INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Direction-aware Speaker Beam for Multi-channel Speaker Extraction

Guanjun Li[1,2], *Shan Liang*[1], *Shuai Nie*[1], *Wenju Liu*[1], *Meng Yu*[3], *Lianwu Chen*[4], *Shouye Peng*[5], *Changliang Li*[6]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[3]Tencent AI Lab, Bellevue, WA, USA
[4]Tencent AI Lab, Shenzhen, China
[5]Xueersi Online School, China
[6]KingSoft AI Lab, China

{guanjun.li, sliang, shuai.nie, lwj}@nlpr.ia.ac.cn, {raymondmyu, lianwuchen}@tencent.com
pengshouye@100tal.com, lichangliang@kingsoft.com

Abstract

SpeakerBeam is a state-of-the-art method for extracting a speech signal of target speaker from a mixture using an adaption utterance. The existing multi-channel SpeakerBeam utilizes the spectral features of the signals with the ignorance of the spatial discriminability of the multi-channel processing. In this paper, we tightly integrate spectral and spatial information for target speaker extraction. In the proposed scheme, a multi-channel mixture signal is firstly filtered into a set of beamformed signals using fixed beam patterns. An attention network is then designed to identify the direction of the target speaker and to combine the beamformed signals into an enhanced signal dominated by the target speaker energy. Further, SpeakerBeam inputs the enhanced signal and outputs the mask of the target speaker. Finally, the attention network and SpeakerBeam are jointly trained. Experimental results demonstrate that the proposed scheme largely improves the existing multi-channel SpeakerBeam in low signal-to-interference ratio or same-gender scenarios.

Index Terms: speaker extraction, multi-channel signal processing, fixed beamforming, jointly training

1. Introduction

Given a speech recording of multiple speakers talking at the same time, extracting a speech signal of a target speaker is desired for numerous applications such as meeting recognition systems and home driven devices. Although humans can easily perform this task, it is still challenging to build an effective system for the machine to model this process [1].

An effective solution is to use speech separation methods whose purpose is to recover all signals in the mixed signal. Traditional speech separation methods mainly include nonnegative matrix factorization (NMF) [2] , independent component analysis (ICA) [3] and spatial clustering [4, 5]. In recent years, the deep learning-based speech separation methods represented by permutation invariant training (PIT) [6, 7], deep clustering (DC) [8, 9] and deep attractor network (DAN) [10] also made significant progress. However, these speech separation methods require the number of speakers in advance, which makes these methods difficult to apply in many real-world scenarios where the number of speakers cannot be accurately estimated.

In contrast to speech separation, speaker extraction is independent of the number of speakers. It only extracts the target speaker from a mixed signal. One solution for speaker extraction is to use a traditional multi-channel beamforming technique that extracts the target signal based on the direction of arrival (DOA) of the target speaker. However, accurately estimating the target DOA in the presence of multiple speakers is still a challenging task.

With the booming of the deep learning, another speaker extraction method is SpeakerBeam [11, 12, 13, 14], where the deep neural network (DNN) is informed about the target speaker from an adaption utterance — a speech segment only containing the target speaker. Thus, DNN can output the mask corresponding to the target speaker using the spectral cues. Although a multi-channel based SpeakerBeam has been proposed [11, 12, 14], DNN only estimates the target mask separately on each channel and these masks are then combined using a median operation to obtain an overall mask. The existing multi-channel SpeakerBeam has not taken advantage of the spatial discriminability of multi-channel signals, which can distinguish signals from different directions. Therefore, multi-channel spatial discriminability can be leveraged for better estimation of the target mask, as speaker sources are directional and usually spatially separated in actual environment.

In this paper, we combine the merits of beamforming and SpeakerBeam so that the spatial and spectral information can be tightly integrated to achieve better speaker extraction performance. More specifically, we first apply fixed beamforming on the observed signals to generate 12 fixed beams, equally sampled in space. When speakers are fully separated in space, there will be a target beam in which the target speaker's energy dominates. Except for the target beam, the remaining 11 beams can be viewed as non-target beams, where the energy of the non-target speaker or background is dominant. Furthermore, we design an attention network, which selects the target beam and non-target beams based on the adaption utterance. Based on the weights of the attention network output, we weight-sum these 12 beams. This weighted summation operation can further eliminate the energy of the non-target speaker from the target beam, and obtain an enhanced single-channel spectrum, which is very similar to the traditional generalized sidelobe canceller (GSC) [15] structure. The enhanced single-channel spectrum is then sent to SpeakerBeam to estimate the target speaker's mask. Finally, the attention network and SpeakerBeam are

http://dx.doi.org/10.21437/Interspeech.2019-1474

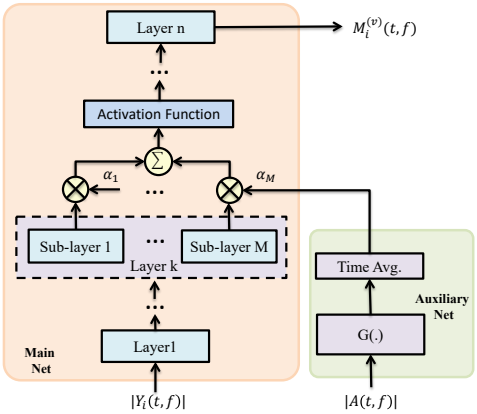
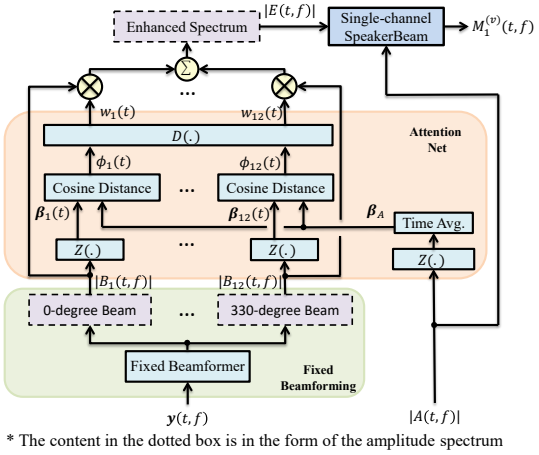


Figure 1: *Single-channel SpeakerBeam.*



* The content in the dotted box is in the form of the amplitude spectrum

Figure 2: *Proposed multi-channel SpeakerBeam scheme.*

jointly trained. Since the enhanced spectrum obtained by the attention network has a higher signal-to-interference ratio (SIR) than the spectrum of the microphone signal, SpeakerBeam can achieve a better mask estimation performance. It is worth noting that the adaption utterance guides not only SpeakerBeam to estimate the mask, but also the attention network to weight 12 fixed beams. Thanks to the attention network, the proposed scheme does not require the DOA of the target speaker. We evaluate the proposed scheme on a spatialized reverberant version of the wsj0-2mix corpus [8]. Larger improvement has been achieved compared to the existing multi-channel SpeakerBeam.

The remainder of the paper is organized as follows. In Section 2, we summarize the structure of SpeakerBeam. Section 3 describes the proposed scheme. Section 4 discusses relations with prior works. We then report experimental results in Section 5 and draw the conclusion in Section 6.

2. SpeakerBeam

In this section, we review the single-channel and multi-channel based SpeakerBeam structures proposed in [11, 12, 13, 14].

2.1. Problem formulation

The mix signal $Y_i(t,f)$ received at the i-th microphone can be modeled in the short-time Fourier transform (STFT) as, $Y_i(t,f) = S_i(t,f) + N_i(t,f)$, where $i = 1 \dots I$ is the index of the microphone, $S_i(t,f)$ is the reverberant speech signal corresponding to the target speaker, $N_i(t,f)$ is the interference signal containing non-target speakers and background noise and t and f denote the time and frequency indices, respectively. SpeakerBeam aims to estimate the target speaker mask from $\mathbf{y}(t,f)$, where $\mathbf{y}(t,f) = [Y_1(t,f), Y_2(t,f), \cdots, Y_I(t,f)]^T$ and $(\cdot)^T$ denotes the transposition operator.

2.2. Single-channel SpeakerBeam

When the number of microphones $I = 1$, we can use a single-channel SpeakerBeam [13], whose structure is shown in Figure 1. The single-channel SpeakerBeam consists of two parts: one main network and one auxiliary network.

The main network is used to estimate the target speaker's mask. In order to effectively inform the main network about the target speaker, the k-th layer in the main network is factorized into M sub-layers. The output of the k-th layer is obtained by weighted combination of the outputs of M sub-layers. The weight vector $\boldsymbol{\alpha} = [\alpha_1, \alpha_2 \dots, \alpha_M]^T$ corresponding to M sub-layers can also be considered as the identity feature of the target speaker.

The weight vector $\boldsymbol{\alpha}$ corresponding to the target speaker is extracted from the auxiliary network which operates on the adaption utterance $A(t,f)$, $\boldsymbol{\alpha} = \frac{1}{T_A}\sum_{t=1}^{T_A} G\left(|A(t,f)|\right)$, where T_A is the length of the adaption utterance and $G(\cdot)$ is a DNN whose input is the amplitude spectrum of the adaption utterance, $|A(t,f)|$.

The outputs of SpeakerBeam are a target speaker mask $M_i^{(S)}(t,f)$ and an interference mask $M_i^{(N)}(t,f)$. The main and auxiliary networks are jointly trained using ideal binary masks (IBM) as targets.

2.3. Multi-channel SpeakerBeam

Multi-channel SpeakerBeam [11, 12, 14] was proposed when $I > 1$. It utilizes single-channel SpeakerBeam to estimate masks on each channel separately, and the final mask is obtained by a median operation, i.e., $M^{(v)}(t,f) = \underset{i}{\text{median}}\left(M_i^{(v)}(t,f)\right)$, where $v \in \{S, N\}$.

3. Proposed multi-channel scheme

In the actual environment, the signals of different speakers often come from different directions. However, in estimating the final mask, the multi-channel SpeakerBeam mentioned in Section 2.3 does not take advantage of the spatial discriminability of multi-channel signals, which can distinguish signals from different directions.

In this section, we introduce our proposed multi-channel scheme (see Figure 2), which can identify the direction of the target speaker to further enhance the performance of SpeakerBeam. It consists of two parts, including a fixed beamforming part and an attention network.

3.1. Fixed beamforming

Given a microphone array, we first design 12 fixed beamformers whose beam patterns are targeted at 12 directions, which are uniformly sampled in space. We use these 12 fixed beamformers to spatially filter the observed multi-channel signals to obtain 12 beams, which are 0° beam, 30° beam, and so on. Each beam is represented by $B_j(t,f)$, where $j = 1, 2, \dots, 12$.

3.2. Attention network

Suppose that each speaker is at least 30 degrees apart in space (the more fixed beamformers are in Section 3.1, the smaller this angle is). Among the 12 beams, there will be a target beam in which the energy of the target speaker is dominant. Except for the target beam, the remaining 11 beams are non-target beams, where the energy of the interference is dominant.

We then design an attention network that uses the target speaker's adaption utterance $A(t,f)$ to select the target beam and the non-target beams from the 12 beams without the target DOA information. In the attention network, a DNN $Z(\cdot)$ is used to extract speaker identity features in different beams and the adaption utterance, respectively. It is worth noting that the output of $Z(\cdot)$ and the output of $G(\cdot)$ in SpeakerBeam can both characterize the speaker, but they guide the network in different forms. Therefore, $Z(\cdot)$ and $G(\cdot)$ are two different networks. We expect that the 12 beams can be mapped to the corresponding frame-level identity features by $Z(\cdot)$ and the frame-level identity features obtained by the transformation of the adaption utterance through $Z(\cdot)$ will undergo a time-averaging operation to obtain the utterance-level identity feature, i.e.,

$$\boldsymbol{\beta}_j(t) = Z(|B_j(t,f)|), \quad j = 1, 2, \ldots, 12, \tag{1}$$

$$\boldsymbol{\beta}_A = \frac{1}{T_A}\sum\nolimits_{t=1}^{T_A} Z(|A(t,f)|), \tag{2}$$

where $\boldsymbol{\beta}_j(t)$ is the frame-level identity feature vector of the j-th beam and $\boldsymbol{\beta}_A$ is the utterance-level identity feature vector of the adaption utterance.

Since the energy of the target speaker in the target beam is dominant, we believe that the target beam and the adaption utterance should be similar to the identity features extracted by $Z(\cdot)$, and the identity features of the non-target beams are as different as the identity feature of the adaption utterance. The cosine distance $\phi_j(t)$ is used to score the similarity between $\boldsymbol{\beta}_j(t)$ and $\boldsymbol{\beta}_A$,

$$\phi_j(t) = \frac{\boldsymbol{\beta}_j(t)\boldsymbol{\beta}_A}{|\boldsymbol{\beta}_j(t)||\boldsymbol{\beta}_A|}, \quad j = 1, 2, \ldots, 12. \tag{3}$$

Inspired by the traditional GSC [15] structure, the 12 similarity scores are input into a DNN $D(\cdot)$ to obtain the weight corresponding to each beam, and then we weight the 12 beams to obtain an enhanced amplitude spectrum $|E(t,f)|$, i.e.,

$$\mathbf{w}(t) = D\left(\phi_1(t), \phi_2(t), \ldots, \phi_{12}(t)\right), \tag{4}$$

$$|E(t,f)| = \mathbf{w}^T(t)\mathbf{B}(t,f), \tag{5}$$

where, $\mathbf{w}(t) = [w_1(t), w_2(t), \ldots, w_{12}(t)]^T$ is a vector containing the weights of the 12 beams and $\mathbf{B}(t,f) = [|B_1(t,f)|, |B_2(t,f)|, \ldots, |B_{12}(t,f)|]^T$ is a vector containing the amplitude spectrum of the 12 beams. Unlike GSC, the proposed attention network will identify the target beam itself without the need for DOA information.

The enhanced amplitude spectrum $|E(t,f)|$ and the adaption utterance's amplitude spectrum $|A(t,f)|$ are further input to the single-channel SpeakerBeam mentioned in Section 2.2 to get the final mask. During the training phase, the attention network and SpeakerBeam are jointly trained using IBM of the first channel as the training target (although it may be better to use the target speaker's spectrum as the training target [6, 7], we use IBM to be consistent with [11, 12, 13, 14]). In the test phase, in order to reduce spectral distortion, we use a mask-based minimum variance distortionless response (MVDR) beamformer [16] to obtain the target speaker's signal.

4. Relation to prior work

To the best of our knowledge, the proposed scheme is the first to combine the spatial discriminability of multi-channel signals with spectral features for speaker extraction.

There are some methods [17, 18, 19, 20] that combine spatial features with spectral features in terms of speech separation task different from speaker extraction task. In [17, 18], similar to the part of the proposed scheme, a set of fixed beamformers is used. [17] performs speech separation on each beam, which suffers from high computational cost. In [18], the target beam is selected for speech separation without combining all the beams as in the attention network we proposed, and the network for selecting the target beam and the speech separation network cannot be jointly trained. [19, 20] add a directional feature to the input of the speech separation network to improve performance, but this directional feature relies on an accurate estimate of each speaker's mask, which is difficult to guarantee.

5. Experiments

5.1. Data

To evaluate the proposed scheme, we convolved the room impulse responses (RIRs) with the utterances in the wsj0-2mix data of 8 kHz [8], which contains 20,000 training, 5,000 validation and 3,000 test single-channel two-speaker mixtures. Image method [21, 22] was used to create the RIRs with a circular microphone array with 8 microphones, 20 cm diameter and moderate reverberation (about 0.2s). The speakers were randomly located in angles from 0° to 360°. For any two speakers, we constrained them to be at least 90° apart. We mixed the images of two speakers with SIR uniformly drawn from -5 dB to 5 dB in the training and the validation sets, and we split the test set into 5 subsets varying only the SIR between -15 dB to 5 dB. Besides, the speakers in the test set did not appear in the training and the validation sets. For each mixture, we randomly chose an adaptation utterance from the target speaker (different from the utterance in the mixture). The average length of the adaptation utterance is 10 s.

5.2. Settings

We utilized a superdirective beamformer [23] as the fixed beamformer we mentioned in Section 3.1, because compared to other beamformers, such as delay-and-sum beamformer, the superdirective beamformer achieves a higher directivity [24].

The STFT frame size was 64 ms with 75 % overlap. The structure of the single-channel SpeakerBeam in the proposed scheme was the same as that of [13] with $M = 30$. $Z(\cdot)$ in the attention network consisted of three fully connected layers, i.e., two layers with a ReLU [25] activation and one layer with a linear activation. The number of neurons in the three layers was 256-128-64, respectively. $D(\cdot)$ in the attention network had two fully connected layers of neurons 24 and 12 respectively. The first layer used a ReLU activation function and the second layer used a linear activation function.

Before jointly training the attention network and SpeakerBeam, we first pre-trained the attention network to minimize the mean-square error (MSE) w.r.t the true target amplitude spectrum. We found experimentally that the pre-training is essential to the network convergence. All the models were trained using the ADAM optimizer [26].

We used signal-to-distortion ratio (SDR) [27] and cepstral distance (CD) [28] as our performance measures for the exper-

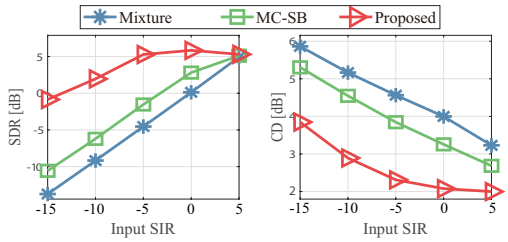


Figure 3: *SDR (left) and CD (right) as a function of the input SIRs for the test set.*

Table 1: *Results for the test set in different mixing conditions*

	Different gender SDR / CD	Same gender SDR / CD	All SDR / CD
Mixture	-4.42 / 4.55	-4.58 / 4.58	-4.50 / 4.57
OMVDR	4.58 / 1.97	4.89 / 1.98	4.73 / 1.97
MC-SB	1.04 / 3.40	-5.77 / 4.55	-2.09 / 3.93
Proposed	**3.52 / 2.58**	**3.47 / 2.68**	**3.50 / 2.63**

iments. The larger SDR, the better the performance, but CD is the opposite.

5.3. Results and discussions

We compared the proposed scheme with the multi-channel SpeakerBeam (MC-SB) mentioned in Section 2.3 instead of the speech separation method, because our task is to perform speaker extraction when the number of speakers is unknown. Figure 3 shows the SDR and CD for the test set as a function of the input SIRs. We can see from Figure 3 that the lower the SIR is, the more advantageous the proposed scheme is. In the low SIR scenario, the interference masks the target speaker. Thus, it is difficult for MC-SB to capture the spectral cues of the target speaker in the mixture. However, the proposed scheme overcomes this problem by first using the attention network to enhance the signal from the target direction. With the increase of SIR, the proposed scheme is still superior to MC-SB on CD, but close to the mixture on SDR, which may be due to the inaccurate mask estimation which causes great damage to the target signal.

Table 1 summarizes the results for the test set in the same-gender and different-gender scenarios. In this experiment, we added an oracle MVDR beamformer (OMVDR) deriving from IBM. It can be seen from Table 1 that MC-SB almost fail in the case of the same-gender scenario, which was also reported in the experiment of [12]. However, the proposed scheme is not sensitive to the gender of the speakers in the mixture, because it does not rely entirely on the spectral cues. Besides, in each scenario, the proposed scheme performs better than MC-SB, which again confirms that spatial information is an important cue in speaker extraction. OMVDR leads to around 1.2 dB improvement to the proposed scheme on SDR, owing to using the oracle data, which indicates that the proposed scheme still has room for improvement.

Figure 4 shows an example of speaker extraction. The mixture contains two female speakers. It can be seen from Figure 4(c) and Figure 4(d) that MC-SB fail at this time, but the proposed scheme can still obtain a more accurate target speaker's mask. Figure 4(e) and Figure 4(f) show the intermediate output of the proposed network. From Figure 4(b) and

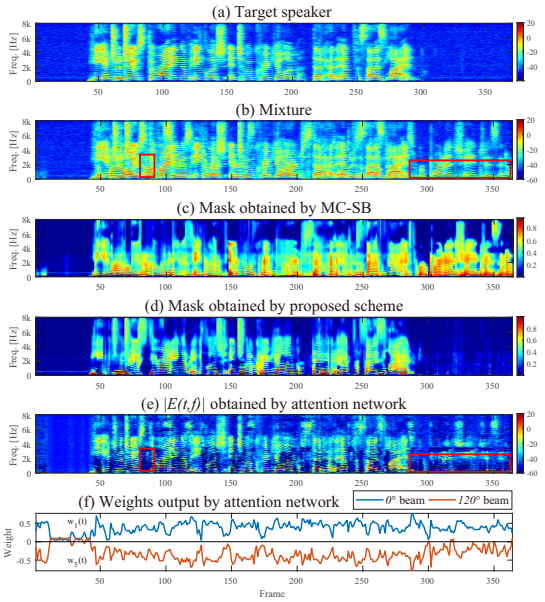


Figure 4: *An example of speaker extraction. The mixture contains two female speakers with SIR = 0 dB. The target speaker is at* 0° *and the interference is at* 120°*. (a)-(b) Spectrograms. (c)-(d) Estimated masks. (e)-(f) Intermediate outputs of the proposed network.*

Figure 4(e), especially in the red box, we see that the attention network allows $|E(t,f)|$ in Eq. (5) to obtain more suppression of interference energy than the spectrum of the mixture. This is also confirmed in Figure 4(f), where the weight $w_1(t)$ of the target beam (0° beam) is positive for almost all frames and the weight $w_5(t)$ of the non-target beam (120° beam), where the interference energy dominates, is the opposite. According to Eq. (5), this will further suppress the interference energy in the target beam, which helps SpeakerBeam to more accurately utilize spectral features when estimating the target mask.

6. Conclusions

In this paper, we combined complementary spectral and spatial information for multi-channel speaker extraction. The attention network in the proposed scheme can identify the direction of the target speaker according to an adaption utterance and enhance the signal of the target direction, thereby helping SpeakerBeam to estimate the target mask from the spectral cues. The performance of the proposed scheme is particularly prominent in low SIR or same-gender scenarios. In the future, we plan to utilize the more accurate speaker representation to guide the attention network and SpeakerBeam to track the target speaker.

7. Acknowledgements

We would like to thank Dr. Dong Yu from Tencent AI Lab for constructive comments. This work was supported in part by the National Key R&D Plan of China (No. 2016YFB1001404) and the China National Nature Science Foundation (No. 61573357, No. 61503382, No. 61403370, No. 61273267, No. 91120303). This work was sponsored by CCF-Tencent Open Fund (No. RAGR20180106) and Xueersi Cooperation Fund.

8. 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] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. John Wiley & Sons, 2004, vol. 46.

[4] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Learning speaker representation for neural network based multichannel speaker extraction," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 8–15.

[12] 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.

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

[14] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, T. Nakatani, and J. Černockỳ, "Optimization of speaker-aware multichannel speech extraction with asr criterion," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6702–6706.

[15] K. Buckley and L. Griffiths, "An adaptive generalized sidelobe canceller with derivative constraints," *IEEE Transactions on antennas and propagation*, vol. 34, no. 3, pp. 311–319, 1986.

[16] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5210–5214.

[17] 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.

[18] 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.

[19] 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.

[20] ——, "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.

[21] 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.

[22] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[23] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone arrays*. Springer, 2001, pp. 19–38.

[24] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2013.

[25] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] 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.

[28] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2008.