Neural Full-Rank Spatial Covariance Analysis for Blind Source Separation

Yoshiaki Bando, *Member, IEEE*, Kouhei Sekiguchi, *Member, IEEE*, Yoshiki Masuyama, *Member, IEEE*, Aditya Arie Nugraha, *Member, IEEE*, Mathieu Fontaine, *Member, IEEE*, and Kazuyoshi Yoshii, *Member, IEEE*

***Abstract*—This paper describes a neural blind source separation (BSS) method based on amortized variational inference (AVI) of a non-linear generative model of mixture signals. A classical statistical approach to BSS is to fit a linear generative model that consists of spatial and source models representing the inter-channel covariances and power spectral densities of sources, respectively. Although the variational autoencoder (VAE) has successfully been used as a non-linear source model with latent features, it should be pretrained from a sufficient amount of isolated signals. Our method, in contrast, enables the VAE-based source model to be trained only from mixture signals. Specifically, we introduce a neural mixture-to-feature inference model that directly infers the latent features from the observed mixture and integrate it with a neural feature-to-mixture generative model consisting of a full-rank spatial model and a VAE-based source model. All the models are optimized jointly such that the likelihood for the training mixtures is maximized in the framework of AVI. Once the inference model is optimized, it can be used for estimating the latent features of sources included in unseen mixture signals. The experimental results show that the proposed method outperformed the state-of-the-art BSS methods based on linear generative models and was comparable to a method based on supervised learning of the VAE-based source model.**

***Index Terms*—Neural source separation, unsupervised training, deep generative models, variational autoencoders.**

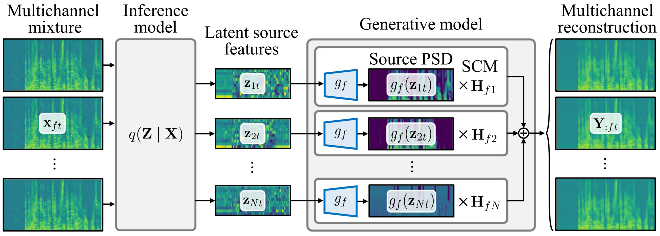


Fig. 1. Overview of proposed unsupervised method called neural FCA.

I. Introduction

SOUND source separation is a fundamental function to understand acoustic scenes computationally [1]–[4]. Blind source separation (BSS), for example, has been actively investigated to separate a multichannel mixture signal into source signals with little prior information about the sources and microphones [5]–[10]. Such a modern statistical multichannel method is based on a generative model consisting of spatial and source models. A standard spatial model assumes full-rank spatial covariance matrices (SCMs) [8] of sources at each frequency bin. This model is originally proposed for full-rank spatial covariance analysis (FCA) [8] to handle small source movements and reverberation. Since SCMs are independently defined for each frequency bin, their source indices are not aligned over frequencies. Various source models have been proposed not only to represent source signals precisely but also to solve the frequency permutation ambiguity [10]–[12].

Deep spectral models have recently gained attention for source models to have powerful expression capability for complex source spectra [13]–[18]. A typical model utilizes the decoder of a variational autoencoder (VAE) [19] as a non-linear generative model of a source signal. The VAE is trained in advance to represent the target source spectra. Its trained decoder is combined with a spatial model to separate source signals by estimating the latent feature vectors of sound sources from a multichannel mixture. It has been reported that the multichannel speech separation with a VAE-based source model outperformed the conventional linear model based on non-negative matrix factorization (NMF) [13], [14].

A main drawback of the existing deep spectral models is that their training requires a sufficient number of isolated source signals. Since most natural audio events are mainly captured only in mixture signals, it is practically hard to prepare such supervised training data. Although several kinds of sources (e.g., speech) have clean source corpora, the domain mismatch at the target environments (e.g., the Lombard effect [20]) could degrade the separation performance of the trained model.

In this paper, we propose an unsupervised method that only requires multichannel mixture signals to train a deep spectral

Manuscript received May 28, 2021; revised July 23, 2021; accepted July 25, 2021. Date of publication August 4, 2021; date of current version August 30, 2021. This work was supported in part by JST ACT-X under Grant JPMJAX200N; in part by NEDO; and in part by JSPS KAKENHI under Grants JP19H04137, JP20K21813, JP20H00602, and JP20K19833. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Maximo Cobos. *(Corresponding author: Yoshiaki Bando.)*

Yoshiaki Bando is with the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan, and also with the RIKEN Center for Advanced Intelligence Project (AIP), Tokyo 103-0027, Japan (e-mail: y.bando@aist.go.jp).

Kouhei Sekiguchi, Aditya Arie Nugraha, and Mathieu Fontaine are with the RIKEN Center for Advanced Intelligence Project (AIP), Tokyo 103-0027, Japan (e-mail: kouhei.sekiguchi@riken.jp; adityaarie.nugraha@riken.jp; mathieu.fontaine@riken.jp).

Yoshiki Masuyama is with the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan, and also with Tokyo Metropolitan University, Tokyo 191-0065, Japan (e-mail: mas-03151102@akane.waseda.jp).

Kazuyoshi Yoshii is with the RIKEN Center for Advanced Intelligence Project (AIP), Tokyo 103-0027, Japan, and also with Kyoto University, Kyoto 606-8501, Japan (e-mail: yoshii@i.kyoto-u.ac.jp).

Digital Object Identifier 10.1109/LSP.2021.3101699

model. The proposed method called neural FCA utilizes a generative model consisting of the full-rank spatial model and the deep spectral model. As shown in Fig. 1, this generative model is regarded as a large decoder of a VAE and jointly trained with an inference model (encoder) that estimates the latent source features from a mixture signal. Once the networks are trained, the inference model is utilized to estimate the latent source features of an unseen mixture signal.

The main contribution of this study is to jointly train a deep spectral model and its inference model with a unified solution to the frequency permutation problem. Several unsupervised methods have been proposed to train a separation network by using multichannel mixture signals and the conventional statistical models [21]–[26]. The existing methods, however, resolve the frequency permutation ambiguity by using external permutation solvers [21]–[23] or the microphone array geometries [24]–[26], which limit the separation performance and/or their applicability. Our neural FCA, in contrast, resolves the permutation ambiguity by encouraging the independence of latent source features. We experimentally demonstrate that the training of the VAE-based source model itself has the ability to solve the frequency permutation even in the unsupervised condition. The experimental results also show that our neural FCA outperformed existing unsupervised methods and is comparable to a supervised VAE-based method.

II. Related Work

This study is related to two research fields: deep spectral models and unsupervised neural source separation.

A. Deep Spectral Models

The deep spectral model has been proposed to precisely represent the power spectral density (PSD) of source signals with a neural network [13]–[16]. In this model, the source signal on the time-frequency domain $\mathbf{S} = \{s_{ft} \in \mathbb{C}\}_{f,t=1}^{F,T}$ is assumed to follow a complex Gaussian distribution characterized by D-dimensional latent vectors $\mathbf{z}_t \in \mathbb{R}^D$:

$$s_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(0, g_{\theta,f}(\mathbf{z}_t)\right), \tag{1}$$

where $g_{\theta,f} : \mathbb{R}^D \to \mathbb{R}_+$ is a neural network with a set of parameters θ to associate $\mathbf{z}_t$ and the PSD of s_{ft}. Assuming the latent vector $\mathbf{z}_t$ to follow a standard Gaussian distribution:

$$\mathbf{z}_t \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right), \tag{2}$$

the network $g_{\theta,f}$ is trained as a decoder of a VAE by maximizing the log-marginal likelihood $\log p_\theta(\mathbf{S})$ for clean source signals. The trained model has been combined with a rank-1 spatial model [13] or a full-rank spatial model [14] for multichannel source separation called a multichannel VAE (MVAE). It has also been proposed for speech enhancement in unseen noisy environments by combining the VAE-based speech model and an NMF-based noise model [15], [16].

B. Unsupervised Neural Source Separation

Unsupervised neural source separation has been investigated to handle sources whose clean signals cannot be collected. One approach is called mixture invariant training [27], which uses the temporal independence of the source signals. While this approach and its variants [28], [29] can work with monaural mixture signals, the performance could deteriorate when the source signals have a temporal correlation in a mixture (e.g., music recordings). Another approach is to use spatial information in multichannel mixture signals [21]–[26], [30]. The source signals estimated by a conventional BSS method can be used as pseudo-supervised data. Togami *et al.* [23] proposed to train a network to predict a multichannel Wiener filter (MWF) estimated by FCA. Drude *et al.* [22] proposed to train a network by directly maximizing a log-marginal likelihood of a BSS model called complex angular central Gaussian mixture model (cACGMM) [31]. Since these methods are based on the conventional linear BSS, their performance was limited by the BSS methods.

III. Neural FCA for Unsupervised Source Separation

Our unsupervised method jointly trains a deep spectral model and its inference model by utilizing an amortized variational inference (AVI) [19]. The inference model firstly encodes the mixture signal into latent features of individual sources, and the encoded features are reversely decoded to the multichannel mixture by using the source model and SCMs. Our method solves the frequency permutation by assuming a prior distribution on the latent features. The AVI enables us to efficiently estimate the intractable posterior distribution of the latent features by the inference model. The training objective is formulated as a sum of a reconstruction loss to the multichannel observation and a regularization loss to the latent variables.

A. Generative Model of Multichannel Mixture Signal

To derive our unsupervised training method, we utilize a generative model of an M-channel mixture signal $\mathbf{x}_{ft} \in \mathbb{C}^M$, which is originally proposed for the supervised MVAE [14]. In this model, the observation $\mathbf{x}_{ft}$ is represented by a sum of N source signals s_{nft} $(n = 1, \ldots, N)$ as follows:

$$\mathbf{x}_{ft} = \sum_{n=1}^{N} \mathbf{a}_{nf} s_{nft}, \tag{3}$$

where $\mathbf{a}_{nf} \in \mathbb{C}^M$ is the steering vector for source n. Each source signal s_{nft} is characterized with frame-wise latent vectors $\mathbf{z}_{nt} \in \mathbb{R}^D$ as in Eqs. (1) and (2). By marginalizing s_{nft}, we obtain the following likelihood function:

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n=1}^{N} g_{\theta,f}(\mathbf{z}_{nt})\mathbf{H}_{nf}\right), \tag{4}$$

where $\mathbf{H}_{nf} = \mathbf{a}_{nf}\mathbf{a}_{nf}^{\mathsf{H}} \in \mathbb{S}_+^M$ is an SCM for source n at frequency f. By relaxing the rank-1 constraint on $\mathbf{H}_{nf}$ (i.e., assuming a full-rank SCM), this model can handle the fluctuation of $\mathbf{a}_{nf}$ caused by the source movements and reverberation [8].

B. Amortized Variational Inference for Unsupervised Training

We train the source model $g_{\theta,f}$ in an unsupervised manner by introducing an inference model $q_\phi(\mathbf{Z} \mid \mathbf{X})$ that predicts the latent source vectors $\mathbf{z}_{nt}$ from a mixture signal $\mathbf{X}$. More specifically, this model approximates the posterior $p_\theta(\mathbf{Z} \mid \mathbf{X}, \mathbf{H}) \propto p_\theta(\mathbf{X} \mid$

$\mathbf{Z}, \mathbf{H})p(\mathbf{Z})$ with the following Gaussian distribution q:

$$q_\phi(\mathbf{Z} \mid \mathbf{X}) = \prod_{n,t,d} \mathcal{N}\left(z_{ntd} \mid \mu_{\phi,ntd}(\mathbf{X}), \sigma^2_{\phi,ntd}(\mathbf{X})\right), \qquad (5)$$

where $\mu_{\phi,ntd}(\mathbf{X}) \in \mathbb{R}$ and $\sigma^2_{\phi,ntd}(\mathbf{X}) \in \mathbb{R}_+$ are the outputs of a neural network with parameters ϕ for the inference model.

Using a set of multichannel mixture signals, we jointly optimize the network parameters θ and ϕ, and SCMs $\mathbf{H}_{nf}$ to maximize the following evidence lower bound (ELBO):

$$\mathcal{L} = \mathbb{E}_{q_\phi}\left[\log p_\theta(\mathbf{X} \mid \mathbf{Z}, \mathbf{H})\right] - \mathcal{D}_{\mathrm{KL}}\left[q_\phi(\mathbf{Z} \mid \mathbf{X}) \mid p(\mathbf{Z})\right], \qquad (6)$$

where $\mathbb{E}_{q_\phi}[\cdot]$ is the expectation by the posterior $q_\phi(\mathbf{Z} \mid \mathbf{X})$, and $\mathcal{D}_{\mathrm{KL}}[q \mid p]$ is the Kullback–Leibler (KL) divergence between q and p. By maximizing this ELBO, the decoder parameters θ and $\mathbf{H}_{nf}$ are trained to maximize $\log p_\theta(\mathbf{X} \mid \mathbf{H})$, and the encoder parameters ϕ are trained to minimize $\mathcal{D}_{\mathrm{KL}}[q_\phi(\mathbf{Z} \mid \mathbf{X}) \mid p_\theta(\mathbf{Z} \mid \mathbf{X}, \mathbf{H})]$. Note that θ and ϕ are optimized for all training data, and $\mathbf{H}_{nf}$ is updated for each training mixture.

As in the training of a VAE, the intractable expectation in the ELBO is approximated by using a sample $\mathbf{z}^*_{nt} \sim q_\phi(\mathbf{z}_{nt} \mid \mathbf{X})$ with the reparametarization trick [19] as follows:

$$\mathbb{E}_{q_\phi}\left[\log p_\theta(\mathbf{X} \mid \mathbf{Z}, \mathbf{H})\right] \approx -\sum_{f,t} \log |\mathbf{Y}_{:ft}| - \sum_{f,t} \mathbf{x}^{\mathsf{H}}_{ft} \mathbf{Y}^{-1}_{:ft} \mathbf{x}_{ft},$$

where $\mathbf{Y}_{:ft} = \sum_{n=1}^{N} \mathbf{Y}_{nft} \in \mathbb{S}^M_+$ is the sum of source images $\mathbf{Y}_{nft} = g_{\theta,f}(\mathbf{z}^*_{nt})\mathbf{H}_{nf} \in \mathbb{S}_+$. The SCM $\mathbf{H}_{nf}$ is obtained by an expectation-maximization algorithm [8], [32] as follows:

$$\mathbf{H}_{nf} \leftarrow \frac{1}{T} \sum_{t=1}^{T} \frac{1}{g_{\theta,f}(\mathbf{z}^*_{nt})} \hat{\mathbf{X}}_{ft}, \qquad (7)$$

$$\hat{\mathbf{X}}_{ft} = \mathbf{Y}_{nft} + \mathbf{Y}_{nft}\left(\mathbf{Y}^{-1}_{:ft}\mathbf{x}_{ft}\mathbf{x}^{\mathsf{H}}_{ft}\mathbf{Y}^{-1}_{:ft} - \mathbf{Y}_{:ft}\right)\mathbf{Y}_{nft}. \qquad (8)$$

Since these equations depend on $\mathbf{H}_{nf}$ itself, we initialize it with an identity matrix and update it multiple times. In this paper, we updated $\mathbf{H}_{nf}$ five times from the identity matrix at each update of the networks. Since the gradients of all the above operations can be calculated analytically, we update the two networks by using stochastic gradient descent (SGD).

C. Frequency Permutation Alignment

We utilize the independence of latent source features for resolving the frequency permutation ambiguity caused by the frequency-wise likelihood of Eq. (4). When the permutation is not correctly aligned, a source signal is split into multiple source classes. In such a situation, latent source vectors $\mathbf{z}_{nt}$ for different classes n have correlated components representing the same source. Conversely, we can resolve the permutation by encouraging each component of $\mathbf{z}_{nt}$ to be independent. This regularization corresponds to the KL term in the ELBO $\mathcal{D}_{\mathrm{KL}}[q_\phi(\mathbf{Z} \mid \mathbf{X}) \mid p(\mathbf{Z})]$, which encourages the variational posterior q to be the standard Gaussian distribution. In this paper, we utilize the cyclic annealing of the KL term [33] to surely align the permutation as detailed in Section IV-B.

D. Source Separation With Trained Networks

Once the networks are trained, the inference model is utilized to estimate the latent source features of unseen mixture signals as $\mathbf{z}_{nt} \leftarrow \boldsymbol{\mu}_{\phi,nt}(\mathbf{X})$. While fixing the network parameters θ and ϕ, we then iteratively and alternately update the latent source vectors $\mathbf{z}_{nt}$ with SGD and SCMs $\mathbf{H}_{nf}$ with Eq. (7) such that Eq. (4) is maximized as in the MVAE. The separated signals are finally obtained by an MWF.

IV. Experimental Evaluation

The proposed method was evaluated by using multichannel speech mixture signals numerically simulated. Audio samples are available at https://ybando.jp/projects/spl2021.

A. Dataset

We utilized the spatialized WSJ0-2mix dataset [34], whose mixture includes two speech signals randomly selected from the WSJ0 English speech corpus. The speech sources are placed at random locations in a simulated room having random dimensions. An 8-channel microphone array is assumed with random geometry. We used its first four channels to reduce the computational complexity. The reverberation time (RT_{60}) of the room was randomly chosen between 200 and 600 ms. This dataset provides 20000, 5000, and 3000 mixtures for training, validation, and test sets, respectively. The mixture signals are generated at 16 kHz. As in the previous studies [22], [23], we performed dereverberation [35] to the mixture signals. For stabilizing the dereverberation, we added white Gaussian noise with the signal-to-noise ratio of 30 dB to the mixture signals.

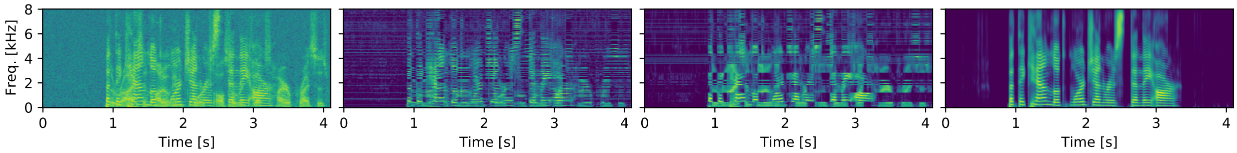
B. Experimental Condition

We designed our networks based on existing separation networks [4], [34]. Following [34], the inference network took as input a log-power spectrogram at the first (reference) microphone and inter-channel phase differences between the reference and other microphones. To reduce the computational cost, the input frames were first transformed into 256-channel vectors with a 1×1-convolutional (1×1-conv) layer. We then stacked four modules with each having eight dilated convolutional layers as in [4]. Each layer was the separable depth-wise convolution with a 512-channel depth-wise layer and parametric rectified linear units (PReLUs). The outputs $\mu_{\phi,ntd}(\mathrm{X})$ and $\sigma^2_{\phi,ntd}(\mathrm{X})$ were obtained by 1×1-conv layers, and the non-negative $\sigma^2_{\phi,ntd}(\mathrm{X})$ was obtained by the softplus activation. The source model $g_{\theta,f}$ consisted of three 1×1-conv layers with residual connections followed by one output layer with the softplus activation. Each layer had 256 channels and PReLUs.

The networks were trained by an Adam optimizer [36] for 200 epochs with the learning rate of 1.0×10^{-3}. The spectrograms were obtained by the short-time Fourier transform with the window size of 512 samples and the hop length of 128 samples. The training was performed by splitting the mixture spectrograms into 500-frame clips, and the batch size was set to 128 clips. The dimension of the latent variable D was set to 50. The number of sources N was set to 3 assuming two target sources and one noise signal. The weight of the KL term was changed by the cyclic annealing in 10-epoch cycles where its maximum value was set to 10.0 for the first 50 epochs and 1.0 thereafter. At the test time, the latent source vectors $\mathbf{z}_{nt}$ were updated by an Adam optimizer with the learning rate of 0.2. These hyperparameters were empirically determined.

TABLE I
SEPARATION PERFORMANCE IN AVERAGES AND STANDARD DEVIATIONS OF SDR, PESQ, AND STOI. SUPERVISED METHODS ARE COLORED IN GRAY

Method	Freq. perm.	# of iters.	Average			Male + Male			Female + Female			Female + Male		
			SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI
cACGMM	✗	200	10.8 ± 3.1	1.75 ± 0.34	0.86 ± 0.07	10.7	1.78	0.86	11.0	1.73	0.85	10.8	1.74	0.85
FCA	✗	200	12.7 ± 4.5	1.90 ± 0.45	0.86 ± 0.07	12.4	1.90	0.87	12.6	1.87	0.85	12.8	1.91	0.86
FastMNMF	✓	200	13.0 ± 5.3	1.85 ± 0.52	0.86 ± 0.10	12.9	1.86	0.87	12.6	1.80	0.84	13.1	1.87	0.86
Pseudo supervised	✗	50	14.7 ± 4.6	2.13 ± 0.51	0.88 ± 0.06	14.2	2.10	0.88	14.8	2.12	0.87	14.9	2.15	0.88
Neural cACGMM	✗	5	12.4 ± 2.9	1.89 ± 0.36	0.88 ± 0.06	12.0	1.88	0.88	12.6	1.88	0.87	12.5	1.90	0.88
Neural FCA (fix z)	✓	200	14.4 ± 4.4	2.27 ± 0.50	0.88 ± 0.06	13.8	2.20	**0.89**	14.4	2.24	0.87	14.8	2.31	**0.89**
Neural FCA	✓	5	12.9 ± 3.7	1.93 ± 0.41	0.87 ± 0.06	12.2	1.88	0.87	13.0	1.88	0.86	13.3	1.97	0.87
Neural FCA	✓	10	13.9 ± 4.1	2.07 ± 0.47	0.88 ± 0.06	13.2	2.02	0.88	13.9	2.02	0.87	14.2	2.12	0.88
Neural FCA	✓	50	15.0 ± 4.5	2.33 ± 0.54	$\mathbf{0.89 \pm 0.06}$	14.4	2.28	**0.89**	15.1	2.31	**0.88**	15.3	2.37	**0.89**
Neural FCA	✓	200	$\mathbf{15.2 \pm 4.5}$	$\mathbf{2.37 \pm 0.54}$	$\mathbf{0.89 \pm 0.06}$	**14.7**	**2.32**	**0.89**	**15.4**	**2.36**	**0.88**	**15.5**	**2.41**	**0.89**
MVAE (random init.)	✓	200	2.9 ± 5.2	1.23 ± 0.26	0.66 ± 0.16	4.6	1.29	0.69	0.8	1.15	0.61	2.7	1.22	0.66
MVAE (FCA init.)	✗	200	15.2 ± 4.5	2.30 ± 0.52	0.89 ± 0.06	14.7	2.25	0.89	15.5	2.32	0.88	15.4	2.33	0.89



(a) Input mixture (b) AE ($D = 256$) (c) AE ($D = 50$) (d) Proposed VAE ($D = 50$)

Fig. 2. Excerpts of PSDs $g_{\theta,f}(\boldsymbol{\mu}_{\phi,nt}(\mathbf{X}))$ estimated by proposed method and its variants, which have differences in solving frequency permutation ambiguity.

Our neural FCA was compared with existing BSS methods, unsupervised neural methods, and supervised MVAE. As BSS methods, we evaluated the cACGMM [31], FCA [8], and FastMNMF [5]. The number of basis vectors for FastMNMF was set to 8. Because the cACGMM and FCA cannot resolve the frequency permutation by themselves, we utilized an external permutation solver[1]. As unsupervised neural methods, we evaluated the pseudo supervised method [23] that approximates the results of the FCA [8] and the direct training method [22] that maximizes the likelihood of the cACGMM (neural cACGMM). The network outputs of these two methods were refined to fit the observation by the FCA and cACGMM. Their numbers of iterations were determined to maximize the performance because too many iterations degraded the performance. We finally evaluated an MVAE whose source model is trained on clean speech signals. The MVAE estimated sources by maximizing the likelihood of Eq. (4). Its latent variable $\mathbf{z}_{nt}$ was initialized by the FCA, which gave better initialization than FastMNMF in our preliminary experiment. The network architectures for these methods were determined to be as similar as possible to our method. We evaluated the separation performance with source-to-distortion ratio (SDR) [37] in dB, perceptual evaluation of speech quality (PESQ) [38],[2] and short term objective intelligibility (STOI) [39].

C. Experimental Results

As summarized in TABLE I, the proposed neural FCA with 200 iterations achieved the best SDR, PESQ, and STOI in all of the unsupervised methods. We can see that only 10 iterations were enough for neural FCA to outperform the BSS methods of cACGMM, FCA, and FastMNMF. In addition, regardless of the genders of the speakers, the performance of the neural FCA was comparable to that of the supervised MVAE whose latent vectors $\mathbf{z}_{nt}$ were initialized by FCA. The appropriate initialization of $\mathbf{z}_{nt}$ is important since the MVAE significantly deteriorated with the random initialization of $\mathbf{z}_{nt}$. In contrast, the proposed inference model initialized the latent vectors $\mathbf{z}_{nt}$ effectively. Even when $\mathbf{z}_{nt}$ (and PSDs) was fixed to the initial value, the neural FCA outperformed the BSS methods. These results show that the proposed framework combining the deep spectral model and inference model is effective in unsupervised training of source separation.

Unlike the conventional unsupervised neural methods, neural FCA itself solved the frequency permutation problem as demonstrated in Fig. 2. Fig. 2 (b) and (c) are the PSDs estimated by the networks trained as a deterministic autoencoder (AE). In this condition, the inference model is trained to estimate $\mathbf{z}_{nt}$ in a maximum likelihood manner without the KL term. We can see that the bottleneck architecture ($D = 50$) itself has some effects on solving the permutation. As shown in Fig. 2 (d), the proposed method completely resolved the permutation by the training based on the variational inference. We conclude that the independence of the latent source vectors encourages solving the frequency permutation ambiguity.

V. CONCLUSION

This paper presented an unsupervised method that trains a neural source separation by using only mixture signals. Our neural FCA jointly trains a neural source model and its inference model by maximizing the ELBO for the training data of multichannel mixture signals. The experimental results show that the proposed method outperformed the existing unsupervised methods and was comparable to the supervised MVAE. Our future work includes handling an unknown number of sound sources to analyze real-world mixture recordings. We also plan to utilize the source latent features obtained by our unsupervised training for down-streaming tasks such as sound event detection and acoustic scene classification.

[1] The implementation is available on https://github.com/fgnt/pb_bss

[2] We excluded 446c0209_1.8488_22hc010z_-1.8488.wav for the evaluation because its first source was almost silent and PESQ could not be measured.

References

[1] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 780–793, Apr. 2017.

[2] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[3] N. Turpault *et al.*, "Improving sound event detection in domestic environments using sound separation," in *Proc. Detection Classification Acoust. Scenes Events Workshop*, 2020, pp. 1–5.

[4] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[5] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2610–2625, 2020.

[6] N. Ito and T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 371–375.

[7] 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 Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, Sep. 2016.

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

[9] A. A. Nugraha, K. Sekiguchi, M. Fontaine, Y. Bando, and K. Yoshii, "Flow-based independent vector analysis for blind source separation," *IEEE Signal Process. Lett.*, vol. 27, pp. 2173–2177, 2020.

[10] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[11] N. Makishima *et al.*, "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 10, pp. 1601–1615, Oct. 2019.

[12] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2011, pp. 189–192.

[13] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Semi-blind source separation with multichannel variational autoencoder," 2018, *arXiv:1808.00892*.

[14] S. Seki, H. Kameoka, L. Li, T. Toda, and K. Takeda, "Generalized multichannel variational autoencoder for underdetermined source separation," in *Proc. Eur. Signal Process. Conf.*, 2019, pp. 1–5.

[15] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Semi-supervised multichannel speech enhancement with a deep speech prior," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2197–2212, Dec. 2019.

[16] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 101–105.

[17] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 371–375.

[18] L. Li, H. Kameoka, and S. Makino, "Determined audio source separation with multichannel star generative adversarial network," in *Proc. IEEE 30th Int. Workshop Mach. Learn. Signal Process.*, 2020, pp. 1–6.

[19] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[20] M. Garnier, N. Henrich, and D. Dubois, "Influence of sound immersion and communicative interaction on the Lombard effect," *J. Speech, Lang., Hear. Res.*, vol. 53, no. 3, pp. 588–608, 2010.

[21] L. Drude, D. Hasenklever, and R. Haeb-Umbach, "Unsupervised training of a deep clustering model for multichannel blind source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 695–699.

[22] L. Drude, J. Heymann, and R. Haeb-Umbach, "Unsupervised training of neural mask-based beamforming," in *Proc. Interspeech*, 2019, pp. 1253–1257.

[23] 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. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 56–60.

[24] Y. Bando, Y. Sasaki, and K. Yoshii, "Deep Bayesian unsupervised source separation based on a complex Gaussian mixture model," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2019, pp. 1–6.

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

[26] P. Seetharaman, G. Wichern, J. Le Roux, and B. Pardo, "Bootstrapping single-channel source separation via unsupervised spatial clustering on stereo mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 356–360.

[27] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, and J. R. Hershey, "Unsupervised sound separation using mixture invariant training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 3846–3857.

[28] T. Fujimura, Y. Koizumi, K. Yatabe, and R. Miyazaki, "Noisy-target training: A training strategy for DNN-based speech enhancement without clean speech," in *Proc. Eur. Signal Process. Conf.*, 2021, [Online]. Available: https://arxiv.org/pdf/2101.08625.pdf.

[29] M. Maciejewski, J. Shi, S. Watanabe, and S. Khudanpur, "Training noisy single-channel speech separation with noisy oracle sources: A large gap and a small step," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 5774–5778.

[30] E. Tzinis, S. Venkataramani, and P. Smaragdis, "Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 81–85.

[31] 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. Eur. Signal Process. Conf.*, 2016, pp. 1153–1157.

[32] H. Sawada, R. Ikeshita, and T. Nakatani, "Experimental analysis of EM and MU algorithms for optimizing full-rank spatial covariance model," in *Proc. Eur. Signal Process. Conf.*, 2020, pp. 885–889.

[33] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin, "Cyclical annealing schedule: A simple approach to mitigating KL vanishing," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 240–250.

[34] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 1–5.

[35] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[37] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[38] 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 2, pp. 749–752.

[39] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 4214–4217.