*INTERSPEECH 2020*
October 25–29, 2020, Shanghai, China

# Mentoring-Reverse Mentoring for Unsupervised Multi-channel Speech Source Separation

*Yu Nakagome*[1], *Masahito Togami*[2], *Tetsuji Ogawa*[1], *Tetsunori Kobayashi*[1]

[1]Department of Communications and Computer Engineering, Waseda University, Tokyo, Japan
[2]LINE Corporation, Tokyo, Japan

nakagome@pcl.cs.waseda.ac.jp

## Abstract

*Mentoring-reverse mentoring*, which is a novel knowledge transfer framework for unsupervised learning, is introduced in multi-channel speech source separation. This framework aims to improve two different systems, which are referred to as a *senior* and a *junior* system, by mentoring each other. The senior system, which is composed of a neural separator and a statistical blind source separation (BSS) model, generates a pseudo-target signal. The junior system, which is composed of a neural separator and a post-filter, was constructed using teacher-student learning with the pseudo-target signal generated from the senior system i.e, imitating the output from the senior system (mentoring step). Then, the senior system can be improved by propagating the shared neural separator of the grown-up junior system to the senior system (reverse mentoring step). Since the improved neural separator can give better initial parameters for the statistical BSS model, the senior system can yield more accurate pseudo-target signals, leading to iterative improvement of the pseudo-target signal generator and the neural separator. Experimental comparisons conducted under the condition where mixture-clean parallel data are not available demonstrated that the proposed mentoring-reverse mentoring framework yielded improvements in speech source separation over the existing unsupervised source separation methods.

**Index Terms**: mentoring-reverse mentoring, unsupervised training, deep neural network, speech source separation

## 1. Introduction

Speech source separation is an essential technique in video conferencing systems, speech diarization systems, and robot audition systems, where a mixture of multiple speech sources and noise sources are simultaneously observed at multi-channel microphones. In particular, blind source separation (BSS) [1–8] has been actively studied and shown to be well optimized in an unsupervised manner with a statistical model. In this case, however, the permutation ambiguity of BSS should be aligned, and the separation performance can be highly affected by the initial parameter of the model.

On the other hand, supervised learning of deep neural network (DNN) has achieved an overwhelming performance in speech source separation e.g., deep clustering (DC) [9, 10], permutation invariant training (PIT) [11, 12], and hybrid modeling of the DNN and statistical BSS model [13, 14]. This approach generally aims to train a DNN that yields a time-frequency (TF) mask for extracting the corresponding sound source. Although this approach can capture complicated spectral characteristics of a speech source, it requires a large amount of paired data composed of the observed mixed-signal and supervisory clean signal. It should be noted that collecting clean signals in a real environment is infeasible. The paired data therefore have been created by simulation, such as the image method [15]. Consideration of all possible types of sound sources and room shapes in our daily lives for robust training, however, is troublesome work. Unsupervised training of speech source separation systems, which assumes to utilize only microphone observations without oracle clean signals, therefore is highly desired for the practical application of speech source separation systems.

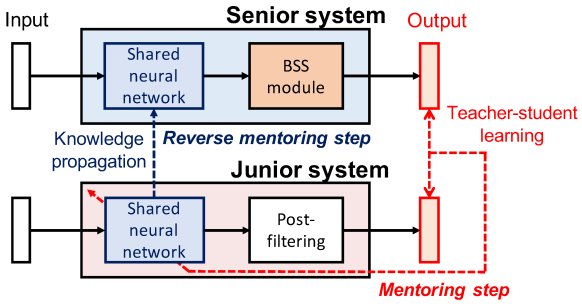


Figure 1: *Conceptual image of mentoring-reverse mentoring framework. Junior system is learned by teacher-student leaning (in mentoring step) and senior system is improved by propagation of shared knowledge (in reverse mentoring step).*

Recently, unsupervised training has been applied to DNN-based speech source separation [16–19]. In general, this approach has exploited a pseudo-target signal (also referred to as a reference signal) that is estimated by using unsupervised BSS techniques, instead of using the oracle clean signal as a target. Here, the following BSS techniques were used as a pseudo-target signal generator: the $k$-means clustering with the inter-channel phase difference [16], the complex angular central Gaussian mixture model (cACGMM) [8, 17], and the local Gaussian modeling (LGM) [19]. In addition, the statistical BSS model initialized with a learned DNN outperformed a randomly-initialized BSS model [17, 19]. In this approach, errors in the pseudo-target signal, however, lead directly to degraded training of a DNN-based separator.

To improve the accuracy of predicting the pseudo-target signal, the present study attempts to introduce the concepts of *mentoring and reverse mentoring* (as illustrated in Fig. 1) to unsupervised multi-channel speech source separation. In general, junior people start by imitating seniors (mentoring), and seniors try to take advantage of the new knowledge that young people have (reverse mentoring). Based on this analogy, two different systems, which are referred to as a *senior* and a *junior* systems, can be improved by mentoring each other. In the mentoring step, the junior system can be built to imitate the output of the senior system: this can be achieved by teacher-student learning. By allowing the senior and junior systems to have common components, reverse mentoring can transfer the components of the grown-up junior system to the senior sys-



http://dx.doi.org/10.21437/Interspeech.2020-2082

tem: this enables the senior system to generate more accurate outputs. The mentoring-reverse mentoring framework is relevant to knowledge distillation techniques [20] such as teacher-student learning [21] and born again neural networks [22, 23]. The teacher-student learning has a clear hierarchy between two systems: this does not aim to improve the teacher model but focuses on building a compressed student model. In contrast, the mentoring-reverse mentoring framework aims to alternately improve two systems under unsupervised setting.

Specifically, the senior system is composed of a neural separator and a statistical BSS model, and the junior system is composed of a neural separator and a post-filter. During mentoring, the junior system is trained with the pseudo-target signal generated from the senior system. During reverse mentoring, the neural separator of the refined junior system is transferred to the senior system. The knowledge obtained from this study could be useful for developing source separation systems under the practical condition where parallel data are not available.

The rest of the present paper is organized as follows. Section 2 presents existing works and their relevance to the proposed method. Section 3 describes the proposed mentoring-reverse mentoring framework. Section 4 demonstrates the effectiveness of the proposed method on unsupervised multi-channel speech separation. Section 5 concludes this paper.

## 2. Related Works

This section briefly reviews the existing methods for building neural speech separators when parallel data are not available.

### 2.1. Unsupervised training of neural separators

Unsupervised training techniques for neural speech source separators have been proposed [24, 25]. In [24], a DNN was directly optimized by using the likelihood function of a spatial model based on cACGMM [8]. This method simultaneously learns a separation network and a localization network [25]: the former and the latter aim to estimate a TF mask and a direction of arrival (DoA) for each source, respectively. This method not only solves the permutation problem in the spatial model, but also estimates the number of sound sources. The estimation accuracy, however, can be degraded because the loss function is affected by a mismatch between the assumed spatial propagation and the actual one.

### 2.2. Pseudo-supervised training of neural separators

Several attempts to exploit pseudo supervision have been made for training neural source separators [16–19]. In this approach, instead of using a clean signal as a target in a supervised approach, an output signal of unsupervised BSS has been used as a pseudo-target signal. In [16], a separated signal was obtained by $k$-means clustering using a phase difference and exploited as a pseudo-target signal for training a single-channel separation network. In [17], the output of cACGMM was exploited as pseudo supervision for training a separation network. Note that the pseudo-target signal contains errors because the unsupervised BSS is not perfect: this can hinder the training of a subsequent separation network.

To avoid overfitting of neural source separators to such *noisy* pseudo-target signal, several loss functions for training neural separators have been proposed [18, 19]. In [18], a loss function was weighted with confidence for the pseudo-target signal. In [19], we have already proposed a loss function that computes Kullback-Leibler divergence (KLD) between the posterior probability density function (PDF) of the output signal from the statistical model and that from the DNN. Thanks to consideration of uncertainty of the signals in the form of distribution, it is expected that a neural separator can be trained while avoiding overfitting to errors in the TF bins with low posterior probabilities given by the statistical model.

In addition to [19], this study introduced the mentoring-reverse mentoring framework to iteratively improve the pseudo-target signal generator and the neural source separator.

## 3. Proposed Method

This section presents a mentoring-reverse mentoring framework for unsupervised multi-channel speech source separation. The following subsections first explain the overview of the proposed method. Then, the details on the formulation and implementation are explained and finally, how this method works is described. Here, the present study assumes that the DoAs are known because face recognition is available in, for example, teleconferencing and human-robot communication.

### 3.1. Overview of proposed method

Figure 2 illustrates an overview of the proposed method. The pseudo-target signal generator, referred to as a *senior* system, is composed of a neural separator and a statistical BSS model. The main system, referred to as a *junior* system, is composed of a neural separator and a post-filter. Note that the neural separator is shared between the senior and junior systems.

The proposed framework aims to refine both the senior and junior systems by mentoring each other. This is conducted by the following four steps: **Step 1** generates a pseudo-target signal from the senior system based on a randomly-initialized LGM [26]; **Step2** (*mentoring*) conducts teacher-student learning to build the junior system to imitate the pseudo-target signal generated from the senior system; by allowing the senior and junior systems to have the common neural separator, **Step 3** (*reverse mentoring*) transfers the neural separator of the grown-up junior system to the senior system; and **Step 4** can generate more accurate pseudo-target signal from the refined senior system, then returns to Step 2.

This alternating procedure is repeated several times, and finally, the separation signal is obtained by using iterative parameter estimation with the posterior PDF obtained in the junior system followed by multi-channel Wiener filtering (MWF).

### 3.2. Pseudo-target signal generator: Senior system

The senior system is composed of the neural separator and an LGM-based speech source separator [7, 26]. To separate multiple speech sources, the LGM-based speech source separation assumes that the PDF of each speech source belongs to a time-varying Gaussian distribution that is represented as:

$$p(\boldsymbol{c}_{i,l,k}) = \mathcal{N}(\boldsymbol{0}, v_{i,l,k}\boldsymbol{R}_{i,k}), \tag{1}$$

where $l$ denotes the frame index, $k$ denotes the frequency index, $\boldsymbol{c}_{i,l,k}$ denotes the $i$-th speech source, $v_{i,l,k}$ denotes the time-frequency variance of the $i$-th speech source, and $\boldsymbol{R}_{i,k}$ denotes the multi-channel spatial covariance matrix of the $i$-th speech source. In addition, DoA information is incorporated into the speech source separation process by utilizing a complex inverse Wishart distribution as a prior distribution of a multi-channel covariance matrix [26] that is defined as:

$$\boldsymbol{R}_{i,k} \sim \mathcal{IW}(u, (u - N_m)(\boldsymbol{a}_{\theta_i,k}\boldsymbol{a}_{\theta_i,k}^H + \epsilon\boldsymbol{I})), \tag{2}$$

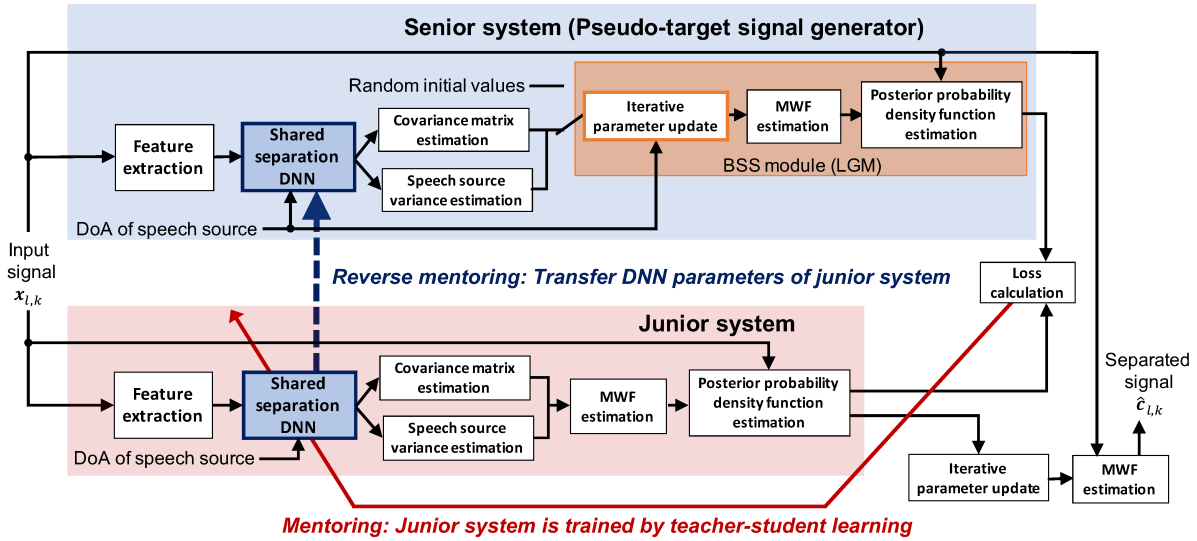


Figure 2: *Overview of proposed mentoring-reverse mentoring framework for unsupervised multi-channel speech separation.*

where $u$ is a degree of freedom of distribution, $N_m$ denotes the number of microphones, and $\boldsymbol{a}_{\theta_i,k}$ denotes the direction of speech source. Besides, $H$ denotes the Hermitian transpose of a matrix and a vector.

Since all the PDFs are Gaussian distributions, the posterior PDF of the $i$-th speech source is estimated to be a Gaussian distribution as [19]:

$$p(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k},\Theta_k) = \mathcal{N}(\boldsymbol{c}_{i,l,k}|\boldsymbol{\mu}_{p,i,l,k},\boldsymbol{V}_{p,i,l,k}), \quad (3)$$

where $\boldsymbol{x}_{l,k} \in \mathbb{C}^{N_m}$ denotes the microphone input signal, $\boldsymbol{\mu}_{p,i,l,k}$ and $\boldsymbol{V}_{p,i,l,k}$ are calculated as $\boldsymbol{\mu}_{p,i,l,k} = \boldsymbol{W}_{i,l,k}\boldsymbol{x}_{l,k}$ and $\boldsymbol{V}_{p,i,l,k} = (\boldsymbol{I} - \boldsymbol{W}_{i,l,k})v_{i,l,k}\boldsymbol{R}_{i,k}$. $\boldsymbol{W}_{i,l,k}$ denotes the time-varying MWF that is defined as $\boldsymbol{W}_{i,l,k} = v_{i,l,k}\boldsymbol{R}_{i,k}\boldsymbol{R}_{x,l,k}^{-1}$. The separation parameter $\Theta_k = \{v_{i,l,k}, \boldsymbol{R}_{i,k}\}$ is iteratively updated to maximize the likelihood function in Eq. (3) with the expectation-maximization (EM) algorithm [27]. Note that it is not necessary to align the permutation at the frequency-level and source-level because the DoA is given in this study.

At the beginning of unsupervised training, a pseudo-target signal is generated from a randomly-initialized LGM and used for training a junior system. After this first mentoring step, the neural separator of the trained junior system is transferred to the senior system and hereafter, the neural separator can provide better initial parameters for the LGM-based PDF estimation. This contributes to generation of more accurate pseudo-target signals.

### 3.3. Main speech source separator: Junior system

The junior system is composed of a neural separator and a post-filter. In the neural separator, the TF mask $\mathcal{M}_{i,l,k}$ and TF variance $v_{i,l,k}$ for each speech source are inferred via the DNN. A spectral amplitude of the microphone observation $\log|\boldsymbol{x}_{l,k}|$ is concatenated with a steered response amplitude $\log|\boldsymbol{a}_{\theta,k}^H\boldsymbol{x}_{l,k}|$, where $\boldsymbol{a}_{\theta,k}$ denotes the steering vector of the direction $\theta$, and then taken as an input to the DNN. The directions of speech sources are incorporated into the direction attractor vectors [28]. The multi-channel spatial covariance matrix for each source $\boldsymbol{R}_{i,k}$ is estimated with the TF mask $\mathcal{M}_{i,l,k}$ as:

$$\boldsymbol{R}_{i,k} = \frac{1}{\sum_l \mathcal{M}_{i,l,k}} \sum_l \mathcal{M}_{i,l,k}\boldsymbol{x}_{l,k}\boldsymbol{x}_{l,k}^H. \quad (4)$$

The posterior PDF of each speech source is calculated as $q(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k},\Phi_k) = \mathcal{N}(\boldsymbol{c}_{i,l,k}|\boldsymbol{\mu}_{q,i,l,k},\boldsymbol{V}_{q,i,l,k})$, where $\Phi_k = \{v_{i,l,k}, \boldsymbol{R}_{i,k}\}$ is the separation parameter. During inference, $\Phi_k$ is inferred via a DNN and iteratively updated with EM algorithm. Finally, the separated signal is obtained by the MWF.

### 3.4. Loss function for training deep neural network

To avoid overfitting of the neural separator to the error in pseudo-target signal, the proposed method attempts to minimize the KLD between the posterior PDF computed with the LGM $p_{i,l,k}$ and that computed with the DNN $q_{i,l,k}$ [19] as:

$$\begin{aligned}&\mathcal{D}(p_{i,l,k}\|q_{j,l,k})\\ &= (\boldsymbol{\mu}_{q,j,l,k} - \boldsymbol{\mu}_{p,i,l,k})^H\boldsymbol{V}_{q,j,l,k}^{-1}(\boldsymbol{\mu}_{q,i,l,k} - \boldsymbol{\mu}_{p,i,l,k})\\ &+ \mathrm{tr}(\boldsymbol{V}_{q,j,l,k}^{-1}\boldsymbol{V}_{p,i,l,k}) + \log\frac{|\boldsymbol{V}_{q,j,l,k}|}{|\boldsymbol{V}_{p,i,l,k}|} - N_m.\end{aligned} \quad (5)$$

### 3.5. Effectiveness of mentoring-reverse mentoring

This subsection describes how the proposed mentoring-reverse mentoring works in multi-channel speech source separation. In the first mentoring step, the junior system is constructed by using teacher-student learning to imitate the pseudo-target signal estimated by the randomly-initialized BSS, which is a frequency-independently formulated separator that results in spontaneous inconsistencies between frequencies. The previous work on teacher-student learning of DNN with pseudo-target signals obtained by the BSS [17, 19] observed that DNN does not overfit to such inconsistencies, but rather learn parameters that take into account the time-frequency correlations. This property seems to be the key to the success of the mentoring step. In the reverse mentoring step, the data-driven knowledge of the learned junior system is transferred to the senior system as a parameter of DNN. In the senior system, the BSS is initialized with the parameters yielded from the DNN to estimate the pseudo-target signal with higher accuracy. Using the improved pseudo-target signal, the mentoring step is performed again to obtain a junior system. The reduced error in the pseudo-target signal contributes to the junior system estimating the parameters of the DNN more accurately.

Table 1: *Performance of existing and proposed unsupervised speech source separation methods.*

| | # iteration of mentoring-revere mentoring | SDR [dB] | SIR [dB] | FWseg.SNR [dB] | CD | PESQ |
|---|---|---|---|---|---|---|
| Unprocessed | - | -0.80 | 2.26 | 7.38 | 4.78 | 1.69 |
| Pseudo-target signal [26] | - | 2.92 | 6.54 | 9.31 | 4.07 | 1.92 |
| Training w/ pseudo-target signal [19] | 0 | 4.84 | 7.92 | 9.53 | 3.95 | 2.03 |
| Proposed method | 1 | 5.73 | 8.42 | 9.60 | 3.81 | 2.04 |
| | 2 | 5.80 | 8.49 | 9.67 | **3.77** | **2.05** |
| | 3 | **5.82** | **8.54** | **9.71** | **3.77** | **2.05** |
| Training w/ oracle target signal [28] | - | 7.79 | 10.10 | 11.32 | 3.46 | 2.15 |

# 4. Experiments

To demonstrate the effectiveness of the proposed method, experimental comparisons were conducted in an environment with multiple speech sources and diffuse noise.

### 4.1. Speech materials

The clean speech and the diffuse noise were selected from the TIMIT corpus [29] and the diverse environments multi-channel acoustic noise database (DEMAND) [30], respectively. To simulate simultaneous speech, a target and an interference speech source were placed respectively at one of the 13 directions (-90° to 90° at 15° intervals) without duplication. Here, the measured impulse responses in the multi-channel impulse response database (MIRD) [31] were convoluted to the aforementioned dry sources at a signal-to-interference ratio (SIR) of the range of -5 dB to 5 dB. Then, the diffuse noise was superposed at a signal-to-noise ratio (SNR) of the range of 20 dB to 30 dB.

A linear microphone array with eight microphones was used. The microphone spacing was 3-3-3-8-3-3-3 cm, 4-4-4-8-4-4-4 cm, or 8-8-8-8-8-8-8 cm. The microphone alignment was assumed to be known to calculate steering vectors used for an input feature of the DNN. The distance between a speech source and a microphone was 1 m, and the reverberation time was randomly set to 0.16 s, 0.36 s, or 0.61 s for each utterance.

The talkers and utterances were different between the training and the testing data. For training, 1000 utterances were used: this is smaller than the conventional study (e.g., 30000 in [17]) because a small number of required utterances is preferable in practice. For testing, 500 utterances were used.

The sampling rate was 8 kHz, the frame size was 256, and the frame-shift was 64. The number of frequency bins was 129.

### 4.2. Neural network architecture

The network architecture for the proposed method was empirically determined. The neural separator has three layers of bidirectional long short-term memory (BLSTM) with 300 units for each direction, and one fully connected layer for estimating a time-frequency mask and a time-frequency activity. The direction attractor net [28] has four fully connected layers for each speech source. All network parameters were optimized using the Adam optimizer [32]. The learning rate of the optimizer was $1.0 \times 10^{-3}$ and the mini-batch size was 32.

For an impartial comparison, the number of epochs was set to 300 in all DNN methods. When the pseudo-target signal was updated $N$ times in the proposed method, the pseudo-target signal was generated and replaced every $300/(N+1)$ epochs with the parameters of DNN at that time.

The hyper-parameter $u$ in LGM [26] was 50 and the parameters $\Theta_k$ were updated 30 times with EM algorithm, which were empirically determined.

### 4.3. Experimental results

The performance of speech source separation was evaluated using the signal-to-distortion ratio (SDR) and the SIR from BSS-EVAL [33], the frequency-weighted segmented SNR (FWseg. SNR), the cepstrum distortion (CD), and the PESQ. The average scores for all evaluation measures are listed in Table 1.

First, the pseudo-target signal estimated by the randomly-initialized LGM [26] was compared with the output signal from the junior system (i.e., Training w/ pseudo-target signal [19]) to examine the effectiveness of the mentoring step. The result demonstrated that the first mentoring step performed well: the junior system outperformed the senior system (i.e., randomly-initialized LGM).

Next, the effectiveness of the proposed mentoring-reverse mentoring framework was investigated. The result demonstrated that performing the mentoring and reverse mentoring once greatly improved the separation performance over the case with only the mentoring step (i.e., Training w/ pseudo-target signal [19]). This indicates that the neural separator transferred from the learned junior system provided better parameters for initialization of the LGM, and decreased residual noise in the pseudo-target signal, leading to improved performance in training the junior system again.

Finally, the effectiveness of repeating the mentoring-reverse mentoring steps was evaluated. The result demonstrated that each iteration of the mentoring and reverse mentoring improved the separation performance and the improvement almost converged in the third update. The proposed method focuses on unsupervised learning and does not reach the performance of the junior system built in a supervised manner [28] (i.e., Training w/ oracle target signal [28]). In our future work, selecting training data based on the reliability of the pseudo-target signal in the mentoring step is expected to improve the performance efficiently.

# 5. Conclusions

The concept of mentoring-reverse mentoring was introduced in unsupervised multi-channel speech source separation. In the mentoring step, the junior system was built by teacher-student learning with the pseudo-target signal generated from the senior system. In the reverse mentoring step, the senior system was improved by taking advantage of the shared neural separator of the refined junior system. The experimental results demonstrated the iterative improvement of the senior and junior systems and the improvement of the proposed method over the existing methods.

# 6. Acknowledgements

The research was supported by NII CRIS collaborative research program operated by NII CRIS and LINE Corporation.

# 7. References


[1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[2] P. Common, "Independent component analysis, a new concept ?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, April 1994.

[3] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proceedings ICA*, Mar. 2006, pp. 601–608.

[4] T. Kim, H. Attias, S.-Y. Lee, and T.-W. Lee, "Independent vector analysis: an extension of ICA to multivariate components," in *Proceedings ICA*, Mar. 2006, pp. 165–172.

[5] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, Sept 2016.

[6] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, *Determined Blind Source separation with Independent Low-Rank Matrix Analysis*. Springer Publishing Company, Incorporated, 2018, ch. 6, pp. 125–155.

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

[8] N. Ito, S. Araki, and T. Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *2016 24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1153–1157.

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

[10] Z. Wang, J. L. Roux, and J. Hershey, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 1–5.

[11] 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)*, March 2017, pp. 241–245.

[12] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5739–5743.

[13] A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 9, pp. 1652–1664, 2016.

[14] A. A. Nugraha, A. Liutkus, and E. Vincent, "Deep neural network based multichannel audio source separation," in *Audio Source Separation*. Springer, 2018, pp. 157–185.

[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] E. Tzinis, S. Venkataramani, and P. Smaragdis, "Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 81–85.

[17] L. Drude, D. Hasenklever, and R. Haeb-Umbach, "Unsupervised training of a deep clustering model for multichannel blind source separation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 695–699.

[18] P. Seetharaman, G. Wichern, J. Le Roux, and B. Pardo, "Bootstrapping single-channel source separation via unsupervised spatial clustering on stereo mixtures," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 356–360.

[19] 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 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 56–60.

[20] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[21] C. Buciluundefined, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 535–541. [Online]. Available: https://doi.org/10.1145/1150402.1150464

[22] T. Furlanello, Z. C. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," *International Conference on Machine Learning*, 2018.

[23] Q. Xie, E. Hovy, M.-T. Luong, and Q. V. Le, "Self-training with noisy student improves imagenet classification," *arXiv preprint arXiv:1911.04252*, 2019.

[24] L. Drude and R. Haeb-Umbach, "Integration of neural networks and probabilistic spatial models for acoustic blind source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 815–826, 2019.

[25] Y. Bando, Y. Sasaki, and K. Yoshii, "Deep bayesian unsupervised source separation based on a complex gaussian mixture model," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2019, pp. 1–6.

[26] N. Q. K. Duong, E. Vincent, and R. Gribonval, "An acoustically-motivated spatial prior for under-determined reverberant source separation," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 9–12.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 39, no. 1, pp. 1–38, 1977.

[28] Y. Nakagome, M. Togami, T. Ogawa, and T. Kobayashi, "Deep speech extraction with time-varying spatial filtering guided by desired direction attractor," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 671–675.

[29] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.

[30] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, p. 3591, 05 2013.

[31] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 313–317, 2014.

[32] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[33] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.