Joint Separation and Tracking of Moving Sources With Distributed Microphone Arrays Based on Time-Varying Inertial Spatial Models

Ryunosuke Nihei* Yoshiaki Bando†‡ Aditya Arie Nugraha‡ Diego Di Carlo‡
Hiroyuki Ueda* Yosuke Ito* Kazuyoshi Yoshii*‡

*Graduate School of Engineering, Kyoto University, Japan
†National Institute of Advanced Industrial Science and Technology (AIST), Japan
‡Center for Advanced Intelligence Project (AIP), RIKEN, Japan
nihei.ryunosuke.38i@st.kyoto-u.ac.jp

***Abstract*—This paper describes the first attempt at separation and tracking (3D localization) of multiple *moving* sound sources using multiple microphone arrays fixed at known locations in an indoor environment. As for *static* sources, location-dependent priors have been incorporated on the time-invariant spatial covariance matrices (SCMs) of sources in the statistical framework of blind source separation based on multichannel nonnegative matrix factorization (MNMF), achieving the maximum likelihood estimation of source locations. One may thus make both the SCMs and their priors vary over time to deal with source movements. This naive extension, however, fails to localize sources when the sources are inactive, yielding non-smooth, non-continuous trajectory estimates. To solve this problem, we formulate a hierarchical probabilistic model for multichannel mixture signals that consists of inertial Markov models for source locations, location-aware moving-average models for source SCMs, and NMF-based low-rank models for the power spectral densities (PSDs) of sources. All the time-varying attributes of sources are jointly estimated under a maximum-a-posteriori (MAP) principle, and the source images are then estimated with a multichannel Wiener filter. The experiment using simulated data with two moving sources and four four-channel arrays showed that the proposed method achieved better separation and smoother localization.**

I. Introduction

Multichannel sound source separation and localization have been widely studied for acoustic scene understanding, including distant speech recognition [1], multi-speaker separation [2], and acoustic event detection [3]. In general, separation and localization have been investigated separately, in particular with supervised methods based on deep learning [4], [5]. To leverage the mutual dependency between the two tasks and avoid the overfitting problem of supervised methods, we here focus on unsupervised joint separation and localization.

Blind source separation (BSS) is relatively robust against the variation of acoustic environments thanks to its unsupervised nature. Modern BSS methods are based on a probabilistic model of multichannel mixture signals that consists of a *source model* representing the power spectral densities (PSDs) of sources and a *spatial model* representing the spatial covariance matrices (SCMs) of the sources related to the directions of arrival (DOAs) [6], [7]. The PSDs and SCMs can be estimated jointly under the maximum likelihood principle.

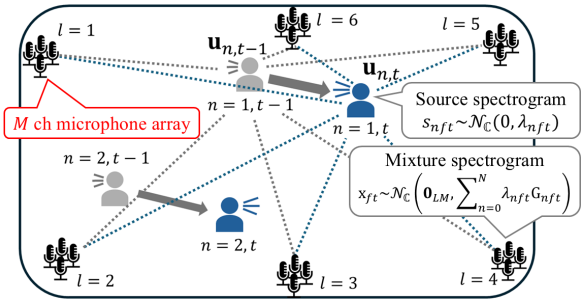


Fig. 1. Joint localization and separation of multiple moving sound sources with distributed microphone arrays in an indoor environment. Given the mixture $\mathbf{X}$, we aim to estimate the images $\mathbf{S}$ and trajectories $\mathbf{U}$ of sources.

Among BSS methods, we focus on multichannel nonnegative matrix factorization (MNMF) [6], which consists of a low-rank source model and a full-rank spatial model. Although full-rank SCMs are highly expressive and capable of handling moderate reverberation, they often converge to suboptimal local optima during iterative optimization due to the high degree of freedom (DOF) in the spatial model. Proper initialization and the use of prior knowledge are thus crucial to draw the full potential of MNMF for better performance.

In practical non-blind conditions, geometric knowledge can be exploited effectively for localization as well as separation. For an array with known geometry, the theoretical steering vectors (rank-1 SCMs) with respect to candidate DOAs can be used as a prior to regularize the full-rank SCMs, enabling joint *DOA estimation* and separation of *moving* sources [8]. Recently, multiple microphone arrays distributed at known positions have been successfully used for *3D localization* and separation of *static* sources [9]. This method is based on MNMF that treats all the arrays as a big array and the static 3D positions of the microphones and sources are used as a prior to regularize the time-invariant full-rank SCMs.

To deal with moving sources, one needs to formulate a time-varying spatial model. A naive extension of the basic spatial model [9], however, suffers from the unreliable frame-wise

979-8-3315-7206-8/25/$31.00

estimates of time-varying source SCMs, resulting in the non-continuous estimates of source trajectories. This problem is especially serious at frames where some sources are inactive, because the SCMs of each frame are estimated from only the observed mixture of that frame.

To address this problem, we propose an extension of MNMF that incorporates time-varying inertial spatial models for joint *3D localization* and separation of *moving* sources (Fig. 1). We assume the trajectory of each source to follow an inertial Markov model. At each frequency, we assume the time-varying SCMs of the source to follow inverse-Wishart priors, with modes aligned to the theoretical SCMs derived from the source trajectory, as in [9]. These SCMs are temporally smoothed using a moving-average filter and integrated with NMF-based source models to formulate a probabilistic model of a mixture. Using the maximum-a-posteriori (MAP) principle, we jointly estimate the PSDs, SCMs, and trajectories of the sources from the observed mixture, leveraging neighboring observations to obtain reliable SCM estimates at each frame.

This study is the first to achieve joint 3D localization and separation of multiple moving sources using a unified hierarchical probabilistic model. This model lays the foundation for multimodal scene understanding within a unified probabilistic framework and can easily be extended to deep generative models.

II. Related Work

This section reviews blind source separation, source localization, and leveraging prior information.

A. Blind Source Separation

A major BSS paradigm consists of a source model representing the time–frequency structure of each source and a spatial model representing the propagation from each source to the microphones. These models can be jointly optimized under the maximum likelihood principle. For instance, frequency-domain independent component analysis (FD-ICA) [10] assumes non-Gaussian source models at each frequency but suffers from the permutation problem across frequencies. To avoid this problem, independent vector analysis (IVA) [11], [12] uses multivariate source models that capture higher-order inter-frequency correlations. Independent low-rank matrix analysis (ILRMA) [7] uses low-rank source models based on NMF for better permutation alignment, leading to better performance.

Whereas these methods employ rank-1 SCMs, full-rank spatial covariance analysis (FCA) [13] and MNMF [6] use full-rank SCMs, providing richer spatial modeling at the cost of larger sensitivity to local optima and higher computational complexity. FastMNMF [14] assumes the joint diagonalizability of the source SCMs, accelerating the parameter updates in MNMF. It still uses a full-rank spatial model while keeping better separation accuracy. More recently, Bando *et al.* [15] introduced a self-supervised framework that uses a physics-based generative model to generate pseudo-labels and trains a neural separator, reducing the initialization sensitivity.

B. Source Localization

Sound source localization is one of the essential technologies for acoustic scene understanding. When the array geometry is known, DOAs can be estimated from pair-wise phase differences between microphones via generalized cross-correlation with phase transform (GCC-PHAT) [16] and integrated using beamforming-based approach [17]. An alternative strategy is multiple signal classification (MUSIC) [18], which exploits the orthogonality between the signal and noise subspaces of the array covariance matrix. For real-time applications, sparsity-aware frameworks [19] and polynomial MUSIC [20] for wide-band signals have emerged. And deep learning has further improved 2-D DOA accuracy, as demonstrated by DOAnet [21].

Localization performance has seen improvements by networking multiple microphone arrays into a distributed configuration. Distributed microphone arrays have enabled multi-speaker tracking via triangulation with local DOA estimates [22]. With synchronized arrays, multiple source positions can be estimated from inter-array time-difference observations [23]. More recently, graph neural networks (GNNs) [24] have aggregated pairwise time-delay features across ad-hoc distributed networks for superior localization accuracy.

C. Leveraging Prior Information

Prior information can enhance separation and localization performance. By exploiting DOA priors, separation performance was enhanced by introducing a Wishart prior on the SCM within a non-negative tensor factorization (NTF) framework [25]. Spatial priors have been exploited for joint DOA estimation and source separation. For example, moving sources can be separated by considering the steering vectors derived from tracked DOAs and array geometry [26]. More recently, a neural-network-based approach that incorporates position-dependent SCM priors was introduced to estimate DOAs and separate moving sources [8]. When the geometrically-computed steering vectors are used as spatial priors, not only the DOAs but also the 3D positions of multiple static sources can be estimated with multiple arrays [9]. We extend this method for 3D tracking of moving sources.

III. Proposed Method

This section describes the proposed method for joint separation and 3D tracking of multiple moving sources

A. Problem Specification

Suppose that L microphone arrays with M channels and N sound sources are located in a room (Fig. 1). Let $\mathbf{r}_{lm} \in \mathbb{R}^3$ be the position of microphone $m \in [1, M]$ in array $l \in [1, L]$. Let $\mathbf{X} \triangleq \{\mathbf{x}_{ft} \in \mathbb{C}^{LM}\}_{f,t=1}^{F,T}$ be the observed multichannel spectrogram, where F and T represent the number of frequency bins and that of time frames, respectively. We assume that all the arrays are synchronized [27], [28]. Our goal is to estimate the *image* $\mathbf{Z}_n \triangleq \{\mathbf{z}_{nft} \in \mathbb{C}^{LM}\}_{f,t=1}^{F,T}$ and the time-varying positions $\mathbf{U}_n \triangleq \{\mathbf{u}_{nt} \in \mathbb{R}^3\}_{t=1}^{T}$ of each source $n \in [1, N]$.

B. Generative Modeling and Source Separation

We formulate a time-varying extension of MNMF. Assuming the low-rank structure of the PSDs $\{\lambda_{nft}\}_{f,t=1}^{F,T}$ of each source n, we first formulate a source model that represents the complex spectrogram $\mathbf{S}_n \triangleq \{s_{nft} \in \mathbb{C}\}_{f,t=1}^{F,T}$ as follows:

$$s_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \lambda_{nft} \triangleq \sum_{k=1}^{K} w_{nkf} h_{nkt}\right), \tag{1}$$

where the PSDs are decomposed by NMF with K bases and $\{w_{nkf}\}_{f=1}^{F}$ and $\{h_{nkt}\}_{t=1}^{T}$ represent the basis and activation of source n and basis k, respectively. We then formulate a *time-varying* spatial model that represents the image (multichannel spectrogram) of source n as follows:

$$\mathbf{z}_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{Y}_{nft} \triangleq \lambda_{nft}\mathbf{G}_{nft}\right), \tag{2}$$

where $\mathbf{G}_{nft} \in \mathbb{S}_{+}^{LM}$ is the full-rank SCM of source n at frequency f and time t. Unlike the original MNMF with time-invariant SCMs [6], our model can deal with the temporal change of the SCMs due to the source movements.

Assuming the source additivity in the complex spectrogram domain, i.e, $\mathbf{x}_{ft} = \sum_{n=1}^{N} \mathbf{z}_{nft}$, we have

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n=1}^{N} \lambda_{nft}\mathbf{G}_{nft}\right). \tag{3}$$

Given the mixture $\mathbf{X}$ with $\mathbf{W} \triangleq \{w_{nkf}\}_{n,k,f=1}^{N,K,F}$, $\mathbf{H} \triangleq \{h_{nkt}\}_{n,k,t=1}^{N,K,T}$, and $\mathbf{G} \triangleq \{\mathbf{G}_{nft}\}_{n,f,t=1}^{N,F,T}$, we can estimate the source image $\mathbf{z}_{nft}$ with a multichannel Wiener filter (MWF) as:

$$\mathbb{E}[\mathbf{z}_{nft}|\mathbf{x}_{ft}, \mathbf{W}, \mathbf{H}, \mathbf{G}] = \mathbf{Y}_{nft}\mathbf{Y}_{ft}^{-1}\mathbf{x}_{ft}, \tag{4}$$

where $\mathbf{Y}_{ft} \triangleq \sum_{n=0}^{N} \mathbf{Y}_{nft}$.

C. Priors on Source SCMs

Assuming smooth continuous source trajectories, we encourage the SCM sequence $\{\mathbf{G}_{nft}\}_{t=1}^{T}$ of each source n at each frequency f to gradually vary over time. To achieve this, we represent the actual SCM sequence $\{\mathbf{G}_{nft}\}_{t=1}^{T}$ as a smoother version of a *raw* SCM sequence $\{\tilde{\mathbf{G}}_{nft} \in \mathbb{S}_{+}^{LM}\}_{t=1}^{T}$ based on a moving-average (MA) filter as follows:

$$\mathbf{G}_{nft} = \sum_{-3\sigma \le \tau \le 3\sigma} \omega_\tau \, \tilde{\mathbf{G}}_{n,f,t+\tau}, \tag{5}$$

$$\omega_\tau \propto \exp\left(-\tau^2/2\sigma^2\right), \tag{6}$$

where ω_τ is a filter coefficient such that $\sum_{\tau=-3\sigma}^{3\sigma} \omega_\tau = 1$ and σ sets the filter length to control the degree of smoothness. We then put a complex inverse Wishart prior on $\tilde{\mathbf{G}}_{nft}$ as in [9]:

$$\tilde{\mathbf{G}}_{nft} \sim \mathcal{IW}_{\mathbb{C}}\left(\nu, (\nu + LM)\hat{\mathbf{G}}_{nft}\right), \tag{7}$$

where ν is the degree of freedom and $(\nu + LM)\hat{\mathbf{G}}_{nft} \in \mathbb{S}_{+}^{LM}$ is a scale matrix such that the mode (the most probable value) of the *raw* SCM $\tilde{\mathbf{G}}_{nft}$ is the *theoretical* SCM $\hat{\mathbf{G}}_{nft}$.

As in [9], we represent the theoretical SCM $\hat{\mathbf{G}}_{nft}$ as a block-diagonal matrix given by

$$\hat{\mathbf{G}}_{nft} = \mathrm{BlockDiag}(\hat{\mathbf{G}}_{nft1}, \cdots, \hat{\mathbf{G}}_{nftL}), \tag{8}$$

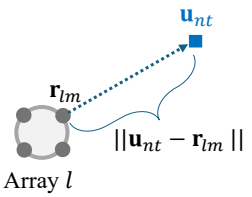


Fig. 2. Computation of the steering vector of source n at time t based on the positions of microphones in array l.

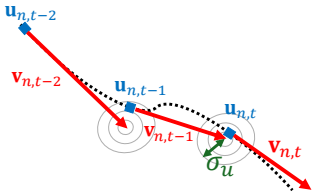


Fig. 3. The inertial Markov model that stochastically generates the trajectories $\mathbf{U}_n$ and velocities $\mathbf{V}_n$ of source n.

where $\hat{\mathbf{G}}_{nftl} \in \mathbb{S}_{+}^{M}$ is the local theoretical SCM of source n for array l. We here assume that the array-wise SCMs $\{\hat{\mathbf{G}}_{nftl}\}_{l=1}^{L}$ are independent because the positions of the L arrays can be measured much less precisely than the geometry of each array [9]. Using the steering vector computed from the propagation delays illustrated in Fig. 2, each $\hat{\mathbf{G}}_{nftl}$ is given by

$$\hat{\mathbf{G}}_{nftl} \propto \mathbf{b}_{nftl}\mathbf{b}_{nftl}^{\mathsf{H}} + \epsilon\mathbf{I} \;\; \text{s.t.} \;\; \mathrm{tr}(\hat{\mathbf{G}}_{nftl}) = M, \tag{9}$$

where $\epsilon > 0$ is a small number to ensure the full-rankness of $\hat{\mathbf{G}}_{nftl}$ and $\mathbf{b}_{nftl} \triangleq [b_{nftl1}, \cdots, b_{nftlM}] \in \mathbb{C}^M$ is the theoretical steering vector of source n for array l. In free field, given source and microphone positions, b_{nftlm} can be computed algebraically as

$$b_{nftlm} = \exp(-\mathring{\jmath}\omega_f \|\mathbf{u}_{nt} - \mathbf{r}_{lm}\|/c) \tag{10}$$

where $\mathring{\jmath}$ is the imaginary unit, ω_f is the angular frequency corresponding to frequency bin f, and c is the speed of sound.

D. Priors on Source Trajectories

To encourage the trajectory $\{\mathbf{u}_{nt}\}_{t=1}^{T}$ of each source n to represent the inertial motion of the source, i.e, prevent physically-unnatural sudden turns and random-walk behaviors, we formulate an inertial Markov prior on $\mathbf{u}_{nt}$ as follows:

$$\begin{cases} \mathbf{u}_{nt} \sim \mathcal{N}(\mathbf{u}_{n,t-1} + \mathbf{v}_{n,t-1}, \sigma_u^2\mathbf{I}), \\ \mathbf{v}_{nt} \sim \mathcal{N}(\mathbf{v}_{n,t-1}, \sigma_v^2\mathbf{I}), \end{cases} \tag{11}$$

where $\{\mathbf{v}_{nt} \in \mathbb{R}^3\}_{t=1}^{T}$ is the velocity history of source n and σ_u^2 and σ_v^2 are the position and velocity variances, respectively. Fig. 3 illustrates the sound source trajectory.

E. Parameter Estimation

Given the mixture $\mathbf{X}$ as observed data, we aim to estimate the parameters $\mathbf{W}$, $\mathbf{H}$, $\mathbf{G}$, $\mathbf{U} \triangleq \{\mathbf{u}_{nt}\}_{n,t=1}^{N,T}$, and $\mathbf{V} \triangleq \{\mathbf{v}_{nt}\}_{n,t=1}^{N,T}$ that maximize the posterior distribution given by

$$\begin{aligned} p(\mathbf{G}, \mathbf{U}|\mathbf{X}, \mathbf{W}, \mathbf{H}, \mathbf{V}) &\propto p(\mathbf{X}, \mathbf{G}, \mathbf{U}|\mathbf{W}, \mathbf{H}, \mathbf{U}, \mathbf{V}) \\ &\propto p(\mathbf{X}|\mathbf{W}, \mathbf{H}, \mathbf{G})p(\tilde{\mathbf{G}}|\mathbf{U})p(\mathbf{U}, \mathbf{V}), \end{aligned} \tag{12}$$

where the three terms of the right-hand side are given by Eqs. (3), (7), and (11), respectively. We thus have

$$\begin{aligned} &\log p(\mathbf{G}, \mathbf{U}|\mathbf{X}, \mathbf{W}, \mathbf{H}, \mathbf{V}) \\ &= \sum_{f=1}^{F}\sum_{t=1}^{T} \log \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{ft}\middle|\mathbf{0}, \sum_{n=1}^{N} \lambda_{nft}\mathbf{G}_{nft}\right) \end{aligned}$$

$$+\alpha\sum_{n=1}^{N}\sum_{f=1}^{F}\sum_{t=1}^{T}\log\mathcal{IW}_{\mathbb{C}}\Big(\tilde{\mathbf{G}}_{nft}\big|\nu,(\nu+LM)\hat{\mathbf{G}}_{nft}\Big)$$
$$+\sum_{n=1}^{N}\sum_{t=1}^{T}\log\mathcal{N}\big(\mathbf{u}_{nt}|\mathbf{u}_{n,t-1}+\mathbf{v}_{n,t-1},\sigma_u^2\mathbf{I}\big)$$
$$+\sum_{n=1}^{N}\sum_{t=1}^{T}\log\mathcal{N}\big(\mathbf{v}_{nt}|\mathbf{v}_{n,t-1},\sigma_v^2\mathbf{I}\big)+\text{const.},\tag{13}$$

where α is a weighting coefficient (hyper-parameter) that balances the Gaussian log-likelihood (first term), related to source separation, and the inverse-Wishart log-likelihood (second term), related to source localization.

We use an iterative optimization method that alternately updates the parameters until convergence. Since the NMF parameters $\mathbf{W}$ and $\mathbf{H}$ appear in only the first term of Eq. (13), we use the multiplicative update rules in the same way as MNMF:

$$w_{nkf}\leftarrow w_{nkf}\sqrt{\frac{\sum_t h_{nkt}\,\mathrm{tr}\big(\mathbf{G}_{nft}\mathbf{Y}_{ft}^{-1}\bar{\mathbf{X}}_{ft}\mathbf{Y}_{ft}^{-1}\big)}{\sum_t h_{nkt}\,\mathrm{tr}\big(\mathbf{G}_{nft}\mathbf{Y}_{ft}^{-1}\big)}},\tag{14}$$

$$h_{nkt}\leftarrow h_{nkt}\sqrt{\frac{\sum_f w_{nkf}\,\mathrm{tr}\big(\mathbf{G}_{nft}\mathbf{Y}_{ft}^{-1}\bar{\mathbf{X}}_{ft}\mathbf{Y}_{ft}^{-1}\big)}{\sum_f w_{nkf}\,\mathrm{tr}\big(\mathbf{G}_{nft}\mathbf{Y}_{ft}^{-1}\big)}}.\tag{15}$$

where $\bar{\mathbf{X}}_{ft}\triangleq\mathbf{x}_{ft}\mathbf{x}_{ft}^{\mathsf{H}}$.

To solve the scale ambiguity between $\mathbf{W}$ and $\mathbf{H}$ we thus insert the following normalization step:

$$\mu_{nk}\triangleq\sum_{f=1}^{F}w_{nkf},\quad\begin{cases}w_{nkf}&\leftarrow\mu_{nk}^{-1}w_{nkf}\\h_{nkt}&\leftarrow\mu_{nk}h_{nkt}\end{cases},\tag{16}$$

The source SCMs $\mathbf{G}$, the trajectories $\mathbf{U}$, and the velocities $\mathbf{V}$ are updated via a gradient ascent method (e.g., Adam [29]) such that the posterior probability given by Eq. (13) is maximized.

IV. Evaluation

This section describes the experiments conducted to evaluate the proposed method for joint separation and tracking of moving sound sources.

A. Data and Metrics

We considered two speakers ($N=2$) moving along parallel trajectories in a room that measures $4.0\,\mathrm{m}\times4.0\,\mathrm{m}\times2.5\,\mathrm{m}$. The room is equipped with four 4-channel microphone arrays ($L=4$, $M=4$), each placed 0.5 m away from the walls in each corner. The height of all speakers and microphone arrays was set to 1.6 m, reflecting the typical human mouth level. The two sources moved along linear paths, shown as dashed lines in Figs. 4 and 5, i.e., from $(3.0, 2.75)$ to $(1.0, 2.75)$ for source 1 and from $(1.0, 1.25)$ to $(3.0, 1.25)$ for source 2. The mixtures observed at the arrays were generated by convolving speech signals taken from the CMU ARCTIC corpus [30] with room impulse responses (RIRs) simulated using the pyroomacoustics library [31]. To simulate a time-varying RIR along a given trajectory, we generated a filter every 50 ms and applied it to the corresponding section of the signal. The wall absorption coefficient was set to 0.80, resulting in a theoretical RT60 of 93 ms according to the Sabine equation. However, in this work, the reflection order was limited to 3, which led to a milder reverberation effect in practice. The mixture duration was approximately 8 s, sampled at a rate of 16 kHz, during which both speakers were primarily active, with silence only occurring naturally between words. The STFT coefficients were computed using a 1024-sample Hann window with 75% overlap.

The source separation performance is evaluated in terms of the signal-to-distortion ratio (SDR) with respect to the ground-truth source images, while the source tracking performance is evaluated in terms of the root mean square error (RMSE) of the estimated positions per time frame.

B. Compared Methods

We carried out a performance comparative analysis of our proposed method, a block-wise MNMF using a sliding window approach, referred to as "sliding MNMF" hereafter, and the vanilla MNMF, which is equivalent to applying the sliding MNMF with a single window. All methods estimate N sources without assigning a dedicated noise source, i.e., there is no $n=0$. The number of NMF bases was set to $K=32$ as in [9], with both the basis spectra $\mathbf{W}$ and the activations $\mathbf{H}$ were initialized by random sampling from $[0,1)$, unless stated otherwise.

The sliding MNMF divides the input mixture signal into 6 overlapping windows, each with a size of 130 frames and an overlap of 30 frames. The time-invariant SCMs $\bar{\mathbf{G}}\triangleq\{\mathbf{G}_{nf}\}_{n,f=1}^{N,F}$ were initially set to an identity matrix for each window, while $\mathbf{W}$ and $\mathbf{H}$ were randomly initialized for the first window. To mitigate the permutation problem that may arise across windows, $\mathbf{W}$ was initialized with values from the previous window and $\mathbf{H}$ was initialized by carrying over values from the overlap time region for the subsequent windows. The parameter optimization was done for 175 iterations for each window, followed by MWF to compute the window-level source image $\mathbf{Z}_n$. The complete utterance-level source image was then obtained by averaging over the overlapping regions between adjacent windows. Although MNMF is a BSS method that does not estimate the source positions, we could infer the positions for each window, rather than each time frame, based on $\bar{\mathbf{G}}$.

C. Configurations

We set the inverse Wishart prior in our proposed method to have a degree of freedom $\nu=LM+1$ and a diagonal loading with $\epsilon=10^{-13}$ for the theoretical SCM. The actual SCMs were computed from the raw SCMs using an MA filter with $\sigma=10$. We set the weighting factor α to 10^{-4}. The velocities $\mathbf{V}$ were initialized following a uniform distribution with a range of $[-0.15, 0.15]$ at each time frame. The noise variances for the trajectories and the velocities were fixed to $\sigma_u=10^{-6}$ and $\sigma_v=10^{-7}$, respectively.

The actual SCMs $\mathbf{G}$ and the trajectories $\mathbf{U}$ were initialized based on the SCMs $\bar{\mathbf{G}}$ estimated using the sliding MNMF. Similarly to $\tilde{\mathbf{G}}_{nft}$ in Eq. (7), we assume that each SCM in $\bar{\mathbf{G}}$ follows an inverse Wishart distribution $\bar{\mathbf{G}}_{nf}\sim\mathcal{IW}_{\mathbb{C}}\big(\nu,(\nu+LM)\hat{\mathbf{G}}_{nf}\big)$, where $\hat{\mathbf{G}}_{nf}$ is the theoretical SCM given a source location. Given $\bar{\mathbf{G}}$ and a set of candidate source locations,

TABLE I
SOURCE SEPARATION AND LOCALIZATION PERFORMANCES.

	SDR [dB] ↑		RMSE [cm] ↓	
	$n=1$	$n=2$	$n=1$	$n=2$
MNMF	0.75	2.01	68.6	90.5
Sliding MNMF	−0.86	−2.63	18.6	19.2
Proposed (init: MNMF)	**7.54**	0.73	34.1	84.2
Proposed (init: sliding MNMF)	3.96	**9.03**	**2.8**	**2.9**

we performed a grid search to find the most likely positions, i.e., the ones that maximize the likelihood function. We then assigned these optimal positions as the initial $\mathbf{U}$ accordingly, taking into account the time frame correspondence. Similarly, we computed the theoretical SCMs based on these positions and assigned them as the initial $\mathbf{G}$.

We performed 800 iterations of parameter updates. In each iteration, we updated $\mathbf{W}$ and $\mathbf{H}$ by multiplicative update rules, followed by Adam-based updates for $\mathbf{G}$, $\mathbf{U}$, and $\mathbf{V}$. Based on preliminary experiments, in a single iteration, $\mathbf{G}$ was updated 20 times with a learning rate of 0.03, while $\mathbf{U}$ and $\mathbf{V}$ were updated 100 each with learning rates of 0.01 and 0.001, respectively.

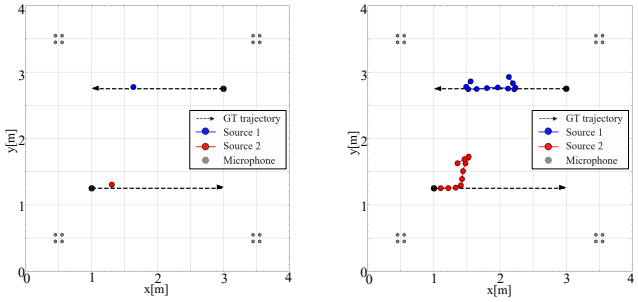
D. Experimental Results

Table I summarizes the source separation performance in terms of SDR and the source tracking performance in terms of RMSE. It presents the results for the baseline methods, i.e., vanilla MNMF and sliding MNMF, as well as the proposed method with initialization of $\mathbf{G}$ and $\mathbf{U}$ based on either MNMF or sliding MNMF. The source locations for the baseline methods were estimated by the grid search given the optimized $\hat{\mathbf{G}}$ as described in Section IV-C. This yields utterance-level source locations for the vanilla MNMF (Fig. 4a) and window-level locations for the sliding MNMF (Fig. 5a). Despite poor initialization from the utterance-level estimated locations of MNMF, our proposed method demonstrated significant improvement in source separation. With improved initialization using the window-level estimated locations of sliding MNMF, our method notably enhanced both source separation and tracking performance.

E. Discussion

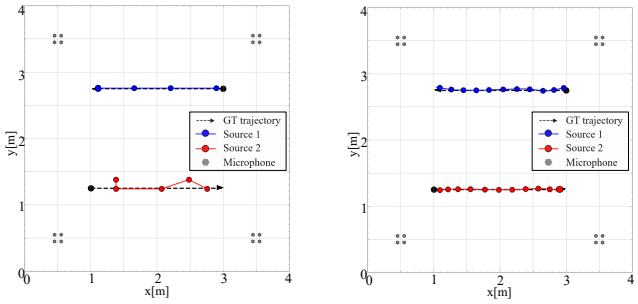
The proposed method exhibited impressive performance in both separation and tracking (Fig. 5b), showcasing the effectiveness of our time-varying inertial spatial models. Fig. 6 visualizes the localization likelihood given by Eq. (7) over the spatial grid at convergence. The brightest pixel at each frame coincides with the estimated source position, confirming localization accuracy. The proposed method effectively localized moving sound sources by smoothly connecting the trajectories, which were initially given at the window level, over time.

The proposed method, however, requires a sufficiently good initialization of $\mathbf{G}$ and $\mathbf{U}$. When the initial trajectories were given at the utterance level, the method could only converge to the correct trajectories for a portion of the time (Fig. 4b), specifically ones close to the given initial values. For this example, the estimated trajectory of source 2 was drawn towards that of source 1. This is likely because the SCM of source 2



(a) Initial (b) Estimated

Fig. 4. Initial and estimated trajectories for the proposed method with initialization of $\mathbf{G}$ and $\mathbf{U}$ based on the MNMF estimation.The ground truth trajectories are the two parallel dashed-arrow lines.



(a) Initial (b) Estimated

Fig. 5. Initial and estimated trajectories for the proposed method with initialization of $\mathbf{G}$ and $\mathbf{U}$ based on the sliding MNMF estimation.

was contaminated with that of source 1, which led to relatively poor separation performance for source 2 (Table I). It is well-known that MNMF, the underlying model for our method, has a high dependency on initial values. Our method similarly suggests this dependency, highlighting the critical need for careful consideration of the initialization method.

Both MNMF and sliding MNMF struggled to separate and track moving sources due to the utterance- and window-level spatial modeling. MNMF attained better separation than sliding MNMF, likely due to the difference in the amount of data processed. MNMF was applied on a 513-frame utterance, whereas sliding MNMF was applied on 130-frame windows. Nonetheless, despite its window-level nature, sliding MNMF still managed to estimate source locations that were close to the actual trajectories of the sources.

V. CONCLUSION

We presented a statistical method for joint separation and localization of multiple moving sources based on a hierarchical generative model for distributed microphone arrays audio recordings. It models time-varying SCMs with a moving-average model and inverse-Wishart prior conditioned on source positions. To ensure smooth source trajectories, we modeled source positions using a Markov model. The source signals, SCMs, and source trajectories are jointly estimated in a MAP manner based on the combination of the majorization-minimization updates and gradient ascent updates. Experiments using simulated data showed significant improvements in both

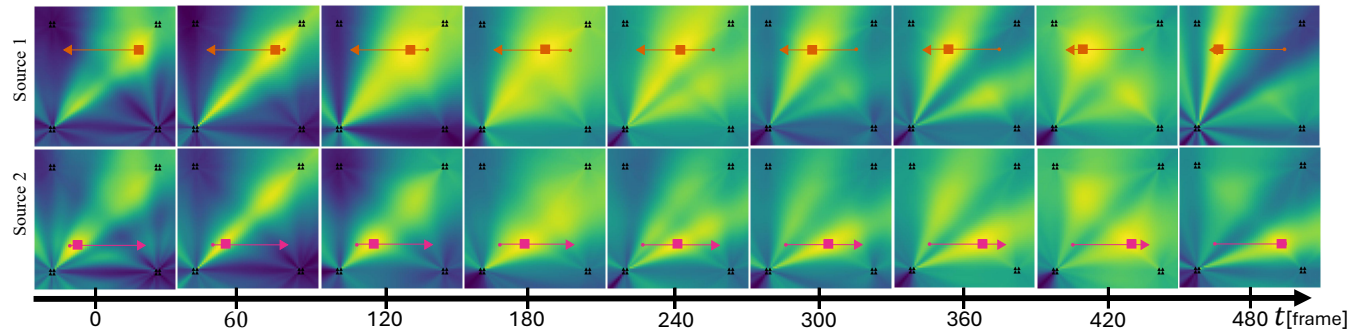


Fig. 6. The estimated trajectories $\mathbf{U}$ of two moving sources. Each heatmap represents the likelihood of the position $\mathbf{u}_{nt}$ of source n at time t given by Eq. (7). The square marks indicate the maximum likelihood estimates.

separation and localization for moving sources. Evaluation on real data with longer reverberation, louder noise, and complex nonlinear source movements remains as future work.

One interesting direction is to estimate the source number varying over time through probabilistic integration of audio-visual modalities. It helps scene understanding in silent periods, broadening applicability.

Acknowledgment

This work was partially supported by JST FOREST Grant No. JPMJFR2270 and JSPS KAKENHI Grant Nos. 23K16912, 23K16913, 24H00742, 24H00748, 25H01142, and 25K22841.

References

[1] S. Cornell, M. S. Wiesner, S. Watanabe, D. Raj, X. Chang, P. Garcia, Y. Masuyama, Z.-Q. Wang, S. Squartini, and S. Khudanpur, "The CHiME-7 DASR challenge: Distant meeting transcription with multiple devices in diverse scenarios," in *CHiME*, 2023, pp. 1–6.

[2] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Comput. Speech Lang.*, vol. 72, 2021.

[3] N. Turpault, R. Serizel, S. Wisdom, H. Erdogan, J. R. Hershey, E. Fonseca, P. Seetharaman, and J. Salamon, "Sound event detection and separation: A benchmark on DESED synthetic soundscapes," in *ICASSP*, 2021, pp. 840–844.

[4] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP*, 2017, pp. 241–245.

[5] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *EUSIPCO*, 2018, pp. 1462–1466.

[6] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE ASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[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 TASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[8] H. Munakata, Y. Bando, R. Takeda, K. Komatani, and M. Onishi, "Joint separation and localization of moving sound sources based on neural full-rank spatial covariance analysis," *IEEE SPL*, vol. 30, pp. 384–388, 2023.

[9] Y. Sumura, D. D. Carlo, A. A. Nugraha, Y. Bando, and K. Yoshii, "Joint audio source localization and separation with distributed microphone arrays based on spatially-regularized multichannel NMF," in *IWAENC*, 2024, pp. 145–149.

[10] C. Jutten and J. Herault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Process.*, vol. 24, no. 1, pp. 1–10, 1991.

[11] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *ICA*, 2006, pp. 165–172.

[12] I. Lee, T. Kim, and T.-W. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *Signal Process.*, vol. 87, no. 8, pp. 1859–1871, 2007.

[13] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE ASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[14] N. Ito and T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *ICASSP*, 2019, pp. 371–375.

[15] Y. Bando, S. Cornell, S. Fukayama, and S. Watanabe, "Investigation of spatial self-supervised learning and its application to target speaker speech recognition," in *ICASSP*, 2025, pp. 1–5.

[16] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in lownoise, reverberative environments?" in *ICASSP*, 2008, pp. 2565–2568.

[17] J. H. Dibiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, Rhode Island, 2000.

[18] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE TAP*, vol. 34, no. 3, pp. 276–280, 1986.

[19] M. Togami and R. Scheibler, "Sparseness-aware DOA estimation with majorization minimization." in *Interspeech*, 2020, pp. 5046–5050.

[20] A. O. T. Hogg, V. W. Neo, S. Weiss, C. Evers, and P. A. Naylor, "A polynomial eigenvalue decomposition music approach for broadband sound source localization," in *WASPAA*, 2021, pp. 326–330.

[21] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *EUSIPCO*, 2018, pp. 1462–1466.

[22] A. Plinge and G. A. Fink, "Multi-speaker tracking using multiple distributed microphone arrays," in *ICASSP*, 2014, pp. 614–618.

[23] J. Yang, X. Zhong, W. Chen, and W. Wang, "Multiple acoustic source localization in microphone array networks," *IEEE/ACM TASLP*, vol. 29, pp. 334–347, 2021.

[24] E. Grinstein, M. Brookes, and P. A. Naylor, "Graph neural networks for sound source localization on distributed microphone networks," in *ICASSP*, 2023, pp. 1–5.

[25] M. Guzik and K. Kowalczyk, "Wishart localization prior on spatial covariance matrix in ambisonic source separation using non-negative tensor factorization," in *ICASSP*, 2022, pp. 446–450.

[26] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel NMF and acoustic tracking," *IEEE/ACM TASLP*, vol. 26, no. 2, pp. 281–295, 2018.

[27] Y. Sumura, K. Sekiguchi, Y. Bando, A. A. Nugraha, and K. Yoshii, "Joint localization and synchronization of distributed camera-attached microphone arrays for indoor scene analysis," in *IWAENC*, 2022, pp. 1–5.

[28] F. Jacob, J. Schmalenstroeer, and R. Haeb-Umbach, "Microphone array position self-calibration from reverberant speech input," in *IWAENC*, 2012, pp. 1–4.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015, pp. 1–15.

[30] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Proc. 5th ISCA Worksh. Speech Synth.*, 2004, pp. 223–224.

[31] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *ICASSP*, 2018, pp. 351–355.