INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Efficient and Stable Adversarial Learning Using Unpaired Data for Unsupervised Multichannel Speech 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

This study presents a framework to enable efficient and stable adversarial learning of unsupervised multichannel source separation models. When the paired data, i.e., the mixture and the corresponding clean speech, are not available for training, it is promising to exploit generative adversarial networks (GANs), where a source separation system is treated as a generator and trained to bring the distribution of the separated (fake) speech closer to that of the clean (real) speech. The separated speech, however, contains many errors, especially when the system is trained unsupervised and can be easily distinguished from the clean speech. A real/fake binary discriminator therefore will stop the adversarial learning process unreasonably early. This study aims to balance the convergence of the generator and discriminator to achieve efficient and stable learning. For that purpose, the autoencoder-based discriminator and more stable adversarial loss, which are designed in boundary equilibrium GAN (BEGAN), are introduced. In addition, generator-specific distortions are added to real examples so that the models can be trained to focus only on source separation. Experimental comparisons demonstrated that the present stabilizing learning techniques improved the performance of multiple unsupervised source separation systems.

Index Terms: unpaired data, boundary equilibrium generative adversarial network, unsupervised training, multichannel speech separation

1. Introduction

Speech source separation [1, 2] has played an important role in communication robots, automatic speech recognition systems, automatic diarization systems, and hands-free communication systems. In recent years, supervised source separation methods using deep neural network (DNN), such as Deep clustering [3] and permutation invariant learning [4, 5], and hybrid methods consisting of DNNs and blind source separation based on probabilistic statistical models [6, 7] have shown to achieve excellent separation performance due to the high expressive power of DNNs. 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 such clean signals in a real environment is infeasible. The paired data therefore have been created by simulation, such as the image method [8]. Consideration of all possible types of sound sources and room shapes in our daily lives for robust training, however, is troublesome work. For practical use of source separation systems, it is desirable to use only the microphone observations (i.e., not require *oracle* clean signals) for training DNNs.

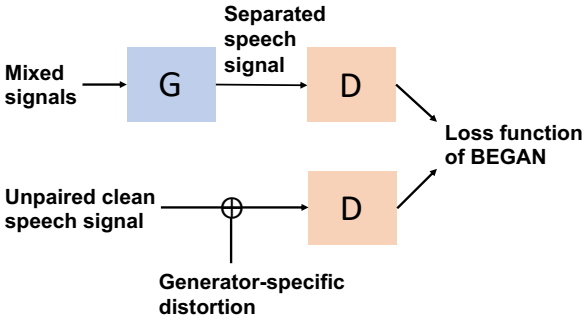

Figure 1: *Overview of proposed stable adversarial training of speech separation network using unpaired data. Unsupervised speech separation system is treated as generator (G), and trained by using adversarial loss designed for boundary equilibrium generative adversarial network (BEGAN) [9], where autoencoder is applied to discriminator (D) instead of general real/fake binary discriminator, for efficient and stable learning.*

The present study therefore focuses on unsupervised training of speech separation models using available clean speech (hereafter referred to as "unpaired data") rather than *oracle* clean speech corresponding to mixed sounds. Adversarial learning, in which the source separation network is trained so that the distribution of the separated (*fake*) signals is close to the distribution of the clean (*real*) speech, is well suited for this purpose. Existing adversarial network-based source separation models have shown to be effective but generally been constructed using paired data [10–16]. The performance of unsupervised source separation systems, however, is not as good as that of supervised source separation systems that can use paired data, and their output is likely to contain many errors. In this case, since the separated (*fake*) speech differs significantly from the clean (*real*) speech, the real/fake binary discriminator can easily discriminate between the two, so the discriminator training stops at its early stage, and as a result, the generator training does not progress. It thus is difficult to balance the convergence of the generator and discriminator, when using the general real/fake discriminator, especially in unsupervised speech source separation.

In contrast, in order to stabilize adversarial learning even when using unpaired data with large differences for unsupervised training, the present study attempts to *i)* introduce a discriminator and a loss function to enable stable learning and to *ii)* add reasonable noise to the real examples (see Fig. 1). First, we use an autoencoder-based discriminator, instead of the real/fake binary discriminator, and more stable adversarial loss,

http://dx.doi.org/10.21437/Interspeech.2021-523

which are designed in boundary equilibrium generative adversarial network (BEGAN) [9] and has been shown to be capable of converging GANs efficiently and stably. In BEGAN, the autoencoder-based discriminator aims to solve the problem of reconstructing the signal. Weighting the losses in the discriminator by the reconstruction error facilitates the elimination of the imbalance in learning between the two. Our second proposal is to add reasonable noise to the real example to further stabilize adversarial learning. The output of the generator (i.e., source separation system) contains the separation errors and the processing distortions inherent to the generator. Here, the generator-specific distortion is added to the unpaired clean speech (i.e., real example) and the result is fed into the discriminator. By doing so, the difference between the real and fake example is reduced and the learning process can be stabilized. More importantly, the model can be trained to focus on speech source separation rather than distortion reduction.

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 speech separation framework based on generative adversarial network with unpaired data. Section 4 demonstrates the effectiveness of the proposed method on multichannel speech separation. Section 5 concludes this paper.

2. Related works

This section surveys the literature on unsupervised training methods of neural speech separators and GAN-based speech enhancement and separation methods.

2.1. Unsupervised training of neural separators

Several attempts have been made to train neural speech separators in unsupervised manners [17–22]. In [17], a DNN was directly optimized by using the likelihood function of a spatial model based on cACGMM [23]. Another method attempts to simultaneously learn a separation network and a localization network [18]: the former and the latter aim to estimate a time-frequency mask and a direction of arrival (DoA) for each source, respectively, to solve the frequency permutation ambiguity. Although these methods have performed well in unsupervised source separation, their performances are limited because they do not use any prior knowledge such as time-frequency characteristics of the target source. The present work incorporates the adversarial loss into the training of unsupervised source separation systems to leverage the knowledge on the characteristics of the target source (i.e., clean speech).

2.2. GAN-based speech enhancement and separation

GAN [24] is a generative model that learns a mapping between a sample $\mathbf{z}$ from a prior distribution $\mathcal{Z}$ and sample $\mathbf{x}$ from another distribution $\mathcal{X}$. GAN has been successful in the field of computer vision and speech synthesis (i.e., realistic image and speech generation). Its great success has been similarly observed in speech enhancement [10–16]. Speech enhancement GAN (SEGAN) [10] is the pioneering work among them. In this method, a denoising network that maps noisy speech to clean speech is regarded as a generator, while a binary classifier that distinguishes between clean and enhanced speech is exploited as a discriminator. SEGAN optimizes the min-max game between the generator (G) and discriminator (D) with the loss function L as:

$$\min_G \max_D L(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \\ + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]. \quad (1)$$

These methods, however, require a large amount of paired data (i.e., the observed signal and the corresponding clean signal) to learn the mapping of the data distribution, which is not feasible in realistic settings. In addition, most of the existing studies have focused on speech enhancement or noise suppression, and not enough attention has been paid to speech separation, which is more challenging in terms of handling signals containing sources in the same domain (i.e., simultaneous speech).

Adversarial learning using unpaired data has been applied to the task of separating singing voice and accompaniment [25]. In this work, a real/fake binary discriminator was exploited and performed well for stable training. Because of the different nature of singing voice and accompaniment, it is possible that unsupervised learning could have performed well. On the other hand, the problem of separating speech from multiple speakers is more difficult, and adversarial learning based on the binary discriminator is not likely to be stable. In fact, in our preliminary experiments, the binary discriminator did not work at all in unsupervised simultaneous speech separation.

3. Stabilizing learning of speech separation models using unpaired data

This section describes two of our attempts to stabilize adversarial learning: *i)* the introduction of the BEGAN concept and *ii)* the addition of reasonable noise to real examples (i.e., unpaired clean signals).

3.1. System overview

Figure 2 illustrates an overview of the developed system. Since the speech separation system obtained by unsupervised learning contains a relatively large amount of separation errors and processing distortion, we aim to improve the performance of the system by bringing its output explicitly closer to the characteristics of clean speech. For this purpose, the speech separation system to be developed is considered as a generator and is trained by using the adversarial loss (i.e., to deceive a discriminator). To stabilize the convergence of the adversarial learning, an autoencoder-based discriminator is used as in BEGAN, instead of the more general real/fake binary discriminator. We also add reasonable noise to the real example (i.e., unpaired clean speech) to further stabilize the learning.

3.2. Generator

The generator is composed of the neural separator and a local Gaussian model (LGM)-based speech source separator [26,27]. To separate multiple speech sources, the LGM-based speech source separation assumes that the probability density function (PDF) of each speech source belongs to a time-varying Gaussian distribution that is represented as:

$$p(\boldsymbol{s}_{i,l,k}) = \mathcal{N}(\mathbf{0}, v_{i,l,k}\boldsymbol{R}_{i,k}), \quad (2)$$

where l denotes the frame index, k denotes the frequency index, $\boldsymbol{s}_{i,l,k}$ denotes the i-th source, $v_{i,l,k}$ denotes the time-frequency variance of the i-th source, and $\boldsymbol{R}_{i,k}$ denotes the multichannel spatial covariance matrix of the i-th source. After estimating the parameter $\theta_k = \{v_{i,l,k}, \boldsymbol{R}_{i,k}\}$ by DNN from multichannel

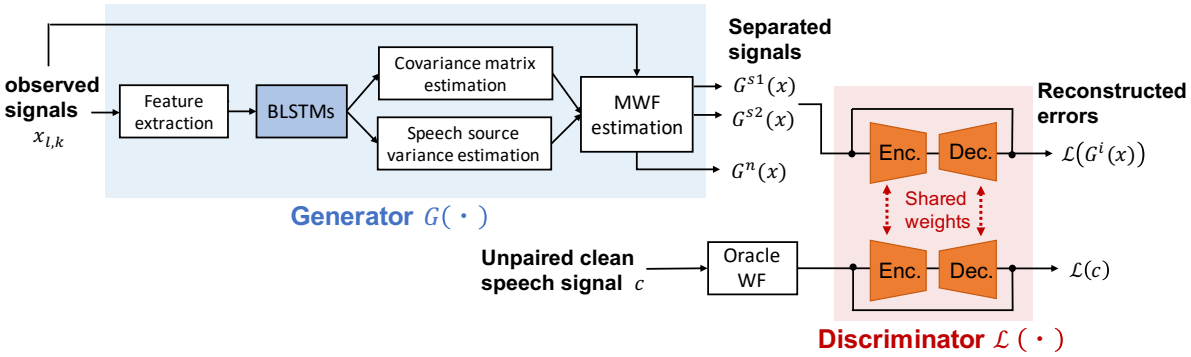

Figure 2: *Structure of proposed stable adversarial training of speech separation network using unpaired data.*

observed singnal $\boldsymbol{x}_{i,l,k}$, a separated source $\boldsymbol{G}_{i,l,k}(x)$ can be obtained using a time-varying multichannel Wiener filter (MWF) as:

$$\boldsymbol{G}_{i,l,k}(x) = \boldsymbol{W}_{i,l,k}\boldsymbol{x}_{i,l,k}, \tag{3}$$

where $\boldsymbol{W}_{i,l,k}$ denotes the time-varying MWF that is defined as:

$$\boldsymbol{W}_{i,l,k} = \boldsymbol{R}_{i,l,k}\left(\sum_{\acute{i}=1}^{N_s}\boldsymbol{R}_{\acute{i},l,k}\right)^{-1}. \tag{4}$$

Here, the number of sources N_s and the number of speech sources in the observed signal are assumed to be known. Among the output signals of the generator, the source with the index of speech source is input to the discriminator.

The separation network in this generator can be trained in the unsupervised manner by directly maximising the likelihood function of the separation parameter [17,18]. The log likelihood function is calculated as $\sum_l \log p(\boldsymbol{x}_{l,k} \mid \theta_k)$.

3.3. Adversarial loss in BEGAN

Using adversarial loss is a promising option to bring the output of a speech separation system closer to clean speech. However, since the output of unsupervised speech separation has a large difference from clean speech, and the real/fake binary discriminator can easily determine the difference, the adversarial learning of the generator can stop early. The present study attempts to exploit the autoencoder-based discriminator and adversarial loss designed in BEGAN to achieve stable learning while balancing the convergence of the generator and discriminator.

The autoencoder-based discriminator takes either the signal of each source separated by the generator (i.e., *fake* example) or a clean signal that is not paired with the mixture (i.e., *real* example). The discriminator is trained to minimize the reconstruction error for the *real* example while maximizing the error for the *fake* example.

Let θ_D and θ_G be the parameters for the discriminator and generator, respectively. These parameters are learned by minimizing the corresponding objective functions $\mathcal{L}_D$ and $\mathcal{L}_G$ as follows:

$$\begin{cases} \mathcal{L}_D = \mathcal{L}(c) - k_t\mathcal{L}(G(x)) & \text{for} \quad \theta_D \\ \mathcal{L}_G = \mathcal{L}(G(x)) & \text{for} \quad \theta_G \\ k_{t+1} = k_t + \lambda_k(\mathcal{L}(c) - \mathcal{L}(G(x))) \end{cases} \tag{5}$$

where $\mathcal{L}(\cdot)$ denotes the reconstruction error in the discriminator; $G(\cdot)$, the output signal of the generator; c, the clean speech signal; x, the multichannel observation signal; and k_t, the balancing term for the losses of the generator and discriminator at the t-th training step. In addition, λ_k denotes the learning rate for k. Here, the reconstruction error $\mathcal{L}(\cdot)$ is the mean square error between the input and output signals. Weighting the loss function of the discriminator with the balancing term k_t avoids that either the generator or the discriminator always wins and stopping the DNN training.

Based on the reconstruction error of the clean and separated speech signals, the convergence of the learning can be determined. When $\mathcal{M}_{\text{global}}$ in the following equation approaches zero, the learning is judged to be converged.

$$\mathcal{M}_{\text{global}} = \mathcal{L}(c) + |\mathcal{L}(c) - \mathcal{L}(G(x))|. \tag{6}$$

3.4. Reasonable noise addition to real examples

Another attempt to stabilize adversarial learning is to add reasonable noise to the real example. One of the goals of doing so is to stabilize learning by intentionally reducing the difference between the real example and the fake example. Here, generator-specific distortions are selected as reasonable noise and added to the unpaired clean signal (i.e., the real example). The aim is to enable efficient training of the generator by focusing on improving the source separation ability rather than reducing the separation processing distortion.

The generator-specific processing distortions occur mainly when the Wiener filter is applied. We therefore consider adding the distortion caused by Wiener filtering to a *real* clean signal by the following procedure: *i)* the *real* clean speech is superimposed with another clean speech and noise; *ii)* An oracle time-invariant Wiener filter that extracts the *real* clean speech is estimated by using an ideal ratio mask; and *iii)* the estimated Wiener filter is applied to the *real* clean speech. The clean speech with generator-specific distortions obtained in this way is taken as an input to the discriminator as a real example.

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 [28] and the diverse environments multichannel acoustic noise database (DEMAND) [29], 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 mea-

Table 1: *Performance of existing unsupervised speech separation methods and effectiveness of proposed stable adversarial learning. Lines with check marks are proposed method. PIT uses paired data, which corresponds to upper bound on performance of systems built using only unpaired data.*

speech separation method	+ BEGAN-based adversarial loss	+ distortion addition	SDR [dB]	SIR [dB]	FWseg.SNR [dB]	CD	PESQ
Unprocessed			-0.80	2.26	7.38	4.78	1.69
[17]			5.12	6.29	8.79	4.40	1.78
	✓		5.51	6.44	8.89	4.36	1.80
	✓	✓	**5.67**	**6.52**	**8.91**	**4.33**	**1.81**
[18]			5.47	8.88	9.15	4.46	2.02
	✓		5.63	9.16	9.29	4.33	2.03
	✓	✓	**5.76**	**9.30**	**9.35**	**4.31**	**2.03**
PIT [5]			8.65	10.92	11.56	3.39	2.18

sured impulse responses in the multichannel impulse response database (MIRD) [30] 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 unpaired clean speech that is input to the discriminator was also selected from the TIMIT corpus. The talkers and utterances were different between the observed speech in training data, the unpaired clean speech in training data and the observed speech in testing data. For training, 3000 utterances were used for each the observed and unpaired signals. 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 in the generator had three layers of bi-directional long short-term memory (BLSTM) with 300 units for each direction, followed by one fully connected layer for estimating a time-frequency mask and a time-frequency activity. The autoencoder in the discriminator had four convolutional layers for each encoder and decoder, and the dimensionality of the intermediate feature was 256. In equation 5, k_0 was 0 and λ_k was 1.0×10^{-3}. All network parameters were optimized using the Adam optimizer [31]. The learning rate of the optimizer was 1.0×10^{-3} and the mini-batch size was 32.

4.3. Developed systems

We compared the sound quality of unprocessed speech with the outputs of the following four speech separation systems:

1. the speech separation system trained in the existing unsupervised manner [17, 18];
2. the speech separation system developed with the BEGAN-based adversarial training of a generator pre-trained in the unsupervised manner;
3. the speech separation system developed with the BEGAN-based adversarial training of a generator pre-trained in the unsupervised manner and the generator-specific distortion addition to the *real* clean signal; and
4. the speech separation system developed with permutation invariant training [5] of a generator using paired data.

In this experiment, we applied two existing unsupervised speech separation schemes: one aims to train a DNN by directly maximizing the likelihood of the separated signal against the observed signal [17], and the other aims to solve the frequency permutation ambiguity by jointly training localization and separation [18].

4.4. Experimental results

The performance of speech source separation was evaluated using the signal-to-distortion ratio (SDR) and the SIR from BSS-EVAL [32], the frequency-weighted segmented SNR (FWseg. SNR), the cepstrum distortion (CD), and the PESQ.

Table 1 lists the effect of stabilizing adversarial learning by *i)* introducing the concept of BEGAN and *ii)* adding processing distortions to unpaired clean speech on the performance of the source separation system. The results show that regardless of the base unsupervised source separation methods, the introduction of BEGAN improved the performance for all evaluation measures, and the introduction of processing distortion further improved the performance. It can also be seen that these systems are obtained by unsupervised learning using unpaired data, and did not reach the performance of source separation systems built using paired data (i.e, PIT [5]).

5. Conclusions

Attempts were made to stabilize adversarial learning of unsupervised multichannel speech separation models. Using the autoencoder-based discriminator instead of the general real/fake binary discriminator enabled stable learning by balancing the convergence of the generator and discriminator. To further stabilize the learning process, the generator-specific distortion was added to the *real* clean speech and fed into the discriminator. The experimental comparisons demonstrated that the proposed stabilizing learning techniques performed well in unsupervised speech separation.

6. Acknowledgements

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

7. References

[1] P. Loizou, *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*, 1st ed. CRC Press, 2007.

[2] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks," *Proc. Neural Inf. Process. Syst.,*, 2017.

[10] S. Pascual, A. Bonafonte, and J. Serrá, "Segan: Speech enhancement generative adversarial network," in *Proc. Interspeech*, Aug. 2017, pp. 5024–5028.

[11] M. Omachi, T. Ogawa, and T. Kobayashi, "Associative memory model-based linear filtering and its application to tandem connectionist blind source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 637–650, 2017.

[12] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5024–5028.

[13] S. Ye, T. Jiang, S. Qin, W. Zou, and C. Deng, "Speech enhancement based on a new architecture of wasserstein generative adversarial networks," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 399–403.

[14] D. Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 106–110.

[15] S. Ye, X. Hu, and X. Xu, "Tdcgan: Temporal dilated convolutional generative adversarial network for end-to-end speech enhancement," *arXiv preprint arXiv:2008.07787*, 2020.

[16] G. Liu, K. Gong, X. Liang, and Z. Chen, "Cp-gan: Context pyramid generative adversarial network for speech enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6624–6628.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Y. Nakagome, M. Togami, T. Ogawa, and T. Kobayashi, "Mentoring-reverse mentoring for unsupervised multi-channel speech source separation," *Proc. Interspeech 2020*, pp. 86–90, 2020.

[23] 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.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B.Xu, D. W.-F. S. O. A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680.

[25] D. Stoller, S. Ewert, and S. Dixon, "Adversarial semi-supervised audio source separation applied to singing voice extraction," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2391–2395.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.