Interspeech 2018
2-6 September 2018, Hyderabad

Multi-talker Speech Separation Based on Permutation Invariant Training and Beamforming

Lu Yin[1,2], *Ziteng Wang*[1,2], *Risheng Xia*[1], *Junfeng Li*[1,2] *and Yonghong Yan*[1,2,3]

[1]Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Xinjiang Laboratory of Minority Speech and Language Information Processing, Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences

{yinlu, wangziteng, xiarisheng, lijunfeng, yanyonghong}@hccl.ioa.ac.cn

Abstract

The recently proposed Permutation Invariant Training (PIT) technique addresses the label permutation problem for multi-talker speech separation. It has shown to be effective for the single-channel separation case. In this paper, we propose to extend the PIT-based technique to the multichannel multi-talker speech separation scenario. PIT is used to train a neural network that outputs masks for each separate speaker which is followed by a Minimum Variance Distortionless Response (MVDR) beamformer. The beamformer utilizes the spatial information of different speakers and alleviates the performance degradation due to misaligned labels. Experimental results show that the proposed PIT-MVDR-based technique leads to higher Signal-to-Distortion Ratios (SDRs) compared to the single-channel speech separation method when tested on two-speaker and three-speaker mixtures.

Index Terms: multi-channel speech separation, beamforming, permutation invariant training, mask estimation

1. Introduction

Although the human auditory system can excellently perceive information in complex acoustic environments, the speech recognition performance of computer remains a challenging task when interferences and noises exist. The overlapping of speech utterances is one critical factor that impairs auditory comprehension of both human beings and machines. This issue is well known as the cocktail party problem [1] and its corresponding solution, speech separation, has been studied for many years. Various techniques have been proposed to address the speech separation task, such as Non-negative Matrix Factorization (NMF) [2, 3, 4, 5], Computational Auditory Scene Analysis (CASA) [6, 7, 8], Hidden Markov Model(HMM) [9, 10], and so on. In recent years, with the development and successful applications of deep neural network in many fields, neural network-based speech separation has attracted increasing attention.

In [11, 12, 13], the authors proposed to train a deep neural network to estimate time-frequency masks, which take either binary or ratio values calculated from the relative energy between the target source and noise. Then, the target source is obtained by multiplying the mask to the mixture signal in the frequency domain. However, only one target source is considered in these works. If there exist multiple target speakers in the signal, these methods tend to fail because of the *label permutation* or *label ambiguity* problem. One approach to deal with this problem is the speaker-aware method proposed in [14, 15], which informs the neural network using the features of the target speaker so that it learns to follow the speaker characteristics throughout the utterance. In the method, speaker-dependent features are extracted and concatenated with spectral features as inputs to the neural network.

In contrast to the particular speaker extraction work, the multi-talker separation task proposes to reconstruct all the speaker sources from the mixture. In [16, 17, 18], the neural networks are designed to output two masks, each corresponding to one target speaker. Then two target speeches can be separately obtained using the estimated masks. Nevertheless, these methods only work under particular conditions. [16, 18] only considers different gender of speakers and [17] assumes specific relative energy ratios of the mixed talkers. These constraints are helpful for the neural network to trace the target speakers, but the label permutation problem is not successfully solved.

To deal with the label ambiguity problem, two kinds of approaches were recently proposed: the deep clustering (DPCL) method [19, 20, 21] and the Permutation Invariant Training (PIT) method [22, 23]. The DPCL based approach first trains a deep recurrent neural network to map the mixture into an embedding space, where k-means clustering is used to assign the time-frequency bins to different speakers. This strategy assumes that only one source dominates each time-frequency bin; and the frequency bins belonging to same source are close to each other in the embedded space; and hence, by using a clustering algorithm, multi-talker speech can be separated. However, each time-frequency bin could simultaneously belong to different targets. Without making these assumptions, the PIT method is proposed to compute two losses by exchanging the target labels when calculating the training loss and uses just the lower one in the back propagation process. The multi-frame features are fed to the neural network and the outputs are multi-frame masks for each targets. Nevertheless, there is one more issue to be tackled, the consistency of masks over time for each speaker. This issue is addressed by using utterance level PIT (uPIT) in [23], and otherwise, a speaker tracing strategy is needed.

Although the above speech separation methods have shown promising results, they are all single-channel based and have inherent limitations compared with multichannel algorithms that can additionally utilize the spatial information of different speakers. There has been some pioneering work integrating single-channel neural networks and beamforming techniques, such as [24][25]. These methods originate from the idea of mask estimation network based beamforming [26][27], that has proved effective in speech enhancement and speech recognition tasks. The neural networks are trained to predict the mask of

10.21437/Interspeech.2018-1739

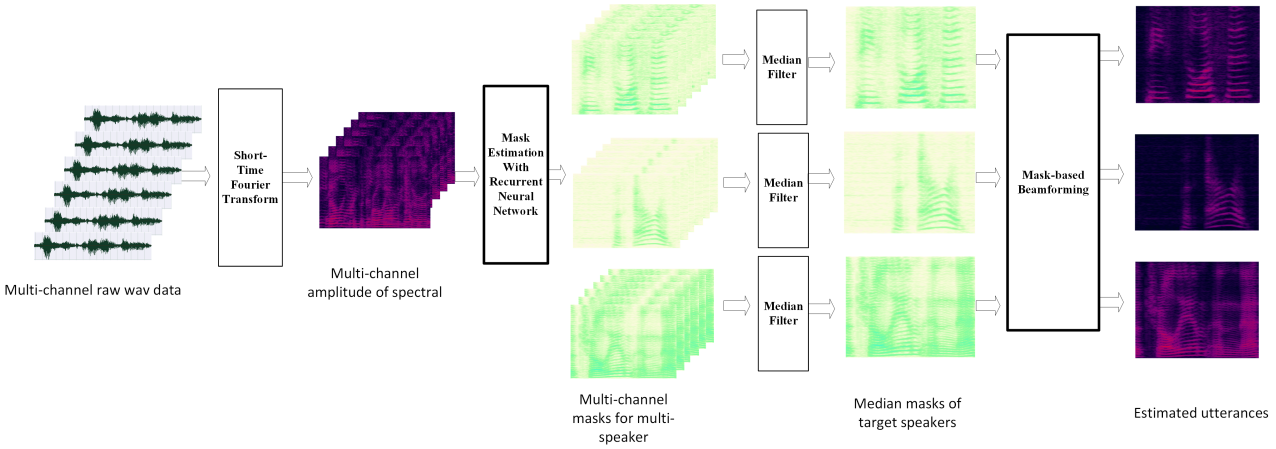

Figure 1: *Overall flowchart of proposed system.*

the target speech. Then a mask-based beamformer is used to estimate the expected signal. In this paper, we follow a similar routine and propose to tackle the multi-talker speech separation task using PIT and a minimum variance distortionless response (MVDR) beamformer. Specifically, utterance based PIT is employed to a single-channel speaker separation network, which provides separate masks for the following beamforming stage. The beamformer is based on spatial covariance matrixes that involve temporal expectation in the estimation process. This turns out helpful in alleviating the misaligned speaker masks. The PIT-MVDR based technique leads to higher signal-to-distortion ration (SDR) compared with the single-channel based separation algorithms.

2. Overview of proposed system

The overall flowchart of our proposed system is illustrated in Figure 1. Consider M-channel observations of S speakers, the systerm consists of four main components: the Short-Time Fourier Transform (STFT) component, the mask estimation component, the median filtering components and the beamforming components. Firstly, the M-channel raw wave data is transformed into time-frequency domain by STFT. Then, the magnitude spectrum is fed to the mask estimation component. The mask estimation component consists of M Recurrent Neural Networks (RNN), one RNN network for each wave channel. These RNN networks have shared weights and are trained to output masks for all the speakers. For each speaker, the RNN network output the M-channel masks. The median of the M-channel masks is taken and fed to the beamforming component. The beamforming component consists of S MVDR beamformers, each MVDR beamformer targeted for one speaker. Finally, the beamforming component outputs the estimated speech for all target speakers.

3. Permutation invariant training for speaker masks estimation

This section presents the permutation invariant trained neural network for multi-speaker mask estimation. The training is performed on one neural network and the networks of all channels share the same framework and weights. For comprehensiveness, we take the 2-speaker scenario as example. The 2-speaker separation model is shown in Figure 2. The components in the dash block is used only for training procedure, while the outside parts contribute for both training and inferring procedures. For training the networks, a phase-sensitive mask is utilized as labels with minimum cross loss function.

3.1. Phase-sensitive mask

For supervised speech separation, masking based targets performs better than spectral amplitude based targets in general [11]. Among these proposed masks, the phase-sensitive mask considers the phase difference between the expected signals and mixture and hence yields better results [28] in terms of signal-to-distortion ratio (SDR). The labels with phase-sensitive mask is defined as follows:

$$\mathcal{M}_{\text{psm},s}(t,f) = \frac{|X_s(t,f)|}{|Y(t,f)|} cos(\theta_y(t,f) - \theta_s(t,f)) \quad (1)$$

where t is the time index, f is the frequency index, and θ_y, θ_s are phases of the mixture $Y(t,f)$ and the target speech $X_s(t,f)$, respectively. Here, s is the index of speaker.

3.2. Minimum cross loss functtion

In contrast to speech enhancement task, one critical challenge for multi-speaker separation is the label ambiguity problem. To solve the problem, Minimum Cross Loss (MCL) function was adopted for neural network training. Minimum cross loss function training is also called permutation invariant training in [22]. By this training strategy, the loss function is selected from Normal Mean Square Errors (NMSE) or Cross Mean Square Errors (CMSE) between the estimated signals and labels. Let $loss_{\text{NMSE}}$ and $loss_{\text{CMSE}}$ denote NMSE and CMSE, respectively. Hence,

$$loss_{\text{NMSE}} = ||\widetilde{\mathcal{M}}_{\text{psm}_1} - \mathcal{M}_{\text{psm}_1}||^2 + ||\widetilde{\mathcal{M}}_{\text{psm}_2} - \mathcal{M}_{\text{psm}_2}||^2 \quad (2)$$

$$loss_{\text{CMSE}} = ||\widetilde{\mathcal{M}}_{\text{psm}_1} - \mathcal{M}_{\text{psm}_2}||^2 + ||\widetilde{\mathcal{M}}_{\text{psm}_2} - \mathcal{M}_{\text{psm}_1}||^2 \quad (3)$$

where $\widetilde{\mathcal{M}}_{\text{psm}_1}$ and $\widetilde{\mathcal{M}}_{\text{psm}_2}$ denote the estimated masks, and $\mathcal{M}_{\text{psm}_1}$ and $\mathcal{M}_{\text{psm}_2}$ denote the two speaker labels. The minimum loss function is defined as:

$$loss_{\text{MCL}} = \lambda \cdot loss_{\text{NMSE}} + (1-\lambda) \cdot loss_{\text{CMSE}} \quad (4)$$

where λ is a chosen determiner defined as:

$$\lambda = \begin{cases} 1, & loss_{\text{NMSE}} \leq loss_{\text{CMSE}} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

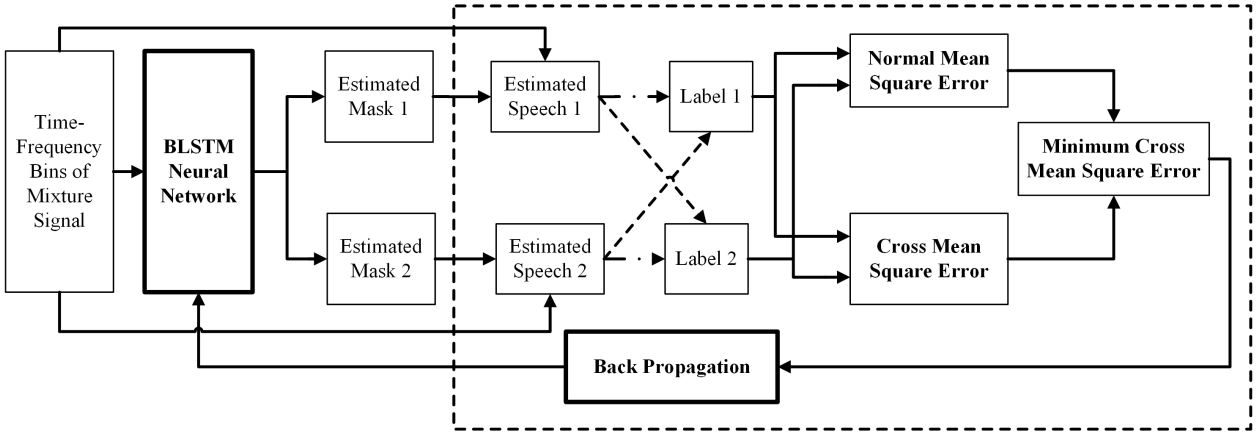

Figure 2: *The model for 2-speaker speech separation.*

λ ensures that using the lower one of $loss_{\text{NMSE}}$ and $loss_{\text{CMSE}}$ as the final training loss in the back propagation stage.

4. Speaker mask based beamforming

The estimated masks can be directly used to obtain the separated speakers by applying the mask to the mixture signal. However, given the limitation of single-channel based approach, we propose to construct mask-based MVDR beamformers for better speaker separation. Generally, beamforming techniques aim to design a complex-valued filter that extract the desired source and suppresses the interference signals and noises at the same time. The proposed MVDR only relies on the spatial covariance matrixes of the target source, interference and noise, that are computed from the estimated masks, and no longer depend on the geometry of the microphone array as in conventional methods.

4.1. MVDR beamforming

The MVDR beamformer is derived to minimize the noise power under the constraint that the target source remain distortionless:

$$\mathbf{w}_{\text{MVDR}} = \underset{\mathbf{w}}{\text{argmin}} \mathbf{w}^H \mathbf{\Phi}_{n+i} \mathbf{w} \quad \mathbf{s.t.}\ \mathbf{w}^H \mathbf{d} = 1 \tag{6}$$

where $(.)^H$ indicates the Hermitian transpose, and $\mathbf{d}$ denotes the steering vector of the target signal. $\mathbf{\Phi}_{n+i}$ is the spatial covariance matrix of the noises and interferences. The solution to this optimization equation is

$$\mathbf{w}_{\text{MVDR}} = \frac{\mathbf{\Phi}_{n+i}^{-1}\mathbf{d}}{\mathbf{d}^H \mathbf{\Phi}_{n+i}^{-1}\mathbf{d}} \tag{7}$$

For multi-talker speech separation, multiple MVDR beamformers are used each for one target speaker. Let s denote the index of the MVDR beamformers as well as the index of speakers. The estimated signal of speaker s is given by

$$\widetilde{X_s}(t,f) = \mathbf{w}_{\text{MVDR},s}^H \mathbf{Y}(t,f) \tag{8}$$

4.2. Steering vector computation

The beamformer coefficients require the knowledge of the steering vector, that is derived from the principle eigenvector of the target speech spatial covariance matrix as:

$$\mathbf{d}_s = \mathcal{P}\{\mathbf{\Phi}_{ss}\} \tag{9}$$

where the operation $\mathcal{P}\{\cdot\}$ returns the eigenvector corresponding to the maximal eigenvalue of a matrix. $\mathbf{\Phi}_{ss}$ denotes the signal spatial covariance matrix of speaker s and can be obtained by

$$\mathbf{\Phi}_{ss}(f) = \frac{\sum_{t=1}^{T} \mathcal{M}_{ss}(t,f)\mathbf{Y}(t,f)\mathbf{Y}(t,f)^H}{\sum_{t=1}^{T} \mathcal{M}_{ss}(t,f)} \tag{10}$$

where $\mathcal{M}_{ss}$ denotes masks belonging to speaker s, which are taken median of the outputs of all channels. Additionally, $\mathbf{\Phi}_{ss}$ is constrained to rank-1 as described in [27].

The spatial variance matrix of noises and interferences can be estimated as follows :

$$\mathbf{\Phi}_{i+n}(f) = \mathbf{\Phi}_i(f) + \mathbf{\Phi}_n(f) \tag{11}$$

where $\mathbf{\Phi}_n(f)$ denotes noise spatial covariance matrix. $\mathbf{\Phi}_i(f)$ denotes the spatial covariance matrix of all interferences that corresponds to target speaker s, that are estimated as follows:

$$\mathbf{\Phi}_n(f) = \frac{\sum_{t=1}^{T} \mathcal{M}_n(t,f)\mathbf{Y}(t,f)\mathbf{Y}(t,f)^H}{\sum_{t=1}^{T} \mathcal{M}_n(t,f)} \tag{12}$$

$$\mathbf{\Phi}_i(f) = \sum_{j \neq s}^{S} \mathbf{\Phi}_{jj}(f) \tag{13}$$

where $\mathcal{M}_n$ denotes the median mask of noises. $\mathbf{\Phi}_{jj}(f)$ denotes the spatial covariance matrix of source j and can be derived similarly as in equation (10).

5. Experiments and results

In this section, we describe the experimental setups and results. The experimental results show that our proposed approach significantly improves performance in terms of SDR[29] compared with uPIT-based single-channel method and deep clustering-based beamforming approach[25].

5.1. Database

We created two data sets for evaluation: 2-speaker mixture set and 3-speaker mixture set. Moreover, each data set was divided into training set, development set, and evaluation set. For each data sets, we first generated mixture lists to indicate the source utterances of each mixture. The lists of training set (20000 utterances) and development set (5000 utterances) were generated from Wall Street Journal (WSJ0) [30] training set si_tr_s. The

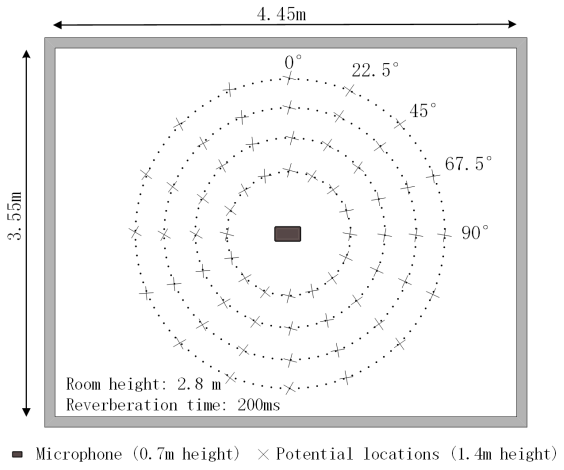

Figure 3: *Simulated Room Conditions.*

lists of evaluation set (3000 utterances) were generated from WSJ0 development set si_dt_05 and evaluation set si_et_05.

The multi-channel data were generated by convolving impulse responses with speech according to the mixture lists. The impulse responses were simulated by image method [31]. The dimensions of simulated room are $4.45m \times 3.55m \times 2.8m$, as shown in Figure 3. The T60 reverberation time is 200 ms. We assumed a rectangular microphone array with 6 sensors, which arranged as CHiME3 [32] while all directions of microphone were upward. The microphone array was located at the center of the room. The position of speakers were randomly selected from potential locations, which are marked by cross symbol in Figure 3. The number of potential locations is 64, and these potential locations are evenly distributed in four concentric circles, for which the radius are 0.4m, 0.7m, 1.0m, 1.3m, respectively. The degree of angle between adjacent location is 22.5°.

5.2. Neural Network

For 2-speaker and 3-speaker separation, we constructed six RNN networks for mask estimation, respectively. Each RNN network has 3 bi-direction long short-term memory (BLSTM) hidden layers followed with a feedfordward (FF) output layer. Each BLSTM layer has 896 cells with tanh activation functions. The feedfordward layer has S*129 units with ReLU activation functions, here S equals the number of sources. The Adam learning algorithm was used with initial learning rate of 0.0005 and dropout rate of 0.5. The utterance-level loss function was used for optimization [23]. If the validation loss decreased, the model will be stored. Otherwise, previous model was restored and the learning rate was scaled by 0.7. The training process was stopped after 5 times consecutive increase of validation loss.

5.3. Results

Experimental results were evaluated in terms of SDR. The results is shown in Table 1. The SDR for 2-speaker and 3-speaker separation of our proposed approach (uPIT-BFM) are 12.38 dB and 10.59 dB, respectively. This significantly improved by 2.25 dB and 2.57 dB compared with uPIT-based single-channel (uPIT-SGL) method. The comparisons between different mask-based beamforming show that our proposed approach

Table 1: *SDR improvements of 2-speaker and 3-speaker separation compared with uPIT-based single-channel separation and other mask-based beamforming separation. The SDR of CGMM-BFM and DPCL-BFM (oracle number of sources) [25] are listed as comparisons*

	CGMM-BFM	DPCL-BFM	uPIT-SGL	uPIT-BFM
2-spk.	11.48	10.36	10.13	12.38
3-spk.	10.95	10.27	8.02	10.59

Table 2: *SDR improvements for same gender and different gender on 2-speaker scenario. The position of speakers in the test set contains all potential locations marked in Figure 3.*

	Avg.	Dif.-Gen.	F - F	M - M
uPIT	10.13	11.61	9.30	7.89
Beamforming	12.38	13.23	11.29	11.49
Increased	2.25	1.62	1.99	3.6

achieves better performance than CGMM-based beamforming (CGMM-BFM)[25] and DCLP-based beamforming (DPCL-BFM)[25] on 2-speaker mixture scenario by 0.9 dB and 1.1 dB, respectively. For the 3-speaker separation, the proposed uPIT-BFM improves SDR by 0.37 dB compared with DPCL-BFM while decreases 0.4 dB compared with CGMM-BFM. Although the performance slightly decreased than CGMM-BFM for 3-speaker separation, it is still comparable between uPIT-BFM and CGMM-BFM.

The SDR for same gender and different gender of 2-speaker separation is shown in Table 2. The result concludes that SDR for different gender speech separation is better than the same gender condition. Moreover, from Table 2, we also conclude that the lower SDR of single-channel based separation, the higher SDR improved by MVDR beamforming. For example, on 2-speaker scenario, the lowest SDR of single-channel separation is 7.89 dB under M-M condition and the corresponding SDR is increased by 3.6 dB, of which gained 45.6% improvement.

6. Conclusion

In this paper, we have proposed an approach for multichannel multi-talker speech separation by using PIT training and MVDR beamforming. Firstly, we trained neural networks for mask estimation with utterance PIT. Then, we constructed mask-based MVDR beamformers for target speeach prediction. This method takes the advantages of the PIT method and the beamforming technique. Experimental result demonstrates that the proposed method significantly improves SDR for both 2-speaker and 3-speaker separation by 2.25 dB and 2.57 dB, respectively. In future work, we will integrate spatial features in training the speaker mask estimation network and evaluate our proposed approach in real-world automatic speech recognition tasks.

7. References

[1] S. Haykin and Z. Chen, "The cocktail party problem," *Current Biology Cb*, vol. 19, no. 22, pp. 1024–7, 2009.

[2] D. Lee, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999.

[3] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *The International Conference on Spoken Language Processing*, 2006, pp. 2614–2617.

[4] F. Weninger, J. L. Roux, J. R. Hershey, and S. Watanabe, "Discriminative nmf and its application to single-channel source separation," in *International Conference on Information Integration and Web-Based Applications and Services*, 2014, pp. 543–547.

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

[6] G. J. Brown and D. L. Wang, *Separation of Speech by Computational Auditory Scene Analysis*. Springer Berlin Heidelberg, 2005.

[7] P. Li, Y. Guan, B. Xu, and W. Liu, "Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech," in *International Conference on Innovative Computing, Information and Control*, 2006, pp. 742–745.

[8] Y. Shao, S. Srinivasan, Z. Jin, and D. L. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Computer Speech and Language*, vol. 24, no. 1, pp. 77–93, 2010.

[9] A. P. Varga and R. K. Moore, "Hidden markov model decomposition of speech and noise," in *International Conference on Acoustics, Speech, and Signal Processing*, 1990, pp. 845–848 vol.2.

[10] A. Ozerov, C. Fvotte, and M. Charbit, "Factorial scaled hidden markov model for polyphonic audio representation and source separation," in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09. IEEE Workshop on*, 2009, pp. 121 – 124.

[11] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[12] X. Zhang and D. L. Wang, "Binaural reverberant speech separation based on deep neural networks," in *INTERSPEECH*, 2017, pp. 2018–2022.

[13] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments." *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 5, pp. 1075–1084, 2017.

[14] K. molkov, 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, pp. 2655–2659.

[15] Y. H. Tu, J. Du, L. R. Dai, and C. H. Lee, "A speaker-dependent deep learning approach to joint speech separation and acoustic modeling for multi-talker automatic speech recognition," in *International Symposium on Chinese Spoken Language Processing*, 2017, pp. 1–5.

[16] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 1562–1566.

[17] C. Weng, D. Yu, M. L. Seltzer, and J. Droppo, "Deep neural networks for single-channel multi-talker speech recognition," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 23, no. 10, pp. 1670–1679, 2015.

[18] Y. Wang, J. Du, L. R. Dai, and C. H. Lee, "Unsupervised single-channel speech separation via deep neural network for different gender mixtures," in *Signal and Information Processing Association Summit and Conference*, 2017, pp. 1–4.

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

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

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

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

[23] M. Kolbaek, D. Yu, Z. H. Tan, J. Jensen, M. Kolbaek, 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. PP, no. 99, pp. 1–1, 2017.

[24] L. Drude and R. Haeb-Umbach, "Tight integration of spatial and spectral features for bss with deep clustering embeddings," in *INTERSPEECH*, 2017, pp. 2650–2654.

[25] T. Higuchi, K. Kinoshita, M. Delcroix, K. molkov, and T. Nakatani, "Deep clustering-based beamforming for separation with unknown number of sources," in *INTERSPEECH*, 2017, pp. 1183–1187.

[26] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "Blstm supported gev beamformer front-end for the 3rd chime challenge," in *Automatic Speech Recognition and Understanding*, 2016, pp. 444–451.

[27] Z. Wang, E. Vincent, R. Serizel, and Y. Yan, "Rank-1 constrained multichannel wiener filter for speech recognition in noisy environments," *Computer Speech & Language*, vol. 49, 2018.

[28] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 708–712.

[29] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans on Audio Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[30] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete ldc93s6a," *philadelphia: Linguistic Data Consortium*, 1993.

[31] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating smallroom acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. S1, pp. 943–950, 1998.

[32] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chime speech separation and recognition challenge: Dataset, task and baselines," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, 2015.