Boosting Spatial Information for Deep Learning Based Multichannel Speaker-Independent Speech Separation In Reverberant Environments

Ziye Yang and Xiao-Lei Zhang*
* Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen, China
Center for Intelligent Acoustics and Immersive Communications,
School of Marine Science and Technology,
Northwestern Polytechnical University, Xi'an, China
E-mail: 2015300797@mail.nwpu.edu.cn, xiaolei.zhang@nwpu.edu.cn

***Abstract*—Recently, supervised speaker-independent speech separation methods, such as deep clustering and permutation invariant training, have demonstrated better performance than conventional unsupervised speech separation methods. However, their performance drops sharply in reverberant environments. To solve the problem, we propose a multi-channel speech separation algorithm that fully explores spatial information. It first extracts a spatial feature, named interaural phase difference (IPD), as one of the input features of the single-channel deep clustering algorithm. Then, it uses the deep clustering as the noise estimation component of the deep-learning-based beamforming. The novelty of the proposed algorithm lies in that it extends the spatial-feature-based deep clustering to a multichannel algorithm which boosts the performance by exploring spatial information at both the input and output of deep clustering. Its advantages have two aspects. First, the spatial feature IPD significantly improves the robustness of deep clustering in reverberant environments. Second, the deep-clusteing-based beamforming, which is a linear algorithm, suffers less nonlinear distortions than the single-channel deep clustering. We have compared the proposed algorithm with the single-channel deep clustering algorithm, spatial-feature-based multi-channel deep clustering with IPD, and deep-clustering-based beamforming without IPD in reverberant environments. Experimental results show that the proposed algorithm performs significantly better than the comparison methods.**

Index Terms: speaker-independent speech separation, deep clustering, time-frequency masking, beamforming.

I. Introduction

Speech separation separate the overlapped speech of multiple speakers to multiple speech streams, each of which belonging to a single speaker. This paper focuses on deep learning based supervised speech separation [1]. According to the number of microphones, speech separation techniques can be divided into two categories—single-channel speech separation and multi-channel speech separation [2]. According to whether the speakers are predefined or known as a prior, speech separation techniques can be divided into three categories—speaker-dependent [3], target-dependent [4], and speaker-independent [5]–[7]. If all test speakers are known in the training stage, then the separation model can be trained speaker-dependently. If only a target speaker is known in the training stage, then the separation model can be trained target-dependently. If all test speakers are unknown in the training stage, then the separation model must be trained speaker-independently. In practice, speaker- and target-dependent models usually produce better performance than speaker-independent models, while speaker-independent models require the minimal prior knowledge. See [1] for an overview of recent supervised speech separation methods. This paper focuses on developing multi-channel speaker-independent speech separation methods.

Speaker-independent speech separation was first developed as single-channel methods. It can be categorized to two main streams—permutation invariant training (PIT) [5] and deep clustering [6], [7], both of which solve the speaker permutation problem well. Specifically, given a frame or an utterance of a pair of speakers at each epoch, PIT calculates the local mean squared errors (MSE) of all permutations of the training speakers at either the frame-level or the utterance-level, and pick the locally optimal permutation corresponding to the minimum MSE to train the separation network. Deep clustering uses a bi-directional long short-term memory network (BLSTM) to produce an embedding vector for each time-frequency pair of a mixture spectrogram. The Frobenius norm between the affinity matrix of the embedding vectors and the affinity matrix of the ideal speaker assignment (also known as the ideal binary mask) is used as the training objective. The main idea behind the training objective is that the within-class distance between the embedding vectors is minimized, and meanwhile the between-class distance is maximized. Although the two algorithms are effective in clean environments, they suffer significant performance degradation in adverse environments.

One way to improve the performance of speaker-independent speech separation in adverse environments is to incorporate spatial information. Several multi-channel speaker-independent speech separation methods have been proposed. They are mainly deep-clustering-based algorithms, which can be categorized to two representative types—beamforming [8] and spatial feature extraction [9]. Deep-clustering-based beamforming first takes deep clustering as the noise estimator to generate speaker masks. Then, for each speaker, it uses the masks to estimate a spatial convariance matrix which is further

applied to estimate the coefficients of the beamformer, such as the minimum variance distortion-less response (MVDR) beamformer [9] and the maximum signal-to-noise ratio (max-SNR) beamformer [10], [11]. Spatial-feature-extraction-based deep clustering concatenates spatial and spectral features [3], [12], [13] together as the acoustic feature for the model training of deep clustering, where the spatial feature is collected by a microphone array. The filters produced by the beamforming methods are linear ones, therefore, they suffer less nonlinear distortions than the spatial feature extraction methods. On the other side, the beamforming methods are less robust than the latter in reverberant environments [14]. To summarize, how to combine the advantages of the two types of multi-channel speech separation methods is an interesting problem.

In this paper, we propose a multi-channel speaker-independent algorithm by combining the above two kinds of multi-channel methods for improving the performance of speech separation in reverberant environments. The algorithm first concatenates a spatial feature, named interaural phase difference (IPD), with the magnitude spectrogram of the short-time Fourier transform coefficients as the acoustic feature for training deep clustering. Then, it uses the deep clustering as the noise estimator for the deep-clustering-based beamforming. The novelty of the algorithm is that it explores spatial information at both the input and output of deep clustering instead of at one side. Theoretically, the deep-clustering-based noise estimator provides high-quality estimated masks, which is the key requirement for improving the performance of beamforming in reverberant environments. Empirically, it outperforms both the deep-clustering-based beamforming [8] and the spatial-feature-extraction-based deep clustering [9] by at least 10% absolute improvement in terms of short-time objective intelligibility (STOI) in reverberant environments.

The rest of the paper is organized as follows. We describe the proposed algorithm in Section 2. Speech separation experiments are presented in Sections 3. Section 4 concludes this paper.

II. Method

As illustrated in Fig. 1, our algorithm consists of two components—spatial-feature-based deep clustering and deep-clustering-based beamforming.

A. Signal Model

All speech separation methods throughout the paper operate in the frequency domain on a frame-by-frame basis. Suppose that a physical space contains U speakers and a microphone array of P microphones. If the U speakers talk simultaneously, then the physical model for the received signals by the microphone array is assumed to be

$$\mathbf{y}(t,f) = \sum_{u=1}^{U} \mathbf{c}_u(f) s_u(t,f) + \mathbf{h}_u(t,f) + \mathbf{n}(t,f) \tag{1}$$

where $s_u(t,f)$ is the short-time Fourier transform (STFT) value of the clean speech of the u-th speaker at time t and frequency f, $\mathbf{c}_u(f)$ is the time-invariant acoustic transfer function from the u-th speaker to the array which is a P-dimensional complex number:

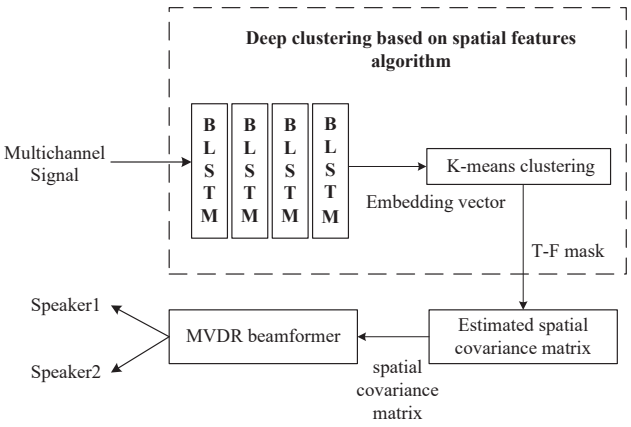

Fig. 1. The diagram of the proposed algorithm

$$\mathbf{c}_u(f) = [c_{u,1}(f), c_{u,2}(f), \ldots, c_{u,P}(f)]^T \tag{2}$$

$\mathbf{c}_u(f)s_u(t,f)$ and $\mathbf{h}_u(t,f)$ are the direct sound and early and late reverberation of the u-th speech source signal, and $\mathbf{n}(t,f)$ and $\mathbf{y}(t,f)$ are the additive noise and received signal at time t and frequency f respectively:

$$\mathbf{n}(t,f) = [n_1(t,f), n_2(t,f), \ldots, n_P(t,f)]^T \tag{3}$$

$$\mathbf{y}(t,f) = [y_1(t,f), y_2(t,f), \ldots, y_P(t,f)]^T. \tag{4}$$

B. Spatial Feature Based Deep Clustering

In the training stage, we first extract P STFT spectrograms from the audio recordings, denoted as $\{y_{i,1}, y_{i,2}, \ldots, y_{i,P}\}_{i=1}^{n}$, where i is a time-frequency (T-F) index (t,f), n is the total number of the T-F units of a STFT spectrogram, and $y_{i,p}$ denotes the i-th T-F unit of the p-th spectrogram. Then, we extract a log-magnitude spectrum by

$$z_{i,p} = \log |y_{i,p}| \tag{5}$$

and a spatial feature IPD by

$$\theta_{i,p,q} = \angle y_{i,p} - \angle y_{i,q}. \tag{6}$$

To handle the 2π ambiguity, we further transform IPD by a cosine function so as to unwrap the phase values into the range $[-1, 1]$ [9]:

$$\delta_{i,p,q} = \cos(\theta_{i,p,q}). \tag{7}$$

Finally, the input acoustic feature of deep clustering at the i-th T-F unit is:

$$\mathbf{m}_i = [y_{i,1}, \ldots, y_{i,p}, \ldots, y_{i,P}, \delta_{i,1,1}, \ldots, \delta_{i,p,q}, \ldots, \delta_{i,P,P}]^T \tag{8}$$

However, if we take into account of all pairs of the microphones in the array, then $\mathbf{m}_i$ is too high-dimensional. In practice, we partition the microphone array into $P/2$ pairs, and train a deep clustering model for each channel. Suppose we are to train a deep clustering model for the p-th channel, and

suppose the p-th channel and the p'-th channel falls into the same pair. Then, the input feature of the p-th deep clustering model is:

$$\mathbf{m}_{i,p} = [y_{i,p}, \delta_{i,p,p'}]^T \tag{9}$$

Deep clustering learns a k-dimensional embedding vector $\mathbf{v}_{i,p}$ from $\mathbf{m}_{i,p}$ by a BLSTM model $g_p(\cdot)$:

$$\mathbf{v}_{i,p} = g_p(\mathbf{m}_{i,p}) \tag{10}$$

It minimizes the following cost function:

$$\mathcal{J}_p = ||\mathbf{V}_p^T\mathbf{V}_p - \mathbf{B}_p^T\mathbf{B}_p||_F^2 \tag{11}$$

where $\|\cdot\|_F$ denotes the Frobenius norm operator, $\mathbf{V}_p = [\mathbf{v}_{1,p}, \ldots, \mathbf{v}_{n,p}]$ is an $n \times k$ embedding matrix, and $\mathbf{B}_p = [\mathbf{b}_{1,p}, \ldots, \mathbf{b}_{n,p}]$ is an $n \times U$ ground-truth indicator matrix with $\mathbf{b}_{i,p} = [b_{i,p,1}, \ldots, b_{i,p,u}, \ldots, b_{i,p,U}]^T$ defined as:

$$b_{i,p,u} = \begin{cases} 1, & \text{if the T-F unit is dominated by speaker } u. \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

where c is the number of speakers.

In the test stage, we use k-means clustering to partition the embedding vectors into U clusters, which generates U estimated binary masks for each channel:

$$\hat{M}_{p,u}(t,f) = \begin{cases} 1, & \text{if the } (t,f)\text{-unit is assigned to speaker } u. \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

C. Deep Clustering Based Beamforming

Deep clustering based beamforming finds U linear estimators $\{\mathbf{w}_u(f)\}_{u=1}^U$ to filter $\mathbf{y}(t,f)$ by the following equation:

$$\hat{x}_u(t,f) = \mathbf{w}_u^H(f)\mathbf{y}(t,f), \quad \forall u = 1, \ldots, U \tag{14}$$

where $(\cdot)^H$ is the conjugate transpose operator and $\hat{x}_u(t,f)$ is an estimate of the direct speech of the u-th speaker at the reference microphone of the array. We take MVDR as the beamformer, which derives the following solution:

$$\mathbf{w}_u(f) = \frac{\widehat{\mathbf{\Phi}}_{\bar{u}\bar{u}}^{-1}(f)\hat{\mathbf{c}}_u(f)}{\hat{\mathbf{c}}_u^H(f)\widehat{\mathbf{\Phi}}_{\bar{u}\bar{u}}^{-1}(f)\hat{\mathbf{c}}_u(f)} \tag{15}$$

where $\widehat{\mathbf{\Phi}}_{\bar{u}\bar{u}}(f)$ is an estimate of the spatial covariance matrix of the interference of the u-th speaker, and $\hat{\mathbf{c}}_u(f)$ is the first principal component of $\widehat{\mathbf{\Phi}}_{\bar{u}\bar{u}}(f)$:

$$\widehat{\mathbf{\Phi}}_{\bar{u}\bar{u}}(f) = \frac{1}{\sum_t \eta_u(t,f)} \sum_t \eta_u(t,f)\mathbf{y}(t,f)\mathbf{y}(t,f)^H \tag{16}$$

$$\hat{\mathbf{c}}_u(f) = \text{principal}\left(\widehat{\mathbf{\Phi}}_{\bar{u}\bar{u}}(f)\right) \tag{17}$$

where $\eta_u(t,f)$ is defined as the product of individual estimated T-F masks:

$$\eta_u(t,f) = \prod_{i=1}^{P} \hat{M}_{p,u}(t,f) \tag{18}$$

III. Experiments

A. Experimental Settings

1) Datasets: We focused on 2-speaker speech separation problems. To simulate real-world environments, we generated a scenario for each training or test mixture. Each scenario needs to simulate a room in which a microphone array and two speakers are further generated. For each scenario, we randomly generated a room that is 5 to 10 meters long, 5 to 10 meters wide, and 3 to 4 meters high. We randomly generated a spherical microphone array with a radius varying from 0.075 to 0.125 meter. The microphone array consists of four microphones, two of which are inside the sphere and the other two are on the surface of the sphere. Its coordinate varies from $(0.2, 0.2, 1)$ to $(0.2, 0.2, 2)$ meters. We randomly generated two speakers that are located in a circle centered at the microphone array with a radius of 1.5 meters. The distance between the microphone array and the speaker is at least 0.5 meter. The distance between the two speakers is at least 1 meter. All simulated environments were generated by the room impulse response function [15].

We used the WSJ0-2mix data [6], [16]–[18] as the speech source, and resampled the speech data to 8 kHz. We generated three datasets for the model training, development, and test. The training set contains 20,000 mixtures. The validation set contains 5000 mixtures. The test set contains 3000 mixtures. The three datasets are about 30, 10, and 5 hours long respectively. For each mixture, we generated its anechoic recording by setting $T_{60} = 0$. We further generated its reverberant recording by selecting T_{60} from the range of $[0.2, 0.6]$ second [19].

2) Parameter Settings: We set the frame length to 32 milliseconds and frame shift to 8 milliseconds. We extracted a 129-dimensional Hamming window reweighted STFT feature from each frame, and further transformed the STFT feature to the feature described in (9). Figures 2 and 3 show the logarithmic magnitude spectrum of a mixture and its components.

In order to prevent the clustering results biased towards silence regions, some inactive T-F units with low energy should not be incorporated into the model training and test process of the spatial-feature-based deep clustering. Based on the above analysis, we conducted an energy-based voice activity detection on the T-F units after the feature extraction.

We built a BLSTM network that consists of four BLSTM layers with 300 hidden units per layer for deep clustering. The network was optimized by stochastic gradient descent. The momentum was set to 0.9. The learning rate was set to 10^{-5}. To avoid the local minima of BLSTM, we added Gaussian noise with a mean of 0 and a variance of 0.6 to the weights. The dimension of the embedding vector was set to 40 which yields the best performance in experience.

3) Comparison Methods: We summarize the comparison methods as follows:

- **LogMag+DC.** This is a single-channel speech separation baseline. It uses the logarithmic magnitude spectrum

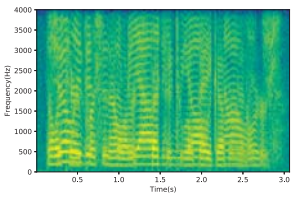

Fig. 2. Logarithmic magnitude spectrum of a mixed speech signal.

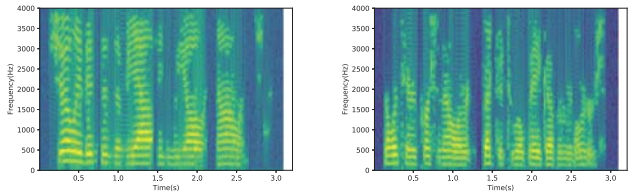

Fig. 3. Separated logarithmic magnitude spectra of the mixed speech signal in Fig. 2.

(LogMag) as the acoustic feature to train a deep clustering (DC) model.

- **LogMag+cosIPD+DC.** This is a multi-channel nonlinear speech separation baseline. It combines LogMag and cosIPD together as the acoustic feature to train a DC model.
- **LogMag+DC+MVDR.** This is a multi-channel linear speech separation baseline. It uses LogMag+DC as the noise estimator of MVDR.
- **LogMag+cosIPD+DC+MVDR (proposed).**

The components of the comparison methods, including the LogMag, cosIPD, DC, and MVDR, have the same parameter setting unless otherwise stated.

4) Evaluation Metrics: We take the source distortion ratio (SDR) as the main evaluation metric, and take the perceptual evaluation of speech quality (PESQ) and STOI as two supplement evaluation metrics.

B. Results

To demonstrate how reverberation degrades the performance of speech separation, we evaluated the single-channel LogMag+DC method in both the anechoic and reverberant environments. The results are listed in Table I. From the table, we see that the performance of LogMag+DC in the reverberant environments is significantly lower that in the anechoic environments, which emphasizes the importance of our research topic— speech separation in reverberant environments.

To show the advantage of the multi-channel speech separation in reverberant environments, we evaluated LogMag+cosIPD+DC, LogMag+DC+MVDR, and the proposed LogMag+cosIPD+DC+MVDR in the reverberant environments. Table II lists the comparison results in terms of SDR, PESQ, and STOI. From the table, we observe the following phenomena. First, LogMag+cosIPD+DC+MVDR significantly outperforms all comparison methods. For example, its STOI scores are over 10% higher than the two multichannel speech separation baselines. Second, all multi-channel methods outperform the single-channel LogMag+DC method significantly, which indicates the importance of exploring spatial information.

TABLE I
SDR PERFORMANCE OF THE LOGMAG+DC BASELINE. THE TERM "M+F" MEANS THAT EACH UTTERANCE IN THE TEST CORPUS IS A MIXED SIGNAL OF A MALE SPEAKER AND A FEMALE SPEAKER, WHERE THE SYMBOLS "M" AND "F" REPRESENT MALE AND FEMALE SPEAKERS RESPECTIVELY. SO AS TO THE TERMS "M+M" AND "F+F".

Genders	anechoic	reverberant
M+M	3.6	1.5
F+F	3.3	1.4
M+F	3.8	2.1

TABLE II
PERFORMANCE OF THE MULTI-CHANNEL COMPARISON METHODS IN REVERBERANT ENVIRONMENTS.

	Genders	SDR	PESQ	STOI
LogMag+DC+MVDR	M+M	8.6	1.65	0.58
	F+F	8.3	1.64	0.55
	M+F	8.8	1.67	0.62
LogMag+cosIPD+DC	M+M	8.9	1.68	0.63
	F+F	8.5	1.67	0.61
	M+F	9.2	1.72	0.69
LogMag+cosIPD+DC+MVDR	M+M	9.1	1.73	0.75
	F+F	8.9	1.71	0.72
	M+F	9.5	1.74	0.77

IV. CONCLUSIONS

In this paper, we have proposed a multi-channel speaker-independent speech separation method by combining the spatial-feature-based deep clustering algorithm with the deep-clustering-based beamforming method. The algorithm first concatenates IPD and the logarithmic magnitude spectrogram as the acoustic feature for training deep clustering. Then, it uses the deep clustering as the noise estimator for MVDR. The novelty of the algorithm is that it explores spatial information at both the input and output of deep clustering. We have compared the proposed method with its two components as well as the single-channel deep clustering algorithm in reverberant environments. Experimental results show that the proposed method significantly outperforms the comparison methods in terms of all three evaluation metrics.

ACKNOWLEDGMENT

This paper was supported in part by the Shenzhen Science and Technology Plan under grant No. JCYJ20170815161820095, in part by the National Natural Science Foundation of China under grant No. 61671381, and in part by the Shaanxi Natural Science Basic Research Program under grant No. 2018JM6035.

REFERENCES

[1] 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.

[2] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[3] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 2112–2121, 2014.

[4] X.-L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 5, pp. 967–977, 2016.

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

[6] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.

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

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

[9] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.

[10] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on antennas and propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[11] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[12] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 116–120.

[13] P. Pertilä and J. Nikunen, "Distant speech separation using predicted time–frequency masks from spatial features," *Speech communication*, vol. 68, pp. 97–106, 2015.

[14] Y. Zhao, Z.-Q. Wang, and D. Wang, "A two-stage algorithm for noisy and reverberant speech enhancement," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5580–5584.

[15] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

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

[17] 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.

[18] M. Kolbæk, 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 (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.

[19] S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 4, pp. 692–730, 2017.