SPEECH SEPARATION USING SPEAKER INVENTORY

Peidong Wang[1,2*] *, Zhuo Chen*[1],
Xiong Xiao[1], *Zhong Meng*[1], *Takuya Yoshioka*[1], *Tianyan Zhou*[1], *Liang Lu*[1], *Jinyu Li*[1]

[1]Microsoft, Redmond, WA, USA
[2]The Ohio State University, Columbus, OH, USA

wang.7642@osu.edu, {zhuc, xioxiao, zhong.meng, tayoshio,
tizhou, liang.lu, jinyli}@microsoft.com

ABSTRACT

Overlapped speech is one of the main challenges in conversational speech applications such as meeting transcription. Blind speech separation and speech extraction are two common approaches to this problem. Both of them, however, suffer from limitations resulting from the lack of abilities to either leverage additional information or process multiple speakers simultaneously. In this work, we propose a novel method called speech separation using speaker inventory (SSUSI), which combines the advantages of both approaches and thus solves their problems. SSUSI makes use of a speaker inventory, i.e. a pool of pre-enrolled speaker signals, and jointly separates all participating speakers. This is achieved by a specially designed attention mechanism, eliminating the need for accurate speaker identities. Experimental results show that SSUSI outperforms permutation invariant training based blind speech separation by up to 48% relatively in word error rate (WER). Compared with speech extraction, SSUSI reduces computation time by up to 70% and improves the WER by more than 13% relatively.

Index Terms— speaker inventory, speech separation, speech extraction, PIT, LibriSpeech

1. INTRODUCTION

Speech overlaps occur commonly in human conversations. They make speech recognition and diarization in conversations difficult. The task of separating overlapped speech is referred to as speech separation and has long been an active research area.

A key challenge in speech separation is the so-called permutation problem as defined in [1]. When multiple speakers are involved in one utterance, the order of the output signals may be arbitrary, which generates conflicting gradients across training utterances. Two families of algorithms were proposed to handle the permutation problem, namely blind speech separation and speech extraction. With blind speech separation, the permutation problem is usually handled by a specially designed training objective that is invariant to the order of the output. Deep clustering (DC) [1, 2] and permutation invariant training (PIT) [3, 4] are two representative approaches. Several studies were conducted to improve these approaches, including new objective functions [5, 6, 7], end-to-end training [8, 9, 10, 11, 12], different architectures [13, 14, 15], and feature spaces [16, 17, 18, 19, 20, 21]. The speech extraction approach tackles the permutation problem by leveraging bias information that helps distinguish the target speaker from others. Such bias information includes vision signals [22, 23, 17], speaker locations [24, 25, 26], and speaker identities (SIDs) [27, 28, 29, 30, 31, 32].

Compared with the vision and location signals, SIDs are easier to acquire since they do not need additional hardware such as cameras or microphone arrays. In fact, they may be readily available in many conversational scenarios such as meetings. Delcroix *et al.* proposed a method called SpeakerBeam to adapt sub-layers to a target speaker in the context adaptive deep neural network (CADNN) framework [27, 28]. VoiceFilter proposed by Wang *et al.* [29] concatenates spectral features with a d-vector generated by an SID model to extract the speech of the target speaker. Wang *et al.* proposed deep extractor network (DENet) [30], which stacks two DANets. The output of an "anchor" (i.e. speaker profile) based DANet is used as additional features for the mixed speech based DANet. Xiao *et al.* proposed an attention based speech extraction model [31], which uses an attention mechanism to generate context-dependent biases for speech extraction. Recently, Ochiai *et al.* proposed ASENet, a unified framework for speech separation and extraction [32]. They use an attention mechanism to combine internal embedding vectors of mixed speech and the embedding of the profile for the target speaker. Speech extraction systems have limitations when applied to conversation processing tasks. First, since only one speaker can be extracted at a time, the computational cost is proportional to the number of speakers. Second, the extraction process is performed independently for each speaker, which may result in insufficient discrimination between some speakers.

We propose a novel speech separation system combining the advantages of speech extraction and speech separation. Using a speaker inventory, i.e. a list of audio snippets of candidate speakers, the proposed system achieves better separation quality than PIT based blind speech separation. Meanwhile, it improves the efficacy and performance over speech extraction since it is not tied with a single target speaker.

The rest of this paper is organized as follows. In Section 2, we describe SSUSI. Section 3 and 4 contain the experimental setup and results. Finally, we make a conclusion in Section 5.

2. SYSTEM DESCRIPTION

2.1. Task Definition

This study addresses the speech separation problem when a list of candidate speakers is available, as with the case of scheduled business meetings, where the candidate speakers correspond to the meeting invitees. The voice profiles of some candidate speakers are given beforehand, forming a speaker inventory. Speakers involved in the

*This work was performed during an internship at Microsoft.

overlapped speech are referred to as relevant speakers. In this paper, the number of relevant speakers is assumed to be two and the speaker inventory only contains voice recordings collected before the meeting. It should be noted that the relevant speakers may not be included in the inventory. In the experiment section, we test our algorithm in this setting.

2.2. Speech Separation Using Speaker Inventory

The speech separation using speaker inventory (SSUSI) model consists of a profile selection system and a speech separation system. The profile selection system selects relevant profiles from the speaker inventory. The speech separation system then uses the selected profiles as additional information to separate the mixed speech. An illustration of SSUSI is shown in Fig. 1.

2.2.1. Profile Selection System

The profile selection system, shown in Fig. 2-(a), aims at selecting relevant profiles from the speaker inventory. It has three main components, an embedding module, a weight calculation module, and a profile selector.

The embedding module maps input features $\boldsymbol{X} \in R^{T\times F}$ to embeddings $\boldsymbol{E} \in R^{T\times E}$, where T refers to frames, F is the input feature dimension, and E denotes the embedding dimension. This embedding module is shared among mixed speech and speaker profiles. For mixed speech $\boldsymbol{X}^m \in R^{T_m\times F}$, the embedding can be expressed as $\boldsymbol{E}^m \in R^{T_m\times E}$. Similarly, for a profile p in speaker inventory $\boldsymbol{P}$, the embedding can be represented as $\boldsymbol{E}^p \in R^{T_p\times E}$.

The weight calculation module measures the correlations between the embedding of mixed speech and those of speaker profiles. We denote the vector in $\boldsymbol{E}^m$ at time i as $\boldsymbol{e}_i^m$ and that in $\boldsymbol{E}^p$ at time j as $\boldsymbol{e}_j^p$. i ranges from 1 to T_m, whereas j from 1 to T_p. The operations in the weight calculation module can be described as equations (1) and (2):

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

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

where $d_{i,j}^p$ denotes the dot product of embedding vectors $\boldsymbol{e}_i^m$ and $\boldsymbol{e}_j^p$. Note the denominator in equation (2) is a summation over both profile time steps j and profiles p.

The profile selector then calculates the average weight w^p for each profile p as follows:

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

We select two profiles c_1 and c_2 that have the first and second largest w^p values to be the relevant profiles for the speech separation system:

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

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

2.2.2. Speech Separation System

Using selected profiles c_1 and c_2, the speech separation system generates estimated masks $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ in three steps, embedding, attention, and separation. A diagram depicting the speech separation system is shown in Fig. 2-(b).

The embedding module in the profile selection system is reused in the speech separation system. Note that the embedding modules for profile selection and speech separation can also be different, as will be discussed in Section 2.3.

The attention mechanism for speech separation is similar to that for profile selection but is applied to form speaker biases. The speaker bias $\boldsymbol{b}_i^{c_1}$ for selected profile c_1 at time i can be calculated according to equations (6) and (7) shown below:

$$\alpha_{i,j}^{c_1} = \frac{\exp(d_{i,j}^{c_1})}{\sum_{j=1}^{T_{c_1}}\exp(d_{i,j}^{c_1})} \tag{6}$$

$$\boldsymbol{b}_i^{c_1} = \sum_{j=1}^{T_{c_1}} \alpha_{i,j}^{c_1}\boldsymbol{e}_j^{c_1} \tag{7}$$

where $\alpha_{i,j}^{c_1}$ denotes element i,j in the attention matrix for selected profile c_1. The bias for c_2 can be calculated similarly.

Note that weight matrix element $w_{i,j}^p$ in equation (2) differs from attention matrix element $\alpha_{i,j}^{c_1}$ in equation (6) in that $w_{i,j}^p$ measures the correlations between the embedding of mixed speech and those of speaker profiles, whereas $\alpha_{i,j}^{c_1}$ softly aligns the embeddings of relevant profiles to that of the mixed speech.

Finally, a PIT based separator takes speaker biases as additional input and generates the separation result. Let $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ be the target clean features. An utterance-wise PIT loss can be expressed by equations (8) and (9) as follows:

$$L(\theta) = \min\{l_{1,1} + l_{2,2}, l_{1,2} + l_{2,1}\} \tag{8}$$

where L denotes the loss of a training sample and θ means learnable parameters. $l_{u,v}$ refers to the loss corresponding to $\boldsymbol{M}_u$ and $\boldsymbol{Y}_v$, which is defined as:

$$l_{u,v} = ||\boldsymbol{M}_u \otimes \boldsymbol{X}^m - \boldsymbol{Y}_v||_F^2, \tag{9}$$

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

2.3. SSUSI with Profile Selection Embedding

In the simplest implementation of SSUSI (vanilla SSUSI) described in subsection 2.2, the embedding module is optimized only for speech separation and may not be well-suited for profile selection. To improve it, we can train a separate embedding module specifically for the profile selection system. This method is denoted as SSUSI with profile selection embedding (SSUSI-PSE).

The loss function for the profile selection embedding module is as follows:

$$L(\theta) = (1 - w^{c_1} - w^{c_2})^2 + \sum_{\bar{c}_k\in\boldsymbol{P}-\{c_1,c_2\}} (w^{\bar{c}_k})^2 \tag{10}$$

We divide speaker inventory $\boldsymbol{P}$ into two subsets, relevant profiles $\{c_1, c_2\}$ and irrelevant profiles $\boldsymbol{P} - \{c_1, c_2\}$. For relevant profiles c_1 and c_2, the training objective is to make their summation equal one, whereas for each irrelevant profile $\bar{c}_k \in \boldsymbol{P} - \{c_1, c_2\}$, the objective is to set it to zero.

2.4. SSUSI with Matched Training

In addition to a specifically trained embedding module for speaker profile selection, another way to improve vanilla SSUSI is to make

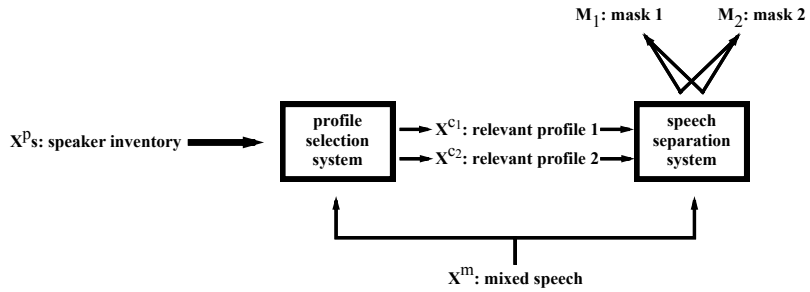


Fig. 1. An illustration of SSUSI. The thick arrows denote the representations corresponding to the speaker inventory and the four arrows between the speech separation system and separated speech refer to PIT.

the speech separation system more robust to wrong profile selections. SSUSI with matched training (SSUSI-MT) achieves this goal by jointly optimizing the profile selection and speech separation systems using a single PIT objective during training. Note that for vanilla SSUSI, we only train the speech separation system and reuse the embedding module in the speech separation system for profile selection at test time.

Note that the profile selection system uses an argmax function to select relevant profiles. Although the gradients w.r.t. the indices selected by argmax are hard to derive, we can still back-propagate errors w.r.t. the selected profiles for SSUSI-MT.

3. EXPERIMENTAL SETUP

3.1. Dataset

Our experiments are conducted on the LibriSpeech corpus [33] following the same recipe in [31]. We generate the training set using both of the clean training sets in LibriSpeech. At test time, the mixed speech is generated using the clean test set. For SSUSI models, the profile orders in the speaker inventory are randomized for both training and test. There are 1172 speakers in the training set and 40 speakers in the test set.

We use log spectrum features as model input in this study. The length of each frame is 32ms (i.e. 512 samples with a sampling rate of 16kHz) and the shift is 16ms. The waveform signals are transformed using a 512 dimensional short-time Fourier transformation (STFT) function. For the training target M, we use the spectral magnitude domain ideal ratio mask.

3.2. Implementation Details

3.2.1. Baseline Systems

An utterance-wise PIT based model [34] and Xiao *et al.*'s speech extraction system are used as the baselines in our experiments.

The PIT based model consists of six bidirectional long short-term memory (BLSTM) layers, each has 512 nodes with 256 forward and 256 backward cells. The mask for each source is estimated with the signal restoration loss function as suggested in (8) and (9).

The speech extraction baseline has the same model architecture and hyper-parameters as Xiao *et al.*'s model [31]. In addition, the number of learnable parameters is the same as that of the PIT based model above. For both the PIT based and speech extraction baselines, the optimizer is Adam and the learning rate is 10^{-4}.

3.2.2. SSUSI

The three types of SSUSI models have the same architecture as shown in Fig. 2. The vanilla SSUSI and SSUSI-MT have the same number of learnable parameters as those in the baselines, whereas SSUSI-PSE needs an additional embedding module consisting of three BLSTM layers.

For SSUSI-PSE and SSUSI-MT, two irrelevant profiles (i.e. four speaker profiles in total) are included during training. The profile selection embedding module in SSUSI-PSE is initialized with the well-trained embedding module in the vanilla SSUSI, whereas the whole SSUSI-MT is initialized with the vanilla SSUSI. The learning rates for the vanilla SSUSI, SSUSI-PSE and SSUSI-MT are 10^{-4}, 10^{-6}, and 10^{-5}, respectively. All the other hyper-parameters are the same as the baselines.

Note that during training, we shuffle the order of speaker profiles. This makes SSUSIs invariant to speaker profile permutations.

3.2.3. ASR Backend

The ASR backend is trained on the clean training set in LibriSpeech. The model consists of three BLSTM layers, each has 512 nodes. We generate forced aligned senone labels using Kaldi [35] and train the model using PyTorch with the maximum mutual information (MMI) criterion. Using this backend, the average WER for non-overlapped LibriSpeech test set is 5.7%.

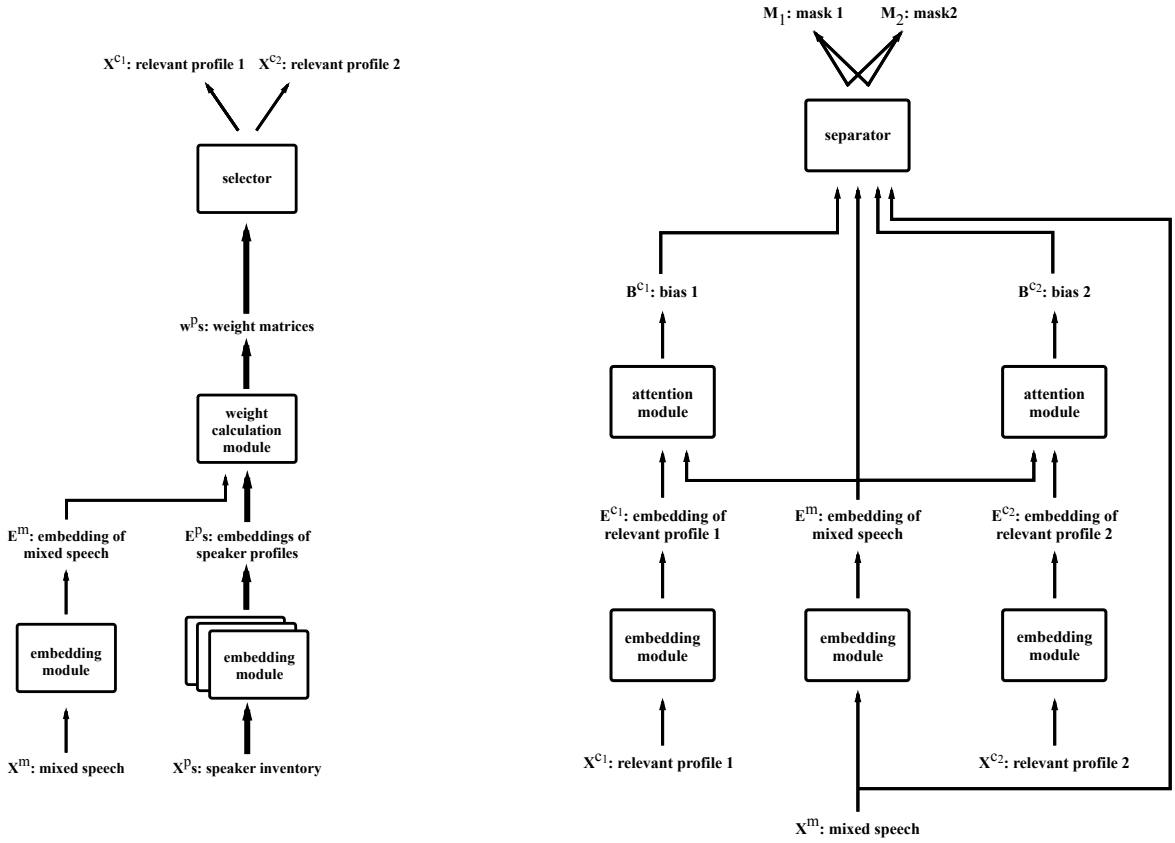
4. EVALUATION RESULTS

We first show the results of the three types of SSUSI models described in Section 2. Then we compare SSUSI with PIT and Xiao *et al.*'s speech extraction method. Finally, we analyze the performance of SSUSI by testing it in cases when one or both speakers are missing from the inventory. The last experiment shows some degree of robustness of the proposed algorithm and that the performance improvement of SSUSI over PIT is a result of the use of the speaker inventory.

4.1. SSUSI

The correct profile selection rates and signal to distortion ratios (SDRs) of the vanilla SSUSI model tested on up to six irrelevant profiles are shown in Table 1:

The vanilla SSUSI model gets an SDR of 12.1 dB when both relevant profiles are correctly selected. With the increase of irrele-



(a) profile selection system

(b) speech separation system

Fig. 2. The profile selection and speech separation systems in SSUSI. See Fig. 1 caption for acronyms.

method	# ir-profiles	$\geq$1 (%)	2 (%)	SDR (dB)
SSUSI	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

Table 1. The correct profile selection rates and SDRs of the vanilla SSUSI model. 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 $\geq$*1* and that of both relevant profiles as *2*.

vant profiles, all three metrics decrease. The correct selection rate of at least one relevant profile drops slightly from 100% to 99.0%, whereas that of both relevant profiles decreases significantly from 100% to 51.4%. SDRs are degraded by wrong profile selections. In the case when there are six irrelevant profiles, the SDR drops to 11.0 dB.

4.2. SSUSI-PSE

We show the correct profile selection rates and SDRs of SSUSI-PSE in Table 2 below:

method	# ir-profiles	$\geq$1 (%)	2 (%)	SDR (dB)
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

Table 2. The correct profile selection rates and SDRs of SSUSI-PSE. See Table 1 caption for acronyms.

Using profile selection embedding, the correct profile selection rate of both relevant profiles improves significantly. Moreover, the improvement gets larger as the number of irrelevant profiles increases. Table 2 also shows that the SDRs benefit from better profile selection. As regards the SDRs, for the six irrelevant profile case, the 11.0 dB result of the vanilla SSUSI increases to 11.3 dB using

SSUSI-PSE.

4.3. SSUSI-MT

The correct profile selection rates and SDR results of SSUSI-MT are shown in Table 3:

method	# ir-profiles	$\geq$1 (%)	2 (%)	SDR (dB)
SSUSI-MT	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

Table 3. The correct profile selection rates and SDRs of SSUSI-MT. See Table 1 caption for acronyms.

SSUSI-MT yields substantial SDR improvements over the vanilla SSUSI. Its SDR in the six irrelevant profile setting is 11.5 dB, outperforming both the vanilla SSUSI and SSUSI-PSE. The SDR of SSUSI-MT on zero irrelevant profile is 12.2dB, slightly surpassing that of the vanilla SSUSI. Note that the correct profile selection rates of SSUSI-MT are worse than those of the vanilla SSUSI. This supports our argument that the gain of the matched training comes mainly from the consistency between training and test rather than the improvement in profile selection.

Since SSUSI-MT performs the best among the three types of SSUSI in this study, we use it for the comparisons with other methods in the following subsections.

4.4. Comparisons Between SSUSI and PIT

The SDR and word error rate (WER) comparisons between SSUSI-MT and PIT are shown in Table 4. For SSUSI-MT, in addition to 0 and 6 irrelevant profiles, we also show the results on 22 and 30 irrelevant profiles.

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

Table 4. The SDR and WER comparisons between SSUSI-MT and PIT. See Table 1 caption for acronyms.

With respect to SDR, SSUSI-MT performs significantly better than PIT. Even when there are 30 irrelevant profiles, SSUSI-MT still yields an SDR of 10.8 dB. Note that SSUSI-MT is trained using only 2 irrelevant profiles. The results on 22 and 30 irrelevant profiles show the robustness and generalization ability of SSUSI-MT.

For WERs, SSUSI-MT outperforms PIT by 48% relatively when there is no irrelevant profile. In the case when there are 30 irrelevant profiles, the relative improvement is still 34%. This clearly shows the effectiveness of SSUSI-MT. The consistency between SDR and WER results is aligned with the observation in Weninger *et al.*'s study [36]. Note that the absolute values of the WERs in Table 4 are relatively high for the LibriSpeech corpus. In addition to the distortion in separated speech, we think another reason may be that the long silent segments in separated speech may result in inaccurate estimations of the statistics for utterance-wise input feature normalization.

4.5. Comparisons Between SSUSI and Speech Extraction

We show the SDR and WER comparisons between SSUSI-MT and the speech extraction system of [31] in Table 5:

method	# ir-profiles	SDR (dB)	WER (%)
Speech Extraction [31]	0	11.5	21.9
	1	11.1	23.3
	2	10.9	24.4
SSUSI-MT	0	12.2	19.1
	1	12.0	19.9
	2	11.9	20.4

Table 5. The SDR and WER comparisons between SSUSI-MT and speech extraction system. See Table 1 caption for acronyms.

SSUSI-MT substantially outperforms the speech extraction system. In terms of SDR, an improvement of more than 0.7 dB is yielded. As for WER, the overall relative improvement is over 13%. Moreover, the improvement in both SDR and WER gets larger with the increase of irrelevant profiles. The reason of the performance improvement may be that the profile selection system in SSUSI-MT filters out the interfering information in irrelevant profiles by a deterministic selection of relevant profiles. Even in the case of zero irrelevant profile, SSUSI-MT provides a meaningful improvement over the speech extraction system. This implies SSUSI-MT's better discrimination capability between similar speakers. This could be because it attempts to separate two speakers simultaneously, unlike speech extraction.

In addition to separation accuracy, SSUSI-MT improves the efficacy over the speech extraction system significantly. In the case of zero irrelevant profiles, the computation time reduction is about 40% relatively. If there are 30 irrelevant profiles, the computation time reduces by 70% relatively. The reason is that using a single thread, speech extraction systems need to be run repetitively for each candidate speaker, whereas the profile selection system in SSUSI-MT filters out all but one pair of speaker profiles for the separation system.

4.6. An Analysis On SSUSI

We analyze the influence of speaker inventory's accuracy on SSUSI-MT by testing it in cases where one or both relevant profiles are missing. The SDR values in these cases are shown in Table 6 below:

In the case when one relevant profile is missing, the SDRs of SSUSI-MT drop to values close to 10.1 dB. When both relevant profiles are missing, the SDRs further drop to 8.5 dB or so. This confirms that SSUSI-MT indeed leverages the information in the speaker inventory. Note that SSUSI-MT performs similarly to PIT when both relevant profiles are missing. This shows that the performance improvement of SSUSI-MT over PIT comes only from the speaker inventory. This also suggests the robustness of the proposed SSUSI-MT to the missing profiles. Even when only one relevant speaker is included in the inventory, SSUSI-MT substantially outperforms PIT.

method	# ir-profiles	standard	m1	m2
SSUSI-MT	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 6. The SDRs of SSUSI-MT in missing profile cases. *standard* denotes both relevant profiles are in the speaker inventory, *m1* refers to cases with one relevant profiles missing, and *m2* means both relevant profiles are missing. See Table 1 caption for other acronyms.

5. CONCLUDING REMARKS

We have proposed SSUSI, a novel speech separation system that is able to leverage the information in the speaker inventory comprehensively. In addition to the vanilla SSUSI, two improved versions are investigated, namely SSUSI-PSE and SSUSI-MT. SSUSI-MT performs the best among them in terms of SDR. The experimental results show that SSUSI-MT outperforms PIT based blind speech separation by up to 48% relatively in WER. Compared with speech extraction, SSUSI yields more than 13% relative improvement in WER and achieves up to 70% computation time reduction. Future work includes extending SSUSI to the multichannel case, evaluating SSUSI in real conversations, and designing a better profile selection strategy.

6. REFERENCES

[1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 31–35.

[2] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. of INTERSPEECH*, 2016, pp. 545–549.

[3] D. Yu, M. Kolbæk, Z. H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 241–245.

[4] 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 Transactions on Audio, Speech and Language Processing (TASLP)*, , no. 10, pp. 1901–1913, 2017.

[5] Z. Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018, pp. 686–690.

[6] Y. Luo and N. Mesgarani, "Augmented time-frequency mask estimation in cluster-based source separation algorithms," in *Acoustics, Speech and Signal Processing (ICASSP), 2019 IEEE International Conference on*. IEEE, 2019, pp. 710–714.

[7] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 246–250.

[8] C. Xu, W. Rao, E. S. Chng, and H. Li, "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss," in *Acoustics, Speech and Signal Processing (ICASSP), 2019 IEEE International Conference on*. IEEE, 2019, pp. 6990–6994.

[9] Z. Chen and J. Droppo, "Sequence modeling in unsupervised single-channel overlapped speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018, pp. 4809–4813.

[10] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 26, no. 4, pp. 787–796, 2018.

[11] Z. X. Li, Y. Song, L. R. Dai, and I. McLoughlin, "Listening and grouping: an online autoregressive approach for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 27, no. 4, pp. 692–703, 2019.

[12] Z. X. Li, Y. Song, L. R. Dai, and I. McLoughlin, "Source-aware context network for single-channel multi-speaker speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018, pp. 681–685.

[13] L. Li and H. Kameoka, "Deep clustering with gated convolutional networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018, pp. 16–20.

[14] K. Tan, J. Chen, and D. L. Wang, "Gated residual networks with dilated convolutions for supervised speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018, pp. 21–25.

[15] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018, pp. 6–10.

[16] 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 (TASLP)*, vol. 27, pp. 1256–1266, 2019.

[17] J. Wu, Y. Xu, S. X. Zhang, L. W. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," *arXiv preprint arXiv:1904.03760*, 2019.

[18] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*, 2018, pp. 1–5.

[19] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018, pp. 5739–5743.

[20] R. Gu, J. Wu, S. X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "End-to-end multi-channel speech separation," *arXiv preprint arXiv:1905.06286*, 2019.

[21] 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," *arXiv preprint arXiv:1902.04891*, 2019.

[22] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *arXiv preprint arXiv:1804.03619*, 2018.

[23] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," *arXiv preprint arXiv:1804.03160*, 2018.

[24] 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 Language Technology Workshop (SLT)*, 2018.

[25] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Multi-channel speech separation with recurrent neural networks from high-order ambisonics recordings," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*, 2018, pp. 36–40.

[26] Y. Zhao, Z. Q. Wang, and D. L. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 1, pp. 53–62, 2018.

[27] 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 *Interspeech*, 2017.

[28] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5554–5558.

[29] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," *arXiv preprint arXiv:1810.04826*, 2018.

[30] J. Wang, J. Chen, D. Su, L. Chen, M. Yu, Y. Qian, and D. Yu, "Deep extractor network for target speaker recovery from single channel speech mixtures," in *Proc. of INTERSPEECH*, 2018, pp. 307–311.

[31] X. Xiao, Z. Chen, T. Yoshioka, H. Erdogan, C. Liu, D. Dimitriadis, J. Droppo, and Y. Gong, "Single-channel speech extraction using speaker inventory and attention network," in *Acoustics, Speech and Signal Processing (ICASSP), 2019 IEEE International Conference on*. IEEE, 2019, pp. 86–90.

[32] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, "A unified framework for neural speech separation and extraction," in *Acoustics, Speech and Signal Processing (ICASSP), 2019 IEEE International Conference on*. IEEE, 2019, pp. 6975–6979.

[33] 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*, 2015, pp. 5206–5210.

[34] M. Kolbk, Z. H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, pp. 153–167, 2017.

[35] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, and others, "The Kaldi speech recognition toolkit," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 1–4.

[36] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J.R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. of International Conference on Latent Variable Analysis and Signal Separation*, 2015, pp. 91–99.