DEEP BAYESIAN UNSUPERVISED SOURCE SEPARATION BASED ON A COMPLEX GAUSSIAN MIXTURE MODEL

Yoshiaki Bando[1], *Yoko Sasaki*[1], *Kazuyoshi Yoshii*[2]

[1]National Institute of Advanced Industrial Science and Technology, Japan,
[2]RIKEN AIP / Kyoto University, Japan

ABSTRACT

This paper presents an unsupervised method that trains neural source separation by using only multichannel mixture signals. Conventional neural separation methods require a lot of supervised data to achieve excellent performance. Although multichannel methods based on spatial information can work without such training data, they are often sensitive to parameter initialization and degraded with the sources located close to each other. The proposed method uses a cost function based on a spatial model called a complex Gaussian mixture model (cGMM). This model has the time-frequency (TF) masks and direction of arrivals (DoAs) of sources as latent variables and is used for training separation and localization networks that respectively estimate these variables. This joint training solves the frequency permutation ambiguity of the spatial model in a unified deep Bayesian framework. In addition, the pre-trained network can be used not only for conducting monaural separation but also for efficiently initializing a multichannel separation algorithm. Experimental results with simulated speech mixtures showed that our method outperformed a conventional initialization method.

Index Terms— Unsupervised source separation, complex Gaussian mixture model, deep Bayesian learning

1. INTRODUCTION

Deep neural networks (DNNs) have demonstrated excellent performance in source separation tasks, such as speech separation [1–3] and music separation [4, 5]. Permutation invariant training (PIT), for example, trains a DNN to output time-frequency (TF) masks for corresponding sources. Such a method requires a large number of clean source signals and their mixtures for supervised training. It is, however, practically difficult to prepare such supervised data in several tasks. Source separation for audio scene analysis, for example, has to separate daily-life audio events, which are generally captured only in mixture recordings. This calls for an unsupervised method that works without any supervised data.

Thanks to JSPS KAKENHI 18H06490 and 19H04137 for funding.

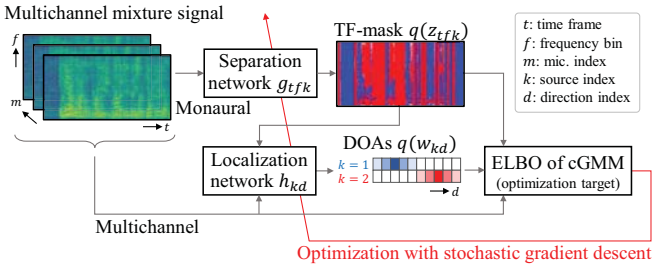

Fig. 1. Overview of cGMM-based unsupervised training.

Unsupervised source separation based on spatial information observed in multichannel recordings has widely been studied [6–9]. A standard approach is to estimate TF masks from phase and power differences among microphones. A complex Gaussian mixture model (cGMM) [9–11], for example, represents such spatial characteristics as spatial covariance matrices (SCMs) and estimates TF masks by clustering TF bins. Since the cGMM is independently formulated at frequency bins, it has permutation ambiguity that the indices of sources are not aligned over frequency bins. This ambiguity can be resolved by aligning estimated sources based on the direction of arrival (DoA) of each source, and several methods have been proposed to jointly estimate the TF masks and DoAs [9, 11]. The directional information also makes it possible to estimate the number of sources [9, 11], which many methods require in advance [6–8]. The multichannel methods, however, are often sensitive to parameter initialization and degraded when the sources are located close to each other.

Unsupervised training for neural source separation using multichannel mixture signals has recently gained a lot of attention [12–15]. One approach is to generate supervised data by using multichannel separation methods [12–14]. This approach suffers from the estimation errors of the multichannel methods mentioned above. To solve this problem, Drude et al. [15] trained a separation network by directly optimizing the likelihood function of a cGMM. They reported that the performance of a conventional multichannel method was improved by initializing it with the network output. To solve the frequency permutation ambiguity by using the correlation of TF masks over frequency bins [16], the method requires the number of latent sources in advance. It is thus difficult

978-1-7281-0824-7/19/$31.00 ©2019 IEEE

to apply this method for recordings of daily-life audio events, which include an unknown number of source signals.

To tackle this problem, we solve the frequency permutation ambiguity by jointly training separation and localization networks instead of using the correlation of masks (Fig. 1). The objective function is derived as an evidence lower bound (ELBO) [17] of a cGMM that has TF masks and DoAs as latent variables. Given the geometry of a microphone array, the two networks are trained to respectively estimate the posterior probabilities of the TF masks and DoAs. Since DoAs can be used for counting the number of sources in a mixture recording, our framework could be extended to deal with training data including an unknown number of sources by utilizing a non-parametric Bayesian model [9, 18].

The main contribution of this paper is to resolve the frequency permutation ambiguity with a unified deep Bayesian framework during the unsupervised training. We show that the separation network can be trained from random weights by maximizing the ELBO without any additional solvers or steps for the permutation problem. The trained network can be used not only for conducting monaural source separation but also for efficiently initializing a multichannel separation algorithm. Experimental results also show that the proposed method outperforms an existing initialization method.

2. RELATED WORK

This section overviews cGMM-based TF clustering and then introduces unsupervised neural source separation.

2.1. Complex Gaussian mixture models

A popular approach to separating a multichannel mixture signal is to mask each TF bin [9–11,19,20]. This mask is conventionally estimated by clustering hand-crafted features at each TF bin [19, 20]. To directly conduct a clustering on a multichannel spectrogram, probabilistic mixture models for a multichannel observation have been studied [9–11]. The cGMM, for example, represents the multichannel spectrogram as a mixture of complex Gaussian distributions with SCMs and power spectral densities of sources [10]. A complex angular central Gaussian mixture model (cACGMM) [21] is defined on a multichannel spectrogram normalized by power at each TF bin. It has been proven that the expectation-maximization (EM) algorithms for the cGMM and cACGMM are equivalent [21]. Since these models are independently formulated at frequency bins, they have the frequency permutation ambiguity. To solve this problem, a cGMM-based method estimates the TF mask and DoA of each source by using an inverse Wishart mixture prior on the SCMs [11]. Wishart distributions of this mixture represent potential DoAs characterized by using premeasured steering vectors. Another cGMM-like spatial model inspired by latent Dirichlet allocation (LDA) [9] jointly estimates the TF mask and the DoA of each source, and the number of sources in a unified framework. This joint estimation is conducted with a collapsed Gibbs sampling by assuming a hierarchical Dirichlet process.

2.2. Unsupervised training of neural source separation

Unsupervised training of neural source separation has been studied by using visual information [22, 23] or multichannel recordings [12–14]. The audio-visual-based methods use video recordings that capture the audio events and corresponding visual events, such as music signals and corresponding performances [22, 23]. These methods are based on the co-occurrence of the audio and visual events and train a network so that the separated signals correlate to the visual events. Multichannel-audio-based methods, on the other hand, can train a DNN to separate sound sources out of view or behind obstacles. Tzinis et al. [14] trained a monaural separation network by using source signals estimated by applying K-means clustering on interchannel phase differences (IPDs) between two microphones. Almost simultaneously, Drude et al. [12] proposed a similar approach that uses signals separated by the cACGMM [21]. They reported that the cACGMM performance was improved by initializing it with the pre-trained separation network. Seetharaman et al. [13] designed a loss function weighted by a confidence measure of the estimated references. Drude et al. [15] also proposed a novel approach that directly trains a separation network from the cACGMM likelihood. They applied the method to noisy speech recordings and reported that the performance of automatic speech recognition was superior to that of their previous approach mentioned above.

3. DEEP BAYESIAN SOURCE SEPARATION

The proposed method trains separation and localization networks by using only multichannel mixture signals and resolves the frequency permutation ambiguity in a unified framework. This training is based on the LDA model [9, 24], which has TF masks and DoAs of sources as latent variables. The objective function is derived as an ELBO of the spatial model, which consists of an expectation of the likelihood function and a Kullback-Leibler (KL) divergence between the network outputs and their prior distributions. Since the existing studies [9, 24] only show Bayesian inference for the LDA model, we also describe an EM algorithm of the model and initialize it with the pre-trained network.

3.1. Probabilistic generative model

To jointly estimate the TF-masks and DoAs of latent sound sources, an observed M-channel spectrogram $\mathbf{x}_{tf} \in \mathbb{C}^M$ is represented as a sum of K source spectrograms $s_{tfk} \in \mathbb{C}$:

$$\mathbf{x}_{tf} = \sum_{k=1}^{K} \sum_{d=1}^{D} z_{tfk} w_{kd} \left(\mathbf{a}_{fd} s_{tfk} \right), \tag{1}$$

where $z_{tfk} \in \{0,1\}$ ($\sum_{k=1}^{K} z_{tfk} = 1$) is a TF mask that indicates which source is relevant at each TF bin, $w_{kd} \in \{0,1\}$ ($\sum_{d=1}^{D} w_{kd} = 1$) is a DoA variable that assigns source k to a DoA candidate $d \in \{1,\ldots,D\}$, and $\mathbf{a}_{fd} \in \mathbb{C}^M$ is a steering vector for direction d. As in other cGMMs [9–11], the TF mask z_{tfk} is introduced by assuming a sparseness that each TF bin has exclusively one relevant source. The potential directions d are, in this paper, assumed as directions with an angular interval of 5° on a horizontal plane ($D = 72$).

The TF masks and DoAs are estimated as their posterior probabilities by putting prior distributions on them. Since the activity of each source changes over time frames, a frame-wise categorical distribution (denoted as Cat) is put on the TF-masks z_{tfk} as follows:

$$[z_{tf1},\ldots,z_{tfK}]^{\mathsf{T}} \mid \boldsymbol{\pi}_t \sim \mathrm{Cat}\left(\pi_{t1},\ldots,\pi_{tK}\right), \tag{2}$$

where $\pi_{tk} \in \mathbb{R}_+$ ($\sum_{k=1}^{K} \pi_{tk} = 1$) is a model parameter to be estimated. On the other hand, the following categorical distribution is put on w_{kd} as follows:

$$\mathbf{w}_k = [w_{k1},\ldots,w_{kD}]^{\mathsf{T}} \sim \mathrm{Cat}\left(\phi_1,\ldots,\phi_D\right). \tag{3}$$

where $\phi_d \in \mathbb{R}_+$ ($\sum_{d=1}^{D} \phi_d = 1$) is a model parameter.

Each source spectrogram s_{tfk} is assumed to follow a zero-mean complex Gaussian distribution:

$$s_{tfk} \sim \mathcal{N}_{\mathbb{C}}\left(0, \lambda_{tfk}\right), \tag{4}$$

where $\mathcal{N}_{\mathbb{C}}\left(\mu,\sigma^2\right)$ is a complex Gaussian distribution with mean μ and variance σ^2, and $\lambda_{tfk} \in \mathbb{R}_+$ represents the power spectral density of source k. Using (1) and (4), an observed mixture signal $\mathbf{x}_{tf}$ is found to follow a multivariate complex Gaussian mixture distribution as follows:

$$\mathbf{x}_{tf} \sim \prod_{k=1}^{K}\prod_{d=1}^{D} \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \lambda_{tfk}\mathbf{H}_{fd}\right)^{z_{tfk}w_{kd}}, \tag{5}$$

where $\mathbf{H}_{fd} = \mathbb{E}[\mathbf{a}_{fd}\mathbf{a}_{fd}^{\mathsf{H}}] \in \mathbb{C}^{M\times M}$ is a SCM of direction d. To estimate $\mathbf{H}_{fd}$ while constraining it to direction d, the following complex inverse Wishart distribution is put on $\mathbf{H}_{fd}$:

$$\mathbf{H}_{fd} \sim \mathcal{IW}_{\mathbb{C}}\left(\nu, (\nu - M)\mathbf{G}_{fd}\right), \tag{6}$$

where $\mathcal{IW}_{\mathbb{C}}(\nu,\mathbf{G}) \propto |\mathbf{H}|^{-(\nu+M)}\exp[-\mathrm{tr}(\mathbf{G}\mathbf{H}^{-1})]$ represents the complex inverse Wishart distribution, $\nu > M$ is a hyperparameter, $\mathbf{G}_{fd} = \mathbf{b}_{fd}\mathbf{b}_{fd}^{\mathsf{H}} + \epsilon\mathrm{I} \in \mathbb{C}^{M\times M}$ is a template SCM for direction d. The $\mathbf{b}_{fd}$ is a template steering vector for direction d and prepared in advance, and $\epsilon\mathbf{I}$ ($\epsilon > 0$) is added to make $\mathbf{G}_{fd}$ positive definite.

3.2. Variational inference framework

Both the proposed unsupervised training and multichannel separation are based on a variational inference that estimates the posterior distribution $p(\mathbf{Z},\mathbf{W}|\mathbf{X},\boldsymbol{\Theta})$, where $\boldsymbol{\Theta} = \{\mathbf{H},\boldsymbol{\lambda},\boldsymbol{\pi},\boldsymbol{\phi}\}$ represents the parameters obtained by point estimation. Since it is difficult to analytically calculate the true posterior distribution $p(\mathbf{Z},\mathbf{W}|\mathbf{X},\boldsymbol{\Theta})$, we approximate it with the following variational posterior distribution:

$$p(\mathbf{Z},\mathbf{W}|\mathbf{X},\boldsymbol{\Theta}) \approx q(\mathbf{Z})q(\mathbf{W}). \tag{7}$$

The variational inference is conducted by maximizing the following lower bound of the log marginal likelihood $p(\mathbf{X}|\boldsymbol{\Theta})$:

$$\begin{aligned}\mathcal{L} = {} & \mathbb{E}_q\left[\log p(\mathbf{X} \mid \boldsymbol{\lambda},\mathbf{H},\mathbf{Z},\mathbf{W})\right] \\ & - \mathbb{KL}\left[q(\mathbf{Z})|p(\mathbf{Z}|\boldsymbol{\pi})\right] - \mathbb{KL}\left[q(\mathbf{W})|p(\mathbf{W}|\boldsymbol{\phi})\right].\end{aligned} \tag{8}$$

The lower bound $\mathcal{L}$ is called an ELBO, and its maximization corresponds to the minimization of KL divergence between the variational and true posterior distributions. This framework iteratively and alternately updates the variational posteriors q and parameters $\boldsymbol{\Theta}$ until convergence.

The SCM $\mathbf{H}$ is updated with maximum a posteriori (MAP) estimation and the other parameters $\boldsymbol{\lambda}$, $\boldsymbol{\pi}$, and $\boldsymbol{\phi}$ are updated with maximum likelihood estimation. Since it is also difficult to analytically calculate these variables, we update them by using the ELBO (8) as follows:

$$\mathbf{H}_{fd} \leftarrow \frac{\mathbf{G}_{fd} + \sum_{t,k=1}^{T,K} \hat{z}_{tfk}\hat{w}_{kd}\frac{1}{\lambda_{tfk}}\mathbf{x}_{tf}\mathbf{x}_{tf}^{\mathsf{H}}}{\nu + \sum_{t,k=1}^{T,K}\hat{z}_{tfk}\hat{w}_{kd} + M}, \tag{9}$$

$$\lambda_{tfk} \leftarrow \frac{1}{M}\sum_{d=1}^{D}\hat{w}_{kd}\mathbf{x}_{tf}^{\mathsf{H}}\mathbf{H}_{fd}^{-1}\mathbf{x}_{tf}, \tag{10}$$

$$\pi_{tk} \leftarrow \frac{1}{F}\sum_{f=1}^{F}\hat{z}_{tfk}, \qquad \phi_d \leftarrow \frac{1}{K}\sum_{k=1}^{K}\hat{w}_{kd}, \tag{11}$$

where $\hat{z}_{tfk}$ is $q(z_{tfk} = 1)$ and $\hat{w}_{kd}$ is $q(w_{kd} = 1)$.

3.3. Training based on amortized variational inference

By using N mixture signals $\mathbf{x}_{tf}^{(n)}$, we train separation and localization networks that respectively estimate the TF mask z_{tfk} and DoA w_{kd} (Fig. 1). The suffix $^{(n)}$ is hereinafter omitted because the objective function is a sum of the local loss value for each mixture signal $\mathbf{x}_{tf}^{(n)}$. The separation network (denoted by g_{tfk}) takes as input a monaural log-magnitude spectrogram and expects the posterior distribution of the TF mask $q_g(z_{tfk} = 1)$:

$$q_g(z_{tfk} = 1) = \hat{z}_{tfk} = g_{tfk}(\log|\mathbf{X}|), \tag{12}$$

where $\log|\mathbf{X}| \in \mathbb{R}^{T\times F}$ denotes a monaural log-magnitude spectrogram. We simply take the recording of the first microphone ($m = 1$) as the input. The localization network (denoted by h_{kd}), on the other hand, expects the probability that direction d is selected for the k-th source $q_h(w_{kd} = 1)$:

$$q_h(w_{kd} = 1) = \hat{w}_{kd} = h_{kd}\left(\boldsymbol{\omega}\right), \tag{13}$$

where $\boldsymbol{\omega} = \{\omega_{kd}\}_{k,d=1}^{K,D} \in \mathbb{R}^{K\times D}$ is an input feature that represents spatial characteristics. Since it is difficult for networks to directly take complex numbers as input, we alternatively use the following Gaussian-mixture log likelihood:

$$\omega_{kd} = \sum_{t=1}^{T}\sum_{f=1}^{F}\hat{z}_{tfk}\log\mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{tf};\mathbf{0},\mathbf{G}_{fd}\right). \tag{14}$$

The training of networks g_{tfk} and h_{kd} is conducted by maximizing the ELBO $\mathcal{L}$ for each mixture signal in the training data. For numerical stability, we fix λ_{tfk} and $\mathbf{H}_{fd}$ to $\tilde{\lambda} =$

$\frac{1}{TFM}\sum_{t,f=1}^{T,F}\mathbf{x}_{tf}^{\mathsf{H}}\mathbf{x}_{tf}$ and $\mathbf{G}_{fd}$, respectively. More specifically, the proposed training is conducted by iteratively executing the following three steps:

1) predict TF masks $\hat{z}_{tfk}$ and DoAs $\hat{w}_{kd}$ with g_{tfk} and h_{kd} for each mixture recording in a mini-batch,
2) update model parameters $\boldsymbol{\Theta} = \{\boldsymbol{\pi}, \boldsymbol{\phi}\}$ with (11), and
3) calculate $\mathcal{L}$ and update the network parameters by using a stochastic gradient descent (SGD) method.

The ELBO $\mathcal{L}$ can be calculated as follows:

$$\mathcal{L} = -\sum_{t,f,k,d=1}^{T,F,K,D}\hat{z}_{tfk}\hat{w}_{kd}\left(\log|\mathbf{G}_{fd}| + \frac{1}{\tilde{\lambda}}\mathbf{x}_{tf}^{\mathsf{H}}\mathbf{G}_{fd}^{-1}\mathbf{x}_{tf}\right) + \sum_{t,f,k=1}^{T,F,K}\hat{z}_{tfk}\log\frac{\pi_{tk}}{\hat{z}_{tfk}} + \sum_{k,d=1}^{K,D}\hat{w}_{kd}\log\frac{\phi_d}{\hat{w}_{kd}} + \text{const.}. \quad (15)$$

The loss value for a mini-batch is a sum of this local ELBO normalized with $\frac{1}{TF}$. Our method trains neural networks to estimate posterior distributions for unseen observed data by using a training data prepared in advance. This kind of training is called amortized variational inference [17, 25].

3.4. Multichannel separation based on an EM-algorithm

Although the trained network g_{tfk} can be used to separate sources from a monaural mixture signal, it can also improve the performance of a multichannel EM algorithm by initializing TF masks with the network output. The EM algorithm for the cGMM (EM-cGMM) alternately iterates the following E-step and M-step. The E-step updates the TF masks $\hat{z}_{tfk}$ and DoAs $\hat{w}_{kd}$ so that the ELBO $\mathcal{L}$ is maximized:

$$\hat{z}_{tfk} \leftarrow \frac{\pi_{tk}\prod_{d=1}^{D}\mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{tf};\mathbf{0},\lambda_{tfk}\mathbf{H}_{fd}\right)^{\hat{w}_{kd}}}{\sum_{K=1}^{K}\pi_{tk}\prod_{d=1}^{D}\mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{tf};\mathbf{0},\lambda_{tfk}\mathbf{H}_{fd}\right)^{\hat{w}_{kd}}}, \quad (16)$$

$$\hat{w}_{kd} \leftarrow \frac{\phi_d\prod_{t,f=1}^{T,F}\mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{tf};\mathbf{0},\lambda_{tfk}\mathbf{H}_{fd}\right)^{\hat{z}_{tfk}}}{\sum_{d=1}^{D}\phi_d\prod_{t,f=1}^{T,F}\mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{tf};\mathbf{0},\lambda_{tfk}\mathbf{H}_{fd}\right)^{\hat{z}_{tfk}}}. \quad (17)$$

The M-step, on the other hand, updates the parameters $\boldsymbol{\Theta}$ by using (9)–(11). Since the EM algorithm alternately updates these variables until convergence, the careful initialization is important to avoid falling into a local optimum.

The TF masks $\hat{z}_{tfk}$ are initialized by using the output of the separation network g_{tfk}. Since the localization network g_{tfk} can potentially overfit to the spatial bias of the training data, we initialize the DoA $\hat{w}_{kd}$ by using the following formula instead of the output of h_{kd}:

$$\hat{w}_{kd} \propto \exp\left(-\sum_{t=1}^{T}\sum_{f=1}^{F}\hat{z}_{tfk}\mathbf{x}_{tf}^{\mathsf{H}}\mathbf{G}_{fd}^{-1}\mathbf{x}_{tf}\right). \quad (18)$$

4. EXPERIMENTAL EVALUATION

We conducted an evaluation with speech mixture signals generated by using simulated room impulse responses (RIRs).

4.1. Dataset

The mixture signals used in this evaluation were generated by convolving RIRs to source signals in the WSJ0-mix dataset [1], which is widely used for neural speech separation [1–3]. Each of the mixture signals in this dataset included two utterances from two randomly selected speakers in the WSJ0 corpus. The two speech signals were mixed with a signal-to-noise ratio randomly chosen between -5 and $+5$ dB. The RIRs applied to the speech signals were simulated by using the image method[1] [26] with the room configuration randomly changed at each mixture signal between 5 m×5 m×3 m and 10 m×10 m×4 m. We assumed a 4-channel microphone array with the diameter of 8 cm located at the center of the room. The source locations of two speech signals were randomly placed in the room. The reverberation time (RT_{60}) was chosen at random between 0.2 and 0.4 s. The training and validation sets had 20,000 and 5,000 mixture signals, respectively. The test set had 3,000 mixture signals whose speakers were separated from the training and validation sets. We generated these signals with a sampling rate of 8 kHz to reduce computational and memory costs.

4.2. Experimental Condition

The network architectures for the proposed method were experimentally determined as follows. The separation network g_{tfk} had three layers of bi-directional long short-term memory (BiLSTM), each with 600 units for each direction, and one fully connected layer followed by a softmax activation. To reduce the parameters of the localization network h_{kd}, the network h_{kd} consisted of three layers of 1D-convolution with the direction axis d as the convolution axis of each layer. The filter size of the convolution layers and the number of the filters were respectively set to 1 and 2 ($= K$). The network h_{kd} outputs $\log\hat{w}_{kd}$ through a residual connection with the network input.

The separation network g_{tfk} and localization network h_{kd} were jointly optimized using the Adam optimizer [27]. The learning rate of the optimizer was initialized to 1.0×10^{-3} and scaled down by 0.7 when the training loss value increased compared to that of the last epoch. The spectrograms $\mathbf{x}_{tf}$ were obtained with the short-time Fourier transform (STFT) with a window length of 512 samples and a shifting interval of 128 samples. The hyperparameters ν and ϵ were set to $M+5.0$ and 1.0×10^{-2}, respectively. The template steering vectors $\mathbf{b}_{fd}$ were theoretically calculated under the planewave assumption. Note that the $\mathbf{b}_{fd}$ and the RIRs used for generating the mixture signals were much different because the sound sources were randomly located on the room under reverberant conditions. We iterated the EM-cGMM 50 times. The source signals were obtained by masking the observation $\mathbf{x}_{tf}$ with the estimated TF mask $\hat{z}_{tfk}$.

[1] https://github.com/ty274/rir-generator

Table 1. Averages and standard deviations of SDRs

Method	Init.	# of mics. M train	# of mics. M test	SDR [dB]
EM-cGMM	g_{tfk}	4	4	**10.6 ± 4.2**
EM-cGMM	(19)–(20)	–	4	9.7 ± 5.0
AVI-cGMM	–	4	1	5.3 ± 4.5
AuxIVA+	–	–	4	9.9 ± 4.4
AuxIVA	–	–	2	5.6 ± 4.0
PIT	–	1	1	7.7 ± 4.5
DPCL	–	1	1	6.9 ± 4.7

The proposed method was compared with an independent vector analysis (AuxIVA) [8], and the supervised methods of PIT and deep clustering (DPCL) [1]. AuxIVA was evaluated with two channels in all the four channels because it assumes that the number of microphones equals that of sources. To use all the four microphones, we also evaluated an extension of AuxIVA (AuxIVA+) that conducts AuxIVA with a 4-channel input and clusters the separated signals into two sources [28]. The dimension of the latent space for DPCL was set to 20. The separation networks for PIT and DPCL had the same condition as g_{tfk} in the proposed method. We compared the proposed neural initialization for EM-cGMM with the initialization method proposed by Otsuka et al. [9, 24]. Given a sufficient number of source classes K, this method splits directions $d = 1, \ldots, D$ into K groups and initializes the TF masks $\hat{z}_{tfk}$ by using the directional information:

$$\hat{w}_{kd} \propto \begin{cases} 1 & (k-1)\frac{D}{K} \leq d < k\frac{D}{K} \\ 0 & \text{otherwise} \end{cases}, \tag{19}$$

$$\hat{z}_{tfk} \propto \exp\left(-\sum_{d=1}^{D} \hat{w}_{kd} \mathbf{x}_{tf}^{\mathsf{H}} \mathbf{G}_{fd}^{-1} \mathbf{x}_{tf} \right). \tag{20}$$

We set the number of source classes $K = 6$ for this method. The separation performance was evaluated using the signal-to-distortion ratio (SDR) [29].

4.3. Experimental Results

The average SDRs for the test set were summarized in Table 1. The EM-cGMM initialized with g_{tfk} outperformed that with the conventional initialization ((19)–(20)). In addition, it outperformed AuxIVA+, which uses the same number of microphones as the EM-cGMMs. Fig 2 shows the relationship between the DoA differences and SDRs. The EM-cGMM initialized with (19)–(20) significantly deteriorated when the DoA difference was less than 60°. The EM-cGMM initialized with g_{tfk} improved the SDRs in such a condition. The monaural separation with g_{tfk} (AVI-cGMM) achieved 5.3 dB in the average SDR. When the mixture signals had speakers of difference genders (m+f in Table 2), AVI-cGMM outperformed AuxIVA with 2-ch observations.

The initialization with g_{tfk} occasionally decreased the performance regardless of the DoA differences, which is shown as the SDR results around 0 dB in Fig. 2. This is because g_{tfk} (AVI-cGMM) deteriorated with the mixture signals of the same gender speakers (m+m and f+f in Table 2), which are difficult to separate from spectral features. Since the performances of PIT and DPCL were higher than that of the AVI-cGMM, the g_{tfk} has a potential to separate such signals. Comparing AVI-cGMM with EM-cGMM initialized with g_{tfk}, AVI-cGMM could be further improved by making it possible to estimate λ_{tfk} and $\mathbf{H}_{fd}$ during the training. This extension will compensate with the mismatch between the fixed parameters $\tilde{\lambda}$ and $\mathbf{G}_{fd}$ and the observation due to reverberations and reflections.

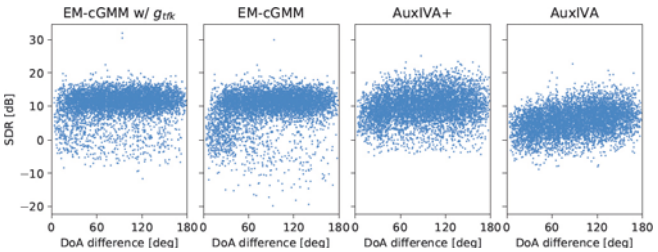

Fig. 2. Scatter plots for the DOA difference of two sources and the corresponding SDR performance.

Table 2. SDRs [dB] averaged by the genders (m: male, f: female) of the speakers in mixture signals.

Method	Init.	m+m	f+f	m+f
EM-cGMM	g_{tfk}	9.2 ± 4.9	**10.2 ± 5.2**	**11.5 ± 3.1**
EM-cGMM	(19)–(20)	9.5 ± 4.6	10.1 ± 5.2	9.7 ± 5.1
AVI-cGMM	–	2.0 ± 4.0	3.0 ± 4.4	7.9 ± 2.8
AuxIVA+	–	**10.2 ± 4.4**	9.3 ± 4.5	9.9 ± 4.4
AuxIVA	–	5.7 ± 3.9	5.4 ± 4.1	5.7 ± 4.0
PIT	–	4.9 ± 4.4	5.2 ± 4.8	10.1 ± 2.7
DPCL	–	3.8 ± 4.6	4.0 ± 4.8	9.6 ± 2.8

5. CONCLUSION

We presented an unsupervised method that trains neural source separation by using only multichannel mixture signals. The proposed method trains separation and localization networks by using a cost function based on a cGMM that has the TF masks and DoAs as latent variables. This joint training enables us to resolve the frequency permutation ambiguity without any additional solvers or steps. In addition, the trained network can also be used for efficiently initializing the cGMM-based multichannel EM algorithm. We experimentally confirmed that the proposed initialization method outperformed a conventional initialization method. To deal with the training data having an unknown number of sources, we plan to train a separation network while estimating the number of sources with the directional information. We also plan to improve the proposed training method with the joint estimation of SCMs and power spectral densities.

6. REFERENCES

[1] J. R. Hershey, et al., "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE ICASSP*, 2016, pp. 31–35.

[2] M. Kolbæk, et al., "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM TASLP*, vol. 25, no. 10, pp. 1901–1913, 2017.

[3] L. Drude, et al., "Deep attractor networks for speaker re-identification and blind source separation," in *IEEE ICASSP*, 2018, pp. 11–15.

[4] A. Jansson, et al., "Singing voice separation with deep U-Net convolutional networks," in *ISMIR*, 2017, pp. 745–751.

[5] Y. Luo, et al., "Deep clustering and conventional networks for music separation: Stronger together," in *IEEE ICASSP*, 2017, pp. 61–65.

[6] A. Ozerov et al., "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE/ACM TASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[7] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Trans. on Circuits and Systems I: Regular Papers*, vol. 57, no. 7, pp. 1431–1438, 2010.

[8] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *IEEE WASPAA*, 2011, pp. 189–192.

[9] T. Otsuka, et al., "Bayesian nonparametrics for microphone array processing," *IEEE/ACM TASLP*, vol. 22, no. 2, pp. 493–504, 2014.

[10] T. Higuchi, et al., "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *IEEE ICASSP*, 2016, pp. 5210–5214.

[11] J. Azcarreta, et al., "Permutation-free CGMM: Complex Gaussian mixture model with inverse Wishart mixture model based spatial prior for permutation-free source separation and source counting," in *IEEE ICASSP*, 2018, pp. 51–55.

[12] L. Drude, et al., "Unsupervised training of a deep clustering model for multichannel blind source separation," in *IEEE ICASSP*, 2019, pp. 695–699.

[13] P. Seetharaman, et al., "Bootstrapping single-channel source separation via unsupervised spatial clustering on stereo mixtures," in *IEEE ICASSP*, 2019, pp. 356–360.

[14] E. Tzinis, et al., "Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information," in *IEEE ICASSP*, 2019, pp. 81–85.

[15] L. Drude, et al., "Unsupervised training of neural mask-based beamforming," *arXiv preprint arXiv:1904.01578 (accepted to Interspeech 2019)*, 2019.

[16] H. Sawada, et al., "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *IEEE ISCAS*, 2007, pp. 3247–3250.

[17] D. P. Kingma et al., "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[18] K. Kurihara, et al., "Collapsed variational Dirichlet process mixture models." in *IJCAI*, vol. 7, 2007, pp. 2796–2801.

[19] S. Araki, et al., "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior," in *IEEE ICASSP*, 2009, pp. 33–36.

[20] M. I. Mandel, et al., "An em algorithm for localizing multiple sound sources in reverberant environments," in *NIPS*, 2007, pp. 953–960.

[21] N. Ito, et al., "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *EUSIPCO*, 2016, pp. 1153–1157.

[22] A. Owens et al., "Audio-visual scene analysis with self-supervised multisensory features," in *ECCV*, 2018, pp. 1–18.

[23] A. Rouditchenko, et al., "Self-supervised audio-visual co-segmentation," in *IEEE ICASSP*, 2019, pp. 2357–2361.

[24] T. Otsuka, et al., "Bayesian unification of sound source localization and separation with permutation resolution," in *AAAI*, 2012, pp. 2038–2045.

[25] R. Ranganath, et al., "Black box variational inference," in *AISTATS*, 2014, pp. 814–822.

[26] J. B. Allen et al., "Image method for efficiently simulating small-room acoustics," *JASA*, vol. 65, no. 4, pp. 943–950, 1979.

[27] D. P. Kingma et al., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] D. Kitamura, et al., "Relaxation of rank-1 spatial constraint in overdetermined blind source separation," in *EUSIPCO*, 2015, pp. 1261–1265.

[29] E. Vincent, et al., "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.