



# Training Data Generation with DOA-based Selecting and Remixing for Unsupervised Training of Deep Separation Models

Hokuto Munakata, Ryu Takeda, and 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 simple and easy-to-apply unsupervised training method for multi-channel deep separation models used in sound source separation. Such models require a large amount of training data, i.e., source signals and their mixtures. A previous method uses pseudo-target source signals, which can be obtained as the outputs of blind source separation (BSS) based on statistical models in place of ground-truth source signals. However, the model performance of the previous method is degraded by some pseudo-targets that are inadequately separated by BSS. To exploit the reliable part of BSS, we select and remix well-separated signals included in the BSS result. In the selection step, we choose well-separated signals using the direction of arrival (DOA). As a criterion that addresses the quality of the separated signals, we adopted the minimum angular difference of DOA between source signals. In the remixing step, we introduce resampling of the DOA, which generates mixtures composed of source signals with both wide and narrow angular differences. These mixtures are not simply given by BSS and allow the deep separation model to learn both spectral and spatial information. In our experiment, our method's model performance was improved for mixture signals composed of the sources from various angles.

**Index Terms:** deep sound source separation model, unsupervised training, direction of arrival, blind source separation, training data generation

## 1. Introduction

### 1.1. Background and Motivation

Sound source separation segregates source signals from observed mixture signals. This technology is important for many sound processing systems running in real environments with multiple speakers. In such environments, we must contend with various situations, for example, unknown noises, reverberations, and locations of the sources.

As a conventional sound source separation method, blind source separation (BSS) has been researched for a long time [1–7]. BSS is based on statistical models, which do not require pre-training. For example, spatial clustering infers separation masks by clustering based on microphone coordinates and the direction of arrival (DOA) [5, 6]. Since BSS mainly uses spatial information that is available from multi-channel observed signals, it does not perform well when the spatial information is poor. When the sources are located nearby and an angular difference of DOA between source signals is narrow, the spatial transfer properties of each sound source are similar and cannot be used as a separation clue.

Recently, deep separation models have been actively researched [8–13]. These models precisely learn complex spectral patterns from many pairs of source signals and their mix-

tures. Their performance outperforms BSS and is impressive even if the spatial information is poor. However, the model performance is degraded when there are mismatches between training and test data, which occur frequently in real environments. A straightforward solution is to train a separation model by preparing all combinations of possible source signals and their mixtures, but it is impractical. It is too expensive to collect every possible sources and prepare their all possible mixtures even with the datasets of various noises and properties [14, 15].

To solve this problem, unsupervised training methods for deep separation models that use mixture signals existing in real environments have been researched [16–20]. These methods are based on the fact that mixture signals are easy to collect in real environments. By utilizing BSS, the model learns the separation ability from the mixture signals even when ground-truth source signals are not prepared. A previous work [16] trained the separation model with pairs of mixtures and pseudo-target source signals, i.e., the outputs of BSS as an alternative to the ground-truth source signals. However, the pseudo-targets include some signals that are badly separated by BSS due to poor spatial information. The poor training data degrades the model performance.

To exploit only the reliable parts of the BSS results, we propose a simple and novel unsupervised training method that utilizes the DOA of the source signals. We generate a training dataset by selecting and remixing the well-separated signals included in the BSS result that contains adequate clues to learn separation ability. First, we select only well-separated signals with the DOA and store them for remixing. As the selection criterion, we adopt the minimum angular difference of DOA between the source signals. This takes advantage of the fact that the BSS performance depends on the angular difference between them. Second, we generate training data by remixing the stored well-separated signals. Then we introduce a resampling of the DOA, which generates a training dataset that includes the mixtures composed of sources with various angular differences. This allows the separation model to learn both spectral and spatial information. Hence the separation model is trained efficiently and becomes less susceptible to BSS bottlenecks, namely, excessive dependence on spatial information.

The main contribution of our work is as follows:

1. We proposed an unsupervised training method that generates training datasets for deep sound source separation models by selecting and remixing well-separated signals that are included in the BSS results. This method is easy to implement and to apply to train existing supervised separation models.
2. In the experiment, the deep separation model trained by our method performed well for the source signals from various angles, even for those with narrow angular differences of DOA.

## 1.2. Related Work

Some proposed methods strongly integrate DNN and BSS [17–20]. Compared to them, our method focuses on generating better training data rather than tightly integrating BSS and DNN to improve performance. Although these models outperformed a previous method [16], they are complex, and integrating them with existing supervised separation models is complicated. In contrast, our method can be applied to supervised models because we focus on generating a training dataset.

In addition, the number of source signals is unknown in most cases in real environments. For this case, a sound source number estimating network was proposed [21]. Although this method requires training data for the network, our proposed method can be applied to this method easily.

## 2. Preliminary

### 2.1. Problem Statement

In this paper, all signals are represented in the time-frequency domain.  $K$  source signals are denoted by  $\mathbf{s}_1, \dots, \mathbf{s}_K \in \mathbb{C}^{F \times T}$ , where  $F$  and  $T$  are the maximum dimensions of the frequency and time frames.  $C$ -channel mixture  $\mathbf{x} = \{\mathbf{x}_{f,t} \in \mathbb{C}^C\}_{f,t=1}^{F,T}$  is observed as a sum of  $K$  sources and noise by time-frequency index  $f, t$  as follows:

$$\mathbf{x}_{f,t} = \sum_{k=1}^K \mathbf{a}_{k,f} s_{k,f,t} + \mathbf{n}_{f,t}. \quad (1)$$

$\mathbf{a}_1, \dots, \mathbf{a}_K \in \mathbb{C}^C$  are the steering vectors of each microphone and represent the spatial transfer properties and  $\mathbf{n}_{f,t}$  is noise. Our objectives are inferring signals  $\mathbf{y}_k = \{\mathbf{a}_{k,f} s_{k,f,t}\}_{f,t}^{F,T}$  ( $k = 1, \dots, K$ ) from the observed mixture.

### 2.2. Deep Separation Model

A deep separation model separates source signals by time-frequency masks  $m_{f,t,k} \in [0, 1]$ , which represent the most dominant source signal per time-frequency slot. The masks inferred by the deep separation model are formulated as follows:

$$\{m_{f,t,k}\}_{f,t,k}^{F,T,K} = \mathbf{f}(\mathbf{x}; \Theta_{DNN}), \quad (2)$$

where  $\mathbf{f}(\cdot; \Theta_{DNN})$  is the non-linear transformation of the model and  $\Theta_{DNN}$  is a set of parameters of the deep separation model. By applying a mask, we obtain the following separated signals as follows:

$$\hat{\mathbf{y}}_{f,t,k} = m_{f,t,k} \mathbf{x}_{f,t}. \quad (3)$$

To train the model to directly infer the masks, we need to solve the label permutation problem, which is caused by ambiguity about the permutation of the output  $K$  masks and the target signals. Permutation invariant training (PIT) solves this problem [12]. The following is PIT's loss function as follows:

$$\mathcal{L} = \frac{1}{TFC} \min_{k' \in \mathfrak{R}} \sum_{t,f,c} (|\hat{y}_{f,t,c,k'}| - |y_{f,t,c,k}|)^2, \quad (4)$$

where  $c$  is a microphone channel index and  $\mathfrak{R}$  is a set of  $K!$  permutations. Since PIT's loss function uses permutations with the lowest loss, the loss does not depend on the permutation of outputs.

Although the input of the single-channel deep separation model is the log magnitude of a mixture, spatial information is used as additional input in a multi-channel condition. Inter-channel phase difference (IPD) was adopted [22]. IPDs are

noted by

$$\cos\text{IPD}_{f,t,c_p,c_q} = \cos(\angle x_{f,t,c_p} - \angle x_{f,t,c_q}) \quad (5)$$

$$\sin\text{IPD}_{f,t,c_p,c_q} = \sin(\angle x_{f,t,c_p} - \angle x_{f,t,c_q}), \quad (6)$$

where  $c_p$  and  $c_q$  are the indices of the two selected microphones and  $\angle x$  is a phase of  $x$ . The separation performance is significantly improved by simultaneously exploiting both the spectral and spatial information.

### 2.3. Unsupervised Training of Deep Separation Models

A previously proposed unsupervised training method [16] uses pseudo-target source signals, which are obtained as BSS outputs. This method trains the deep separation model with the pseudo-target source signals and their mixture in place of ground-truth source signals. A previous method [16] used a complex angular gaussian mixture model (cACGMM) [6] as a BSS module. cACGMM clusters an input mixture by each time-frequency slot using spatial information. In this method, directional statistics vector  $\mathbf{z}_{f,t} = \mathbf{x}_{f,t} / \|\mathbf{x}_{f,t}\|$  as being generated from a mixture distribution as follows:

$$p(\mathbf{z}_{f,t}; \Theta_f) = \sum_{k=1}^K \alpha_{f,k} \mathcal{A}(\mathbf{z}_{f,t}; \mathbf{B}_{f,k}), \quad (7)$$

where  $\Theta_f$  is a set of the parameters of  $K$  mixtures equal to the number of sound sources defined as  $\Theta_f = \{\alpha_{f,k}, \mathbf{B}_{f,k}\}_{k=1}^K$ .  $\alpha_{f,k}$  is a mixture weight, and  $\mathbf{B}_{f,k}$  is a concentration parameter.  $\mathcal{A}(\mathbf{z}_{f,t}; \mathbf{B}_{f,k})$  is denoted by complex angular central gaussian (cACG) distribution as follows:

$$\mathcal{A}(\mathbf{z}; \mathbf{B}) = \frac{(C-1)!}{2\pi^C \det \mathbf{B}} \frac{1}{(\mathbf{z}^H \mathbf{B}^{-1} \mathbf{z})^C}. \quad (8)$$

The parameters are updated by the EM algorithm, and  $\mathbf{z}_{f,t}$  are clustered by direction patterns. After estimating the parameters, the posterior of the class affiliation for each  $\mathbf{z}_{f,t}$  is considered mask  $\mathbf{m}_{f,t,k}$ . Since the parameters are updated independently for frequency, there is ambiguity about the frequency of the so-called frequency permutation problem. For this reason, an external permutation solver is required.

The previous method can train only single-channel models, which cannot use the spatial information available in the training step. On the other hand, if we train a multi-channel model, the model cannot learn the spectral information because it learns not only the well-separated signals but also the inadequately-separated signals. As the result, the model performance is degraded when the spatial information is poor, for example, when the angular difference of DOA between source signals is narrow as in cACGMM.

## 3. Proposed Method

In this section, we propose a simple and easy-to-apply unsupervised training method for multi-channel deep separation models. Our method just generates a training dataset from well-separated signals that contain sufficient clues to learn the separation ability. Fig. 1 overviews our proposed method. Our method includes two steps: the selection and the remixing, which utilizes the minimum angular difference of DOA between source signals. Through these steps, we generate training data that allow the deep separation models to efficiently learn the spectral and spatial information and are less susceptible to a BSS bottleneck, namely, excessive dependence on spatial information.

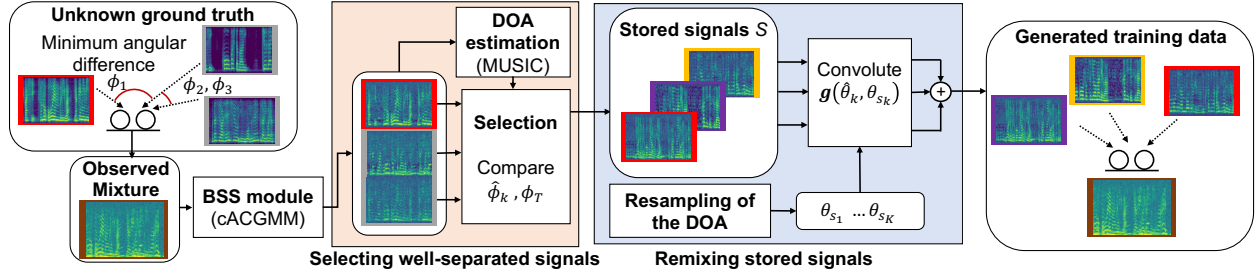


Figure 1: Overview of our proposed method: In selection step, only well-separated signals are stored by minimum angular difference of DOA  $\phi$ . In remixing step, stored signals are convoluted by DOA convert function with resampled DOA. This method outputs pairs of pseudo-target source signals and their mixtures.

### Algorithm 1 Selecting well-separated signals

```

1:  $S = \{\emptyset\}$ 
2: for  $\mathbf{x} \in \mathbf{X}$  do
3:    $\{\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_K\} \leftarrow \text{BSS}(\mathbf{x})$ 
4:   for  $k = 1, \dots, K$  do
5:      $\hat{\theta}_k \leftarrow \text{DOA estimation}(\hat{\mathbf{y}}_k)$ 
6:   end for
7:   for  $k = 1, \dots, K$  do
8:      $\hat{\phi}_k \leftarrow \min_{l \neq k} \|\hat{\theta}_k - \hat{\theta}_l\|$ 
9:     if  $\hat{\phi}_k > \phi_T$  then
10:       $S \leftarrow S \cup \{\hat{\mathbf{y}}_k\}$ 
11:    end if
12:  end for
13: end for
14: return  $S$ 

```

### 3.1. Selection step

The selection step consists of three parts: BSS, DOA estimation, and filtering (Algorithm 1). First, we separate the source signals from mixture signal  $\mathbf{x}$  from set of mixture signals  $\mathbf{X}$  using BSS. We adopt cACGMM as BSS. Second, we estimate the DOAs  $\hat{\theta}_1, \dots, \hat{\theta}_K$  for all the separated signals. Third, we define the minimum angular difference of DOA of the  $k$ -th separated signal as follows:

$$\phi_k = \min_{l \neq k} \|\theta_k - \theta_l\|. \quad (9)$$

We calculate estimated minimum angular difference  $\hat{\phi}_k$  from  $\hat{\theta}_k$  instead of  $\theta_k$  and adopt it as a criterion of how well the signals are separated by BSS. Then it exploits the fact that BSS's performance depends on  $\phi$ . If  $\hat{\phi}_k$  is greater than threshold value  $\phi_T$ ,  $\hat{\mathbf{y}}_k$  is stored in set of selected signals  $S$  as a well-separated signal.

As a DOA estimator, we adopted MUSIC [23], which utilizes the subspace of a spatial covariance matrix  $\mathbf{R}_{f,k} = \sum_{t=1}^T \hat{\mathbf{y}}_{f,t,k} \hat{\mathbf{y}}_{f,t,k}^H$  spanned by source signals and noise signals, where  $H$  represents the hermitian transpose. By scanning  $\theta$ , DOA can be estimated as follows:

$$\hat{\theta}_k = \arg \max_{\theta} \sum_f \frac{\sqrt{\lambda_{1,f,k}} \mathbf{a}_f(\theta)^H \mathbf{a}_f(\theta)}{\sum_{i=2}^C |\mathbf{e}_{i,f,k}^H \mathbf{a}_f(\theta)|^2}, \quad (10)$$

where  $\mathbf{a}(\theta)$  is a steering vector transmitted from direction  $\theta$ ,  $\lambda_{1,f,k}$  is the eigenvalue of the separated signal and  $\mathbf{e}_{2,f,k}, \dots, \mathbf{e}_{C,f,k}$  are the eigenvectors of the noise subspace.

We adopted steering vector  $\tilde{\mathbf{a}}(\theta)$  instead of  $\mathbf{a}(\theta)$  derived from a physical model in a previous work [24] because there is a mismatch in the spatial properties of an unknown environment even if we record the impulse responses.

### Algorithm 2 Remixing stored pseudo-target source signals

```

1:  $\text{Dataset} = \{\emptyset\}$ 
2: while  $|\text{Dataset}| \leq N$  do
3:    $\{\mathbf{y}'_1, \dots, \mathbf{y}'_K\} \leftarrow \text{Sampling signal}(S)$ 
4:   for  $k = 1, \dots, K$  do
5:      $\theta_{s_k} \sim \text{Uniform}(-90, 90)$ 
6:      $\mathbf{y}'_k \leftarrow \mathbf{g}(\hat{\theta}_k, \theta_{s_k}) \otimes \mathbf{y}'_k$ 
7:   end for
8:    $\mathbf{x}' \leftarrow \sum_{k=1}^K \mathbf{y}'_k$ 
9:    $\text{Dataset} \leftarrow \text{Dataset} \cup \{\{\mathbf{x}', \mathbf{y}'_1, \dots, \mathbf{y}'_K\}\}$ 
10: end while
11: return  $\text{Dataset}$ 

```

### 3.2. Remixing step

The remixing step consists of three parts: sampling the stored signals, resampling DOA, and remixing the mixture signals (Algorithm 2). In this step, we generate  $N$  pairs of pseudo-target source signals and their mixtures from the selection step's result. First, we sampled  $k$  well-separated signals  $N$  time as the pseudo-targets from the selected signals  $S$ . Second, by resampling the DOA described below, DOA convert functions  $\mathbf{g}(\hat{\theta}_k, \theta_{s_k}) \in \mathbb{C}^C$  are convoluted with each signal. Here  $\hat{\theta}_k$  and  $\theta_{s_k}$  are the estimated DOA and its resampled value, and  $\otimes$  represented in Algorithm 2 is an element-wise product. This function converts the DOA of  $\mathbf{y}'$  from  $\hat{\theta}_k$  to  $\theta_{s_k}$ . Finally, the remixed mixture is generated as a sum of  $K$  signals. The pairs of  $K$  pseudo-targets and the mixtures are used as a training dataset.

If sound only comes from certain directions, fewer signals with a particular DOA are stored in the selection step. To increase the variety of the DOA of the source signals that compose the remixed mixture signal, we introduce resampling of the DOA and DOA convert function  $\mathbf{g}(\hat{\theta}_k, \theta_{s_k})$ . This process allows the generated training dataset to include a mixture composed of sources with various angular differences. It allows the deep separation models to efficiently learn the spectral and spatial information. This method can be regarded as an application of the spatial normalization that was introduced [25] to data augmentation.  $\mathbf{g}(\hat{\theta}_k, \theta_{s_k})$  is denoted as follows:

$$\mathbf{g}(\hat{\theta}_k, \theta_{s_k}) = \tilde{\mathbf{a}}_f(\theta_{s_k}) \oslash \tilde{\mathbf{a}}_f(\hat{\theta}_k), \quad (11)$$

where  $\oslash$  is an element-wise division and  $\tilde{\mathbf{a}}_f(\theta)$  is a same steering vector of MUSIC. Since this steering vector is derived only from phase differences,  $\mathbf{g}(\hat{\theta}_k, \theta_{s_k})$  does not significantly change spatial properties of the room impulse response convoluted with the source signals except the DOA.  $\mathbf{g}(\hat{\theta}_k, \theta_{s_k})$  normalizes the original DOA by the denominator and converts DOA to  $\theta_{s_k}$  by the numerator.

Table 1: Separation performance of each condition: All columns of evaluation metrics are averages of test data. Each column of evaluation metrics represents average performance for  $\phi$ . The fourth column represents amount of selected signals in selection step.

Method	$\phi_T$	Resampling	Amount of selected signals	SDR [dB]			PESQ		
				$\phi \leq 45^\circ$	$\phi > 45^\circ$	Total	$\phi \leq 45^\circ$	$\phi > 45^\circ$	Total
Proposed	$60^\circ$	✓	44.5%	6.27	6.91	6.63	1.86	1.95	1.91
	$75^\circ$	✓	37.4%	7.06	<b>10.04</b>	<b>8.77</b>	<b>1.97</b>	<b>2.24</b>	<b>2.12</b>
	$90^\circ$	✓	28.9%	6.74	8.90	8.21	1.92	2.15	2.03
	$105^\circ$	✓	19.2%	6.31	7.95	7.25	1.89	2.08	2.00
	$75^\circ$	---	---	---	<b>7.20</b>	8.17	7.76	1.90	2.01
Single-channel [16]	-	-	100.0%	3.48	3.60	3.54	1.72	1.72	1.72
Multi-channel [16]	-	-	100.0%	1.96	7.53	5.15	1.71	2.10	1.93
cACGMM [6]	-	-	-	2.11	7.52	5.21	1.72	2.11	1.94
Supervised	-	-	100.0%	9.87	12.03	11.56	2.18	2.47	2.36

## 4. Experiments

### 4.1. Configuration

We conducted experiments with mixtures composed of two sources. We made a dataset by convoluting the dry source signals of Japanese Newspaper Article Sentences (JNAS)<sup>1</sup> and the impulse responses of the Multi-channel Impulse Response Database<sup>2</sup>. JNAS includes speech from newspapers read by people of various ages and both genders. All signals were downsampled to 8000 Hz. The DOA degree ranged from  $-90^\circ$  to  $90^\circ$  with a resolution of  $15^\circ$ . The impulse response pairs were sampled randomly and not in completely the same direction. We used the impulse responses of two center microphones. The reverberant time was 160 ms. The distance between the microphones and the speakers was 1 meter. The whole mixture was split into 23 hours, 5.9 hours, and a 3.5-hour ratio for training, valid, and test data. Each signal that composed a mixture was randomly weighted from +5 to -5 dB. Gaussian noise was randomly added to all the mixtures from 20 to 30 dB. In any process with a Short Time Fourier Transform (STFT), we used a 32-ms Hann window and 8-ms hops.

We adopted a simple LSTM-based deep separation model. The input was a multi-channel log magnitude and cosIPD and sinIPD of two microphones. The loss function was PIT (Eq. 4). We had two Bidirectional Long-Short Term Memory (BLSTM) layers, each of which had 600 units for each direction. BLSTM was followed by a linear layer and a softmax function. The optimizer was Adam [26]. The learning ratio was  $1.0 \times 10^{-4}$ , and the minibatch size was 64. The training was stopped when the minimum validation loss was not updated for ten consecutive epochs.

We set several threshold values  $\phi_T = 60^\circ, 75^\circ, 90^\circ, 105^\circ$  for selecting well-separated signals. We generated the same amount of training data as the original training mixtures for our proposed method. Note the difference in the amount of pseudo-targets used for learning. In resampling the DOA, the degree was sampled from the same distribution as the original DOA. The resolution of MUSIC was  $1^\circ$ . To solve cACGMM's frequency permutation problem, we used an external permutation solver<sup>3</sup>. The number of cACGMM's EM iterations was 40. To identify the effect of resampling the DOA, we trained a deep separation model without resampling with  $\phi_T = 75^\circ$ .

The evaluation metrics are the Signal-to-Distortion Ratio (SDR) [27] and the Perceptual Evaluation of Speech Qual-

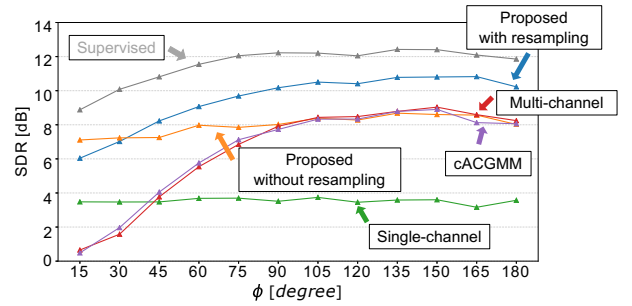


Figure 2: SDR by  $\phi$ : Blue and yellow lines represent our proposed method of  $\phi_T = 75^\circ$ . Performances of supervised and our method have similar tendency.

ity (PESQ) [28]. The baseline is cACGMM, a single-channel model, and a multi-channel model trained by a previous method [16].

### 4.2. Result and Discussion

Table 1 shows the experimental results of each method for SDR and PESQ. In terms of the total average of PESQ, our proposed method of  $\phi_T = 75^\circ$  was the best. Compared with  $\phi_T = 60^\circ$ , we found that the separated signals with  $\phi \leq 60^\circ$  were inappropriate for training. In contrast, the results of  $\phi_T = 90^\circ, 105^\circ$  show that even useful signals for training were removed. In terms of  $\phi > 45^\circ$  of SDR and PESQ, we found that resampling DOA allowed the deep separation model to learn spatial information by comparing rows 2 and 5.

Figure 2 shows the relationship between SDR and  $\phi$ . Although the performance of the baseline multi-channel model is lower than the single-channel model in  $\phi < 90^\circ$ , our proposed method outperformed the single-channel model, even if  $\phi = 15^\circ$ , since the selection step allows the separation model to learn the spectral information. Our proposed method's model performance was monotonically increased to  $\phi = 165^\circ$  because the model learned spatial information and outperformed cACGMM and the baseline multi-channel model by about 2 dB for any  $\phi$ . This difference was derived from whether the model uses spectral information.

## 5. Conclusion

We proposed a simple and easy-to-apply unsupervised training method for multi-channel deep separation models. Our future work apply our proposed method to a real environment including long reverberations or complex noise patterns, which degrade the separation performance.

<sup>1</sup><http://research.nii.ac.jp/src/JNAS.html>

<sup>2</sup><https://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/multi-channel-impulse-response-database/>

<sup>3</sup>It was implemented in [https://github.com/fgnt/pb\\_bss](https://github.com/fgnt/pb_bss)

## 6. References

- [1] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proc. ICASSP*, Apr. 2009, pp. 3437–3440.
- [2] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto, "Infinite positive semidefinite tensor factorization for source separation of mixture signals," in *Proc. ICML*, June 2013, pp. 576–584.
- [3] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, May 2013.
- [4] 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.
- [5] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using gaussian mixture model fitting with dirichlet prior," in *Proc. ICASSP*, Apr. 2009, pp. 33–36.
- [6] N. Ito, S. Araki, and T. Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. EUSIPCO*, Aug. 2016, pp. 1153–1157.
- [7] N. Duong, E. Vincent, and R. Gribonva, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, Sept. 2010.
- [8] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1702–1726, Oct 2018.
- [9] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, May 2016, pp. 31–35.
- [10] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. ICASSP*, Mar. 2017, pp. 246–250.
- [11] Z.-Q. Wang, J. L. Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. ICASSP*, Apr. 2018, pp. 686–690.
- [12] 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 Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 1901–1913, Mar. 2017.
- [13] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 1256–1266, Aug. 2019.
- [14] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *Proc. ICASSP*, May 2020, pp. 696–700.
- [15] J. B. Allen and D. A. Berkley, "The lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, Apr. 1979.
- [16] L. Drude, D. Hasenklever, and R. Haeb-Umbach, "Unsupervised training of a deep clustering model for multichannel blind source separation," in *Proc. ICASSP*, May 2019, pp. 695–699.
- [17] L. Drude, J. Heymann, and R. Haeb-Umbach, "Unsupervised training of neural mask-based beamforming," in *Proc. INTERSPEECH*, Sept. 2019, pp. 1253–1257.
- [18] 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/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1670–1674, Aug. 2021.
- [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 *Proc. ICASSP*, May 2020, pp. 56–60.
- [20] Y. Nakagome, M. Togami, T. Ogawa, and T. Kobayashi, "Mentoring-reverse mentoring for unsupervised multi-channel speech source separation," in *Proc. INTERSPEECH*, Oct. 2020, pp. 86–90.
- [21] N. Takahashi, S. Parthasarathy, N. Goswami, and Y. Mitsufuji, "Recursive speech separation for unknown number of speakers," in *Proc. INTERSPEECH*, Sept. 2019, pp. 1348–1352.
- [22] Z.-Q. Wang, J. L. Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. ICASSP*, Apr. 2018, pp. 1–5.
- [23] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas and Propagation*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.
- [24] H. Sawada, R. Mukai, and S. Makino, "Direction of arrival estimation for multiple source signals using independent component analysis," in *Proc. ISSPA*, vol. 2, Jul. 2003, pp. 411–414.
- [25] R. Takeda, K. Nakadai, and K. Komatani, "Spatial normalization to reduce positional complexity in direction-aided supervised binaural sound source separation," in *Proc. APSIPA*, Dec. 2021, pp. 248–253.
- [26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, May 2015, pp. 1–15.
- [27] R. Scheibler, "SDR – medium rare with fast computations," in *Proc. ICASSP*, May 2022, pp. 701–705.
- [28] 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, May 2001, pp. 749–752.