INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Recursive Sound Source Separation with Deep Learning-based Beamforming for Unknown Number of Sources

Hokuto Munakata, Ryu Takeda, Kazunori Komatani

SANKEN, Osaka University, Osaka, Japan

h_munakata@ei.sanken.osaka-u.ac.jp, {rtakeda,komatani}@sanken.osaka-u.ac.jp

Abstract

We propose a recursive separation model for an unknown number of sound sources based on deep learning-based beamforming. Recursive separation models have been investigated as a way to separate a mixture signal composed of an unknown number of sources in a single-channel condition. The mixture signal is separated with residual information in a recursive manner. Although the recursive separation model can be extended to a multi-channel condition using a beamforming-based filter, the separation performance is degraded because the beamforming-based filter tends to accumulate estimation errors in the recursions. To address this problem, we introduce a local Gaussian model (LGM)-based recursive separation model. The proposed method mitigates the accumulation of errors by reusing estimated parameters and applying only one filter to the mixture signal. Experimental results show that our proposed method outperforms a separation model using an accumulative filter.
Index Terms: Supervised sound source separation, Beamforming, Unknown number of sources

1. Introduction

Sound source separation segregates source signals from an observed mixture signal. This technology is a fundamental technology for machine listening, for example, automatic speech recognition (ASR) or sound event detection [1–3]. Owing to the development of deep learning, supervised sound source separation methods have shown powerful separation performance [4–9]. These methods train a separation model using a deep neural network (DNN) with pairs of isolated source signals and their mixture signals. The separation model learns spectral patterns of the sources from a large amount of training data.

When the observed mixtures are multi-channel signals, beamforming using outputs of the DNN has been investigated [7–10]. Linear filtering based on beamforming mitigates nonlinear distortion and improves the word error ratio of ASR systems [7]. In particular, separation models based on the local Gaussian model (LGM) are promising for their expandability [11]. LGM introduces a generative model of the source signal. The separation is formulated as an estimation of the multi-channel Wiener Filter (MWF) based on the generative model. The generative model of LGM is widely used, for example, in joint separation and dereverberation [12, 13] and unsupervised training [14, 15]. The separation model with DNN is trained to estimate spatial covariance matrices (SCMs) and power spectral densities (PSDs) in order to calculate the MWF.

Most deep learning-based separation methods, including LGM-based beamforming, assume that the number of sound sources is given, although it is actually unknown in most cases of real environments. A straightforward approach is to train the models for each number of sources after estimating the number of sources by source counting methods [16–19]. However, this approach has two problems. First, it requires expensive training costs for individual training with each number of sources. Second, the maximum number of sources that can be separated is limited by the network architecture.

A recursive approach has been investigated to address these problems [20–22]. A separation model in this approach has two output layers for a single separated signal and residual information. After the separation, the mixture signal with residual information is input into the model to obtain the next outputs in a recursive manner. This approach uses a single model that can separate the mixture composed of an arbitrary number of sources by giving the number of sources. In addition, this approach simplifies source counting to a binary classification that determines whether there exists any source signal or not in the residual. This simplification makes it easier to determine a threshold of the source counting than activity detection-based counting methods [23–25].

Although recursive approaches have been proposed to estimate time-frequency (TF) masks for a single-channel condition, their possible extension to multi-channel using beamforming remains unclear. An accumulative filter approach (AFA), which applies multiple separation filters accumulatively to the mixture signal is effective for mask-based separation [21]. However, beamforming uses not only the amplitude but also the phase for the separation, and then beamforming of AFA tends to accumulate estimation errors and degrade performance.

In this paper, we propose a multi-channel recursive separation model for an unknown number of sources. We formulate a multi-channel mixture signal as the sum of source signals and a residual signal based on LGM for the recursive separation model. Based on this formulation, the separation model we propose estimates SCMs and PSDs for a source signal and a residual signal using the original mixture signal and the previous residual signal. The key idea of our proposed method is to reuse the estimated parameters. The parameters of a source signal and a residual signal are estimated in each recursion, and the MWFs are calculated using parameters up to their recursion.

The main contribution of this study is to extend the recursive separation model to the multi-channel condition. Although the AFA accumulates estimation errors and makes training more difficult, our proposed method applies a single filter to the mixture signal and then mitigates the accumulation. The experimental results with 2 to 4 source mixtures show that our proposed method outperformed multi-channel separation models with a given number of sources and with an unknown number of sources based on AFA.

10.21437/Interspeech.2023-488

2. Preliminary

In this section, we explain prior works, a deep learning-based supervised sound source separation with LGM for a given number of sources and recursive sound source separation for an unknown number of sources.

2.1. Deep learning-based Beamforming for Given Number of Sources

In this paper, all processing is in the TF domain. M-channel observed mixture signal $\mathbf{x}_{ft} \in \mathbb{C}^M$ is given as a sum of N target source signals $\mathbf{s}_{nft} \in \mathbb{C}^M$ $(n = 1, \ldots, N)$ in the TF domain as follows:

$$\mathbf{x}_{ft} = \sum_{n=1}^{N} \mathbf{s}_{nft}, \tag{1}$$

where $t = 1, \ldots, T$ and $f = 1, \ldots, F$ are time and frequency indices, respectively. LGM assumes that the source signal $\mathbf{s}_{nft}$ follows a prior complex Gaussian distribution:

$$\mathbf{s}_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \lambda_{nft}\mathbf{H}_{nf}\right), \tag{2}$$

where $\mathbf{H}_{nf} \in \mathbb{S}_{+}^{M\times M}$ is the SCM and $\lambda_{nft} \in \mathbb{R}_{+}$ is the PSD. The distribution of the source signals conditioned by the observed mixture signal is derived from Eq. (1) and (2) as

$$\mathbf{s}_{nft} \mid \mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{W}_{nft}\mathbf{x}_{ft}, \mathbf{V}_{nft}\right), \tag{3}$$

$$\mathbf{W}_{nft} = \lambda_{nft}\mathbf{H}_{nf}\left(\sum_{n=1}^{N}\lambda_{nft}\mathbf{H}_{nf}\right)^{-1}, \tag{4}$$

$$\mathbf{V}_{nft} = \left(\mathbf{I} - \mathbf{W}_{nft}\right)\lambda_{nft}\mathbf{H}_{nf}. \tag{5}$$

The separated signal of the MWF $\mathbf{y}_{nft} = \mathbf{W}_{nft}\mathbf{x}_{ft}$ can be seen as a maximum a posteriori (MAP) estimation of Eq. (3).

The pre-trained deep learning-based separation model estimates the parameters of the source signal (i.e., λ_{nft} and $\mathbf{H}_{nf}$) from the observed mixture signal $\mathbf{x}_{ft}$. While the DNN directly outputs estimated PSD λ_{nft}, the SCM is estimated via the estimation of TF masks $m_{nft} \in [0, 1]$. The TF mask represents the dominant source in each TF bin. The SCM is estimated approximately by TF masks as:

$$\mathbf{H}_{nf} = \sum_{t=1}^{T} m_{nft}\mathbf{x}_{ft}\mathbf{x}_{ft}^{\mathsf{H}}. \tag{6}$$

After this process, $\mathbf{H}_{nf}$ is normalized to $\mathrm{tr}(\mathbf{H}_{nf})$ by M for numerical stability.

In the training of the separation model, the loss is calculated with separated signals and reference source signals. We denote the n-th separated signal and reference source signal as $\mathbf{y}_n, \mathbf{s}_n \in \mathbb{C}^{F\times T\times M}$, respectively. In calculating the loss function, there is permutation ambiguity for the source index n, for example, the output order of the separated signal can be $2, 1$ while the output order of the reference is $1, 2$. Utterance-level permutation invariant training (uPIT) [5] solves this ambiguity using the minimum loss among $N!$ permutations as follows:

$$\mathcal{L}_S = \min_{\phi}\left(\sum_{n=1}^{N}\mathcal{L}(\mathbf{y}_n, \mathbf{s}_{\phi(n)})\right), \tag{7}$$

where ϕ indicates all permutations of the sources.

2.2. Recursive Sound Source Separation for Arbitrary Number of Sources

The separation model described above cannot separate mixture signals when the number of sources exceeds the number of output layers. To handle an arbitrary number of sources, a recursive separation model has been investigated [20, 21, 26]. This model separates mixture signals in recursions corresponding to the number of sources. The model has two output layers for a separated signal and residual information. The recursive sound source separation is formulated with the n-th residual information $\mathbf{r}_n$ as follows:

$$(\mathbf{y}_n, \mathbf{r}_n) = f(\mathbf{x}, \mathbf{r}_{n-1}; \Phi), \tag{8}$$

where $f(\cdot)$ is a transform of the separation model with parameter Φ. The residual information is used to convey the separation result of the previous recursion to the model. The observed mixture signal is separated into only one source signal and residual information, and then the model again separates the mixture with the residual information into another source and the next residual signal.

Previous recursive separation methods have been proposed for the single-channel condition. In this condition, the TF mask is directly used for the separation as $\mathbf{y}_{nft} = m_{nft}\mathbf{x}_{ft}$. In one study [20], the output mask is accumulated as the residual mask for the residual information. The residual mask is initialized with 1 and subtracted by the estimated mask of the source for each recursion in an accumulative manner. In contrast, residual information of LGM is difficult to define because the MWF based on LGM is composed of both the PSD and SCM. In another work [21], only the residual signal is used for the next recursion This method estimates accumulative filters and does not use residual signal and mixture signal simultaneously. Although this method can be applied to beamforming, the estimation errors are accumulated through the recursion.

2.3. Source Counting for Unknown Number of Sources

Counting the number of sources is required for the recursive separation model when the number of sources is unknown. Although conventional source counting methods estimate the number of sources directly from the input mixture, the recursive separation model simplifies it to binary classification through the recursions. In each recursion, a flag $c \in \{0, 1\}$ to stop the recursion is estimated. When there exists any source signal in the input residual information, c is 1, and otherwise, it is 0. This simplification of the source counting improves upon the source counting accuracy reported previously [21].

Previous works [20, 21, 26] use a DNN to estimate a stop flag. The model outputs an estimated stop flag $\hat{c} \in [0, 1]$ in each recursion. The loss function for the training is the binary cross entropy loss as follows:

$$\mathcal{L}_C = -c\log(\hat{c}) - (1 - c)\log(1 - \hat{c}). \tag{9}$$

The DNN for the counting is integrated into the separation DNN [20, 26] or independent [21].

3. Proposed

We propose a deep learning-based beamforming method with the recursive separation model for an unknown number of sources. The separation model is based on LGM with the residual signal.

3.1. Signal Model with Residual Signal

We assume that the mixture signal is composed of n source signals and a residual signal in n-th recursion similarly to Eq. (1):

$$\mathbf{x}_{ft} = \sum_{n'=1}^{n}\mathbf{s}_{n'ft} + \mathbf{r}_{nft}. \tag{10}$$

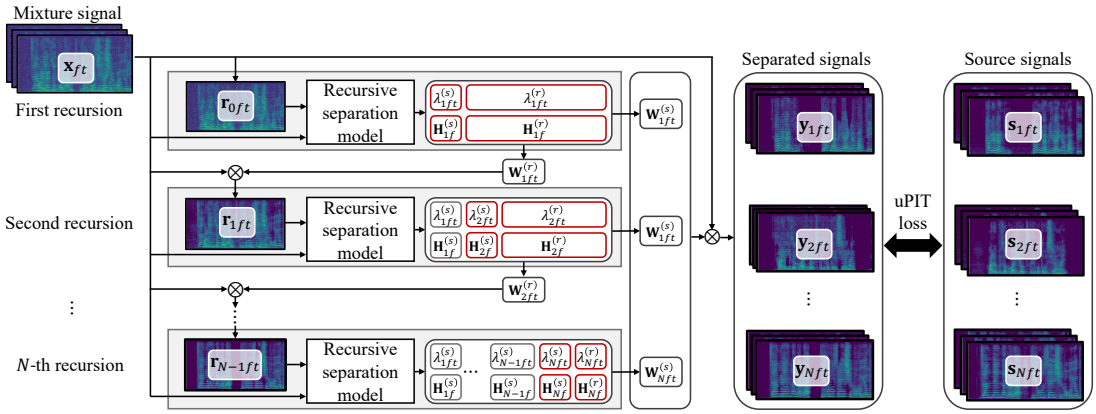

Figure 1: *Overview of training flow of proposed model. MWFs for the separated signal and resignal signal are estimated by each recursion. Parameters estimated in each recursion are encircled by red lines. Reused parameters are encircled by gray lines.*

In addition, we assume that the residual signal follows the complex Gaussian distribution as the same as Eq. (2):

$$\mathbf{r}_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \lambda^{(r)}_{nft}\mathbf{H}^{(r)}_{nf}\right), \tag{11}$$

where $\lambda^{(r)}_{nft}$ and $\mathbf{H}^{(r)}_{nf}$ are PSD and SCM for the residual signal, respectively. We denote $\lambda^{(s)}_{nft}$ and $\mathbf{H}^{(s)}_{nf}$ as the PSD and SCM for the n-th source signal. MWFs of the n-th source signal $\mathbf{W}^{(s)}_{nft}$ and for the residual signal $\mathbf{W}^{(r)}_{nft}$ are derived in the same way as Eq. (3):

$$\mathbf{W}^{(s)}_{nft} = \lambda^{(s)}_{nft}\mathbf{H}^{(s)}_{nf}\left(\left(\sum_{n'=1}^{n}\lambda^{(s)}_{n'ft}\mathbf{H}^{(s)}_{n'f}\right) + \lambda^{(r)}_{nft}\mathbf{H}^{(r)}_{nf}\right)^{-1},$$

$$\mathbf{W}^{(r)}_{nft} = \lambda^{(r)}_{nft}\mathbf{H}^{(r)}_{nf}\left(\left(\sum_{n'=1}^{n}\lambda^{(s)}_{n'ft}\mathbf{H}^{(s)}_{n'f}\right) + \lambda^{(r)}_{nft}\mathbf{H}^{(r)}_{nf}\right)^{-1}.$$

The separated signals of the MWFs correspond to the MAP estimation of n-th source signal and the residual signal.

3.2. Recursive Separation Model Based on LGM

We propose a recursive separation model based on LGM that estimates the PSDs and SCMs for the source signal and the residual signal in each recursion. An overview of the proposed model is shown in Figure 1. A key idea of our approach is to apply only one filter to the mixture signal by reusing estimated parameters. This approach mitigates the accumulation of errors.

In the n-th recursion, the model estimates the PSD and SCM for the n-th source signal $\lambda^{(s)}_{nft}$ and $\mathbf{H}^{(s)}_{nf}$, and for the residual signal $\lambda^{(s)}_{nft}$ and $\mathbf{H}^{(s)}_{nf}$, respectively. As input, our model uses the mixture signal and the residual signal of a previous recursion. The residual signal includes both information on the PSD and the SCM. The initial residual signal $\mathbf{r}_{0ft}$ in the first recursion is the mixture signal. To reduce calculation cost, the first channel of the residual signal is used because the multichannel information on the residual signal is similar to the observed mixture.

The separated signal and residual signal are obtained by the MWFs mentioned above in the training. The MWFs for the n-th separated signal and residual signal are calculated with parameters of the first to the n-th source signals $\lambda^{(s)}_{n'ft}$ and $\mathbf{H}^{(s)}_{n't}$, and n-th residual signal $\lambda^{(r)}_{nft}$ and $\mathbf{H}^{(r)}_{nf}$. The SCMs are calculated with the masked mixture signal based on Eq. (6). The training of the separation model is based on the uPIT using only the separated signals obtained in each recursion. By calculating the loss with separated signals using $\lambda^{(r)}_{nft}$ and $\mathbf{H}^{(r)}_{nf}$ $(n = 1, \ldots, N)$, the model learns to estimate the parameters for residual signals.

In contrast, the separated signals are obtained by the MWF based on Eq. (4) after all recursions in the test. In other words, all separated signals are obtained by the parameters used in the last recursion (described at the bottom of Fig. 1.) We experimentally confirmed that the separation performance improved by using the MWFs based on Eq. (4) compared to using the MWFs used in the training.

4. Experimental Evaluation

The proposed method was evaluated for both separation and source counting performances. In addition, we compared the performances among the number of sources of training data.

4.1. Dataset

We generated three datasets of multi-channel mixtures to evaluate the optimal number of sources composing the mixture for the training of our proposed method. Each dataset had 2 sources, 3 sources, and 2 to 3 sources mixtures. The number of sources for the mixture of 2 to 3 sources was set at a ratio of one to one. The mixture signals were generated as observations of four-channel microphone arrays. Each mixture consisted of multiple source signals randomly selected from the Japanese news article sentence corpus [1]. The source signals were generated by convoluting room impulse responses (RIRs) [27]. The array with random geometry was placed randomly around the center of a room having dimensions of 5 m $\times$ 5 m $\times$ 3 m and each source was located randomly. The angular difference between sources always had at least 15°. The reverberation time (RT_{60}) was fixed to 200 ms. The average power of the source signals was normalized randomly at a level between -2.5 and $+2.5$ dB. The mixture signals were generated at 16 kHz, and Gaussian noise was added with an SNR of 30 dB. The dataset consisted of 20,000 and 5000 mixtures for training and validation sets, respectively. Three test datasets each consisted of 3000 mixtures with 2 to 4 sources.

[1] https://research.nii.ac.jp/src/JNAS.html

Table 1: *Separation performances in averages and standard deviations in SDR and PESQ. In this evaluation, the number of sources was given from the oracle.*

Method	Training data	SDR (dB) ↑ 2 sources	3 sources	4 sources	PESQ ↑ 2 sources	3 sources	4 sources
Unprocessed	-	0.09 ± 2.15	-2.98 ± 1.89	-4.73 ± 1.78	1.17 ± 0.09	1.09 ± 0.05	1.07 ± 0.04
LGM / non-AFA (Proposed)	2 mix	$\mathbf{14.64 \pm 3.44}$	3.44 ± 4.60	-0.37 ± 4.27	$\mathbf{2.06 \pm 0.35}$	1.20 ± 0.20	1.10 ± 0.10
	3 mix	13.34 ± 4.10	$\mathbf{9.27 \pm 3.77}$	$\mathbf{4.06 \pm 4.53}$	1.84 ± 0.34	$\mathbf{1.51 \pm 0.24}$	$\mathbf{1.25 \pm 0.17}$
	2 to 3 mix	14.30 ± 3.36	9.03 ± 3.74	3.97 ± 4.50	2.00 ± 0.34	1.49 ± 0.24	1.24 ± 0.17
LGM / fixed number	2 mix	13.47 ± 3.67	-	-	1.84 ± 0.32	-	-
	3 mix	-	6.80 ± 4.07	-	-	1.34 ± 0.19	-
LGM / AFA	2 to 3 mix	13.94 ± 3.46	7.99 ± 3.93	2.91 ± 4.46	1.93 ± 0.33	1.43 ± 0.22	1.20 ± 0.14
Mask / AFA	2 to 3 mix	9.05 ± 3.95	4.48 ± 4.01	0.34 ± 4.05	1.87 ± 0.42	1.37 ± 0.25	1.11 ± 0.11

4.2. Conditions

The proposed model consisted of a temporal convolutional network [6, 15, 28]. The input features were first transformed into 256-channel vectors with a 1×1-convolutional (1×1-conv) layer. Then three modules having eight dilated convolutional layers were stacked. Each layer was the separable depth-wise convolution with a 512-channel depth-wise layer and parametric rectified linear units (PReLUs). The outputs of the model were obtained by 1×1-convolutional layers. The output layers for the TF masks and PSD were followed by sigmoid and softplus activation, respectively. The input features consisted of a log-power spectrogram, three-dimensional directions of arrival (DoAs) of the mixture signal, and the residual signal. The DoAs at each TF bin were calculated based on an array geometry [29].

The loss function of the training was the mean squared error of the magnitude in the TF domain because of comparison with a mask-based model [30]. The networks were trained by an Adam optimizer [31] for 200 epochs with a learning rate of 0.001. Spectrograms were obtained by short-time Fourier transform with a window size of 512 samples and a hop length of 128 samples. The batch size for training was 16. These hyperparameters were empirically determined.

We trained an external network for the source counting. The input features were the same as the separation network, i.e., the powers, the DoAs, and the residual signal of the separation network. Differences in the network between the separation network and the counting network were the depth of the network and the output layer. The counting network has one module of the dilated convolutional layer. The output layer of the stop flag was followed by the average pooling over the entire time and frequency and sigmoid activation. The counting network was trained based on Eq. (9) in all recursions. This network was trained after the training of the separation network for 10 epochs. The threshold value for the classification was 0.5.

We evaluated the separation performances for each test dataset with the signal-to-distortion ratio (SDR) in dB [32] and the perceptual evaluation of speech quality (PESQ) [33], and the source counting performance with the accuracy, respectively. The baselines are as follows:

1. LGM-based model for a fixed number of sources
2. LGM-based model with the recursive separation (AFA)
3. Mask-based model with the recursive separation (AFA)

The input features and loss function of these baselines are the same as the proposed method. The second baseline estimated the MWFs of the separated signal and residual signal based on the AFA, in other words, the n-th separated signal and residual signal were calculated only using parameters of the n-th recursion, and the filters were applied accumulatively. The third baseline estimated the TF masks. The residual mask was calculated by $1 - m_{nft}$. The difference in the network architecture between our model and the baselines was only the output layer.

Table 2: *Source counting accuracy by test dataset*

Method	Training data	Accuracy (%) 2 src.	3 src.	4 src.
Proposed	2 to 3 mix	**91.8**	**83.7**	24.4
Mask / AFA	2 to 3 mix	80.3	78.7	**47.6**

4.3. Results

Table 1 shows the separation performance of each method. The proposed method trained with 2 to 3-source mixtures outperformed the baselines for all test datasets in both the SDR and PESQ. Recursive models outperformed a model for a fixed number of sources. Compared with the mask-based model, the LGM-based model improved the performance. These results show that the recursive separation with beamforming was effective for the multi-channel condition. In addition, the proposed method outperformed the LGM-based recursive model based on the AFA. This result implies the proposed method mitigates the accumulation of estimation errors. The proposed model trained with 2 to 3-source mixtures was comparable to the proposed model trained with mixtures composed of the number of sources corresponding to the test dataset.

Table 2 shows the source counting accuracy of each method. The proposed method outperformed the mask-based model for 2 and 3-source mixtures. The accuracy for unseen 4-source mixtures, which were not in the training conditions, was significantly degraded in spite of the separation performance. This result implies that source counting performance for an unseen number of sources depends on not only the separation performance but also how we constitute the residual information.

5. Conclusion

In this paper, we proposed a recursive sound source separation method extended for the multi-channel condition. This model is based on the LGM for the source signal and the residual signal. Our method can separate a mixture signal composed of an unknown number of sources using beamforming. The reusing of parameters mitigates the accumulation of estimation errors and improves the separation performance. Our future work includes residual information to improve the source counting accuracy.

Acknowledgements: This work was partly supported by JSPS KAKENHI Grant Numbers JP22H00536 and JP23H03457.

6. References

[1] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj *et al.*, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. CHiME Workshop*, 2020.

[2] N. Turpault, S. Wisdom, H. Erdogan, J. Hershey, R. Serizel, E. Fonseca, P. Seetharaman, and J. Salamon, "Improving sound event detection in domestic environments using sound separation," in *Proc. DCASE Workshop*, 2020.

[3] R. Scheibler, T. Komatsu, Y. Fujita, and M. Hentschel, "Sound event localization and detection with pre-trained audio spectrogram transformer and multichannel separation network," in *Proc. DCASE Workshop*, 2022.

[4] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[5] 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 TASLP*, vol. 25, pp. 1901–1913, 2017.

[6] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM TASLP*, vol. 27, pp. 1256–1266, 2019.

[7] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, 2016, pp. 196–200.

[8] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, "Beam-TasNet: Time-domain audio separation network meets frequency-domain beamformer," in *Proc. ICASSP*, 2020, pp. 6379–6383.

[9] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All deep learning mvdr beamformer for target speech separation," in *Proc. ICASSP*. IEEE, 2021, pp. 6089–6093.

[10] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multichannel end-to-end speech recognition," in *Proc. ICML*, 2017, pp. 2632–2641.

[11] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE/ACM TASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[12] S. Inoue, H. Kameoka, L. Li, S. Seki, and S. Makino, "Joint separation and dereverberation of reverberant mixtures with multichannel variational autoencoder," in *Proc. ICASSP*. IEEE, 2019, pp. 96–100.

[13] M. Togami, "Joint training of deep neural networks for multi-channel dereverberation and speech source separation," in *Proc. ICASSP*. IEEE, 2020, pp. 3032–3036.

[14] M. Togami, Y. Masuyama, T. Komatsu, and Y. Nakagome, "Unsupervised training for deep speech source separation with kullback-leibler divergence based probabilistic loss function," in *Proc. ICASSP*, 2020, pp. 56–60.

[15] Y. Bando, K. Sekiguchi, Y. Masuyama, A. A. Nugraha, M. Fontaine, and K. Yoshii, "Neural full-rank spatial covariance analysis for blind source separation," *IEEE SPL*, vol. 28, pp. 1670–1674, 2021.

[16] S. Araki, T. Nakatani, H. Sawada, and S. Makinoe, "Blind sparse source separation for unknown number of sources using gaussian mixture model fitting with dirichlet prior," in *Proc. ICASSP*, 2009, pp. 33–36.

[17] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE TASLP*, vol. 21, no. 10, pp. 2193–2206, 2013.

[18] O. Walter, L. Drude, and R. Haeb-Umbach, "Source counting in speech mixtures by nonparametric bayesian estimation of an infinite gaussian mixture model," in *Proc. ICASSP*, 2015, pp. 459–463.

[19] F.-R. Stöter, S. Chakrabarty, B. Edler, and E. A. Habets, "Classification vs. regression in supervised learning for single channel speaker count estimation," in *Proc. ICASSP*. IEEE, 2018, pp. 436–440.

[20] K. Kinoshita, L. Drude, M. Delcroix, and T. Nakatani, "Listening to each speaker one by one with recurrent selective hearing networks," in *Proc. ICASSP*, 2018, pp. 5064–5068.

[21] N. Takahashi, S. Parthasaarathy, N. Goswami, and Y. Mitsufuji, "Recursive speech separation for unknown number of speakers," in *Proc. INTERSPEECH*, 2019, pp. 1348–1352.

[22] T. von Neumann, C. Boeddeker, L. Drude, K. Kinoshita, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "Multi-talker asr for an unknown number of sources: Joint training of source counting, separation and ASR," pp. 3097–3101, 2020.

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

[24] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *Proc. ICML*, vol. 119, 2020, pp. 7164–7175.

[25] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. Garcia, "Encoder-decoder based attractors for end-to-end neural diarization," *IEEE/ACM TASLP*, vol. 30, pp. 1493–1507, 2022.

[26] T. von Neumann, K. Kinoshita, L. Drude, C. Boeddeker, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "End-to-end training of time domain audio separation and recognition," in *Proc. ICASSP*, 2020, pp. 7004–7008.

[27] J. B. Allen and D. A. Berkley, "The lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, 1979.

[28] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves *et al.*, "WaveNet: A generative model for raw audio," in *Proc. ISCA Speech Synthesis Workshop*, 2016, p. 125.

[29] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proc. ICASSP*, vol. 5, 2006, pp. 33–36.

[30] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM TASLP*, vol. 22, no. 12, pp. 1849–1858, 2014.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.

[32] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[33] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, 2001, pp. 749–752.