MULTI-MICROPHONE NEURAL SPEECH SEPARATION FOR FAR-FIELD MULTI-TALKER SPEECH RECOGNITION

Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, Fil Alleva

Microsoft AI and Research, One Microsoft Way, Redmond, WA

ABSTRACT

This paper describes a neural network approach to far-field speech separation using multiple microphones. Our proposed approach is speaker-independent and can learn to implicitly figure out the number of speakers constituting an input speech mixture. This is realized by utilizing the permutation invariant training (PIT) framework, which was recently proposed for single-microphone speech separation. In this paper, PIT is extended to effectively leverage multi-microphone input. It is also combined with beamforming for better recognition accuracy. The effectiveness of the proposed approach is investigated by multi-talker speech recognition experiments that use a large quantity of training data and encompass a range of mixing conditions. Our multi-microphone speech separation system significantly outperforms the single-microphone PIT. Several aspects of the proposed approach are experimentally investigated.

Index Terms— Speech separation, multi-talker speech recognition, far-field audio, cocktail party problem, neural networks, acoustic beamforming

1. INTRODUCTION

Machine implementation of cocktail party listening, the ability of selectively recognizing one or each speaker in far-field multi-talker environments, will be a milestone toward realizing computers that can understand complex auditory scenes as accurately as humans. Cocktail-party listening systems that can separate and recognize overlapped voices will significantly enhance the transcription accuracy for human conversations recorded with distant microphones. While it may be possible to build such systems with only a single microphone, approaches using multiple microphones are more practical because information on speaker locations that may be inferred from multi-microphone signals is usually helpful in separating speech.

Most existing multi-microphone approaches to speech separation share a serious drawback. They require prior knowledge of the number of the speakers participating in the input speech mixture. This is a prerequisite for popular approaches such as independent component analysis (ICA) [1–3] and time-frequency (TF) bin clustering [4–7]. Speaker number estimation is a challenging task by itself and far from being solved although a lot of efforts have been made to address it [8, 9]. Thus, the speech separation technology has found little application so far in commercial products, especially in speech recognition space.

In this paper, we take a different approach. We build a neural network that produces a fixed number, I, of outputs where, when there are K speakers, only K of the output channels contain separated speech signals while the remaining $I - K$ channels produce zero values. This is made possible by extending the permutation invariant training (PIT) framework [10], which has recently been developed for single-microphone speech separation, to the multi-microphone scenario. Unlike most neural speech separation methods [11, 12], networks obtained with PIT are speaker independent, i.e., they are not hardwired to specific target speakers.

The contribution of this work is three-fold.

- The PIT framework is extended to multi-microphone speech separation. We explore different multi-microphone features to make the best use of the information that multi-channel audio capture can provide.
- To further capitalize on the spatial information obtained from the multi-channel audio capture, mask-driven beamforming is adopted instead of TF masking. This results in significant gains in multi-talker speech recognition accuracy especially when overlap is severe.
- Extensive performance analysis is conducted at a large scale. Our evaluation is performed across five different mixing conditions from complete overlap to no overlap where people speak one after another. The speech recognition system employed in our experiments makes use of a far-field acoustic model trained on thousands of hours of noise-corrupted speech. We also examine the impact that the quantity of the speech separation training data has on the recognition performance.

The next section reviews the PIT framework. Section 3 describes the proposed speech separation system, which consists of a multi-microphone speech separation neural network and mask-driven beamformers. Section 4 reports our experimental results. Section 5 concludes the paper.

2. PERMUTATION INVARIANT TRAINING

PIT is a method for training a neural network that accepts a sequence of feature vectors of mixed speech as input and generates a fixed number of TF mask sequences. The generated TF masks are applied to the mixture signal to extract each of the constituent speech signals. The network typically has multiple output channels, each of which generates a TF mask sequence for one speaker, as illustrated in Fig. 1. It typically consists of bidirectional long short term memory (BLSTM) layers to take account of a long-term acoustic context, which is essential for the network to be able to track individual speakers. Let us introduce a few notations. The input feature vector is denoted by $\boldsymbol{y}_t$ with t being a short time frame index. The TF mask of the ith output sequence is represented as $m_{i,tf}$, where f is a frequency bin index ranging from 1 to F. We also employ a vector-form notation: $\boldsymbol{m}_{i,t} = [m_{i,t1}, \cdots, m_{i,tF}]^T$. The speech separation neural network takes in an entire input sequence, $(\boldsymbol{y}_t)_{t\in\mathcal{T}}$, and emits $(\boldsymbol{m}_{i,t})_{t\in\mathcal{T}}$ from each of the I output channels, where $\mathcal{T}$ represents a signal observation period and i is an output channel index.

When we train the network, we do not know which output channel corresponds to which speaker. To take account of this ambiguity, PIT

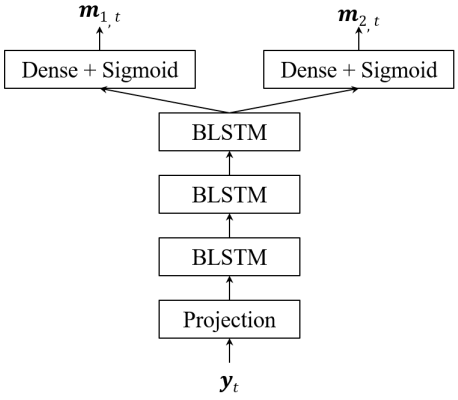

Fig. 1. Typical neural network configuration for $I = 2$. The network has two or more output channels, each producing zero or TF masks for one of the speakers constituting an input speech mixture.

examines all possible permutations of output channels and chooses the best one to invoke gradient descent learning. Therefore, the loss function for a trainer to minimize is formulated as

$$L = \min_{J \in \text{perm}(I)} \sum_{i=1}^{I} \sum_{t \in \mathcal{T}} l\big(\boldsymbol{m}_{j_i,t} \odot \boldsymbol{Y}_t, \boldsymbol{X}_{i,t}\big), \; J = (j_1, \cdots, j_I), \quad (1)$$

where $\text{perm}(I)$ produces all possible permutations for a sequence $(1, \cdots, I)$, $\odot$ denotes element-wise multiplication, and $\boldsymbol{Y}_t$ and $\boldsymbol{X}_{i,t}$ are the power spectra of the mixture signal and the ith speaker signal, respectively. Function $l(\boldsymbol{X}, \boldsymbol{Y})$ measures the degree of discrepancy between two power spectra, $\boldsymbol{X}$ and $\boldsymbol{Y}$, and is usually chosen to be the squared error, i.e.,

$$l(\boldsymbol{X}, \boldsymbol{Y}) = |\boldsymbol{X} - \boldsymbol{Y}|^2. \quad (2)$$

Note that the permutation determination takes place at a sequence level. This discourages the network from swapping speakers at every short time frame, thereby letting the network learn to jointly separate and track individual speaker signals. Unlike other neural separation methods which ask the network to emit a target speaker's signal from a specific output channel (e.g., the first output channel), PIT allows the network to learn to figure out the best ouput channel to use for each speaker. Therefore, the resultant network is not hardwired to certain speakers, i.e., it is speaker-independent. When the number of speakers, K, is smaller than that of the output channel, I, we can simply set $\boldsymbol{X}_{i,t}$ to zero for $i > K$.

One advantage of PIT compared with similar speech separation methods such as deep clustering [13] and deep attractor networks [14] is that PIT does not require an additional clustering or expectation-maximization step at test time.

3. MULTI-MICROPHONE NEURAL SPEECH SEPARATION

Figure 2 shows a schematic diagram of the processing flow of our proposed speech separation system. It comprises a PIT network using multi-microphone input, minimum variance distortionless response (MVDR) beamformers, and gain adjustment mechanisms. These components are detailed in Sections 3.1–3.3.

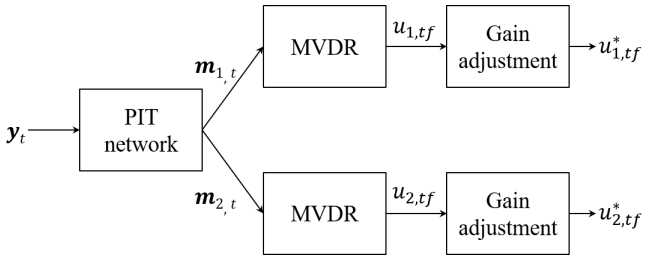

Fig. 2. Processing flow of proposed speech separation system.

3.1. Multi-microphone input network

For single-microphone PIT, short time power spectra are usually used as the input to a network, i.e., $\boldsymbol{y}_t = \boldsymbol{Y}_t$. In our internal test where we compared different features such as log power spectra, linear mel spectra, log mel spectra, and so on, the choice of features had little impact on the separation performance. We also found that utterance-level mean and variance normalization helped a lot for the trained network to generalize to unseen environments.

When multiple microphones are available, spatial features, indicative of speaker locations, can be used. One simple way of utilizing such multi-microphone inputs is to feed the network with both magnitude and phase spectra obtained from all microphones. Alternatively, we may exploit established spatial features such as inter-microphone phase differences (IPDs) with respect to a reference microphone. When we use the first microphone as the reference, the IPDs can be calculated as

$$a_{i,tf} = \angle\left(\frac{y_{i,tf}}{y_{1,tf}}\right), \; i = 2, \cdots, M, \quad (3)$$

where M denotes the number of the microphones being used. This kind of normalization with respect to the reference microphone eliminates phase variations inherent in source signals and hence allows room's acoustic characteristics to be directly captured. The IPD features can be concatenated with magnitude spectra to leverage both spectral and spatial cues. Although one might expect that neural networks can learn such phase normalization effects from data, our results show that explicitly normalizing the phase spectra significantly improve the separation accuracy (see Section 4). After some preliminary experiments, we settled down to using the magnitude spectra from all the microphones and the IPDs between the first microphone and each of the other microphones.

A care must be taken on feature normalization. Based on some preliminary experiments, we eventually decided to apply only mean normalization to the IPD features while doing both mean and variance normalizations on the power spectra. As suggested by existing studies on multi-microphone speech separation [15], clusters in the IPD feature vector space represent speakers or other spatially isolated sound sources. Because variance normalization alters feature vector distributions, it might hinder the neural network from finding clusters corresponding to speakers (even though the network is not explicitly asked to perform clustering). On the other hand, as mean normalization just shifts the feature vector distributions, it reduces feature variations and therefore facilitates network training.

3.2. Mask-driven beamforming

At test time, the TF masks generated by a PIT-trained network are used to create beamformers for separating individual speaker signals.

While it is possible to separate the speaker signals by directy applying the TF masks to the microphone signals, processing artifacts that TF masking generates tend to degrade speech recognition performance [16]. In this paper, we examine two variants of mask-driven MVDR beamforming [6, 17].

Beamforming algorithms compute the (complex-valued) STFT coefficient, $u_{i,tf}$, of speaker i as

$$u_{i,tf} = \boldsymbol{w}_{i,f}^H \boldsymbol{V}_{tf}, \tag{4}$$

where $\boldsymbol{w}_{i,f}$ is a beamformer coefficient vector for speaker i and frequency bin f while $\boldsymbol{V}_{tf}$ is a vector comprising the STFT coefficients of individual microphones. By using the MVDR formulation of [18], the beamformer coefficient vector can be obtained as

$$\boldsymbol{w}_{i,f} = \frac{\boldsymbol{\Phi}_{\bar{i},f}^{-1} \boldsymbol{\Phi}_{i,f} \boldsymbol{e}}{\mathrm{tr}\left(\boldsymbol{\Phi}_{\bar{i},f}^{-1} \boldsymbol{\Phi}_{i,f}\right)}, \tag{5}$$

where $\boldsymbol{e} = [1, 0, \cdots, 0]^T$. The two matrices, $\boldsymbol{\Phi}_{i,f}$ and $\boldsymbol{\Phi}_{\bar{i},f}$, represent estimates of the spatial covariance matrix of the ith speaker signal and that of the mixture of all the other speaker signals.

The spatial covariance matrices may be calculated by using the TF masks in two different ways. One scheme, dubbed as mask-cov, picks up TF bins that are dominated by the target or interfering speakers and calculate the covariance matrices by using these TF points. Specifically, mask-cov uses the following estimators:

$$\boldsymbol{\Phi}_{i,f} = \frac{1}{\sum_{t \in \mathcal{T}} m_{i,tf}} \sum_{t \in \mathcal{T}} m_{i,tf} \boldsymbol{V}_{tf} \boldsymbol{V}_{tf}^H \tag{6}$$

$$\boldsymbol{\Phi}_{\bar{i},f} = \frac{1}{\sum_{t \in \mathcal{T}} (1 - m_{i,tf})} \sum_{t \in \mathcal{T}} (1 - m_{i,tf}) \boldsymbol{V}_{tf} \boldsymbol{V}_{tf}^H. \tag{7}$$

One potential drawback of the mask-cov scheme is that it results in biased estimates of the covariance matrices because the statistics are computed from nonrandom samples. The other scheme, dubbed as sig-cov, calculates estimates of individual speaker signals by applying TF masks to each microphone signal and then computes the spatial covariance matrices from these signal estimates. This may yield less biased estimates especially when the squared error criterion given by equation (2) is used.

3.3. Gain adjustment

One disadvantage of MVDR beamforming compared to TF masking is that it maintains a unit gain toward a certain direction. Therefore, even when the TF masks of a particular output channel are zero almost everywhere, implying the channel does not contain speech, the beamformer cannot filter out the speech signals especially under reverberant conditions. This can be alleviated by modifying the gain of the beamformed signal with that of the masked signal as follows:

$$u_{i,tf}^* = u_{i,tf} \frac{E_i}{\max_{j \in [1,I]} E_j}, \text{ where } E_i = \sqrt{\sum_{t,f} \left\| m_{i,tf} V_{1,tf} \right\|^2}. \tag{8}$$

$u_{i,tf}^*$ is our final estimate of the STFT coefficient of one of the speaker signals. This is converted to a time-domain signal and fed into a speech recognition system.

4. EXPERIMENTS

Far-field multi-talker speech recognition experiments were conducted to evaluate the effectiveness of the speech separation approach described in the previous section.

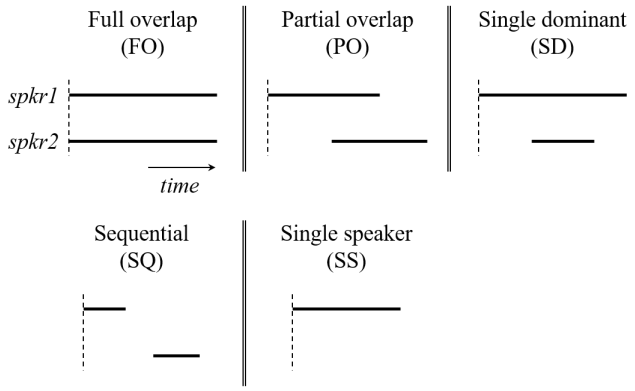

Fig. 3. Five mixing configurations considered in our experiments. Solid horizontal lines represent speaker activity.

4.1. Setup

4.1.1. Test data and tasks

Five different test sets were created as illustrated in Fig. 3. Each test set relates to a distinct mixing configuration: four of them were concerned with two-speaker scenarios; one with a single-speaker scenario. For each mixing configuration, a one-hour test set was created by artificially reverberating anechoic speech signals and mixing them. The anechoic speech signals were taken from our internal gender-balanced clean speech collection, spoken by 44 speakers. The impulse responses used for reverberation simulation were generated with the image method for a hypothetical seven-element circular microphone array with a radius of 4.25 cm. The hypothesized array had six microphones equally spaced along its perimeter and one additional microphone at the circle's center. The room dimensions, the reflection coefficients of the room surfaces, and the microphone and speaker locations were randomly determined for each utterance pair.

A word error rate (WER) was used to measure the speech separation quality. In the "full overlap" (FO), "partial overlap" (PO), and "sequential" (SQ) setups, the two speaker signals were mixed without being truncated. Therefore, speech recognition was run for both separated speech signals and WERs were averaged. In the "single dominant" (SD) setup, the second (i.e., interfering) speaker's signal was truncated to make it sufficiently shorter than the first (i.e., target) speaker's signal. Thus, the task for SD was to recognize the target speaker's speech. The output signal that corresponds to the target speech was determined based on the signal-to-distortion ratio (SDR) with respect to the clean target signal. The same evaluation scheme was adopted for the "single speaker" (SS) case.

In addition to the WER, we also assessed the degree to which one channel was silent. This was measured by the energy ratio between the two output channels, which we call an inter-channel energy ratio (ICER). A very large ICER indicates that the channel with a smaller energy can be regarded as silent. A successful separation system should produce a large ICER for the SS condition.

4.1.2. Systems

All speech separation networks used in our experiments consisted of three 1024-cell BLSTM layers, followed by two parallel dense layers with sigmoid nonlinearity as depicted in Fig. 1. For networks with multi-microphone input, we added a projection layer at the bottom to reduce the input dimensionality to 1024.

Table 1. %WERs of different speech separation systems. ICERs in dB are also shown for proposed system trained on x5.

Separation system	Perf. Metrics	Mixing configurations FO	PO	SD	SQ	SS
Oracle	WER	16.6	17.7	16.4	18.8	16.8
Mixed speech		83.0	83.8	56.8	107.3	16.8
1-mic PIT, x1		63.0	50.6	48.5	31.0	19.3
Proposed, x1		30.6	31.8	24.9	32.5	24.0
Proposed, x5		26.3	31.3	24.0	31.1	19.6
	ICER	0.20	0.14	2.21	0.56	46.2

The speech separation networks were trained on artificially created reverberant speech mixtures, which covered all the five mixing configuraions listed in Fig 3. Source speaker signals were taken from the Wall Street Journal (WSJ) SI-284 utterances. For each utterance pair, a mixing configuration was selected randomly with the probabilitites of picking FO, PO, SD, SQ, and SS being 57.0%, 14.25%, 14.25%, 9.5%, and 5.0%, respectively. (These numbers were arbitrarily determined.) The generated mixed speech signals were clipped to 10 seconds due to GPU memory constraints. Two training sets with different sizes were created. One, called x1, used each utterance only once, resulting in a 43.7-hour set. The other one, called x5, used each utterance five times, where each utterance was mixed with a different utterance at each time. The size of the resulting training set was 219 hours.

For speech recognition, we built a far-field acoustic model by using an artificially reverberated and noised version of speech audio collected from Microsoft Cortana traffic. This model was built by using a teacher-student (TS) adaptation technique [19] as follows. First, a near-field acoustic model using four LSTM layers was trained on the original near-field 3.4K-hour training set. Then, the obtained model was adapted to the far-field data with TS adaptation, which uses pairs of near-field and far-field utterances. For each utterance pair, the near-field speech was forwarded through the original acoustic model, or the teacher model, to generate senone posterior probabilities, which were subsequently used as soft supervision labels for the student model to predict. The student model trained in this way was used as the far-field acoustic model.

The neural networks were implemented by using Microsoft Cognitive Toolkit (formerly known as CNTK), where we utilized 1-bit stochastic gradient descent [20] for efficient distributed learning.

4.2. Results

Table 1 lists the WERs for different speech separation systems. The oracle numbers refers to far-field single-speaker WERs and were around 17% across mixing configurations. The WERs significantly increased when two speakers were mixed as shown in the "Mixed speech" row. The extremely high WER for SQ was due to too many insertion errors. While the conventional single-microphone PIT ("1-mic PIT, x1") improves the recognition accuracy for overlapped speech signals, the WER was still as high as 63.0% for the fully overlapped case. The relative gain obtained from single-microphone PIT was substantially smaller than those reported previously because the prior work on PIT used anechoic mixtures. Our proposed separation system, which uses multiple microphones in mask prediction and beamforming, significantly reduced the WERs. When trained on the larger training set, x5, the proposed system further reduced the WER to 26.3% for the fully overlapped case. The bottom row of Table 1 shows the ICERs for the proposed system trained on the x5 data. It can be clearly observed that the ICER was very high only when the input had a single speaker, which means that one of the output signals was almost zero. This indicates that the separation system learned to figure out whether the input consists of multiple speakers or not.

Table 2. %WER comparison for different network inputs. Separation networks were trained on x1.

Network input	Mixing configurations FO	PO	SD	SQ	SS
1 mic	42.4	42.0	34.6	36.3	25.1
7 mics, raw	45.2	43.0	35.2	36.1	24.1
7 mics, magnitude+IPD	30.6	31.8	24.9	32.5	24.0

Table 3. %WER comparison for different enhancement schemes. Separation networks were trained on x5.

Enhancement	Mixing configurations FO	PO	SD	SQ	SS
TF masking	45.6	34.6	35.5	18.4	17.5
MVDR, mask-cov	30.2	33.8	24.8	31.6	17.2
MVDR, sig-cov	26.3	31.3	24.0	31.1	19.6

Table 2 shows the WERs obtained using different input features. The networks were trained on x1 and the output signals were calculated with MVDR using the sig-cov method. Comparing the first and second rows, we can see that simply feeding the raw multi-microphone features to the network resulted in no improvement for any of the mixing configurations. The use of mean-normalized IPD features improved the WERs a lot across the mixing configurations. This implies the difficulty for the network trained on the raw multi-microphone features to learn the phase normalization effect.

The impact that the choice of enhancement schemes has on recognition performance was also investigated. Table 3 compares three enhancement schemes, i.e., TF masking, MVDR with the mask-cov method, and MVDR with the sig-cov method. When two speakers were overlapped, the beamforming approach outperformed TF masking, which is consistent with previous findings [16]. On the other hand, the beamforming results were worse than the TF masking numbers in the SQ scenario when there was one speaker at a time. This is because the masking approach was more effective at suppressing interfering speakers and thus was able to generated fewer insertion errors. As regards the comparison between the two covariance estimation schemes, sig-cov was slightly better than mask-cov as briefly discussed earlier in Section 3.2.

5. CONCLUSION

In this paper, we described a speech separation approach that combines a multi-microphone neural network for TF mask estimation and a mask-based beamformer. The proposed approach was tested in a far-field multi-talker speech recognition task involving 44 speakers, where the separation network and the acoustic model were trained on 219 hours of mixed speech and 3.4K hours of far-field speech, respectively. Good recognition performance was achieved for all the mixing conditions considered. It was also shown that the separation system produced nearly zero signals from one of the two output channels when and only when the input signals consisted of a single speaker. Further experimental results, using real data and covering more diverse conditions, will be reported in a follow-up paper.

6. REFERENCES

[1] S. Makino, T. W. Lee, and H. Sawada, *Blind speech separation*, Springer, 2007.

[2] F. Nesta, P. Svaizer, and M. Omologo, "Convolutive bss of short mixtures by ica recursively regularized across frequencies," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 3, pp. 624–639, 2011.

[3] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 69–84, 2011.

[4] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 241–244.

[5] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 3, pp. 516–527, 2011.

[6] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 9, pp. 1913–1928, 2013.

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

[8] L. Drude, A. Chinaev, D. H. T. Vu, and R. Haeb-Umbach, "Source counting in speech mixtures using a variational em approach for complex watson mixture models," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 6834–6838.

[9] T. Higuchi, K. Kinoshita, M. Delcroix, K. Žmolíková, and T. Nakatani, "Deep clustering-based beamforming for separation with unknown number of sources," 2017.

[10] 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 Trans. Audio, Speech, Language Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.

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

[12] Y. Wang, J. Du, L.-R. Dai, and C.-H. Lee, "A maximum likelihood approach to deep neural network based nonlinear spectral mapping for single-channel speech separation," in *Proc. Interspeech*, 2017, pp. 1178–1182.

[13] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: discriminative embeddings for segmentation and separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 31–35.

[14] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 246–250.

[15] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2006, pp. V–33–V–36.

[16] Takuya Yoshioka, Nobutaka Ito, Marc Delcroix, Atsunori Ogawa, Keisuke Kinoshita, Masakiyo Fujimoto, Chengzhu Yu, Wojciech J. Fabian, Miquel Espi, Takuya Higuchi, Shoko Araki, and Tomohiro Nakatani, "The NTT CHiME-3 system: advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. Worksh. Automat. Speech Recognition, Understanding*, 2015, pp. 436–443.

[17] H. Erdogan, J. R. Hershey, S. Watanabe, M. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, 2016, pp. 1981–1985.

[18] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 260–276, 2007.

[19] J. Li, M. L. Seltzer, X. Wang, R. Zhao, and Y. Gong, "Large-scale domain adaptation via teacher-student learning," in *Proc. Interspeech*, 2017, pp. 2386–2390.

[20] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, "1-bit stochastic gradient descent and application to data-parallel distributed training of speech DNNs," in *Interspeech*, 2014, pp. 1058–1062.