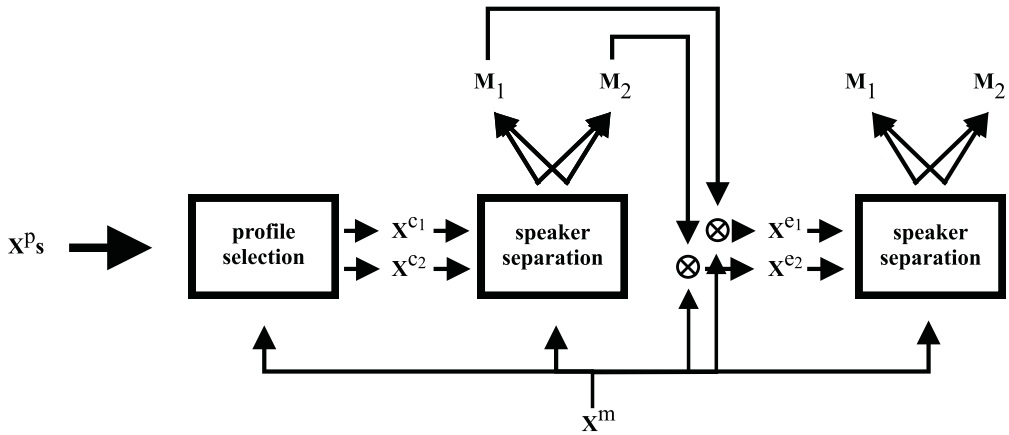


Fig. 5. Illustration of SSUES-JT.

There are 1172 speakers in the training set and 40 other speakers in the test set.

We use globally mean-variance normalized log spectral magnitude features as input. The length of each frame is 32 ms (i.e. 512 samples with a sampling rate of 16 kHz) and the shift between frames is 16 ms. The waveform signals are transformed using 512 dimensional short-time Fourier transformation. For training targets Ms, we use the spectral magnitude mask [28].

B. Baseline Systems

There are two baseline systems in this study, an utterance-wise PIT-based speaker separation model [12] and Xiao *et al.*'s speech extraction system [33]. The PIT-based model consists of six bidirectional long short-term memory (BLSTM) layers, each of which has 512 nodes. The speech extraction system has the same number of learnable parameters as the PIT-based model. For both the PIT-based model and speech extraction system, the optimizer is Adam and the learning rate is 10^{-4}.

C. Ssusi

As mentioned in Section II-C, there are three learnable modules in SSUSI, i.e. two embedding modules and the separator in the speaker separation stage. The two embedding modules have the same architecture, which consists of three BLSTM layers. The separator also has three BLSTM layers. All the BLSTM layers contain 512 nodes. SSUSI-SEP and SSUSI-JT have the same number of learnable parameters as those in the baselines, whereas SSUSI-PSE has an additional three-layer embedding module for profile selection.

The SSUSI-SEP is trained using oracle relevant profiles. For SSUSI-PSE and SSUSI-JT, four speaker profiles, two relevant and two irrelevant, are used as the speaker inventory during training. The embedding module in the profile selection stage of SSUSI-PSE is initialized with the well-trained embedding module in SSUSI-SEP. For SSUSI-JT, the whole model is initialized with SSUSI-SEP. At the speaker separation stage, speaker biases are concatenated with the embeddings and the original features of overlapped speech along the feature dimension. The learning rate for SSUSI-SEP, SSUSI-PSE, and SSUSI-JT are 10^{-4}, 10^{-6}, and 10^{-5}, respectively. All the other hyper-parameters are the same as the baselines. To avoid over-fitting, we apply online simulation, which generates the overlapped speech during model training. The model checkpoint used for evaluation is thus selected based on training loss.

Note that during training, we shuffle the order of speaker profiles. This makes the separation performance of SSUSI uninfluenced by the order of speaker profiles in the speaker inventory.

D. Ssues

SSUES has the same number of learnable parameters as the separator in SSUSI. It contains 3 BLSTM layers, each of which consists of 512 nodes.

As mentioned in Section IV-A, we use globally normalized log spectrum magnitude features in this study. Therefore, during SSUES-JT training, we perform logarithm and mean-variance normalization in addition to the spectral magnitude masking shown in equation (12).

E. ASR Backend

Our ASR backend is a DNN-HMM hybrid model trained on the clean training set of LibriSpeech. The model has three BLSTM layers, each of which contains 512 nodes. We generate forced aligned senone labels using Kaldi [23] and train the model under the maximum mutual information (MMI) criterion using PyTorch. The word error rate (WER) of this model on non-overlapped LibriSpeech test set is 5.7%.

V. Evaluation Results

We first present the results of SSUSI and then show how SSUES improves both PIT and SSUSI.

A. Ssusi

Table I contains the signal to distortion ratios (SDRs) of the three training approaches, SSUSI-SEP, SSUSI-PSE, and SSUSI-JT. We also list correct profile selection rates, which measure the performance of the profile selection stage. The SDR of unprocessed mixtures is 0.0 dB. SSUSI-SEP gets an SDR of

TABLE I
SDRs AND CORRECT PROFILE SELECTION RATES OF SSUSI-SEP, SSUSI-PSE, AND SSUSI-JT. THE NUMBER OF PROFILES CORRESPONDING TO IRRELEVANT SPEAKERS IS DENOTED AS *# IR-PROFILES*. THE TOTAL NUMBER OF PROFILES IN THE SPEAKER INVENTORY IS *# IR-PROFILES* PLUS 2. THE CORRECT SELECTION OF AT LEAST ONE RELEVANT PROFILE IS DENOTED AS ≥ 1 AND THE CORRECT SELECTION OF BOTH RELEVANT PROFILES IS *2*

method	# ir-profiles	$\geq$1 (%)	2 (%)	SDR (dB)
SSUSI-SEP	0	100	100	12.1
	1	100	82.1	11.8
	2	99.9	71.6	11.6
	3	99.8	64.1	11.4
	4	99.5	58.2	11.2
	5	99.2	54.9	11.1
	6	99.0	51.4	11.0
SSUSI-PSE	0	100	100	12.1
	1	100	86.7	11.9
	2	100	78.5	11.7
	3	99.8	72.5	11.6
	4	99.7	67.8	11.5
	5	99.4	63.8	11.3
	6	99.3	61.1	11.3
SSUSI-JT	0	100	100	12.2
	1	100	81.0	12.0
	2	99.8	69.6	11.9
	3	99.6	61.9	11.8
	4	99.4	56.5	11.6
	5	99.0	52.8	11.6
	6	98.7	49.7	11.5

12.1 dB when both relevant profiles are correctly selected. With the increase of irrelevant profiles, all three metrics decrease. From 0 to 6 irrelevant profiles, the correct profile selection rate of at lease one relevant profile drops slightly from 100% to 99.0%, whereas the correct selection rate of both relevant profiles decreases significantly from 100% to 51.4%. SDRs are degraded by wrong profile selections. The SDR on 6 irrelevant profiles drops to 11.0 dB. SSUSI-PSE improves correct profile selection rates substantially by using the additional profile selection embedding module. The improvement gets larger as the number of irrelevant profiles increases. SDRs benefit from better profile selection. Compared with that of SSUSI-SEP, the SDR of SSUSI-PSE on 6 irrelevant profiles increases by 0.3 dB. SSUSI-JT is able to achieve substantial SDR improvement over SSUSI-SEP without increasing the model size. Its SDR with 6 irrelevant profiles is 11.5 dB, outperforming both SSUSI-SEP and SSUSI-PSE. The SDR of SSUSI-JT on 0 irrelevant profile is slightly better than those of SSUSI-SEP and SSUSI-PSE. Note that the correct profile selection rates of SSUSI-JT are worse than those of SSUSI-PSE and even those of SSUSI-SEP. This shows that SSUSI-JT is robust to wrong profile selections. Since SSUSI-JT yields the best SDR results without using extra learnable parameters, we denote it as SSUSI in the remainder of this paper.

Table II shows the SDR and WER comparisons between SSUSI and PIT. Because of the ability to leverage speaker information, SSUSI performs significantly better than PIT in both SDR and WER. In the case of 30 irrelevant profiles (i.e. 32 candidate profiles in the speaker inventory), SSUSI still yields an SDR of 10.8 dB, which is substantially better than

TABLE II
SDR AND WER COMPARISONS BETWEEN SSUSI AND PIT. SEE TABLE I CAPTION FOR ACRONYMS

method	# ir-profiles	SDR (dB)	WER (%)
PIT	-	8.7	36.5
SSUSI	0	12.2	19.1
	6	11.5	21.8
	22	11.0	23.4
	30	10.8	24.1

TABLE III
SDR AND WER COMPARISONS BETWEEN SSUSI AND A SPEECH EXTRACTION SYSTEM. SEE TABLE I CAPTION FOR ACRONYMS

method	# ir-profiles	SDR (dB)	WER (%)
Speech Extraction [33]	0	11.5	21.9
	1	11.1	23.3
	2	10.9	24.4
SSUSI	0	12.2	19.1
	1	12.0	19.9
	2	11.9	20.4

the 8.7 dB SDR of PIT. Note that SSUSI is trained using only 2 irrelevant profiles. The results with 6, 22, and 30 irrelevant profiles demonstrate the robustness of SSUSI. In terms of WERs, SSUSI outperforms PIT by 48% relatively in the case of 0 irrelevant profile. When there are 30 irrelevant profiles, the relative improvement is still 34%. Note that all the WERs in Table II are relatively high for the LibriSpeech corpus. This is due to the distortions in estimated speech.

Table III presents the SDR and WER comparisons between SSUSI and the speech extraction system [33]. SSUSI substantially outperforms the speech extraction baseline. In terms of SDR, an improvement of more than 0.7 dB is yielded. For WER, the overall relative improvement is over 13%. This suggests that SSUSI is better at discriminating speaker pairs by the awareness of a competing speaker. In addition to the improvement in separation performance, SSUSI is significantly more efficient than the speech extraction system. In the case of 0 irrelevant profile, the computation time reduction during test is about 40% relatively. When there are 30 irrelevant profiles, the computation time reduction is about 70% relatively. The reason of this efficiency improvement is that speech extraction needs to launch one model instance for each candidate speaker, whereas SSUSI filters out all but one pair of speaker profiles for speaker separation.

Table IV provides the SDRs of SSUSI in cases when one or both relevant profiles are missing from the speaker inventory. This corresponds to the real-world scenario when one or more unregistered speakers attend the meeting and some of them are involved in overlapped speech. In the case when one relevant profile is missing, the SDRs of SSUSI drop to values close to 10.1 dB, which is still substantially better than the 8.7 dB result of PIT. This suggests that SSUSI is able to leverage the information in the remaining relevant profile even when the other one is missing. When both relevant profiles are missing, the SDRs of SSUSI are around 8.5 dB, which is similar to the SDR

TABLE IV
SDRs OF SSUSI IN CASES WHEN ONE OR BOTH RELEVANT PROFILES ARE NOT IN THE SPEAKER INVENTORY. *STANDARD* DENOTES BOTH RELEVANT PROFILES ARE IN THE SPEAKER INVENTORY, *M1* REFERS TO THE CASE WHEN ONE RELEVANT PROFILE IS MISSING, AND *M2* MEANS BOTH RELEVANT PROFILES ARE MISSING

method	# ir-profiles	standard	m1	m2
SSUSI	0	12.2	-	-
	1	12.0	10.0	-
	2	11.9	10.2	8.6
	3	11.8	10.2	8.6
	4	11.6	10.1	8.5
	5	11.6	10.1	8.5
	6	11.5	10.1	8.3

TABLE V
SDRs OF SSUES-NT AND SSUES-JT USING SSUSI AS FIRST-PASS SEPARATION. *NO-ITER* REFERS TO FIRST-PASS SEPARATION. *ITER1*, *ITER2*, AND *ITER3* DENOTE THE FIRST, SECOND, AND THIRD SSUES BASED SEPARATION ITERATION, RESPECTIVELY

method	# ir-profiles	no-iter	iter1	iter2	iter3
SSUES-NT	0	12.2	12.4	12.4	12.4
	1	12.0	12.3	12.3	12.3
	2	11.9	12.2	12.2	12.2
	3	11.8	12.1	12.2	12.2
	4	11.6	12.0	12.1	12.1
	5	11.6	12.0	12.1	12.1
	6	11.5	11.9	12.0	12.0
	22	11.0	11.5	11.7	11.7
	30	10.8	11.4	11.6	11.7
SSUES-JT	0	12.2	12.3	12.3	12.4
	1	12.0	12.2	12.3	12.3
	2	11.9	12.1	12.2	12.3
	3	11.8	12.1	12.2	12.2
	4	11.6	12.0	12.1	12.2
	5	11.6	12.0	12.1	12.1
	6	11.5	11.9	12.0	12.1
	22	11.0	11.6	11.8	11.8
	30	10.8	11.5	11.7	11.8

of PIT. This indicates that when both speakers in the overlapped speech are unregistered speakers, SSUSI performs similarly to PIT. Note that speech extraction cannot work in above cases when there are unregistered speakers.

B. Ssues

Table V shows the comparison between SSUES-NT and SSUES-JT. Both of them use SSUSI as first-pass separation. For SSUES-NT, the improvement over SSUSI is substantial, especially when the number of irrelevant profiles is large. With 30 irrelevant profiles, the improvement after three iterations is about 1 dB. More specifically, the SDR of SSUES-NT with 30 irrelevant profiles is comparable to that of SSUSI with 3 irrelevant profiles. These results show that SSUES-NT is able to alleviate the wrong profile selection problem of SSUSI and consistently improve the separation performance. Note that the performance of SSUES-NT with 0 irrelevant profile is also better than that of SSUSI. The reason is that, in addition to speaker information, SSUES-NT can leverage the contextual information in estimated speech. SSUES-NT performs similarly to SSUES-JT even without joint training. This shows that trained with a large number of speaker profiles, SSUES-NT is able to generalize to estimated speech. Considering the fact that SSUES-JT is better than SSUES-NT by at most 0.1 dB and that SSUES is designed to work with various first-pass separation approaches, we use SSUES-NT as SSUES in the remainder of this paper.

TABLE VI
SDR AND WER COMPARISONS BETWEEN PIT AND PIT + SSUES. *# ITER* DENOTES THE NUMBER OF SSUES BASED SEPARATION ITERATIONS

method	# iter	SDR (dB)	WER (%)
PIT	-	8.7	36.5
+ SSUES	1	10.5	24.8
	2	10.8	23.2
	3	10.9	22.9

TABLE VII
SDR AND WER COMPARISONS BETWEEN SSUSI AND SSUSI + SSUES WITH 30 IRRELEVANT PROFILES

method	# iter	SDR (dB)	WER (%)
SSUSI	-	10.8	24.1
+ SSUES	1	11.4	21.1
	2	11.6	20.4
	3	11.7	20.3

Table VI presents the SDR and WER comparisons between PIT and PIT + SSUES. With only one iteration of SSUES, the SDR improvement is already 1.8 dB and the WER improvement is 31% relatively. After three iterations, the WER improvement is increased to 37% relatively. These comparisons clearly show the efficacy of SSUES in improving the performance of PIT.

Table VII shows the SDR and WER comparisons between SSUSI and SSUSI + SSUES with 30 irrelevant profiles. With three iterations of SSUES, the SDR improvement is 0.9 dB and the WER reduction is 16% relatively. Note that the SDR and WER results of SSUSI with 2 irrelevant profiles are 11.9 dB and 20.4%, as shown in Table III. The similar results of SSUSI with 2 irrelevant profiles and SSUSI + SSUES with 30 irrelevant profiles show that SSUES can substantially improve the performance of SSUSI.

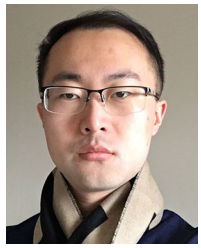
VI. CONCLUDING REMARKS

We have proposed SSUSIES, a speaker separation framework that is capable of leveraging external information such as speaker profiles and estimated speech. Compared with speech extraction, SSUSIES achieves more than 13% relative improvement in WER and up to 70% relative improvement in computational efficiency. In addition, SSUSIES outperforms PIT by 13% relatively in WER. Future research will extend SSUSIES to multi-channel conditions and evaluate SSUSIES in real conversations.

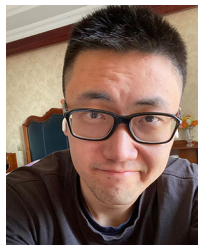
REFERENCES

[1] Z. Chen and J. Droppo, "Sequence modeling in unsupervised single-channel overlapped speech recognition," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4809–4813.
[2] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 246–250.

[3] Z. Chen, X. Xiao, T. Yoshioka, J. Li, H. Erdogan, and Y. Gong, "Multi-channel multi-speaker overlapped speech recognition with location guided speech extraction network," in *Spoken Lang. Technol. Workshop*, 2018, pp. 558–565.
[4] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5554–5558.
[5] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graphics*, vol. 37, no. 4, pp. 1–11, 2018.
[6] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 708–712.
[7] R. Gu *et al.*, "End-to-end multi-channel speech separation," 2019, arXiv:1905.06286.
[8] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 31–35.
[9] K. Hu and D. L. Wang, "An iterative model-based approach to cochannel speech separation," *EURASIP Journal on Audio, Speech, and Music Processing*, 2013.
[10] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. INTERSPEECH*, 2016, pp. 545–549.
[11] M. Kolbæk, D. Yu, Z. H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.
[12] M. Kolbk, Z. H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 1, pp. 153–167, Jan. 2017.
[13] L. Li and H. Kameoka, "Deep clustering with gated convolutional networks," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 16–20.
[14] Z. X. Li, Y. Song, L. R. Dai, and I. McLoughlin, "Source-aware context network for single-channel multi-speaker speech separation," in *Proc. Int. Conf. Acoust., Speech Signal Process*, 2018, pp. 681–685.
[15] Z. X. Li, Y. Song, L. R. Dai, and I. McLoughlin, "Listening and grouping: An online autoregressive approach for monaural speech separation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 27, no. 4, pp. 692–703, Apr. 2019.
[16] Y. Liu and D. L. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, pp. 2092–2102, 2019.
[17] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 4, pp. 787–796, Apr. 2018.
[18] Y. Luo and N. Mesgarani, "Augmented time-frequency mask estimation in cluster-based source separation algorithms," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 710–714.
[19] Y. Luo and N. Mesgarani, " Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
[20] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, "A unified framework for neural speech separation and extraction," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6975–6979.
[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5206–5210.
[22] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 36–40.
[23] D. Povey *et al.*, " The kaldi speech recognition toolkit," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2011, pp. 1–4.
[24] Z. Shi, H. Lin, L. Liu, R. Liu, and J. Han, "Furcanext: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," in *Proc. Int. Conf. Multimedia Model.*, 2020, pp. 653–665.
[25] J. Wang *et al.*, "Deep extractor network for target speaker recovery from single channel speech mixtures," in *Proc. INTERSPEECH*, 2018, pp. 307–311.
[26] P. Wang *et al.*, " Speech separation using speaker inventory," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2019, pp. 230–236.
[27] Q. Wang *et al.*, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. INTERSPEECH*, 2019, pp. 2728–2732.
[28] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.
[29] Z. Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 686–690.
[30] Z. Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 1–5.
[31] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2014, pp. 577–581.
[32] J. Wu *et al.*, "Time domain audio visual speech separation," in *Proc. IEEE Automatic Speech Recognit. Understanding Workshop, ASRU*, 2019, pp. 667–673.
[33] X. Xiao *et al.*, "Single-channel speech extraction using speaker inventory and attention network," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 86–90.
[34] C. Xu, W. Rao, E. S. Chng, and H. Li, "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6990–6994.
[35] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid lstm," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6–10.
[36] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5739–5743.
[37] D. Yu, M. Kolbæk, Z. H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 241–245.
[38] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proc. Eur. Conf. Comput. Vision, ECCV*, 2018, pp. 570–586.
[39] Y. Zhao, Z. Q. Wang, and D. L. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 53–62, Jan. 2019.
[40] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *Proc. INTERSPEECH*, 2017, pp. 2655–2659.

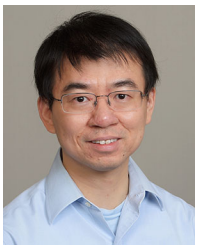


Peidong Wang received the B.E. degree in electronic information engineering from the University of Science and Technology of China, Hefei, China, in 2015. He is currently working toward the Ph.D. degree in artificial intelligence with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA. His research interests include robust automatic speech recognition, speech enhancement, speaker separation, end-to-end speech recognition, and efficient neural networks. He has been a Reviewer in various conferences and journals in speech community.



Zhuo Chen is a Senior Applied Scientist with Microsoft cloud & AI. His main research focuses on conversational speech recognition, that involves the speech separation, multi-channel signal processing, end to end speech diarisation and recognition. Prior to his career in Microsoft, he had the Ph.D. degree from Columbia University in 2017, where he developed several foundational research works for neural network based speech separation such as deep clustering, deep attractor networks etc., and his work in attentional hearing aid was awarded by NETEXPLO as top 10 innovations in 2018. Zhuo has been the Reviewer and Technique Committee Member in various conferences and journals in speech community, and he actively participates public challenges in speech processing, developing award-winning systems in challenges such as MIREX, CHiME 3, IAPRA ASpIRE, VOXSRC 2019 & 2020 etc.

DeLiang Wang, photograh and biography not available at the time of publication.



Jinyu Li received the Ph.D. degree from the Georgia Institute of Technology, Atlanta, in 2008. From 2000 to 2003, he was a Researcher with the Intel China Research Center and Research Manager in iFlytek Speech, China. He is currently a Partner Applied Scientist and Technical Lead with Microsoft Corporation, Redmond, USA. He leads a team to design and improve speech modeling algorithms and technologies that ensure industry state-of-the-art speech recognition accuracy for Microsoft. His major research interests cover several topics in speech recognition, including end-to-end modeling, deep learning, noise robustness, etc. He is the leading Author of the book "*Robust Automatic Speech Recognition – A Bridge to Practical Applications*" (Academic Press, October 2015). He is the member of IEEE Speech and Language Processing Technical Committee since 2018. He also served as the Associate Editor for IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING from 2015 to 2020.

Yifan Gong (Fellow, IEEE) received the Ph.D. degree in computer science from the Department of Mathematics and Computer Science, University of Nancy I, France. He leads a Speech Modeling Team with the Microsoft AI Cognitive Services, developing machine learning and speech and speaker recognition and key-word/diarization modeling technologies and tools to improve speech recognition accuracy across scenarios/tasks, languages and acoustic environments for both server and mobile devices. Products he and his team have delivered include Microsoft Azure Speech API, meeting transcriptions, speech-to-speech translations, Cortana digital voice assistant. Prior to joining Microsoft in 2004, he worked as Senior Research Scientist with the National Center of Scientific Research (CNRS, France) and then Senior Member of Technical Staff at Texas Instruments.

He authored and coauthored more than 250 publications in books, journals, conferences, and more than 50 patents. Yifan is the Senior Editorial Board of the *IEEE Signal Processing Magazine*, and has been an Elected Member of Speech and Language Processing Technical Committee, IEEE Signal Processing Society, for several terms (1998–2002, 2012–2016, 2017–2019).