Joint Separation and Localization of Moving Sound Sources Based on Neural Full-Rank Spatial Covariance Analysis

Hokuto Munakata, Yoshiaki Bando, *Member, IEEE*, Ryu Takeda, Kazunori Komatani, and Masaki Onishi

***Abstract*—This paper presents an unsupervised multichannel method that can separate moving sound sources based on an amortized variational inference (AVI) of joint separation and localization. A recently proposed blind source separation (BSS) method called neural full-rank spatial covariance analysis (FCA) trains a neural separation model based on a nonlinear generative model of multichannel mixtures and can precisely separate unseen mixture signals. This method, however, assumes that the sound sources hardly move, and thus its performance is easily degraded by the source movements. In this paper, we solve this problem by introducing time-varying spatial covariance matrices and directions of arrival of sources into the nonlinear generative model of the neural FCA. This generative model is used for training a neural network to jointly separate and localize moving sources by using only multichannel mixture signals and array geometries. The training objective is derived as a lower bound on the log-marginal posterior probability in the framework of AVI. Experimental results obtained with mixture signals of moving sources show that our method outperformed an existing joint separation and localization method and standard BSS methods.**

***Index Terms*—Amortized variational inference, multichannel signal processing, source separation and localization.**

I. Introduction

SOUND source separation is a fundamental function for various machine listening systems including distant speech recognition and hearing aids [1], [2], [3], [4]. One approach to source separation is to train a separation neural network (e.g., Conv-TasNet [5]) on a large number of pairs of isolated source signals and their mixtures [5], [6], [7], [8]. Unsupervised separation methods, on the other hand, have also been investigated to address the lack of such a supervised dataset and domain mismatch at the target environments [9], [10], [11], [12], [13], [14], [15].

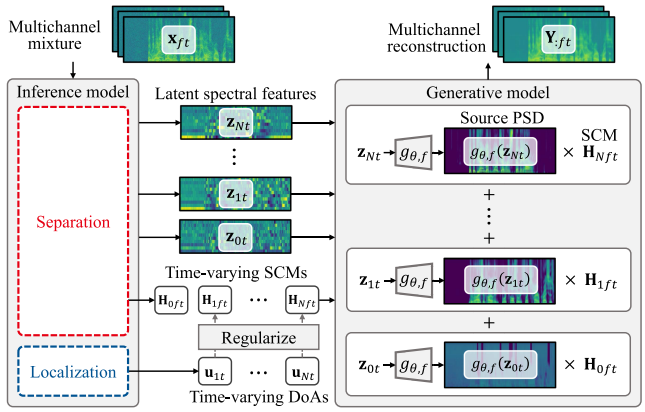


Fig. 1. Overview of our time-varying neural FCA for joint source separation and localization.

Blind source separation (BSS) has been actively studied to separate source signals with little prior information based on a generative model of a multichannel mixture signal [14], [15], [16], [17], [18], [19], [20]. A fundamental method called full-rank spatial covariance analysis (FCA) [14] represents each time-frequency (TF) bin of a mixture as a sum of spatial covariance matrices (SCMs) of sources. Multichannel nonnegative matrix factorization (MNMF) [15], [16], [17] improves the performance of FCA by assuming the source spectra to be low-rank. Its nonlinear extension called neural FCA [21], [22] can precisely represent the source spectra with a neural network called a deep spectral model [23], [24], [25], [26], [27]. This model can be trained blindly by maximizing the log-marginal likelihood for the training data of multichannel mixtures in advance and was reported to be comparable with a multichannel supervised model [21], [28].

Most BSS methods assume that sound sources hardly move, and thus their performance is easily degraded by the movement of the target sources. One solution is to allow source steering vectors (the rank-1 special forms of SCMs) to be time-variant by assuming a Markov process on the vectors [29]. If the geometry of microphones is available, we can efficiently constrain the steering vectors with the directions of arrival (DoAs) of sources estimated by source localization [30], [31]. Furthermore, the separation and localization can be performed jointly, based on a unified generative model [32], to complementarily compensate for their estimation errors.

Manuscript received 15 January 2023; revised 24 March 2023; accepted 28 March 2023. Date of publication 5 April 2023; date of current version 14 April 2023. This work was supported in part by the JST ACT-X under Grant JPMJAX200N and in part by the NEDO. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Aiping Liu. *(Corresponding author: Hokuto Munakata.)*

Hokuto Munakata is with the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan, and also with the SANKEN, Osaka University, Osaka 567-0047, Japan (e-mail: h_munakata@ei.sanken.osaka-u.ac.jp).

Yoshiaki Bando and Masaki Onishi are with the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan (e-mail: y.bando@aist.go.jp; onishi-masaki@aist.go.jp).

Ryu Takeda and Kazunori Komatani are with the SANKEN, Osaka University, Osaka 567-0047, Japan (e-mail: rtakeda@sanken.osaka-u.ac.jp; komatani@sanken.osaka-u.ac.jp).

Digital Object Identifier 10.1109/LSP.2023.3264570

In this paper, we present an unsupervised neural method to perform joint separation and localization for moving sources. As shown in Fig. 1, we extend the neural FCA to handle the source movements by introducing the time-varying SCM and DoA for each source in the nonlinear generative model of a multichannel mixture signal. The joint separation and localization is performed by an inference model that predicts the parameters of each source from an input mixture. These inference and generative models are jointly trained in an unsupervised manner by maximizing a log-marginal posterior probability given only multichannel mixtures and array geometries.

The main contribution of this study is to combine a time-varying nonlinear (neural) BSS model with neural probabilistic inference. While the full-rank SCMs generally improve the performance from that obtained with the rank-1 SCMs, it was difficult to introduce the Markov process or temporal smoothness by modeling only with conjugate priors. We solve this problem by introducing the temporal smoothness of SCMs with the constraints on the neural inference model instead of the generative model. The experimental results with the mixture signals of moving sources show that our method significantly outperformed an existing joint localization and separation method as well as standard BSS methods.

II. Background

This section introduces a nonlinear BSS method called neural FCA as a preliminary for the proposed method.

A. Blind Source Separation

Existing BSS models typically represent an M-channel mixture signal $\mathrm{x}_{ft} \in \mathbb{C}^M$ as a sum of N target (and noise) source signals $s_{nft} \in \mathbb{C}$ $(n = 1, \ldots, N)$ in the TF domain as follows:

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

where $\mathbf{a}_{nf} \in \mathbb{C}^M$ is the time-invariant steering vector for source n, and $f = 1, \ldots, F$ and $t = 1, \ldots, T$ are the frequency and time frame indices, respectively. The source signal s_{nft} is represented by a complex Gaussian distribution as follows:

$$s_{nft} \mid \lambda_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \lambda_{nft}\right), \tag{2}$$

where $\lambda_{nft} \in \mathbb{R}_+$ is the power spectral density (PSD) of source n. By marginalizing source signals s_{nft}, we obtain the following likelihood function of the multichannel mixture x_{ft}:

$$\mathrm{x}_{ft} \mid \boldsymbol{\lambda}, \mathbf{H} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n=1}^{N} \lambda_{nft} \mathbf{H}_{nf}\right), \tag{3}$$

where $\mathbf{H}_{nf} = \mathbf{a}_{nf}\mathbf{a}_{nf}^{\mathsf{H}} \in \mathbb{S}_+^{M \times M}$ is the SCM of source n. By allowing the full-rankness of $\mathbf{H}_{nf}$, this model called FCA [14] can handle small source movements and reverberation.

B. Deep Spectral Model

A nonlinear (neural) source model has been utilized for precisely representing complex source spectra [23], [24], [25], [26], [27]. This model called a deep spectral model assumes that the PSD λ_{nft} is represented by D-dimensional feature vectors $\mathrm{z}_{nt} \in \mathbb{R}^D$ $(t = 1, \ldots, T)$ as follows:

$$\lambda_{nft} = g_{\theta,f}(\mathrm{z}_{nt}), \tag{4}$$

where $g_{\theta,f} : \mathbb{R}^D \to \mathbb{R}_+$ is a neural network with parameters θ for transforming the feature vector to the PSD. Assuming z_{nt} follows the standard Gaussian distribution:

$$\mathrm{z}_{nt} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right), \tag{5}$$

this model can be trained as the decoder of a variational autoencoder by using clean source signals [23]. This supervised source model was reported to outperform NMF-based linear models in speech separation or enhancement [26], [27], [28].

C. Neural Full-Rank Spatial Covariance Analysis

The deep spectral model can be trained in an unsupervised manner by using only multichannel mixtures based on amortized variational inference [21], [33]. This method called neural FCA utilizes an inference (encoder) network to predict the speech features $\mathbf{Z} \triangleq \{\mathrm{z}_{nt}\}_{n,t=1}^{N,T}$ from an input mixture $\mathbf{X} \triangleq \{\mathrm{x}_{ft}\}_{f,t=1}^{F,T}$ as the posterior distribution $q_\phi(\mathbf{Z} \mid \mathbf{X})$, where ϕ represents the network parameters. By using the generative model of a mixture signal ((3)–(5)) as a decoder, the encoder and decoder are jointly trained to maximize the following evidence lower bound (ELBO) [33]:

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

where $\mathbb{E}_{q_\phi}[\cdot]$ is the expectation by the posterior q_ϕ, and $\mathcal{D}_{\mathrm{KL}}[\cdot|\cdot]$ is the Kullback-Leibler (KL) divergence. The network parameters θ and ϕ are updated by stochastic gradient ascent, and the SCMs $\mathbf{H} \triangleq \{\mathbf{H}_{nf}\}_{n,f}$ are updated by an expectation-maximization (EM) algorithm [14]. This training can be considered nonlinear BSS performed on the training mixtures.

III. Time-Varying Neural FCA

We extend the original neural FCA to perform joint localization and separation for moving sound sources.

A. Generative Model of Multichannel Mixture Signals

We assume that a mixture signal x_{ft} consists of N directional moving sources s_{nft} and a diffuse noise n_{ft} as follows:

$$\mathrm{x}_{ft} = \sum_{n=1}^{N} \mathbf{a}_{nft} s_{nft} + \mathrm{n}_{ft}, \tag{7}$$

where $\mathbf{a}_{nft} \in \mathbb{C}^M$ is a time-varying steering vector for source n at time frame t. Assuming both the source and noise signals follow the time-varying version of (2) and (3), we obtain the following likelihood function:

$$\mathrm{x}_{ft} \mid \mathbf{Z}, \mathbf{H} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n=0}^{N} g_{\theta,f}(\mathrm{z}_{nt}) \mathbf{H}_{nft}\right), \tag{8}$$

where $\mathbf{H}_{nft} \in \mathbb{S}_+^{M \times M}$ are time-varying SCMs for the sources $(n = 1, \ldots, N)$ and noise $(n = 0)$.

To exploit the localization results, we assume that the SCM $\mathbf{H}_{nft}$ for each source $(n = 1, \ldots, N)$ follows a conjugate prior

conditioned by a unit vector $\mathrm{u}_{nt} \in \mathbb{R}^3$ ($\|\mathrm{u}_{nt}\| = 1$) representing the DoA of each source:

$$\mathbf{H}_{nft} \mid \mathrm{u}_{nt} \sim \mathcal{IW}_{\mathbb{C}}\left(\nu, (\nu + M)\mathbf{G}_f(\mathrm{u}_{nt})\right), \tag{9}$$

where $\mathcal{IW}_{\mathbb{C}}(\nu, \mathbf{\Gamma}) \propto |\mathbf{H}|^{-(\nu+M)}|\mathbf{\Gamma}|^{\nu}\exp(-\mathrm{tr}(\mathbf{\Gamma}\mathbf{H}^{-1}))$ is the complex inverse Wishart distribution, and $\nu > M$ is a hyperparameter controlling degrees of freedom of $\mathbf{H}_{nft}$. The mode of this prior is equal to the prior SCM $\mathbf{G}_f(\mathrm{u}_{nt})$ defined as:

$$\mathbf{G}_f(\mathrm{u}_{nt}) = \mathbf{b}_f(\mathrm{u}_{nt})\mathbf{b}_f(\mathrm{u}_{nt})^{\mathsf{H}} + \epsilon\mathbf{I}, \tag{10}$$

where $\epsilon > 0$ is a small number to make $\mathbf{G}_f(\mathrm{u}_{nt})$ positive definite, and $\mathbf{b}_f(\mathrm{u}_{nt}) \in \mathbb{C}^M$ is the steering vector for direction u_{nt} calculated from the geometrically calculated time delays of the microphones. The SCM for noise $\mathbf{H}_{0ft}$, on the other hand, is assumed to be diffuse by replacing $\mathbf{G}_f(\mathrm{u}_{nt})$ in (9) with an identity matrix. Note that we did not formulate any temporal smoothness of $\mathbf{H}_{nft}$ because it is difficult to make such a constraint in a generative model. As described in the next section, we alternatively introduce it by the inductive bias of the inference model.

B. Inference Model

Our inference model predicts the latent feature z_{nt}, SCMs $\mathbf{H}_{nft}$, and DoAs u_{nt} from a multichannel mixture signal x_{ft} (Fig. 1). As the estimates of z_{nt}, following the original neural FCA [21], the inference model predicts the posterior distribution $q_\phi(\mathbf{Z} \mid \mathbf{X})$ as follows:

$$q_\phi(\mathbf{Z} \mid \mathbf{X}) = \prod_{n,t,d} \mathcal{N}\left(z_{ntd} \mid \mu_{\phi,ntd}(\mathbf{X}), \sigma^2_{\phi,ntd}(\mathbf{X})\right), \tag{11}$$

where $\mu_{\phi,ntd}(\mathbf{X}) \in \mathbb{R}$ and $\sigma^2_{\phi,ntd}(\mathbf{X}) \in \mathbb{R}_+$ are the outputs of the inference network. Since the time-varying SCM $\mathbf{H}_{nft}$ is difficult to estimate analytically, the inference model estimates it as the moving average of masked observation added to the prior SCM $\mathbf{G}_f(\mathrm{u}_{nt})$ with a weight hyperparameter $\gamma_0 \in \mathbb{R}_+$:

$$\mathbf{H}_{nft} \leftarrow \gamma_0\mathbf{G}_f(\mathrm{u}_{nt}) + \sum_{t'=0}^{T} \gamma^{|t-t'|} w_{\phi,nft'}(\mathbf{X}) \frac{\mathbf{X}_{ft'}\mathbf{X}_{ft'}^{\mathsf{H}}}{\|\mathbf{X}_{ft'}\|^2}, \tag{12}$$

where $w_{\phi,nft}(\mathbf{X}) \in [0, 1]$ is a TF mask predicted by the inference network, and $\gamma \in (0, 1]$ is a decay hyperparameter controlling the smoothness of $\mathbf{H}_{nft}$. For numerical stability, $\mathbf{H}_{nft}$ is normalized to $\mathrm{tr}(\mathbf{H}_{nft})$ be M. Lastly, DoA u_{nt} was predicted with unit vectors $\tilde{\mathbf{u}}_{\phi,nt}(\mathbf{X}) \in \mathbb{R}^3$ output by the network:

$$\mathrm{u}_{nt} \leftarrow \sum_{t'=0}^{T} \eta^{|t-t'|} \tilde{\mathbf{u}}_{\phi,nt'}(\mathbf{X}), \tag{13}$$

where $\eta \in (0, 1]$ is a decay hyperparameter. The DoAs u_{nt} are also normalized to be unit vectors. These moving averages introduce the temporal smoothness of $\mathbf{H}_{nft}$ and u_{nt}.

C. Amortized Variational Inference for Unsupervised Training

We train the inference and generative models by using only multichannel mixtures. The training objective for each mixture is the ELBO $\mathcal{L}'_{\theta,\phi}$ with a regularization term of $\mathbf{H}_{nft}$:

$$\mathcal{L}'_{\theta,\phi}(\mathbf{X}) = \mathcal{L}_{\theta,\phi}(\mathbf{X}) + \log p(\mathbf{H}_\phi \mid \mathbf{U}_\phi), \tag{14}$$

where $\mathbf{H}_\phi$ and $\mathbf{U}_\phi$ are the sets of inference results obtained by (12) and (13), respectively. This ELBO is equivalent to a lower bound on the following log-marginal posterior function:

$$\log p_{\theta,\phi}(\mathbf{H}_\phi \mid \mathbf{X}, \mathbf{U}_\phi) \overset{c}{=} \log p_\theta(\mathbf{X} \mid \mathbf{H}_\phi) + \log p(\mathbf{H}_\phi \mid \mathbf{U}_\phi),$$

where $\overset{c}{=}$ denotes equality up to an additive constant. The network parameters θ and ϕ are updated by using stochastic gradient ascent. This method can be considered as training $q_\phi(\mathbf{Z} \mid \mathbf{X})$ and $\mathbf{H}_\phi$ by the original ELBO $\mathcal{L}_{\theta,\phi}(\mathbf{X})$, while the DoAs $\mathbf{U}_\phi$ regularize the SCMs $\mathbf{H}_\phi$ and are optimized to maximize the log-likelihood $\log p(\mathbf{H}_\phi \mid \mathbf{U}_\phi)$. After training these networks, they can be used to separate and localize moving sources in an unseen mixture. The source signals are obtained with the source images $\mathbf{Y}_{nft} \triangleq g_{\theta,f}(\boldsymbol{\mu}_{\phi,nt}(\mathbf{X}))\mathbf{H}_{nft}$ by a multichannel Wiener filter [14], [21].

IV. Experimental Evaluation

We evaluated our method on simulated speech mixtures due to the need for reference signals. A demonstration with real recordings can be found at https://ybando.jp/projects/spl2023.

A. Dataset

We generated a dataset of multichannel mixtures of moving sources. The mixture signals were generated as observations of six-channel microphone arrays in a way similar to the way the spatialized WSJ0-2mix dataset was generated [34]. Each mixture consisted of two source signals randomly selected from the WSJ0 English speech corpus. The moving source signals were generated by convoluting time-varying room impulse responses (RIRs) [35] generated every 0.1 s. The array with random geometry was placed randomly around the center of a room having dimensions of 5 m $\times$ 5 m $\times$ 3 m. Each source was initially located randomly and moved around the array with a constant speed drawing a horizontal circular arc. We sampled the angular velocities of sources uniformly between $0°$/s and $45°$/s. The angular difference between sources always had at least $45°$ through the movement. The reverberation time (RT_{60}) was fixed to 200 ms. The source signals were mixed with a signal-to-noise ratio (SNR) randomly chosen between -5 and $+5$ dB. The mixture signals were generated at 16 kHz, and Gaussian noise was added with an SNR of 30 dB. The dataset consisted of 20000, 5000, and 3000 mixtures for training, validation, and test sets, respectively. For comparison, we also generated a static dataset in which no sources moved.

B. Experimental Condition

We used almost the same network configuration as that of the original neural FCA [21], whose inference and generative models consisted of temporal convolutional networks. We added the dropout ($p = 0.1$) to avoid bad local optima and two output layers to the inference model for estimating the TF mask and DoA. To utilize the spatial information and array geometries, the input feature consisted of a log-power spectrogram and a DoA spectrogram [36] calculated with 1000 uniformly distributed three-dimensional directions.

The networks were trained by an Adam optimizer [37] for 200 epochs with a learning rate of 0.001. The hyperparameters D, ν, ϵ, γ_0, and η were set to 50, $M + 1$, 0.001, 0.1, and 0.99, respectively. We set γ to 0.99 for sources and 1 (time-invariant)

TABLE I
SEPARATION AND LOCALIZATION PERFORMANCE IN SDR AND DOA ERROR. THE MOVING CONDITION WAS DIVIDED INTO THREE SUBSETS ACCORDING TO SOURCE ANGULAR VELOCITIES: SLOW (0/S–15°/S), MID (15°/S–30°/S), AND FAST (30°/S–45°/S). TV STANDS FOR "TIME-VARYING."

Method	TV model	Static condition		Moving condition							
		SDR↑	DoA Err.↓	SDR↑				DoA Err.↓			
				Avg.	Slow	Mid	Fast	Avg.	Slow	Mid	Fast
cACGMM	×	9.92	–	4.35	6.66	3.40	1.72	–	–	–	–
FCA	×	11.93	–	4.21	6.64	3.25	1.42	–	–	–	–
FastMNMF2	×	12.81	–	4.01	6.66	2.74	1.27	–	–	–	–
DoA-HMM-based clustering	✓	7.83	2.81	7.96	8.05	7.94	7.82	3.89	2.77	3.86	5.85
MUSIC-based localization	✓	–	2.85	–	–	–	–	4.09	2.87	4.02	6.27
Neural FCA	×	16.06	–	8.27	11.26	7.30	4.53	–	–	–	–
+ Mask-based SCMs ($\gamma = 1$)	×	15.95	–	−1.68	−1.71	−1.71	−1.57	–	–	–	–
+ Time-varying SCMs	✓	14.38	–	10.19	12.35	10.08	6.65	–	–	–	–
+ Joint localization (ours)	✓	14.21	2.48	12.53	12.71	12.46	12.32	3.04	2.83	3.00	3.47

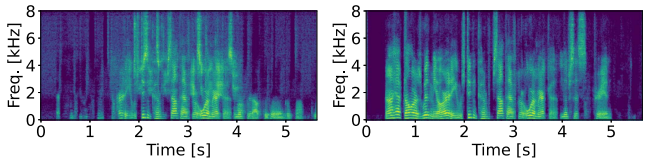
for noise. We scaled the $\log p(\mathbf{H}_\phi|\mathbf{U}_\phi)$ in (14) with 0.001 to avoid over-constraining the SCMs. Following [21], we also performed the cyclic annealing [38] for scaling the KL term in (6). The spectrograms were obtained by the short-time Fourier transform with a window size of 512 samples and a hop length of 128 samples. The mixture spectrograms were fed to the network by splitting them into 500-frame clips. The batch size for training was 128 clips. These hyperparameters were determined empirically.

Our time-varying neural FCA was compared with existing BSS methods, a joint separation and localization method, and the original neural FCA. As BSS methods, we evaluated cACGMM [18], FCA [14], FastMNMF2 [17]. An external frequency permutation solver was used for the cACGMM and the FCA as in [21], and the number of basis vectors for FastMNMF2 was set to 8. We evaluated a joint method based on the clustering of TF bins with a DoA-HMM [32]. This method can separate moving sound sources and was initialized by the localization results using MUSIC [39]. The original (time-invariant) neural FCA was trained with the same input features as the proposed method. The number of the iteration for estimating the SCM $\mathbf{H}_{nf}$ was 5. We evaluated the separation performance with the average signal-to-distortion ratio (SDR) in dB [40] and the localization performance with the average DoA error in degrees [41]. The DoA error was averaged on non-silent frames whose powers of oracle source signals were larger than −20 dB from the average.

C. Experimental Results

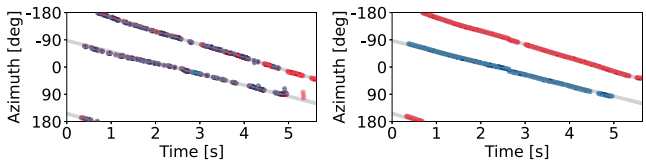
The separation and localization performance is summarized in Table I. We can first see that the SDRs of the original neural FCA and the standard BSS methods were significantly degraded by the source movements. In contrast, our method (the bottom row) improved the average SDR by more than 4 dB for the moving condition from these methods. Although its SDR for the static condition was 1.8 dB worse than that of the original neural FCA, our method was still better than FastMNMF2. Furthermore, our method outperformed the DoA-HMM-based clustering in both the SDRs and DoA errors and MUSIC in the DoA errors for both static and moving conditions. As shown in Fig. 2, while the original neural FCA output unseparated sources or silence, our method successfully estimated speech sources over almost all the time frames.

Our method consists of three extensions of the neural FCA: estimating $\mathbf{H}_{nf}$ with TF masking, estimating time-varying $\mathbf{H}_{nft}$ by (12), and performing joint localization and separation. As



(a) Original neural FCA (b) Proposed method

Fig. 2. Results for separation of a moving source signal by the original and proposed time-varying neural FCAs.



(a) w/o smoothness ($\eta = 0$) (b) w/ smoothness ($\eta = 0.99$)

Fig. 3. DoAs estimated by our method. DoAs whose source was silent are omitted. Gray lines are ground-truth DoAs.

in the bottom two rows of Table I, the time-varying extension degraded the performance for the static condition because it cannot exploit the statistics of entire time frames. In contrast, this extension is key to improving the performance in the moving condition. The temporal smoothness of DoAs introduced by (13) was also an important key as demonstrated in Fig. 3. Our method without the smoothness failed to track each source and estimated two DoAs that were almost the same. The localization results with the smoothness, on the other hand, were estimated correctly.

V. CONCLUSION

We presented an unsupervised multichannel method that can separate and localize moving sound sources without any supervision. Our method trains a joint separation and localization model only from multichannel mixture signals and array geometries. This training is based on an extension of neural FCA to incorporate the time-varying DoAs of each source. The experimental results with moving sound sources demonstrated that our method outperformed existing BSS methods and a joint source separation and localization method. Our future work includes further extending the neural FCA to handle variable numbers of sound sources and long reverberation, which will enable the separation of real-world recordings.

REFERENCES

[1] S. Watanabe et al., "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. Workshop Speech Process. Everyday Environments*, 2020, pp. 1–7.

[2] N. Turpault et al., "Improving sound event detection in domestic environments using sound separation," in *Proc. Detection Classification Acoust. Scenes Events Workshop*, 2020, pp. 205–209.

[3] T. V. Neumann et al., "End-to-end training of time domain audio separation and recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7004–7008.

[4] D. Marquardt, V. Hohmann, and S. Doclo, "Interaural coherence preservation in multi-channel Wiener filtering-based noise reduction for binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2162–2176, Dec. 2015.

[5] 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.

[6] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 21–25.

[7] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, "Beam-TasNet: Time-domain audio separation network meets frequency-domain beamformer," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6379–6383.

[8] 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 Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[9] K. Saijo and R. Scheibler, "Spatial loss for unsupervised multi-channel source separation," in *Proc. Interspeech*, 2022, pp. 241–245.

[10] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 3846–3857.

[11] 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.

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

[13] 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.

[14] N. Q. 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.

[15] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 971–982, May 2013.

[16] 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.

[17] 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.

[18] 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.

[19] K. Yatabe and D. Kitamura, "Determined blind source separation via proximal splitting algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 776–780.

[20] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 236–240.

[21] 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 Signal Process. Lett.*, vol. 28, pp. 1670–1674, 2021.

[22] Y. Bando, T. Aizawa, K. Itoyama, and K. Nakadai, "Weakly-supervised neural full-rank spatial covariance analysis for a front-end system of distant speech recognition," in *Proc. Interspeech*, 2022, pp. 3824–3828.

[23] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 716–720.

[24] L. Li, H. Kameoka, and S. Makino, "Fast MVAE: Joint separation and classification of mixed sources based on multichannel variational autoencoder with auxiliary classifier," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 546–550.

[25] 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.

[26] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Comput.*, vol. 31, no. 9, pp. 1891–1914, Sep. 2019.

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

[28] 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.

[29] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of time-varying convolutive audio mixtures," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 8, pp. 1408–1423, Aug. 2016.

[30] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel NMF and acoustic tracking," *IEEE/ACM Trans. on Audio, Speech, Lang. Process.*, vol. 26, no. 2, pp. 281–295, Feb. 2018.

[31] M. Taseska and E. A. Habets, "Blind source separation of moving sources using sparsity-based source detection and tracking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 657–670, Mar. 2018.

[32] T. Higuchi, N. Takamune, T. Nakamura, and H. Kameoka, "Underdetermined blind separation and tracking of moving sources based on DOA-HMM," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 3191–3195.

[33] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–14.

[34] 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 1–5.

[35] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustical Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[36] M. Togami, "Spatial constraint on multi-channel deep clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 531–535.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[38] 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.

[39] D. Salvati, C. Drioli, and G. L. Foresti, "Incoherent frequency fusion for broadband steered response power algorithms in noisy environments," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 581–585, May 2014.

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

[41] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 34–48, Mar. 2019.