A Target Speaker Separation Neural Network with Joint-Training

Wenjing Yang[*†], Jing Wang[*], Hongfeng Li[†], Na Xu[†], Fei Xiang[†], Kai Qian[*], and Shenghua Hu[*†]
[*]Beijing Institute of Technology , Beijing, China
E-mail: ywj@bit.edu.cn, wangjing@bit.edu.cn
[†]Xiaomi, Beijing, China

***Abstract*— Target speaker separation aims to separate a target speech from multiple interference voices, which is promising for solving conventional difficulties in speech separation, such as arbitrary source permutation and unknown number of sources, and is useful for personal applications, like online meeting and personal phone calls. Recently, the application of deep-learning based models provided more alternatives for target speaker separation tasks. In this paper, we proposed a target speaker separation neural network with joint-training that separates the target voice in the spectrogram domain with the proposed combinative loss function. Experimental results show that compared with the baseline, our proposed method yields better performance on both test data and real data. Meanwhile, the proposed combinative loss function is more effective in addressing this issue.**

I. INTRODUCTION

Our human auditory system has the ability for separating mixed signals and speech separation aims at separating speech of different speakers from a speech mixture, which is often referred to the cocktail-party problem [1]. Speech separation is meaningful for auditory applications like online meeting and automatic speech recognition (ASR) [2]. There are some traditional multichannel signal processing methods like beamforming [3] and independent component analysis (ICA) [4]. Since the performance of these methods is directly related to the number of microphones, a large number of microphones are needed to achieve good performance, making it ineffective on monaural devices.

Recently, with the development of machine learning, many methods based on neural network are proposed and perform better than conventional speech separation algorithms, especially for single-channel speaker separation [5]. According to the input of neural network, deep-learning based speech separation networks can be divided into two main streams, the deep embedding networks and the raw wav networks. The deep embedding networks take speaker feature extraction as input and implement separation by computing a mask for each source, typical examples are deep clustering [6], permutation invariant training (PIT) [7], utterance-level permutation invariant training (uPIT) [8] and Deep Attractor Network (DANet) [9]. Most raw wav networks for speech separation come after the TasNet [10], these networks explore the application of speech separation in time-domain like Conv-TasNet [11]. However sometimes the separation of all speakers may cause long computing period and unnecessary waste of resources, especially when only the target speaker is useful such as teleconference and personal phone calls.

In this paper, we focus on target speaker separation, which is one of the methods that address the problems mentioned above. Given a reference utterance of the target speaker and a mixed utterance containing the target speaker and interference, the target speaker separation system aims at filtering out the target speech from the speech mixture. We refer to the baseline VoiceFilter [12] system proposed by Quan Wang, a deep embedding system that achieved target speaker separation by separately training two neural networks: (1) A speaker registration model that extracts deep embeddings for target speaker; (2) A speaker separation model that takes both mix spectrogram and target embedding as input, and produces a mask for target speech. However, the separate training of these two models may be not effective enough when there is a domain gap for application and to retrain two models is a time-consuming job. Additionally, we think that the mean square error (MSE) loss computed in the spectrogram domain used in [12], which only pays attention to the absolute value of magnitude, may not be the best loss function for target speaker separation.

To address the problems above, we proposed a target speaker separation neural network with joint-training, making the system automatically fit in the current speakers that need to be extracted. Additionally, we also proposed a combinative loss in our system which consists of relative mean square error (RMSE) loss [13] and scale invariant signal-to-distortion (SI-SNR) loss [14] to replace MSE loss used in conventional deep learning based target speaker separation system. Our experiments showed that compared with MSE loss the proposed combinative loss is more effective for addressing target speaker separation tasks and our proposed method yields better performance than the baseline on Librispeech [15] and real data at the same time.

II. TARGET SPEAKER SEPARATION SYSTEM

The target speaker separation system consists of two parts: (1) speaker registration system, which extracts the deep embedding of the target speaker using a reference speech signal in the enrollment stage; (2) speaker separation system, where

we take the target embedding and the mix magnitude spectrogram as input to compute a mask for generating the target speech. In most deep-learning based target speaker separation project like [12,16,17,18], these two systems are trained separately. The flow chart of these systems is shown in Figure 1.

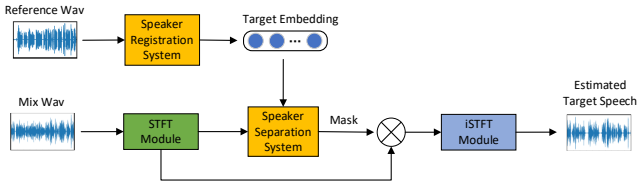

Fig. 1 The flow chart of classical target speaker separation system.

A. *Speaker Registration System*

The speaker registration system works in the enrollment stage to extract the target embedding from the reference wav. Deep-learning based registration system architectures include TDNN [19] based models and its deeper counterparts [20], as well as network architectures conventionally used in the computer vision such as the ResNet [21]. Additionally, a range of pooling methods have been proposed to aggregate frame-level features into utterance-level embeddings, like simple average pooling [22], statistical pooling [23] and dictionary-based encodings [24] usually the output target embedding is described as $e_{target} \in R^{1\times D}$, where D is the dimension of the extracted embedding.

B. *Speaker Separation System*

The speaker separation system is a neural network which takes the speech mixture and the target embedding as input and performs speaker separation by computing a time-frequency domain mask of the target speech. Take n speakers for example, assume that $s_1(t),\ldots,s_n(t)$ are the sources, where $s_i \in R^{1\times T}$ and T is the duration of the utterance, the speech mixture could be described as

$$x(t)=\sum_{i=1}^{n} s_i(t), \tag{1}$$

and after the short-time Fourier transform (STFT) the sum of these source spectrums $S_i(k,f)$ would be $X(k,f) \in R^{K\times F}$, where k and f are the indexes corresponding to the time frame and the frequency bin respectively.

$$X(k,f)=\sum_{i=1}^{n} S_i(k,f) \tag{2}$$

Then we extend the target speaker embedding e_j $(1\le j\le n)$ for K times to get the extended embedding $E_j \in R^{K\times D}$ before inputting the embedding into the separation system. Then we input the magnitude spectrum $|X(k,f)|$ and the extended embedding E_j together into the model to compute the T-F mask M_j :

$$M_j(k,f)=h(|X(k,f)|,E_j), \tag{3}$$

where h is the speaker separation system. By calculating the element-wise product between the mixed spectrogram and the output T-F mask we can get the estimated target magnitude spectrum $|\hat{X}(k,f)|$

$$|\hat{X}(t,f)|=|X(t,f)|\odot M_j(k,f), \tag{4}$$

where $\odot$ denotes the element-wise product operation.

To reconstruct the estimated speech in time domain, the phase θ of the mixture is needed when performing inverse short-time Fourier transform (ISTFT):

$$\hat{s}_j(t)=ISTFT(|\hat{X}(k,f)|\times e^{\angle\theta}) \tag{5}$$

III. The Proposed Method

In most of the target speaker separation projects, only the speaker separation model is trained in the training phase, keeping the pre-trained speaker registration model unchanged throughout both train and test stages. Motivated by jointly-train speech separation methods like [27], we proposed a target speaker separation neural network with joint-training in which two main models working for target embedding extraction and target speaker separation respectively are trained as a whole at the same time, making the target speaker separation system available to adapt to the present input automatically.

The architecture of our proposed target speaker separation neural network with joint-training is shown in Figure 2. Moreover, we equipped our system with the proposed combinative loss function which has been proved to be more effective on target speaker separation tasks according to our experiments.

A. *The Proposed System*

In the speaker registration system of the proposed method, a text-independent speaker registration model serves as a feature extraction that generates a target embedding before speaker separation. By using a reference utterance of the target speaker in the enrollment stage, we extract 40-dimensional log-Mel Spectrum features with 25 ms frame length and 10 ms frameshift where non-speech part of the utterance is removed by an energy based voice activity detection (VAD). After that we input these features to the speaker registration model and get the target speaker embedding $e_{target} \in R^{1\times 256}$.

Referring to the framework in [28], our training model for speaker registration contains a feature extraction module and a classifier set with ge2e [29] loss function. As to the feature extractor, we use Resnet34 [30] with SE-block [31] and the ASP [24] layer is used for generating frame-level weights by

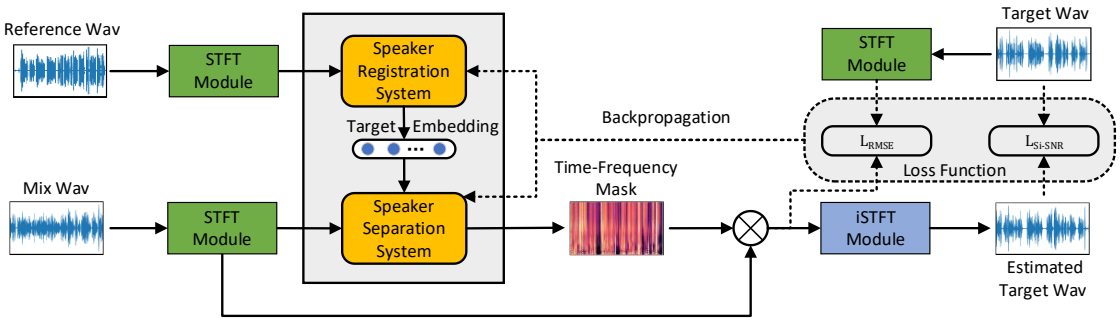

Fig. 2 The architecture of our proposed system. Reference wav and target wav are two different utterance of the target speaker that used for registration and mixture generation respectively. The loss function is only computed during training.

attention mechanism to get a weighted utterance-level target embedding e_{target}. Note that the speaker registration model is trained in advance and a pre-trained model is loaded when training the whole system.

At the front and back end of the separation system, 2 fully connected layers are set the separation model respectively as the input and output transform layer for dimension transformation. The inner network consists of 8 convolution layers, 1 Bi-LSTM layer and 2 fully connected layers. Each layer is equipped with ReLU activations except that the last layer is equipped with a sigmoid activation. To get comparable results, we set the parameters of the separation network the same as VoiceFilter [12]. The architecture of the separation system is shown in Figure.3.

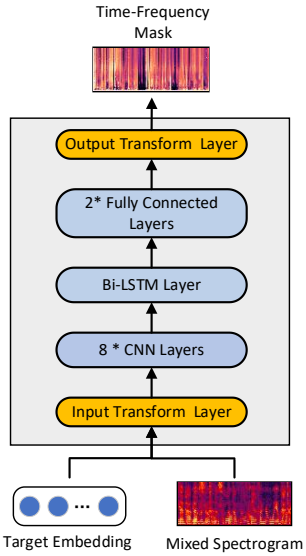

Fig. 3 Model architecture of the separation system.

B. *Combinative Loss Function*

Although MSE is commonly used in separation task as loss function and has been proved to be efficient, we think that it may not be the best loss function for target speaker separation as it only concerns about the absolute value of magnitude in the spectrogram, making the MSE susceptible in computation. In our experiments, we explore the possibility of improving the performance by replacing the MSE with the recently proposed RMSE [13].

Let y be the ground truth and $\hat{y}$ be the estimated value. MSE is defined as

$$L_{MSE} = (y - \hat{y})^2, \tag{6}$$

which is actually an absolute value depending on the distance between y and $\hat{y}$, while RMSE introduces a relative factor to overcome this deficiency, whose expression is

$$L_{RMSE} = \left(\frac{y - \hat{y}}{|y| + |\hat{y}| + \varepsilon}\right)^2, \tag{7}$$

where ε is a positive constant term to prevent division by small values which is set to 0.1 in our experiments. Therefore, the RMSE is more sensitive to the points in spectrogram with smaller values and our experimental results showed that RMSE is more effective than MSE as loss function for target speaker separation tasks.

We further improved the performance by introducing a time domain loss function Si-SNR loss [14] to get the proposed combinative loss, making a more comprehensive evaluation in spectrogram and time domain. So that the combinative loss used in our neural network could be described as

$$L = \alpha L_{RMSE} + \beta L_{Si-SNR}, \tag{8}$$

where α and β are the weights of these two loss functions and we set α=β=0.5 in the experiments which is adjustable in further exploration.

IV. Experiment Setup

In this section, we describe the setup of our experiments including the data used to train the registration system and the

separation model as well as the metrics for accessing the whole target speaker separation system.

A. Pre-Trained Data Description

In the experiments, the data used for training the pre-trained speaker registration system are selected from LibriSpeech [15], aishell2 [1], vox-celeb1, vox-celeb2 [33], and ST-CMDS .There are 10920 train speakers in total while the test data is vox1-test which contains 39 speakers.

The utterances are normalized and augmented with random noise and reverberation. At the classifier, the ge2e loss is calculated for backpropagation. We finally get a 3.41% equal error rate (EER) on the test dataset.

B. Separation System Data Description

The LibriSpeech dataset is used for the separation task in the experiments like [12] did where the training set contains 2338 speakers while both the test set and validation set contain 73 speakers. Besides, we additionally applied real data recorded at the anechoic lab which consists 6 female speakers and 10 male speakers with 18 utterances per speaker in the test stage as a challenging situation. All data are trimmed and reconnected to make sure that each utterance is longer than 10s and the utterances used for registration and separation are 10s and 3s respectively.

The mixtures are generated on-the-fly with 2 steps: (1) randomly select a speaker as target speaker from the speaker list and select a registration utterance and a train utterance of the target speaker; (2) randomly select another speaker as interference speaker with an utterance as interference utterance and mix it with the train utterance using randomly sampled signal-to-interference ratio (SIR) in {-5dB, 0dB, 5dB, 10dB}.

C. Training Setup

In the training phase, since the sampling rate of all data is 16kHz, we set the frame length and hop length of STFT to 400 and 160 respectively and a 512-point Fast Fourier Transform (FFT) is performed. All the networks in our experiments are trained on PyTorch 1.7.1 using the Adam optimizer with a learning rate of 0.001. The batch size for train is 16 and we train the models for 150 epochs with 48000 train records per-epoch to get convergence models.

D. Evaluation Metrics

To evaluate the performance of different systems, we use 2 evaluation metrics: segment signal-to-noise ratio (SSNR) and perceptual evaluation of speech quality (PESQ), and for both evaluation metrics the higher number indicates the better performance.

V. Result

In our experiments, we build the baseline [12] and first compare the performance of the baseline model with different loss functions on Librispeech with 128 test records. As shown in Table.1, the initial SNR and PESQ are 1.61 and 1.26 respectively in our experiments and it can be seen that all modifications have positive effects on the baseline while combinative loss got the best performance (12.17 dB) which improved the SSNR by 10.56 dB, nearly 19% higher than the baseline model which improved by 8.92 dB.

Furthermore, as shown in Table.2 and Table.3, our proposed method got better performance on both Librispeech and real data than baseline. Our proposed method improved SSNR by 11.27 dB and 7.68 dB respectively while the baseline model VoiceFilter improved SSNR by 8.92 dB and 3.71 dB respectively. Therefore, our proposed target speaker separation system performs much better than the baseline even in challenging situations.

Table 1: *The SSNR and PESQ performance on Librispeech under different loss function modifies on baseline.*

Loss Function	SSNR	PESQ
MSE	10.53	2.06
RMSE	11.28	2.22
Si-SNR	11.81	2.30
Combinative Loss	**12.17**	**2.44**

Table 2: *The SSNR and PESQ performance on Librispeech.*

Method	SSNR		PESQ	
	Before	After	Before	After
VoiceFilter	1.61	10.53	1.26	2.06
Our Method	1.61	**12.88**	1.26	**2.53**

Table 3: *The SSNR and PESQ performance on real data*.

Method	SSNR		PESQ	
	Before	After	Before	After
VoiceFilter	2.21	5.92	1.25	1.53
Our Method	2.21	**9.89**	1.25	**2.06**

VI. Conclusion

In this paper, we propose a target speaker separation neural network with joint-training which achieves better SSNR and PESQ performance compared with the baseline. Besides, we conduct a set of experiments to prove that our proposed combinative loss is more effective than single use of MSE, RMSE and Si-SNR loss. Our future work will be towards more effective loss functions and robust target speaker separation with joint-training.

Acknowledgment

This work was supported by National Natural Science Foundation of China (Grant No. 62071039 and No. 6162016002).

References

[1] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” The Journal of the Acoustical Society of America, vol. 25, no. 5, pp. 975–979, 1953.

[2] Yu Dong and Li Deng. AUTOMATIC SPEECH RECOGNITION. Springer London limited, 2016.

[3] Gannot, Sharon, et al. "A consolidated perspective on multimicrophone speech enhancement and source separation." IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.4 (2017): 692-730.
[4] Araki, Shoko, et al. "Underdetermined blind speech separation with directivity pattern based continuous mask and ICA." 2004 12th European Signal Processing Conference. IEEE, 2004.
[5] Wang, DeLiang, and Jitong Chen. "Supervised speech separation based on deep learning: An overview." IEEE/ACM Transactions on Audio, Speech, and Language Processing 26.10 (2018): 1702-1726.
[6] Hershey, John R., et al. "Deep clustering: Discriminative embeddings for segmentation and separation." 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.
[7] Yu, Dong, et al. "Permutation invariant training of deep models for speaker-independent multi-talker speech separation." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.
[8] Kolbæk, Morten, et al. "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks." IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.10 (2017): 1901-1913.
[9] Chen, Zhuo, Yi Luo, and Nima Mesgarani. "Deep attractor network for single-microphone speaker separation." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.
[10] Luo, Yi, and Nima Mesgarani. "Tasnet: time-domain audio separation network for real-time, single-channel speech separation." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.
[11] Luo, Yi, and Nima Mesgarani. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation." IEEE/ACM transactions on audio, speech, and language processing 27.8 (2019): 1256-1266.
[12] Wang, Quan, et al. "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking." Interspeech. 2019.
[13] Li, Hongfeng, et al. "Improving speech enhancement by focusing on smaller values using relative loss." IET Signal Processing 14.6 (2020): 374-384.
[14] Bahmaninezhad, Fahimeh, et al. "A comprehensive study of speech separation: spectrogram vs waveform separation." Interspeech. 2019.
[15] Panayotov, Vassil, et al. "Librispeech: an asr corpus based on public domain audio books." 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015.
[16] Wang, Quan, et al. "VoiceFilter-Lite: Streaming Targeted Voice Separation for On-Device Speech Recognition." Interspeech. 2020.
[17] Du, Jun, et al. "Speech separation of a target speaker based on deep neural networks." 2014 12th International Conference on Signal Processing (ICSP). IEEE, 2014.
[18] Wang, Jun, et al. "Deep extractor network for target speaker recovery from single channel speech mixtures." Interspeech. 2018.
[19] Snyder, David, et al. "X-vectors: Robust dnn embeddings for speaker recognition." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.
[20] Snyder, David, et al. "Speaker recognition for multi-speaker conversations using x-vectors." 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.
[21] Zeinali, Hossein, et al. "But system description to voxceleb speaker recognition challenge 2019." arXiv preprint arXiv:1910.12592 (2019).
[22] Nagrani, Arsha, Joon Son Chung, and Andrew Zisserman. "Voxceleb: a large-scale speaker identification dataset." Interspeech. 2017.
[23] Okabe, Koji, Takafumi Koshinaka, and Koichi Shinoda. "Attentive statistics pooling for deep speaker embedding." Interspeech. 2018.
[24] Xie, Weidi, et al. "Utterance-level aggregation for speaker recognition in the wild." ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.
[25] Xuan Ji et al. " Speaker-aware target speaker enhancement by jointly learning with speaker embedding extraction." 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
[26] Shulin He et al., "Speakerfilter: Deep learning-based target speaker extraction using anchor speech." 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
[27] Zeghidour, Neil, and David Grangier. "Wavesplit: End-to-end speech separation by speaker clustering." arXiv preprint arXiv:2002.08933 (2020).
[28] Chung, Joon Son, et al. "In defence of metric learning for speaker recognition." Interspeech. 2020.
[29] Wan, Li, et al. "Generalized end-to-end loss for speaker verification." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.
[30] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[31] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[32] Du, Jiayu, et al. "Aishell-2: Transforming mandarin asr research into industrial scale." arXiv preprint arXiv:1808.10583 (2018).
[33] Chung, Joon Son, Arsha Nagrani, and Andrew Zisserman. "Voxceleb2: Deep speaker recognition." Interspeech. 2018.