Neural Spatial Filter: Target Speaker Speech Separation Assisted with Directional Information

Rongzhi Gu[1], *Lianwu Chen*[2], *Shi-Xiong Zhang*[3], *Jimeng Zheng*[2], *Yong Xu*[3], *Meng Yu*[3], *Dan Su*[2], *Yuexian Zou*[1], *Dong Yu*[3]

[1]Peking University Shenzhen Graduate School, Shenzhen, China
[2]Tencent AI Lab, Shenzhen, China
[3]Tencent AI Lab, Bellevue WA, USA

{zouyx, 1701111335}@pku.edu.cn,
{lianwuchen, auszhang, jimzzheng, lucayongxu, raymondmyu, dansu, dyu}@tencent.com

Abstract

The recent exploration of deep learning for supervised speech separation has significantly accelerated the progress on the multi-talker speech separation problem. The multi-channel approaches have attracted much research attention due to the benefit of spatial information. In this paper, integrated with the power spectra and inter-channel spatial features at the input level, we explore to leverage directional features, which imply the speaker source from the desired target direction, for target speaker separation. In addition, we incorporate an attention mechanism to dynamically tune the model's attention to the reliable input features to alleviate spatial ambiguity problem when multiple speakers are closely located. We demonstrate, on the far-field WSJ0 2-mix dataset, that our proposed approach significantly improves the performance of speech separation against the baseline single-channel and multi-channel speech separation methods.

Index Terms: target speaker separation, directional features, attention mechanism, permutation invariant training

1. Introduction

Speech separation, which is to isolate an observed mixture speech signal to an individual, contiguous and intelligible stream for each speaker, has been widely studied for decades [1, 2, 3, 4]. Recently, the deep learning based techniques such as deep clustering (DC) [5], deep attractor network (DANet) [6] and permutation invariant training (PIT) [7] address the label permutation problem for separating multi-talker speech. In most of these work, speaker-independent features such as spectral features (e.g., log power spectra) and spatial features (e.g., inter-channel phase difference) are fed into the network, and the time-frequency (T-F) masks of all speakers are estimated. The output speakers' identities remain unknown. However, in most of real applications, only one or a few desired speakers are of interest. The speaker-dependent information is thereby needed for separating interested speakers from the mixture.

A simple yet effective feature is the voice characteristics. The target speaker representation has been proposed to work with the spectra features. The network is either adapted to specific speakers [8], or attend and filter the target speech through interaction between the joint trained speaker features and mixture input[9]. For example, deep extractor network (DENet) [10] used a short audio clip from the target speaker as an anchor, and formed a target extractor in a canonical high-dimensional embedding space. [11] proposed a VoiceFilter to separate the voice of a target speaker from multi-speaker signals, by making use of a reference signal from the target speaker. Although the voice characteristics based methods have been proven successful for extracting target speaker, a common limitation is that they all require the prior knowledge about the reference signal.

Another speaker-dependent information is sound location, which can be either estimated from acoustic/visual signals, or predefined based on real usage scenario. Chen et al. proposed a location-based angle feature which computes the cosine distance between the steering vector and inter-channel phase difference (IPD) for each speaker in mixture [12]. The phase ambiguities creates difficulties for precisely discriminate one speaker from another in certain frequency bands. With the direction of arrival (DOA) information, beamforming techniques [13, 14] can be applied to enhance the speaker from the desired direction. In [15], Wang et al. developed two directional features to improve speech separation. One is the compensated IPD, the other derives from beamforming outputs as beamforming constructively combines target signals and destructively for non-target signals. Although impressive improvement is achieved with the directional features, it is a 2-stage system with high computational complexity.

In this paper, we perform target speaker separation by making use of directional features in a neural network model, named Neural Spatial Filter. Two novel directional features are properly designed based on fixed beamformer outputs, which are then integrated with the conventional multi-channel speech separation training features (e.g. power spectra and inter-channel spatial features) at the input level for the target speaker separation training. Under the supervision of the ideal T-F mask, the network can learn better mask estimation with the assistance from directional features. A limitation of the conventional neural network based speech separation methods is that the maximum number of mixing streams the model can handle is determined by the network architecture, e.g., the PIT with two output segments will not work for three-talker speech separation. The directional features aim to inform the neural network of the target speaker direction and therefore no prior knowledge on the number of mixing speakers is required for this model architecture. Furthermore, we introduce an attention mechanism to alleviate the spatial ambiguity issue discussed in [16] that the performance of multi-channel speech separation drastically degrades when the speakers locate close to each other. In this case, the spatial and directional features become less discriminative.

The rest of paper is organized as follows. Section 2 de-

This work was done when Rongzhi Gu was an intern at Tencent.

scribes the proposed neural spatial filter in detail. The training paradigms are presented in Section 3. Experimental setups and results are summarized in Section 4. We conclude this paper in Section 5.

2. Neural Spatial Filter

2.1. Multi-channel speech separation

In this paper, we follow the multi-channel setup in [12] where IPDs are used as spatial cues and incorporated with spectral features at the input level. For a M-channel mixture signal $\mathbf{y} \in \mathbb{R}^{M \times S}$, we extract K microphone pairs of IPDs, noted as $(k1, k2)$, where $k1$ and $k2$ represents the first and second microphone index of the k-th pair, respectively. The first channel's logarithm power spectrum (LPS) and K pairs of IPDs are concatenated as input of the neural network, and the T-F masks for target speakers are estimated in the output of network.

2.2. Speaker-dependent directional features

Spatial feature such as IPD successfully extracts the spatial information of all the sources in the mixture signal. Moreover, with some or all of the target speaker directions, features of specific speaker-dependent direction could be extracted to improve the performance of separation further. In this paper, it is assumed that the oracle location of each speaker is known by the separation system, this is an reasonable assumption in some real applications, for example, the speaker location could be detected by face detection techniques with very high accuracy.

A location-guided feature for speech separation was introduced in [12]. This feature measures the cosine distance between the steering vector, which is formed according to the direction of target speaker, and IPD:

$$\mathrm{AF}_{\theta}(t,f) = \sum_{k=1}^{K} \frac{\mathbf{e}_{\theta,k1}(f)\frac{\mathbf{Y}_{k1}(t,f)}{\mathbf{Y}_{k2}(t,f)}}{\left|\mathbf{e}_{\theta,k1}(f)\frac{\mathbf{Y}_{k1}(t,f)}{\mathbf{Y}_{k2}(t,f)}\right|} \tag{1}$$

where $\mathbf{e}_{\theta,k1}(f)$ is the steering vector coefficient for target speaker from θ at frequency f for first microphone of k-th pair, and $\mathbf{Y}_{k1}(t,f)/\mathbf{Y}_{k2}(t,f)$ is the IPD between $k1$ and $k2$. Also, the pre-masking step in [12] is also applied to add the discrimination of AF. Note that Eq. 1 can be applied to general microphone array topology rather than the special seven-element microphone array used in [12]. This angle feature provides the desired speaker's directional information to the network so that the network is expected to attend to the target speech.

The beamforming, along with its spatial separation capability, has been well studied in the array processing literature. We propose two new directional features, Directional Power Ratio (DPR) and Directional Signal-to-Noise Ratio (DSNR), based on the output power of multi-look fixed beamformers. For a given microphone array and a pre-defined direction grid $\{\theta_1, \theta_2, ...\theta_P\}$, a set of fixed filters, e.g. Super Cardioid fixed beamformer [17], is designed and denoted as $\mathbf{w}_p(f) \in \mathbb{C}^M$, which aims to enhance sound sources from direction θ_p for f-th frequency bin. Assuming these fixed filters can provide well enough spatial separation and the multiple speakers are not closely located in the space, we can use the processing output power of $\mathbf{w}_p(f)$ as a reasonable estimation of the signal power from direction θ_p. Therefore, the DPR can be considered as an indicator of how well is a T-F bin (t,f) dominated by the signal from direction θ_p, defined as follows:

$$\mathrm{DPR}_{\theta_p}(t,f) = \frac{\left\|\mathbf{w}_p^H(f)\mathbf{Y}(t,f)\right\|_2^2}{\sum_{k=1}^{P}\left\|\mathbf{w}_k^H(f)\mathbf{Y}(t,f)\right\|_2^2} \tag{2}$$

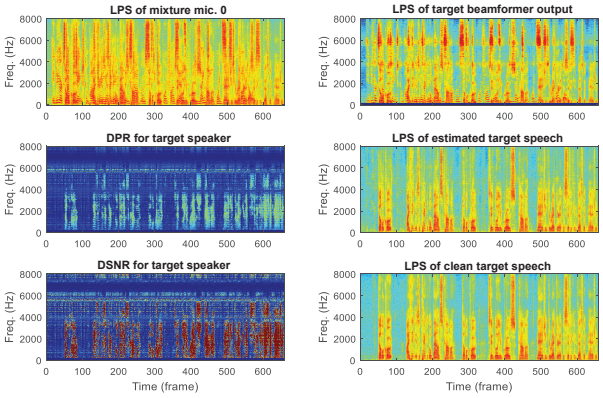

Figure 1: *A example in WSJ0 2-mix of the mixture logarithm power spectrum, output power of the fixed beamformer focuses at target direction, directional features and recovered target speech. The angle difference between 2 speakers is 102 degrees.*

where $\mathbf{Y}(t,f)$ is the complex spectral vector of multi-channel mixture signal in T-F bin (t,f). Furthermore, in most of the beam-pattern design techniques, there are multiple nulling areas by each fixed spatial filter. For example, signals from the neighborhood of θ_p are well preserved by $\mathbf{w}_p(f)$ while severely attenuated by $\mathbf{w}_k(f), \theta_k \in \Omega_p$. Here, Ω_p is the set of directions whose beam-patterns have null at the direction θ_p. It can be well-determined during beamformer design stage. Therefore, if the direction grid covers the whole space, the DSNR can be interpreted as the ratio of signal power from θ_p over the strongest interference:

$$\mathrm{DSNR}_{\theta_p}(t,f) = \frac{\left\|\mathbf{w}_p^H(f)\mathbf{Y}(t,f)\right\|_2^2}{\max\limits_{k\in\Omega_p}\left(\left\|\mathbf{w}_k^H(f)\mathbf{Y}(t,f)\right\|_2^2\right)} \tag{3}$$

Figure 1 illustrates the proposed DPR and DSNR features when applied to a sample in dataset WSJ0 2-mix. Although the designed beamformer that focuses at target direction does not provide significant separation performance, the proposed DPR and DSNR can clearly provide cues for separating target speech from the interference.

2.3. Attention mechanism

An attention mechanism is introduced to alleviate spatial overlap issue discussed in [16], where the performance of multi-channel speech separation drastically degrades when the speakers locate close to each other or the angle difference (AD) between speakers is small. This issue is mainly caused by the increasing dependency that network has on spatial features, since spatial features are more discriminative than spectral features under large AD. Therefore, the network compromises on performances of small angle samples and put too much weight on spatial features in order to achieve overall improvement.

To tackle this issue, we apply an attention mechanism to guide the network to selectively focus on spectral, spatial or directional features under different ADs. The attention is a function of the angle difference ad:

$$att(ad) = 2 * \max\left(\sigma(ad) - 0.5, 0\right) \tag{4}$$

where $\sigma(ad) = 1/(1+exp(-w(ad-b)))$ is the sigmoid score denotes how much emphasis should be put on spatial and direc-

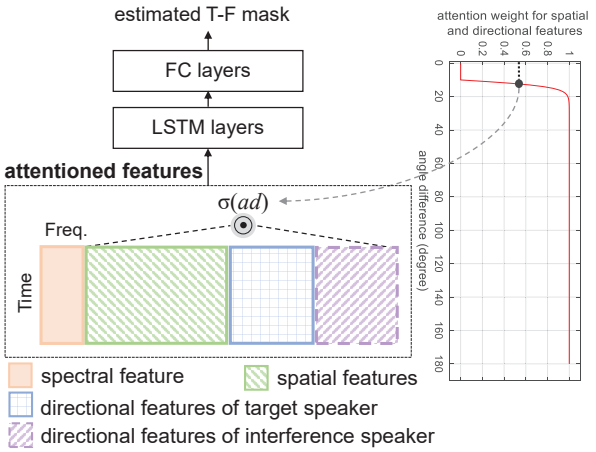

Figure 2: *The overview of our proposed neural spatial filter with attention mechanisum.*

tional features, w and b are trainable parameters. Figure 2 illustrates our proposed neural spatial filter and feature formulation with attention mechanism. On the right is the attention curve under different ADs. Spectral, spatial and directional features are concatenated along frequency axis. The attention weight is multiplied to all spatial and directional features before feeding them to the upper layers. The neural network weights might go large during the training in case that the input features are attenuated by the attention weights, so we add L2 regularization on the input layer and several preceding layers of the separation network, respectively.

3. Training Paradigms

3.1. Permutation Invariant Training

Formulated the speech separation in T-F domain, the T-F representation of a C-mixed speech mixture can be interpreted as $\mathbf{Y}(t,f) = \sum_{c=1}^{C} \mathbf{X}_c(t,f)$, where $\mathbf{X}_c(t,f)$ is the complex spectrogram of speaker c. Under supervised learning framework, when the network has multiple outputs, DNN-based speech separation systems often suffer from label permutation problem [7]. It's difficult to assign reference labels for the network outputs, and the order of which can be arbitrary. Permutation Invariant Training (PIT) [7, 18] tackled this problem by calculating spectrogram estimation errors between all pairs ($C!$) of reference signals and estimated signals and always choosing the minimum error for backpropagation:

$$\mathcal{L}_{PIT} = \min_{\rho \in \mathcal{P}} \sum_{c=1}^{C} \sum_{t} \left\| \left(\hat{\mathbf{M}}_{\rho(c)}(t) - \mathbf{M}_c(t) \right) \circ |\mathbf{Y}(t)| \right\|_2^2 \quad (5)$$

where $\mathcal{P}$ contains all possible permutations for $C!$ output order, M and $\hat{M}$ are ideal and estimated T-F mask respectively, and $|\mathbf{Y}|$ is the mixture magnitude spectrogram. Eq. 5 is known as a spectrum approximation (SA) loss, which is commonly used in speech separation and enhancement tasks [19].

3.2. Target Extraction Training

For target extraction training, each output of network is bounded to a particular target speaker i:

$$\mathcal{L}_{TGT} = \sum_{t} \left\| \left(\hat{\mathbf{M}}_i(t) - \mathbf{M}_i(t) \right) \circ |\mathbf{Y}(t)| \right\|_2^2 \quad (6)$$

PIT is used for multi-talker speech separation when there is no input directional feature. While for single target extraction, we exploit target extraction training (TET). When there are more than one targets, the order of directional features can indicate the output order of target speakers. Two types of inputs are investigated: 1) *tgt*: only the target speaker's directional features are provided for the separation network; 2) *tgt & int*: both target and interference speakers' directional features are used as the input. If there are more than one interference speaker, their directional features will be averaged as one interference input. The reason we add the directional features for the interference speaker is that, given distinct interference information as hints, the network will learn to have more confidence to filter out interference from a latent direction.

4. Experiments and Results

4.1. Dataset simulation and feature extraction

We simulated a spatialized reverberant dataset derived from Wall Street Journal 0 (WSJ0) 2-mix corpus, which is an open and well-studied dataset used in monaural and multi-channel speech separation [5, 7, 20, 21, 22]. There are 20,000, 5,000 and 3,000 multi-channel, reverberant, two-speaker mixed speech in training, development and test set respectively. The performance evaluation is all done on test set, the speakers in which are all unseen during training. We consider a 6-microphone circular array of 7cm diameter with speakers and the microphone array randomly located in the room. The two speakers and microphone array are on the same plane and all of them are at least 0.3m away from the wall. The image method [23] is employed to simulate RIRs randomly from 3000 different room configurations with the size(length-width-height) ranging from 3m-3m-2.5m to 8m-10m-6m. The reverberation time T60 is sampled in a range of 0.05s to 0.5s. Samples with angle difference of 0°-15°, 15°-45°, 45°-90°and 90°-180°respectively account for 16%, 29%, 26% and 29% in the dataset.

For short time Fourier transform (STFT) setting, the window size is 32 ms and the hop size is 16 ms. 512-point FFT is used to extract 257-dimensional LPS. The LPS is computed from the first channel waveform of speech mixture. IPDs are extracted between microphone pairs (1, 4), (2, 5), (3, 6), (1, 2), (3, 4) and (5, 6). These pairs are selected considered that the distance between each pair is either the furthest or nearest. For DPR and DSNR computation, we use 36 fixed spatial filters and the p-th filter is steered at azimuth $10p°$.

4.2. Network structure and evaluation metrics

All the methods share the same network configuration, containing 3 LSTM layers each with 512 nodes, followed by a 512-node fully-connected layer. The output layer consists of 257 nodes for each output speaker. Batch size is set as 64. Adam optimizer is utilized in training. We always use the whole utterance during training and evaluation [18]. For attention experiments, the regularization coefficient is set as 2e-5 empirically.

Results are evaluated on two metrics: scale-invariant signal-to-noise ratio improvement (SI-SNRi), which is commonly used in recent speech separation tasks [5, 7, 20, 24] and signal-to-distortion rate improvement (SDRi) computed with MATLAB bss_eval toolbox [25]. The reverberant speech of each source is used as reference to compute the metric. The performances are evaluated under different range of ADs between speakers.

Table 1: *SDRi (dB) and SI-SNRi (dB) performances of target separation systems on far-field WSJ0 2-mix.*

# of target	Features & Setup	Training loss	SI-SNRi (dB)					SDRi (dB)
			<15°	15°-45°	45°-90°	>90°	Ave.	
2	LPS	PIT	4.71	5.35	5.15	5.30	5.18	5.62
2	LPS, 6IPD	PIT	3.00	6.71	7.90	8.20	6.88	7.35
2	LPS, 6IPD, DPR (2-tgt)	TET	3.31	8.08	9.09	9.38	7.98	8.42
2	LPS, 6IPD, DSNR (2-tgt)	TET	2.20	7.87	8.91	9.21	7.72	8.07
2	LPS, 6IPD, DPR + DSNR (2-tgt)	TET	2.83	8.34	9.28	9.56	8.06	8.49
2	LPS, 6IPD, DPR + DSNR + AF (2-tgt)	TET	4.87	8.77	9.71	10.04	**8.78**	**9.17**
1	LPS, 6IPD, DPR + DSNR (tgt)	TET	0.70	7.32	8.75	9.10	7.27	7.35
1	LPS, 6IPD, DPR + DSNR (tgt & int)	TET	2.08	8.90	9.81	10.02	8.49	8.87
1	LPS, 6IPD, DPR + DSNR + AF (tgt)	TET	4.83	8.92	9.91	10.26	8.93	9.21
1	LPS, 6IPD, DPR + DSNR + AF (tgt & int)	TET	4.84	9.17	10.15	10.50	**9.14**	**9.56**

Table 2: *SI-SNRi (dB) performance of two target speaker separation with attention mechanism on far-field WSJ0 2-mix.*

Features	Setup	SI-SNRi (dB)				
		<15°	15°-45°	45°-90°	>90°	Ave.
LPS, 6IPD	-	3.00	6.71	7.90	8.20	6.88
LPS, 6IPD	wL_{PIT}	2.70	4.19	5.20	5.78	4.69
LPS, 6IPD	Learnable att.+all LSTM L2 reg.	3.31	6.59	8.03	8.17	6.92
LPS, 6IPD	Fixed att.+all LSTM L2 reg.	3.51	6.52	7.79	8.37	6.93
LPS, 6IPD	Fixed att.+1st LSTM L2 reg.	**3.82**	**7.14**	**8.09**	**8.34**	**7.22**
LPS, 6IPD, DPR+DSNR (2-tgt)	-	2.83	8.34	9.28	9.56	8.06
LPS, 6IPD, DPR+DSNR (2-tgt)	Fixed att.+all LSTM L2 reg.	3.31	8.21	9.30	9.59	8.17
LPS, 6IPD, DPR+DSNR (2-tgt)	Fixed att.+1st LSTM L2 reg.	**3.99**	**8.77**	**9.54**	**9.73**	**8.51**

4.3. Results and analysis

Target separation with directional features. Table 1 reports the SI-SNRi and SDRi results for systems with three input setups: 1) spectral feature only (LPS); 2) spectral and spatial features (LPS, 6IPD); 3) spectral, spatial and directional features.

The single-channel 2-speaker separation network performs poorly with only SI-SNRi of 5.18dB. Adding IPDs elevates the overall performance to 6.88dB, especially under large AD. Directional features (AF, DPR and DSNR) of two target speakers further improve the performance to 8.06dB. Furthermore, different directional features DPR, DSNR and AF are combined and achieves 8.78dB.

We also trained a single target speaker network to separate the speaker of interest. By running the single target separation network twice, each time selecting one speaker in mixture signal as target, a better performance is achieved compared to 2-speaker separation network (8.49dB v.s. 8.06dB for DPR+DSNR and 8.78dB v.s. 9.14dB for DPR+DSNR+AF). With only target directional features provided, the performance drops 0.2dB compared to that when both speakers' directional features are available (8.93dB v.s. 9.14dB for DPR+DSNR+AF). However, this architecture does not necessarily require interference speaker's directional information, which makes it more practical in reality.

Spatial overlap issue. All experiments in Table 1 used PIT while training to avoid the speaker ambiguity when directional features are less discriminative in small AD. Comparing to the separation system with spectral feature only, the performance of systems with spatial and directional features degrades when the AD is smaller than 15°. The results of the proposed attention methods are summarized in Table 2, and L2 regularization for first LSTM layer and all LSTM layers are also evaluated.

To verify that the performance degradation for smaller AD is not introduced by the small proportion of data in training set, we increased the weight of small angle data in loss function by four times and no improvement was achieved in small AD (wL_{PIT}). For attention experiments, we firstly fixed the attention, empirically set w to 0.5 and b to 10. With LPS and 6IPD as input feature, fixed attention and regularization can boost the performance from 3.00dB to 3.82dB for small AD, adding regularization term only on the first LSTM layer is better than add on all LSTM layers (3.82dB v.s. 3.51dB). As we expect, with trainable attention, the network learned reasonable parameters as $w = 0.9$, $b = 9.6$ and achieves comparable result with fixed attention. For systems with directional features, the fixed attention methods also boost performances for all of AD ranges.

5. Conclusion

In this paper, directional information is used to separate the target voice given its direction. Two directional features are designed and incorporated with spatial and spectral features to provide more complementary information for training our separation network. Furthermore, an attention mechanism is proposed to improve performance when multiple speakers are closely located. Experimental results on WSJ0 2-mix validate the effectiveness of our proposed neural spatial filter. In future, we will generalize this work to more mixed speakers condition.

6. Acknowledgements

This paper was partially supported by Shenzhen Science & Technology Fundamental Research Programs (No: JCYJ20170817160058246 & JCYJ20180507182908274).

7. References

[1] C. Cherry and J. A. Bowles, "Contribution to a study of the cocktail party problem," *Journal of the Acoustical Society of America*, vol. 32, no. 7, pp. 884–884, 1960.

[2] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[3] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[4] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 11, p. 569270, 2003.

[5] 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.

[6] 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.

[7] 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.

[8] K. Zmolikova, 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.

[9] J. Xu, J. Shi, G. Liu, X. Chen, and B. Xu, "Modeling attention and memory for auditory selection in a cocktail party environment," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[10] J. Wang, J. Chen, D. Su, L. Chen, M. Yu, Y. Qian, and D. Yu, "Deep extractor network for target speaker recovery from single channel speech mixtures," *arXiv preprint arXiv:1807.08974*, 2018.

[11] 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," *arXiv preprint arXiv:1810.04826*, 2018.

[12] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 558–565.

[13] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[14] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.

[15] Z. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 457–468, 2019.

[16] L. Chen, M. Yu, D. Su, and D. Yu, "Multi-band pit and model integration for improved multi-channel speech separation," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[17] J. Benesty, "Study and design of differential microphone arrays," *Springer Topics in Signal Processing*, vol. 6, 2013.

[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] 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.

[20] 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.

[21] Z. Chen, J. Li, X. Xiao, T. Yoshioka, H. Wang, Z. Wang, and Y. Gong, "Cracking the cocktail party problem by multi-beam deep attractor network," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 437–444.

[22] Z. Chen, T. Yoshioka, X. Xiao, L. Li, M. L. Seltzer, and Y. Gong, "Efficient integration of fixed beamformers and speech separation networks for multi-channel far-field speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5384–5388.

[23] 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.

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

[25] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.