J. Shanghai Jiao Tong Univ. (Sci.), 2024
https://doi.org/10.1007/s12204-024-2739-7

Simultaneous Speech Extraction for Multiple Target Speakers Under Meeting Scenarios

ZENG Bang[1,2] (曾 邦), *SUO Hongbin*[3] (索宏彬), *WAN Yulong*[3] (万玉龙), *LI Ming*[1,2*] (李 明)
(1. School of Computer Science, Wuhan University, Wuhan 430027, China; 2. Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Duke Kunshan University, Kunshan 215316, Jiangsu, China;
3. Data & AI Engineering System, OPPO, Beijing 100125, China)

Abstract: The common target speech separation directly estimates the target source, ignoring the interrelationship between different speakers at each frame. We propose a multiple-target speech separation (MTSS) model to simultaneously extract each speaker's voice from the mixed speech rather than just optimally estimating the target source. Moreover, we propose a speaker diarization (SD) aware MTSS system (SD-MTSS). By exploiting the target speaker voice activity detection (TSVAD) and the estimated mask, our SD-MTSS model can extract the speech signal of each speaker concurrently in a conversational recording without additional enrollment audio in advance. Experimental results show that our MTSS model achieves improvements of 1.38 dB signal-to-distortion ratio (SDR), 1.34 dB scale-invariant signal-to-distortion ratio (SISDR), and 0.13 perceptual evaluation of speech quality (PESQ) over the baseline on the WSJ0-2mix-extr dataset, separately. The SD-MTSS system makes a 19.2% relative speaker dependent character error rate reduction on the Alimeeting dataset.

Keywords: target speech separation, interrelationship, speaker diarization (SD), target speaker voice activity detection, multiple-target speech separation (MTSS) model
CLC number: TN912.34 **Document code:** A

0 Introduction

In the real world, noise and speaker interference can degrade the system performance of back-end speech applications. Speech separation effectively solves this problem by extracting the target speech from the mixed utterance. Early methods called blind speech separation, such as deep clustering (DPCL)[1], deep attractor network (DANet)[2], and permutation invariant training (PIT)[3-4], can separate each source from a mixed speech. These algorithms formulated in the time-frequency domain have an upper bound on reconstructing waves[5]. Recent solutions in the time-domain, such as time-domain audio source separation (Tas-Net)[5-6] and dual-path recurrent neural network (DPRNN)[7] break through the constraints and achieve state-of-the-art performance in the separation task. Despite this, the unknown number of speakers and the global permutation problem are still two challenges for blind speech separation.

To address the above two problems, a framework called speaker extraction[8-9] or target speech separation[10-12] can extract a target speaker' speech from the mixed audio by utilizing an auxiliary reference speech of the target speaker. However, it is required to filter out multiple target speakers in certain tasks, e.g., meeting scenarios. The common approach is to infer the mixed speech several times and each process is independent of the other, ignoring the interrelationship between the speech of different speakers at each frame. In addition, obtaining the reference speech of multiple target speakers in advance is difficult to achieve. Considering the aforementioned problems, repeatedly processing the mixture speech towards different target speakers separately may not be a feasible solution.

It is worth noting that speech in the meeting scenario usually has a long duration and contains both single-talker and overlapped voice segments. Thus, it is possible to use the single-talker segments as the reference speech for participants instead of obtaining additional speech for enrollment. Speaker diarization (SD)[13] technology is very suitable for this role. SD aims to slice different speaker segments in a continuous multiple speaker conversation and determine which speaker each segment belongs to. More recently, multi-channel target speaker voice activity detection

Received: 2023-12-19 **Accepted:** 2024-01-05
Foundation item: the National Natural Science Foundation of China (No. 62171207), and the Science and Technology Program of Suzhou City (No. SYC2022051) and OPPO
*E-mail: ming.li369@dukekunshan.edu.cn

(MC-TSVAD)[14], which selects target speaker voice activity detection (TSVAD) as the post-processing module and employs cross-channel self-attention, achieved the best result in the multi-party meeting transcription challenge (M2Met)[15].

In this work, we propose the multiple-target speech separation (MTSS) model, which is a speech extraction method for multiple target speakers. The MTSS model infers each speaker's mask simultaneously and limits their estimated masks to be sum to 1. We consider that the energy values of different speakers at each frame are not independent to each other. Moreover, we propose the SD-MTSS framework, which associates target speech separation with speaker diarization. We select the TSVAD system as the speech diarization network. Based on the decisions from TSVAD[16], we can obtain each speaker's reference speech directly from the mixed audio. Then, each speaker's reference speech is fed into the MTSS module in the separation stage.

1 Methods

1.1 Multiple Target Speech Separation Model

1.1.1 *Backbone*

The backbone of the MTSS model is SpEx+[17] which consists of two twin speech encoders, a speaker encoder, a speaker extractor, and a speech decoder. The twin speech encoder models the input sequence and auxiliary speech in a common latent space through sharing the structure and parameters. The speaker encoder model is a ResNet-based speaker classifier used to generate the speaker embedding of the reference speech. The speaker extractor takes both the speaker embedding and the output of the twin speech encoder as the inputs, and then produces masks in three different scales. The speech decoder outputs the estimation by multiplying the input sequence and the multi-scale masks.

1.1.2 *MTSS Model*

Herein, we propose a speech extraction model MTSS for multiple target speakers, which can simultaneously separate the speech of each speaker present in the conversation. The schematic diagram of the MTSS model is shown in Fig. 1. Unlike the fact that the original SpEx+ neural network takes only one speaker's reference speech, MTSS takes two speakers' reference speeches as the inputs and processes them separately. Moreover, we replace the ReLU with softmax to establish the relationship between the masks of each speaker in the same utterance. We believe that taking the interrelation into account will improve the final separation performance of the model. Because in the definition of binary masks, each time-frequency cell belongs to a speaker with stronger energy. Specifically, the responses of MTSS $\mathbf{est}_{s_1}$, $\mathbf{est}_{s_2}$ can be formulated as follows:

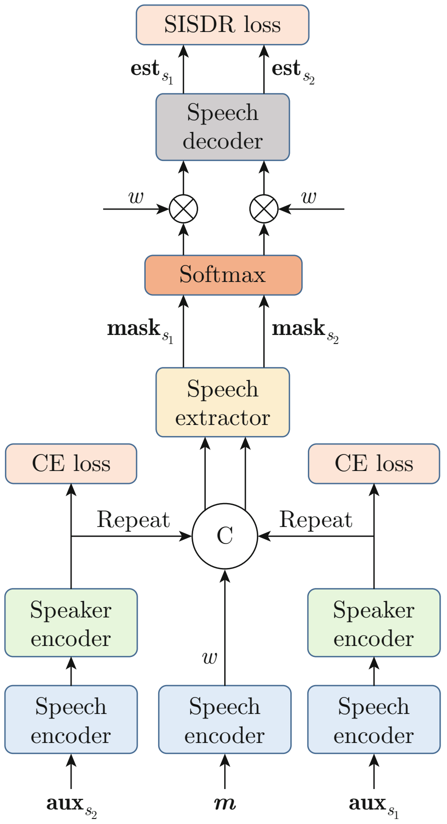

Fig. 1 Details of the MTSS model. Scale-invariant signal-to-distortion ratio (SISDR). Cross-entrop (CE). s_1 and s_2 represent the two speakers presented in the mixture. $\mathbf{est}_{s_1}$, $\mathbf{est}_{s_2}$ denote the estimations of two speakers. $\mathbf{aux}_{s_1}$ and $\mathbf{aux}_{s_2}$ are the reference waves of two speakers. $\boldsymbol{m}$ is the mixed wave. w is the encoder output of $\boldsymbol{m}$. "C" denotes the operation of concatenate. ⊗ is an operation for element-wise product.

$$(\mathbf{est}_{s_1}, \mathbf{est}_{s_2}) = \\ \boldsymbol{m} \otimes \{\text{softmax}(\text{cat}(\mathbf{mask}_{s_1}, \mathbf{mask}_{s_2}))\}, \qquad (1)$$

where, $\mathbf{mask}_{s_1}$, $\mathbf{mask}_{s_2} \in \mathbb{R}^{N\times1\times T}$, N is the feature dimention, and T is the time length; ⊗ is an operation of element-wise product; softmax and cat indicate that a softmax function and concatenation operate on the penultimate dimension, respectively. We also implement a multi-task learning framework for the target speech separation.

1.2 SD-MTSS System

Considering that it is feasible to apply speaker diarization techniques to target speech separation, we expend the MTSS to speaker diarization (SD) aware MTSS (SD-MTSS) system. The SD-MTSS system architecture is shown in Fig. 2. Rather than requiring additional registration, SD-MTSS directly obtains reference speech from the long utterance itself through the

SD module. In real applications, the single-channel SD approach[18] can be used here. The SD-MTSS system consists of an SD module and an MTSS module. The SD module produces the TSVAD decision for multiple speakers, which are the probabilities of each speaker's presence at the frame level. The MTSS module adopts each speaker's reference speech from the SD module and the mixture audio as inputs, and then outputs the estimation for multiple target speakers.

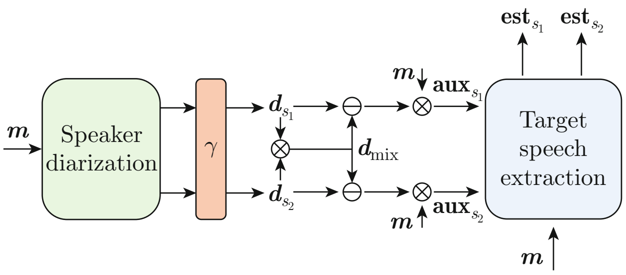

Fig. 2 Schematic of the SD-MTSS system. γ is a threshold (0.5 often). $\boldsymbol{d}_{s_1}$ and $\boldsymbol{d}_{s_2}$ indicate the binarized TSVAD decision. $\boldsymbol{d}_{\text{mix}}$ indicates the TSVAD decision where the values of $\boldsymbol{d}_{s_1}$ and $\boldsymbol{d}_{s_2}$ are both 1. $\ominus$ is element-wise substraction operation.

Using the TSVAD decision, we can get the single-talker audio segments as the reference speech for each speaker. The scheme of obtaining single-talker segments in Fig. 2 is organized as follows. We use $\boldsymbol{m} \in \mathbb{R}^{1\times T}$ indicating the input sequence; s_1 and s_2 indicating two different speakers in the mixture. First, the TSVAD decision is passed through a threshold mechanism and produces the binarized results $\boldsymbol{d}_{s_1}$ and $\boldsymbol{d}_{s_2}$ whose values consist of 0 and 1. Then, the reference speech can be formulated as follows:

$$\mathbf{aux}_{s_1} = \boldsymbol{m} \otimes \tilde{\boldsymbol{d}}_{s_1}, \tag{2}$$

$$\tilde{\boldsymbol{d}}_{s} = \boldsymbol{d}_{s_1} - \boldsymbol{d}_{s_1} \otimes \boldsymbol{d}_{s_2}, \tag{3}$$

$$\mathbf{aux}_{s_2} = \boldsymbol{m} \otimes \tilde{\boldsymbol{d}}_{s_2}, \tag{4}$$

$$\tilde{\boldsymbol{d}}_{s_2} = \boldsymbol{d}_{s_2} - \boldsymbol{d}_{s_1} \otimes \boldsymbol{d}_{s_2}, \tag{5}$$

where, $\tilde{\boldsymbol{d}}_{s_1}$ and $\tilde{\boldsymbol{d}}_{s_2}$ indicate the mono-speaker activity parts of $\boldsymbol{d}_{s_1}$ and $\boldsymbol{d}_{s_2}$, respectively; $\boldsymbol{d}_{s_1}$, $\boldsymbol{d}_{s_2} \in \mathbb{R}^{1\times T}$; $\otimes$ indicates the element-wise product. Selected continuous audio segments of $\mathbf{aux}_{s_1}$ and $\mathbf{aux}_{s_2}$ will be fed into the MTSS module as the reference speech for the subsequent separation task.

1.3 Speaker Diarization Module

The SD module in this work consists of a clustering-based module for target speaker embedding extraction and a TSVAD system for diarization results refinement[14].

1.3.1 *Clustering-Based Module*

The affinity matrix extraction model of TSVAD is based on the neural network in Ref. [19], using an LSTM-based model in similarity measurement for speaker diarization. It consists of two bidirectional long short-term memory network (Bi-LSTM) layers and two fully connected layers. The LSTM-based model first splits the entire audio into short speech clips and extracts the speaker embedding of all segments. Then, it takes these segments as inputs and produces the initialized diarization result through adopting spectral clustering.

1.3.2 *TSVAD System*

The architecture of the TSVAD[14] system is shown in Fig. 3, which consists of three parts.

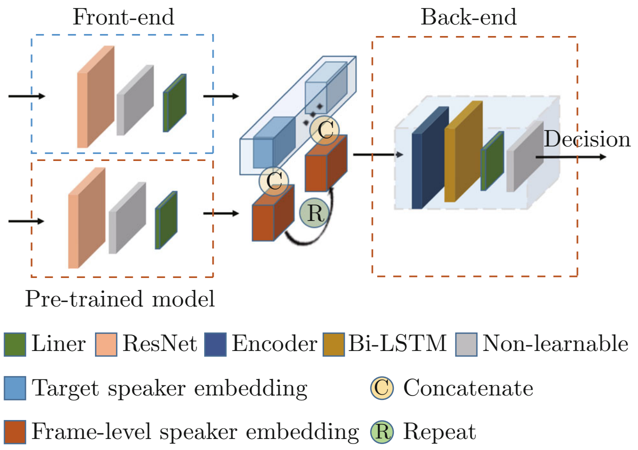

Fig. 3 Structure of the TSVAD system. The front-end shares the same architecture with the pre-trained speaker embedding model. The target speaker embedding concatenates with the frame-level speaker embedding repeatedly and then is fed into the back-end.

(1) A pre-trained speaker embedding model ResNet+[20] based on ArcFace[21] and cosine similarity scoring. The dimension of the speaker embedding layer is 128, and the margin and softmax prescaling of the ArcFace are 0.2 and 32 respectively.

(2) A front-end model with the same architecture as the pre-trained model is used to extract the frame-level speaker embedding.

(3) A back-end model consists of an encoder layer, a Bi-LSTM layer, a linear layer, and a sigmoid function.

First, the pre-trained speaker embedding model extracts the target speaker embeddings. Meanwhile, the front-end network loads its parameters to extract the frame-level speaker embeddings. The target speaker embeddings are repeatedly concatenated with the frame-level speaker embeddings and then fed into the back-end. Next, the encoder layer of the back-end model produces each target speaker's detection state. The Bi-LSTM layer inputs these detection states and models the relationship between speakers. Finally, the linear layer coupled with a sigmoid function generates each speaker's final decision, i.e., TSVAD decision.

More details can be found in Ref. [14].

2 Experiment Setup

2.1 Dataset

Datasets for MTSS Model We simulated a commonly used two-speakers mixture datasets WSJ0-2mix-extr[①] (20 000 utterances in training set, 5 000 utterances in validation set, and 3 000 utterances in test set, respectively), the sampling rate of which is 8 kHz. The simulation process is the same as SpEx+[17], and the only difference is that we produce a couple of target speakers speech ($\mathbf{tgs}_{s_1}$, $\mathbf{tgt}_{s_2}$) and reference speech ($\mathbf{aux}_{s_1}$, $\mathbf{aux}_{s_2}$) for each mixture utterance, while SpEx+[17] only selects the first talker as the target speaker. The utterances from $\mathbf{tgt}_{s_1}$ and $\mathbf{tgt}_{s_2}$ are set in a relative signal-to-distortion ratio between 0 to 5 dB. The average scale-invariant signal-to-distortion ratio (SISDR) of mixed speech is 2.50 dB and −2.50 dB when it takes $\mathbf{tgt}_{s_1}$ and $\mathbf{tgt}_{s_2}$ as the reference.

Datasets for SD-MTSS Model For the SD module, we use the training set of Alimeeting[15] to train the clustering-based affinity matrix extraction neural network. Alimeeting contains 118.75 h of speech data, including 104.75 h (426 speakers) of the training set, 4 h (25 speakers) of the validation set, and 10 h of the test set. For the TSVAD model in the SD module, we create simulated datasets based on the Alimeeting training set. The simulation scheme is the same as Ref. [14]. For MTSS module, we use the Libri-2mix[22] as the training set, with a sampling rate of 16 kHz. We select the signal channel signal on channel 0 of the two-speakers samples from the Eval-Ali-far and Test-Ali-far subsets of Alimeeting to evaluate the performance of SD-MTSS model.

2.2 Implementation Details

To compare with the baseline, the hyperparameters and learning schedule of MTSS module are set the same as SpEx+[17]. The number of filters in the encoder is 256, the number of convolutional blocks in each repeat is 8, the number of repeat is 4, the number of channels in the convolutional blocks is 512, and the kernel size of the convolutional blocks is 3. The hyperparameters of the network are shown in Table 1.

The initial learning rate is 1×10^{-3} and decays by 0.5 if the accuracy of validation set is not improved in 2 epochs. Early stopping is applied if the accuracy of validation set has not improved after 6 epochs. Like that in SpEx+[17], we use the multi-task learning implementation for training with two objectives. We use the SISDR[23] as the loss of output speech quality and a CE loss for speaker classification:

$$\mathrm{SISDR} = 10\lg\frac{\|\boldsymbol{e}_{\mathrm{target}}\|^2}{\|\boldsymbol{e}_{\mathrm{res}}\|^2} = 10\lg\frac{\left\|\dfrac{\hat{\boldsymbol{s}}^{\mathrm{T}}\boldsymbol{s}}{\|\boldsymbol{s}\|^2}\boldsymbol{s}\right\|^2}{\left\|\dfrac{\hat{\boldsymbol{s}}^{\mathrm{T}}\boldsymbol{s}}{\|\boldsymbol{s}\|^2}\boldsymbol{s}-\hat{\boldsymbol{s}}\right\|^2}, \qquad (6)$$

$$\hat{\boldsymbol{s}} = \boldsymbol{e}_{\mathrm{target}} + \boldsymbol{e}_{\mathrm{res}}, \qquad (7)$$

$$L_{\mathrm{SISDR}} = -[(1-\alpha-\beta)\mathrm{SISDR}_{l_1} + \alpha\mathrm{SISDR}_{l_2} + \beta\mathrm{SISDR}_{l_3}], \qquad (8)$$

$$L_{\mathrm{CE}} = -\sum_{i=1}^{N_s} I_i \log\sigma_i(\boldsymbol{W}\cdot\boldsymbol{v}), \qquad (9)$$

where, $\boldsymbol{s}$ and $\hat{\boldsymbol{s}}$ represent the label and estimated speeches, respectively; SISDR_{l_1}, SISDR_{l_2} and SISDR_{l_3} represent three different multi-scale estimations; α and β are the weights to different scales; $\boldsymbol{e}_{\mathrm{target}}$ and $\boldsymbol{e}_{\mathrm{res}}$ indicate the estimated speech's orthogonal projection and residual w.r.t. the reference speech, respectively; N_{s} is the number of speakers in the training datasets; I_i represents the class label of the ith speaker; $\boldsymbol{W}$ represents a weight matrix; $\boldsymbol{v}$ represents the speaker encoder output of the reference speech ($\mathbf{aux}_{s_2}$); $\sigma(\cdot)$ represents a softmax function. The multi-task objective function for single speaker is

$$L_{\mathrm{multi}} = L_{\mathrm{SISDR}} + L_{\mathrm{CE}}. \qquad (10)$$

The overall objective function for our MTSS model is

$$L(\theta|\boldsymbol{m}, \mathbf{aux}_{s_{1,2}}, \mathrm{spk}_{1,2}, I_{s_{1,2}}) = \lambda_1 L_{\mathrm{multi}_{\mathrm{spk}_1}} + \lambda_2 L_{\mathrm{multi}_{\mathrm{spk}_2}}, \qquad (11)$$

where $\boldsymbol{m}$ is the input sequence, $\mathbf{aux}_{s_{1,2}}$ are the reference speeches of two speakers, spk_1 and spk_2 are the target speeches of two speakers, $I_{s_{1,2}}$ are the speaker class labels of two speakers, and λ_1 and λ_2 are the weights of SISDR loss and CE loss, respectively. Herein, we set $\lambda_1 = 0.5$ and $\lambda_2 = 0.25$ as the default values.

The SD module chooses the Adam and binary cross-entropy loss as the optimizer. The input chunk size is 16 s, and the acoustic feature is 80-dimensional log Mel-filterbank energy with a frame length of 25 ms and a frame shift of 10 ms. The training details can be found in Ref. [24].

The training steps of the SD modules are as follows:

(1) Transfer the pre-trained speaker embedding model's parameters to the front-end model in the TSVAD model. Maintain the front-end model in a fixed state while focusing our training efforts on the back-end model.

(2) Subsequently, once the back-end model reaches convergence, we proceed to unfreeze the front-end model and embark on a joint training phase for the entire model, spanning an additional 10 epochs.

[①] https://github.com/xuchenglin28/speaker_extraction

Table 1 Hyperparameters of the MTSS module

Symbol	Setting	Description
L1, L2, L3	20, 80, 160	Lengths of the encoder filter
N	256	Number of filters in encoder
X	8	Number of convolutional blocks
B	256	Number of channels in bottleneck conv blocks
H	512	Number of channels in convolutional blocks
P	3	Kernel size in convolutional blocks
Spk_emb_dim	256	Dimension of the speaker embedding

(3) In the final stage, we initiate fine-tuning of the model using the Alimeeting training set, extending this process over 200 epochs while employing a learning rate of 1×10^{-5}.

The diarization error rates (DERs) of the single-channel SD module[18] on the test set of Alimeeting are shown in Table 2. We use the offline model as a SD system in our proposed SD-MTSS model. The SD module has a 4.12% DER on the evaluation set of the Alimeeting dataset.

Table 2 DERs of the single-channel offline and online SD systems on AliMeeting Eval set

%

Model	2-spk	3-spk	4-spk	Total
Offline	0.89	6.63	5.47	4.12
Online	1.90	8.36	12.12	8.14

We evaluate our proposed models for two steps: ① Examine the performance of MTSS on WSJ0-2mix-extr dataset. We train the MTSS model with a pre-trained model on the training set of WSJ0-2mix-extr. Then, we compare MTSS-Softmax and MTSS-ReeLU in terms of signal-to-distortion ratio (SDR), scale-invariant signal-to-distortion ratio (SISDR), and perceptual evaluation of speech quality (PESQ). ② Examine the performance of SD-MTSS system on Alimeeting. We compare SpEx+[②] (implemented by ourselves with using Libri-2mix dataset as training set) and SD-MTSS model in terms of speaker dependent character error rate (CER)[25].

3 Results and Discussion

3.1 Results on WSJ0-2mix-extr

The results of our proposed MTSS model and the baseline system are shown in Table 3. Since we used the same simulation test set as SpEx+[17], we directly use the evaluation results of SpeakerBeam, SBF-MTSAL-Concat, TseNet, SpEx, and SpEx+[17]. As shown in Table 3, SpEx+[17] is the baseline that we implemented, and MTSS is the model we proposed. Our proposed MTSS model achieves significantly better results across all the metrics. The samples of separated audio are available at this link[③].

Specifically, MTSS-Softmax outperforms SpEx+ with relative improvements of 7.4% in terms of SDR, 7.3% in terms of SISDR, and 3.2% in terms of PESQ, separately. In addition, we can get better improvement on each speaker s_1 and s_2 while extracting their target speech simultaneously. Comparing the results of MTSS-ReLU and MTSS-Softmax, we can conclude that setting the constraint for each speaker's mask mainly contributs to the improvements, and the interrelationship between different speakers at each frame benefits the model to extract the target source.

3.2 Results on Alimeeting

The speech recognition results of our proposed SD-MTSS system are shown in Table 4. Here, we report the average speaker independent CER results on the Eval_Ali_far and Test_Ali_far subsets of Alimeeting. It is important to note that we have adopted minimum variance distortionless response[④] (MVDR) beamformer on the mixture in advance. Due to multi-speaker interference, many insertion errors are generated in recognizing the mixed speech. The difference between SpEx+ and our proposed SD-MTSS system is that the SD-MTSS can extract the speech of each speaker simultaneously in one inference and does not need an enrollment wave in advance. Compared with the SpEx+ model, the SD-MTSS model achieves 21.4% and 18.3% relative average speaker dependent CER reductions on Eval_Ali_far and Test_Ali_far subsets of the Alimeeting, respectively. Since we only need to evaluate the effectiveness of SD-MTSS model, we did not train the recognition and separation models jointly. As far as we know, joint training and fine-tuning with Alimeeting datasets can improve the final recognition performance[25]. The results of our proposed SD-MTSS system are shown in Table 4. Since we evaluate the system on the far-field data and use the corresponding close-talking data as the ground truth, the model

[②] https://github.com/gemengtju/SpEx_Plus

[③] https://github.com/ZBang/SD-MTSS

[④] https://github.com/funcwj/setk

Table 3 SDR, SISDR, and PESQ of separated speech using the MTSS method on the WSJ0-2mix-extr dataset

Method	N	SDR/dB		SISDR/dB		PESQ	
		s_1	s_2	s_1	s_2	s_1	s_2
Mixture	—	2.60	−2.14	2.50	−2.50	2.31	1.86
SpeakerBeam[26]	1	9.62	—	9.22	—	2.64	—
SBF-MTSAL-Concat[27]	1	11.39	—	10.60	—	2.77	—
TseNet[28]	1	15.24	—	14.73	—	3.14	—
SpEx[29]	1	17.15	—	16.68	—	3.36	—
SpEx+[17]	1	18.54	—	18.20	—	3.49	—
Pre-trained	1	18.15	16.42	17.55	15.89	3.44	3.28
MTSS-ReLU	2	19.18	17.29	18.72	16.84	3.56	3.39
MTSS-Softmax	2	**19.92**	**17.42**	**19.54**	**16.99**	**3.62**	**3.41**

Note: N indicates the number of outputs per inference. s_1 and s_2 indicate the different speaker of the mixture. MTSS-ReLU: Using ReLU as the activation function and do not impose constraints on masks. MTSS-Softmax: Using softmax function to limit the sum of masks to 1.

does not performance well in terms of SDR and SISDR improvements. Nevertheless, from Table 4, we can conclude that our proposed multiple target speech separation model surpasses the pre-trained model (SpEx+) with a large margin in terms of SISDR improvement.

Table 4 Average speaker dependent CERs results of SD-MTSS on Eval_Ali_far and Test_Ali_far sets %

Method	N	Eval	Test	Avg
Mixture	—	96.70	95.51	95.83
SpEx+ (Re)	1	45.80	43.79	44.34
SD-MTSS	2	**35.97**	**35.78**	**35.83**

Note: N indicates the number of outputs per inference. Re indicates that the model is implemented by ourselves. We use MTSS-Softmax model as the MTSS module of the SD-MTSS system. We use WeNet[30] as the speech recognition model in this experiments. The WeNet model is trained by WeNetSpeech[31] and our inhouse data together with approximately 1.5×10^4 hours.

4 Conclusion

We propose a multiple target speech separation (MTSS) model which can simultaneously extract each speaker's voice from the mixed speech. To establish a relationship between different speakers in each frame, we constrain the sum of each speaker's estimated mask to 1 when extracting their speeches simultaneously. Moreover, we propose a speaker diarization multiple target speech separation system (SD-MTSS). By associating the speaker diarization task and the target speech separation task together, we do not require the additional reference speech for enrollment. The experimental results show that our proposed MTSS model significantly improves the separation performance on WSJ0-2mix-extr datasets. In addition, the SD-MTSS model outperforms the baseline by a large margin in terms of speaker independent CER on the Alimeeting datasets. For future work, we will implement our method with different state-of-the-art networks and improve the system's performance in the far-field scenarios.

Acknowledgment The authors thank the computational resource provided by the Advanced Computing East China Sub-Center.

Conflict of Interest The authors declare that they have no conflict of interest.

References

[1] HERSHEY J R, CHEN Z, LE ROUX J, et al. Deep clustering: Discriminative embeddings for segmentation and separation [C]//*2016 IEEE International Conference on Acoustics, Speech and Signal Processing*. Shanghai: IEEE, 2016: 31-35.

[2] CHEN Z, LUO Y, MESGARANI N. Deep attractor network for single-microphone speaker separation [C]//*2017 IEEE International Conference on Acoustics, Speech and Signal Processing*. New Orleans: IEEE, 2017: 246-250.

[3] YU D, KOLBÆK M, TAN Z H, et al. Permutation invariant training of deep models for speaker-independent multi-talker speech separation [C]//*2017 IEEE International Conference on Acoustics, Speech and Signal Processing*. New Orleans: IEEE, 2017: 241-245.

[4] KOLBÆK M, YU D, TAN Z H, et al. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks [J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017, **25**(10): 1901-1913.

[5] LUO Y, MESGARANI N. TaSNet: Time-domain audio separation network for real-time, single-channel speech separation [C]//*2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. Calgary: IEEE, 2018: 696-700.

[6] LUO Y, MESGARANI N. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation [J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019, **27**(8): 1256-1266.

[7] LUO Y, CHEN Z, YOSHIOKA T. Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation [C]//*2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. Barcelona: IEEE, 2020: 46-50.

[8] GE M, XU C L, WANG L B, et al. Multi-stage speaker extraction with utterance and frame-level reference signals [C]//*2021 IEEE International Conference on Acoustics, Speech and Signal Processing*. Toronto: IEEE, 2021: 6109-6113.

[9] DELCROIX M, ZMOLIKOVA K, OCHIAI T, et al. Speaker activity driven neural speech extraction [C]//*2021 IEEE International Conference on Acoustics, Speech and Signal Processing*. Toronto: IEEE, 2021: 6099-6103.

[10] WANG Q, MUCKENHIRN H, WILSON K, et al. VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking [C]//*Interspeech 2019*. ISCA: Graz, 2019: 2728-2732.

[11] LI T L, LIN Q J, BAO Y Y, et al. Atss-net: Target speaker separation via attention-based neural network [C]//*Interspeech 2020*. Shanghai: ISCA, 2020: 1411-1415.

[12] CHEN J, RAO W, WANG Z L, et al. MC-SpEx: Towards effective speaker extraction with multi-scale interfusion and conditional speaker modulation [C]//*Interspeech 2023*. Dublin: ISCA, 2023: 4034-4038.

[13] WANG Q, DOWNEY C, WAN L, et al. Speaker diarization with LSTM [C]//*2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. Calgary: IEEE, 2018: 5239-5243.

[14] WANG W Q, QIN X Y, LI M. Cross-channel attention-based target speaker voice activity detection: Experimental results for the M2met challenge [C]//*2022 IEEE International Conference on Acoustics, Speech and Signal Processing*. Singapore: IEEE, 2022: 9171-9175.

[15] YU F, ZHANG S, FU Y, et al. M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge [C]// *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*. Singapore: IEEE, 2022: 6167-6171.

[16] DING S J, WANG Q, CHANG S Y, et al. Personal VAD: Speaker-conditioned voice activity detection [C]//*The Speaker and Language Recognition Workshop* (*Odyssey 2020*). Tokyo: ISCA, 2020: 433-439.

[17] GE M, XU C L, WANG L B, et al. SpEx+: A complete time domain speaker extraction network [C]//*Interspeech 2020*. Shanghai: ISCA, 2020: 1406-1410.

[18] WANG W Q, LI M, LIN Q J. Online target speaker voice activity detection for speaker diarization [C]//*Interspeech 2022*. Incheon: ISCA, 2022: 1441-1445.

[19] LIN Q J, YIN R Q, LI M, et al. LSTM based similarity measurement with spectral clustering for speaker diarization [C]//*Interspeech 2019*. Graz: ISCA, 2019: 366-370.

[20] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//*2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas: IEEE, 2016: 770-778.

[21] DENG J K, GUO J, XUE N N, et al. ArcFace: Additive angular margin loss for deep face recognition [C]//*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 4685-4694.

[22] COSENTINO J, PARIENTE M, CORNELL S, et al. LibriMix: An open-source dataset for generalizable speech separation [DB/OL]. (2020-05-22) [2023-12-19]. http://arxiv.org/abs/2005.11262

[23] LE ROUX J, WISDOM S, ERDOGAN H, et al. SDR–half-baked or well done? [C]//*2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. Brighton: IEEE, 2019: 626-630.

[24] WANG W Q, CAI D W, LIN Q J, et al. The DKU-DukeECE-lenovo system for the diarization task of the 2021 VoxCeleb speaker recognition challenge [DB/OL]. (2021-09-05) [2023-12-19]. http://arxiv.org/abs/2109.02002

[25] YU F, DU Z H, ZHANG S L, et al. A comparative study on speaker-attributed automatic speech recognition in multi-party meetings [C]//*Interspeech 2022*. Incheon: ISCA, 2022: 560-564.

[26] DELCROIX M, ZMOLIKOVA K, KINOSHITA K, et al. Single channel target speaker extraction and recognition with speaker beam [C]//*2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. Calgary: IEEE, 2018: 5554-5558.

[27] XU C L, RAO W, CHNG E S, et al. Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss [C]//*2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. Brighton: IEEE, 2019: 6990-6994.

[28] XU C L, RAO W, CHNG E S, et al. Time-domain speaker extraction network [C]//*2019 IEEE Automatic Speech Recognition and Understanding Workshop*. Singapore: IEEE, 2019: 327-334.

[29] XU C L, RAO W, CHNG E S, et al. SpEx: Multi-scale time domain speaker extraction network [J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020, **28**: 1370-1384.

[30] YAO Z Y, WU D, WANG X, et al. WeNet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit [C]//*Interspeech 2021*. Brno: ISCA, 2021: 4054-4058.

[31] ZHANG B B, LV H, GUO P C, et al. WENETSPEECH: A 10000 hours multi-domain mandarin corpus for speech recognition [C]//*2022 IEEE International Conference on Acoustics, Speech and Signal Processing*. Singapore: IEEE, 2022: 6182-6186.