Speaker-Aware Monaural Speech Separation

Jiahao Xu[1]*, Kun Hu*[1]*, Chang Xu*[1]*, Duc Chung Tran*[2]*, Zhiyong Wang*[1]

[1]School of Computer Science The University of Sydney Sydney, NSW, Australia
[2]Computing Fundamental Department, FPT University, Hoa Lac Hi-Tech Park, Hanoi, Vietnam

{jixu7952,kuhu6123}@uni.sydney.edu.au, c.xu@sydney.edu.au,
chungtd6@fe.edu.vn, zhiyong.wang@sydney.edu.au

Abstract

Predicting and applying Time-Frequency (T-F) masks on mixture signals have been successfully utilized for speech separation. However, existing studies have not well utilized the identity context of a speaker for the inference of masks. In this paper, we propose a novel speaker-aware monaural speech separation model. We firstly devise an encoder to disentangle speaker identity information with the supervision from the auxiliary speaker verification task. Then, we develop a spectrogram masking network to predict speaker masks, which would be applied to the mixture signal for the reconstruction of source signals. Experimental results on two WSJ0 mixed datasets demonstrate that our proposed model outperforms existing models in different separation scenarios.

Index Terms: Speech separation, disentangled representations, speaker identity, Time-Frequency mask.

1. Introduction

Effective speech separation has been a critical prerequisite for robust performance of many speech processing tasks, especially in real-world environments. A typical example is multi-speaker speech recognition under noisy settings, which would depend on the outcome of separating individual speakers from a mixture speech signal [1]. Speech separation is also widely known as the "*cocktail party problem*", which has always been challenging for researchers aiming to reproduce the mechanism of how the human hearing system tackles the problem [2]. Although some efforts have been made to solve this cocktail party problem by utilizing information from multi-channels [3, 4] or other modalities [5] which may not be always available, many studies focus solely on monaural mixture signals.

Many approaches have been developed to tackle monaural speech separation, such as computational auditory scene analysis (CASA) [6, 7], non-negative matrix factorization (NMF) [8, 9, 10] and improved loss functions[11, 12]. Due to the great success of deep learning techniques in recent years, many deep learning algorithms have been proposed to address the speech separation problem, most of which were devised to work in the frequency domain [13, 14, 15, 16, 17, 18, 19], which generate Time-Frequency (T-F) masks to separate the voices of different speakers. In [13, 15, 16], a deep clustering network was utilized to project the T-F bins of the mixture signals into high-dimensional spaces for separation. Deep attractor network (DANet) [14] was proposed to utilize attractor points to associate T-F units to individual speakers. To address the mismatch problem in DANet, an improved version named Anchored DANet (ADANet) was further proposed [20], which introduces referencing anchor points in embedding space to enable direct mask generation during training and testing phases. Alternative loss functions were also proposed for deep clustering network to achieve better separation performance [16]. More recently, several methods were proposed to take waveform data directly as input, instead of involving the conventional T-F transformation. For example, TasNet [21] and its convolutional variant Conv-TasNet [22] were two pioneering studies undertaking speech separation in the time domain, which achieved comparable performance to those T-F transformation based approaches. Similarly, Wave-U-Net [23] employed 1D Convolution on waveform data while incorporating a decoder for the reconstruction of separated signals. Dilated convolutions on both temporal and frequency domains with Gated Residual Network (GRN) were also investigated for speech separation [24]. Nevertheless, convolutional operations on raw waveform data would lead to significantly higher computational complexity than conventional T-F transformations.

Despite the considerable efforts and promising progress in the field, the existing studies have not paid adequate attention to the speaker identity information in the separation process. Early studies on utilizing Speaker Identity (SID) information mainly focused on extracting or filtering the speech signal of a target speaker from a mixture speech signal [25, 26, 27, 28]. SpeakerBeam [26] was proposed to include target speaker adaptive sub-layers in a context-adaptive deep neural network framework. VoiceFilter [29] coordinated convolutional spectral features with speaker identity embedding to predict the ideal mask for extracting the speech by a target speaker. Similarly, ASENet [27] was proposed as a joint framework for speech separation and extraction, combining mixture embedding vectors and speaker identity embedding with attention mechanism for the speech extraction of a target speaker. Nevertheless, speech extraction models cannot be directly utilized to solve the multi-speaker separation problem. In addition, these methods would require some pre-enrolled recordings of target speakers or a speaker inventory [28], which contains identity information of all the speakers to be targeted. Particularly, identity information of target speakers may not be available in advance in many real-world scenarios such as meetings, which further limit the applications of these methods. Note that these methods could only extract one target speaker at a time, the computational cost would be proportional to the number of speakers.

Inspired by the success of speaker-specific speech extraction, in this paper, we propose a novel speaker-aware monaural speech separation model by utilizing a mask inferring neural network with the help of speaker identity information. Specifically, the proposed model shown in Figure 1 aims to disentangle the identity information of different speakers using an LSTM encoder under the supervision of a speaker identity verification sub-task, and speaker-discriminative features will be used together with the original mixture spectrogram in the masking network for mask prediction. We utilize the Permutation Invariant Training (PIT) [30] technique to tackle the label permutation

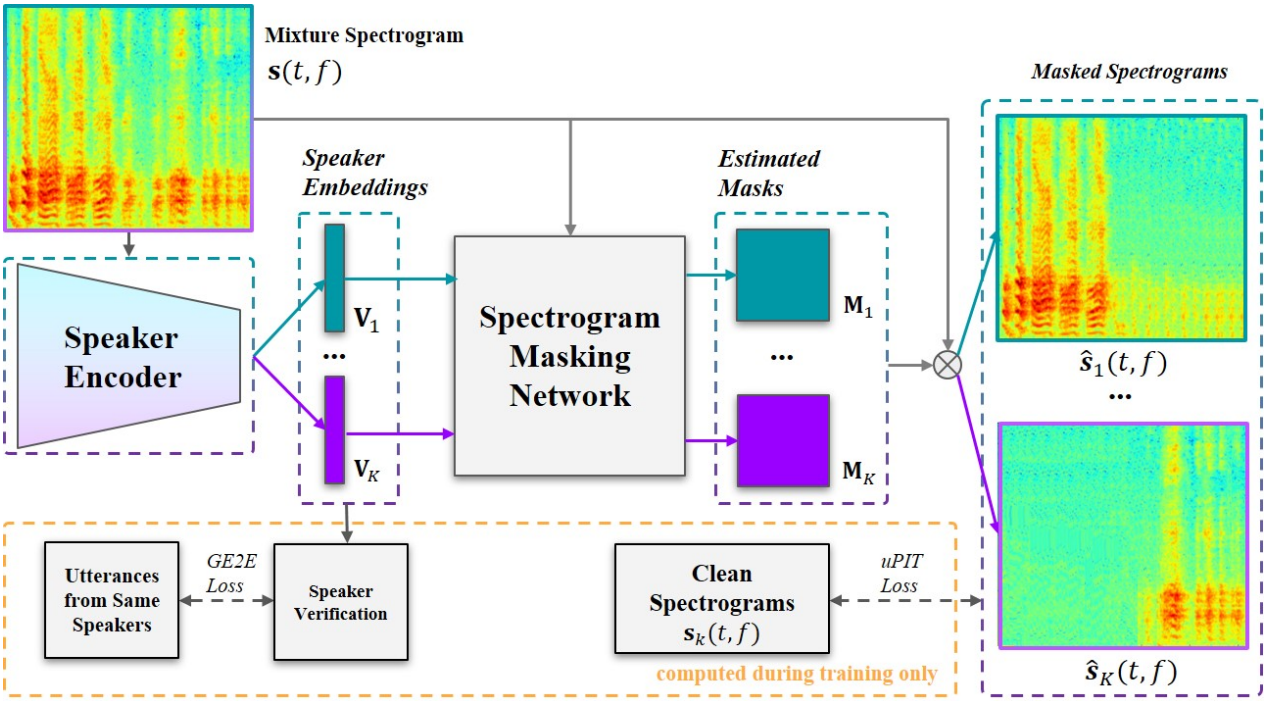

Figure 1: *Illustration of the proposed method, which includes the disentangling encoder and spectrogram masking network.*

problem when separating multiple speakers. As a result, our proposed method can deal with an arbitrary number of speakers without changing the network structure. We have demonstrated the effectiveness of the proposed method on various mixture signals containing different numbers of speakers with superior performance over the existing methods.

The rest of paper is organized as follows. Section 2 describes the proposed speech separation model in detail. The datasets, experimental settings and results are presented in Section 3, followed by the conclusion in Section 4.

2. Proposed Method

In this section, details of our proposed method are explained. Firstly, the preliminary of single channel speech separation is revisited. Secondly, the proposed method including the encoding and mask inferring network is introduced. Lastly, the training objective and label permutation training scheme for the proposed method is elaborated.

2.1. Monaural speech separation

The aim of monaural source separation is to reconstruct K different source signals $\{\mathbf{s}_k \in \mathbb{R}^T\}$, where $k \in \{1, 2, ...K\}$, from a single channel mixture signal $\mathbf{s} \in \mathbb{R}^T$. The mixture signal is assumed to be constructed as:

$$\mathbf{s} = \sum_{k=1}^{K} w_k \mathbf{s}_k, \tag{1}$$

where w_k is a scaling factor representing the intensity of the k-th source. In this paper, the task is formulated as a supervised learning problem.

The objective is to build a model, for an unseen mixture signal $\mathbf{s}' = \sum_{k=1}^{K} w'_k \mathbf{s}'_k$, to accurately estimate the source signals $\mathbf{s}'_k$. In particular, denote the estimation of $\mathbf{s}'_k$ as $\hat{\mathbf{s}}'_k$. Note that the order of the source signals predicted could be arbitrary as the summation of source signals is order-free in our problem formulation.

By following the T-F masking separation approach, our proposed modeling can be conducted on the spectrograms $\mathbf{s}(t,f)$ and $\mathbf{s}_k(t,f)$ for each time-frame t and frequency bin f, which are obtained by the Short-Time Fourier Transformation (STFT) of $\mathbf{s}$ and its source $\mathbf{s}_k$, respectively. With the estimation $\hat{\mathbf{s}}_k(t,f)$ of the source signals, inverse Discrete Fourier Transform (iDFT) is adopted to construct the estimated time-domain analysis windows.

2.2. Speaker disentangling encoder

As shown in Figure 1, the speaker disentangling encoder is devised to generate speaker-discriminative representations from a given mixture signal $\mathbf{s}(t,f)$. The speaker disentangling encoder consists of multi-layer LSTM network supervised by the auxiliary speaker verification task. Separate K last layers of the LSTM network would be trained for the K speakers in the mixture. In detail, the encoder takes the mixture audio spectrogram $\mathbf{s}(t,f)$ as its input, and outputs the speaker embeddings for all speakers in the mixture as $\mathbf{V} \in \mathbb{R}^{D \times K}$, where D is the dimension of the speaker embeddings. For the k-th speaker in a mixture, the speak embedding would be further normalized with ℓ_2-norm of the LSTM output:

$$\mathbf{V}_k = \frac{g(\mathbf{s}(t,f))_k}{\|g(\mathbf{s}(t,f))_k\|_2} \tag{2}$$

where g indicates the LSTM network.

The number of speakers in the mixture, K, needs to be specified before the training. With the specified number of speakers, the encoder aims to produce separate representations of different speakers. We employ permutation invariant training technique to address the order-free property of speakers in

the mixture. As the encoder takes the number of speaker as a hyper-parameter, a set of isolated speaker encoders is required to be trained, of which each targets one specified number of speakers in the mixture signals. Note that the speaker verification would only be involved in the training stage, to ensure the discriminative power of the disentangled representations. For inference, similar to other related studies, the proposed method only requires the number of speakers in the mixture to choose the corresponding encoder.

During the training process, we use randomly selected utterances of the corresponding speakers from the pre-mixed corpus as enrollment utterances, and apply the GE2E loss [31] for speaker verification. The GE2E loss would guide the encoder to generate speaker embeddings closer to embeddings of same speaker reference utterances in the high-dimensional space. Readers can refer to [31] for more details of the GE2E loss we used for speaker verification.

2.3. Spectrogram masking network

The design of our masking network is inspired by a recent work named VoiceFilter [29], which extracts the speech signal of a target speaker. As illustrated in Figure 1, our spectrogram masking network would take both disentangled representations from the LSTM encoder and the mixture audio spectrogram as inputs. Then the masking network computes the soft masks for all speakers $\mathbf{M} = (\mathbf{M}_1, \mathbf{M}_2, ..., \mathbf{M}_K)$ as

$$\mathbf{M}_k = h(\mathbf{s}(t,f), \mathbf{V}_k) \tag{3}$$

where h is the mapping function of the masking network. By applying these masks in an element-wise manner over the mixture spectrogram, the estimated spectrogram for the k-th speaker can be formulated as:

$$\hat{\mathbf{s}}_k(t,f) = \mathbf{s}(t,f) \odot \mathbf{M}_k. \tag{4}$$

The training objective of the masking network is to minimize the divergence between the estimated spectrogram $\hat{\mathbf{s}}_k(t,f)$ and the ground truth spectrogram $\mathbf{s}_k(t,f)$.

2.4. Loss function

For mask inferring separation methods, the Mean Square Error (MSE) is a commonly used cost function, which measures the Euclidean difference between the ground truth and the estimated output. However, as Liu *et al.*investigated in [12], the divergence-based cost functions are more suitable for this type of speech separation models. Therefore, in this work, the Jensen-Shannon divergence is employed as the backbone of the loss function:

$$JS(x,y) = \frac{1}{2}(x \log \frac{2x}{x+y} + y \log \frac{y}{x+y}), \tag{5}$$

where x and y here would be the clean spectrograms $\mathbf{s}_k(t,f)$ and the masked spectrograms $\hat{\mathbf{s}}_k(t,f)$, respectively.

With the base loss function, we utilize the PIT technique to handle the label permutation problems because of its effectiveness in solving such problems. The intuition of PIT is very straightforward, which iterates all the possible permutations and takes the minimum loss as the final loss for training. By following this training strategy and combining with the JS divergence, a permutation invariant training loss for one single frame can be formulated as

$$\mathcal{J}^{PIT} = \min_{\theta \in \mathcal{P}} \sum_{k=1}^{K} JS(\mathbf{s}_k(t,f), \hat{\mathbf{s}}_k(t,f)), \tag{6}$$

where θ indicates a prediction-label permutation and $\mathcal{P}$ is all the possible permutations for a specific frame in the signal. PIT would process each frame and make the separation learning at frame-level. An improved version of PIT, namely uPIT, is utilized in this paper, which enables permutation invariant learning at the utterance-level. Readers interested in the details of utterance-level PIT can refer to [30] for more details.

3. Experimental Results and Discussions

3.1. Datasets

To evaluate our proposed model, we adopted two commonly used benchmark datasets, i.e. WSJ0-2mix and WSJ0-3mix [13]. Both datasets were derived from the WSJ0 corpus following the same procedure in the literature. The 40-hour (30 for training and 10 for validation) WSJ0-2mix mixture audio were generated by merging randomly selected utterances in Wall Street Journal (WSJ0) training set with different SNRs between 0 dB and 5 dB. The test set was generated by following the same approach but with a group of 18 unseen speakers. All the audio signals were re-sampled to 8 kHz for consistency and fair comparison with other studies. The creation of WSJ0-3mix dataset was similar to WSJ0-2mix, except that it worked on mixtures utterances from three speakers rather than two.

3.2. Experimental settings

A 3-layer LSTM network was trained as the speaker disentangling encoder, of which the input is log-mel spectrogram extracted from 2-second windows. We used sliding windows with 50% overlap and took the average of all windows encoder mapping as the output. The speaker discriminative embedding output was of 256 dimensions per speaker in the mixture.

The spectrogram masking network consisted of 11 layers: 8 convolutional layers followed by 1 LSTM layer and 2 fully connected layers at the end. The activation functions for the first 10 layers were ReLu, while the last layer used sigmoid activation. Disentangled speaker representations were concatenated to the convolutional features for each time frame, and fed altogether into the LSTM layer. Data augmentation was applied during the training by randomly shifting the signals to enlarge the training data, thus resulted in more robust separation performance.

3.3. Evaluation metrics

Since the distortion and noise in speech signals would significantly impact the perception of speeches, many studies have adopted distortion-related evaluation metrics such as signal-to-distortion ratio (SDR) and the scale-invariant SDR (SI-SDR) [13, 15, 14, 30, 32]. SDR indicates the ratio between clean signal energies distortion introduced by the output signal, for our case, which can be written as

$$\text{SDR}(\mathbf{s}_k, \hat{\mathbf{s}}_k) = 10 \log_{10} \frac{\|\mathbf{s}_k\|^2}{\|\hat{\mathbf{s}}_k - \mathbf{s}_k\|^2}. \tag{7}$$

SI-SNR is defined as

$$\begin{aligned} \mathbf{s}_k^{target} &= \frac{<\hat{\mathbf{s}}_k, \mathbf{s}_k>\mathbf{s}}{\|\mathbf{s}\|^2}, \\ \mathbf{e} &= \hat{\mathbf{s}}_k - \mathbf{s}_k^{target}, \\ \text{SI-SDR}(\mathbf{s}_k, \hat{\mathbf{s}}_k) &= 10 \log_{10} \frac{\|\mathbf{s}_k^{target}\|^2}{\|\mathbf{e}\|^2}. \end{aligned} \tag{8}$$

where $\mathbf{s}_k$ and $\hat{\mathbf{s}}_k$ in both equations indicate the clean signal and separated signal, respectively. The ℓ_2-norm of the signal is used

to measure the power of the signal. Both $\hat{s}$ and s need to be normalized to zero-mean and ensure scale-invariance. To evaluate the performance of our proposed model in different scenarios, we use the improvements on the Source to Distortion Ratio (SDRi) and the Scale-Invariant Source-to-Noise Ratio (SI-SNRi) as our evaluation metrics, which are commonly used in related studies.

3.4. Result Comparison

We compared the performance of our proposed methods to the existing models, which were based on T-F masking separation, including DPCL++ [15], ADANet [14], uPIT [30], OR-PIT [32] and some end-to-end models working on raw waveform data [21, 33].

Table 1: *SI-SDR and SDR improvements (dB) on WSJ0-2mix and WSJ0-3mix datasets*

	WSJ0-2mix		WSJ0-3mix	
Model	SI-SNRi	SDRi	SI-SNRi	SDRi
DPCL++ [15]	10.8	-	7.1	-
uPIT-BLSTM-ST [30]	-	10.1	-	7.8
DANet [14]	10.5		8.6	8.9
ADANet [20]	10.4	10.8	9.1	9.4
Conv-TasNet [22]	15.3	**15.6**	12.7	13.1
OR-PIT [32]	14.8	15.0	12.6	12.9
FurcaNeXt*[33]	**18.4**	-	N/A	N/A
Ours	15.2	15.4	**13.4**	**13.8**

* "N/A" means model not capable of 3 speaker mixture separation.

The experimental results of our proposed model are presented in Table 1, compared with the existing models. Although our model did not outperform the state-of-the-art models for 2-speaker mixtures, on WSJ0-3mix dataset our proposed model achieved the largest improvements on both metrics. Another advantage of our model is the ability to handle different numbers of speakers flexibly, whilst the state-of-the-art models such as FurcaNeXt [33] were only capable of 2 speaker mixture signals. In addition, when compared with other models which were capable of dealing with an arbitrary number of speakers, our proposed model is more robust to the extra speaker in the mixture (smaller decrease in the performance on WSJ0-3mix comparing with the performance on WSJ0-2mix).

Table 2: *Impacts of different disentangled speaker embedding size on speaker verification task and separation performance*

Metrics	EER		SDRi	
Embed	# of speakers			
Size	2	3	WSJ0-2mix	WSJ0-3mix
64	4.21	4.79	12.5	10.9
128	3.99	4.30	14.7	12.1
256	3.78	4.16	15.4	13.8

We also investigated the impacts of using different embedding dimensions for disentangled speaker representations generated by the encoder. We studied the equal error rate (EER) of the speaker verification task, which was used to train the speaker disentangling encoder. As shown in Table 2, an increase in the number of speakers in the mixture signals would make it more difficult for both the speaker verification and the speech separation tasks. Meanwhile, the higher dimensional embedding features generally led to more accurate verification and better separation results. These findings also demonstrated the effectiveness of involving speaker identity information during the separation process, from another perspective.

Table 3: *Comparison on model size with existing methods*

Model	# of parameters
DPCL++ [15]	13.6M
uPIT-BLSTM-ST [30]	92.7M
DANet [14]	9.1M
ADANet [20]	9.1M
TasNet [21]	23.6M
Conv-TasNet [22]	5.1M
FurcaNeX [33]	51.4M
Ours	**2.8M**

In addition to the separation performance, we also compared the complexity of the aforementioned models in terms of model sizes. Table 3 lists the number of parameters for a number of speech separation models. The models enlisted are almost the same group as in the overall performance studies, however some researchers did not provide their system implementation details, which made it not viable to estimate the sizes of their models. We can see from Table 3 that our model has the smallest number of parameters without compromising the separation performance. To be more specific, the number of parameters in our model was only about 5% of that in the FurcaNeXt [33] model. Compared with Conv-TasNet [22], our model achieved comparable separation performance on 2-speaker mixture and better results on 3-speaker mixture, with 45% less parameters in the model. The compact design of our proposed model benefit from utilizing the same parameters while recovering the soft masks for different speakers. Meanwhile, with assistance from the user identity information, our model can effectively separate source signals at a lower computational cost.

4. Conclusions and Future Work

In this paper, we presented a novel speaker-aware speech separation method which takes speaker identity information into account to tackle the multi-speaker speech separation problem. A novel neural network architecture with masking network was devised, which jointly disentangled speaker identity representations with a speaker encoder and estimated spectrogram masks for source signals. Experimental results demonstrated that, by incorporating speaker identity information in mask prediction, our proposed model achieved significant improvements in terms of both SDRi and SI-SNRi for multi-speaker monaural speech separation, especially considering the compact size of the model.

Despite the improvements on the monaural speech separation task, our proposed method could be improved by considering some further steps: (1) adopting recursive separation strategy to handle an unspecified number of speakers in the mixture [32]; and (2) incorporating phase information in the reconstruction of the signals to alleviate the phase incurred distortion. Another potential direction for future work would be to train the separation model jointly with automatic speech recognition (ASR) system in an end-to-end setting to supervise the separation with a human perception metric.

5. References

[1] Y. Qian, X. Chang, and D. Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *Speech Communication*, vol. 104, pp. 1–11, 2018.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multi-channel speech enhancement with variational autoencoders and non-negative matrix factorization," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 101–105.

[4] Z.-Q. Wang and D. Wang, "Integrating spectral and spatial features for multi-channel speaker separation," in *INTERSPEECH*, 2018, pp. 2718–2722.

[5] 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," *ACM Transactions on Graphics*, vol. 37, no. 4, 2018.

[6] K. Hu and D. Wang, "An unsupervised approach to cochannel speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 122–131, 2012.

[7] Y. Liu and D. Wang, "A casa approach to deep learning based speaker-independent co-channel speech separation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5399–5403.

[8] Z. Wang and F. Sha, "Discriminative non-negative matrix factorization for single-channel speech separation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3749–3753.

[9] F. Weninger, J. L. Roux, J. R. Hershey, and S. Watanabe, "Discriminative nmf and its application to single-channel source separation," in *Annual Conference of the International Speech Communication Association*, 2014.

[10] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep nmf for speech separation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 66–70.

[11] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[12] Y. Liu, H. Zhang, X. Zhang, and Y. Cao, "Investigation of cost function for supervised monaural speech separation," in *INTERSPEECH*, 2019, pp. 3178–3182.

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

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

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

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

[17] Z.-Q. Wang, K. Tan, and D. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 71–75.

[18] 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 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 86–90.

[19] F. Bahmaninezhad, J. Wu, R. Gu, S.-X. Zhang, Y. Xu, M. Yu, and D. Yu, "A comprehensive study of speech separation: spectrogram vs waveform separation," in *INTERSPEECH*, 2019, pp. 4574–4578.

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

[21] Y. Luo and N. Mesgarani, "TasNet: time-domain audio separation network for real-time, single-channel speech separation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[22] ——, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[23] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *International Society for Music Information Retrieval(ISMIR)*, 2018.

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

[25] J. Du, Y. Tu, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *International Conference on Signal Processing (ICSP)*. IEEE, 2014, pp. 473–477.

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

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

[28] P. Wang, Z. Chen, X. Xiao, Z. Meng, T. Yoshioka, T. Zhou, L. Lu, and J. Li, "Speech separation using speaker inventory," in *Automatic Speech Recognition and Understanding Workshop*. IEEE, December 2019.

[29] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *INTERSPEECH*, 2019, pp. 2728–2732.

[30] 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*, vol. 25, no. 10, pp. 1901–1913, 2017.

[31] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[32] N. Takahashi, S. Parthasaarathy, N. Goswami, and Y. Mitsufuji, "Recursive speech separation for unknown number of speakers," in *INTERSPEECH*, 2019, pp. 1348–1352.

[33] L. Zhang, Z. Shi, J. Han, A. Shi, and D. Ma, "Furcanext: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 653–665.