# Fast Multichannel Source Separation Based on Jointly Diagonalizable Spatial Covariance Matrices

Kouhei Sekiguchi*†     Aditya Arie Nugraha*     Yoshiaki Bando‡     Kazuyoshi Yoshii*†

*Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo 103-0027, Japan
Email: {kouhei.sekiguchi, adityaarie.nugraha, kazuyoshi.yoshii}@riken.jp
†Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
‡National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, 135-0064, Japan

*Abstract*—This paper describes a versatile method that accelerates multichannel source separation methods based on full-rank spatial modeling. A popular approach to multichannel source separation is to integrate a spatial model with a source model for estimating the spatial covariance matrices (SCMs) and power spectral densities (PSDs) of each sound source in the time-frequency domain. One of the most successful examples of this approach is multichannel nonnegative matrix factorization (MNMF) based on a full-rank spatial model and a low-rank source model. MNMF, however, is computationally expensive and often works poorly due to the difficulty of estimating the unconstrained full-rank SCMs. Instead of restricting the SCMs to rank-1 matrices with the severe loss of the spatial modeling ability as in independent low-rank matrix analysis (ILRMA), we restrict the SCMs of each frequency bin to jointly-diagonalizable but still full-rank matrices. For such a fast version of MNMF, we propose a computationally-efficient and convergence-guaranteed algorithm that is similar in form to that of ILRMA. Similarly, we propose a fast version of a state-of-the-art speech enhancement method based on a deep speech model and a low-rank noise model. Experimental results showed that the fast versions of MNMF and the deep speech enhancement method were several times faster and performed even better than the original versions of those methods, respectively.

*Index Terms*—Multichannel source separation, speech enhancement, spatial modeling, joint diagonalization

## I. INTRODUCTION

Multichannel source separation plays a central role for computational auditory scene analysis. To use an automatic speech recognition system in a noisy environment, for example, it is necessary to separate speech signals from noise-contaminated signals. A standard approach to multichannel source separation is to use a non-blind method (*e.g.*, beamforming and Wiener filtering) based on the spatial covariance matrix (SCM) of a target source (*e.g.*, speech) and those of the other sources (*e.g.*, noise). In beamforming-based speech enhancement, deep neural networks (DNNs) are often used for classifying each time-frequency bin into speech or noise [1]–[3]. The performance of such a supervised approach, however, is often considerably degraded in an unseen environment. We thus focus on general-purpose blind source separation (BSS) and its extension for environment-adaptive semi-supervised speech enhancement.

The goal of BSS is to estimate both a mixing process and sound sources from observed mixtures. To solve such an ill-
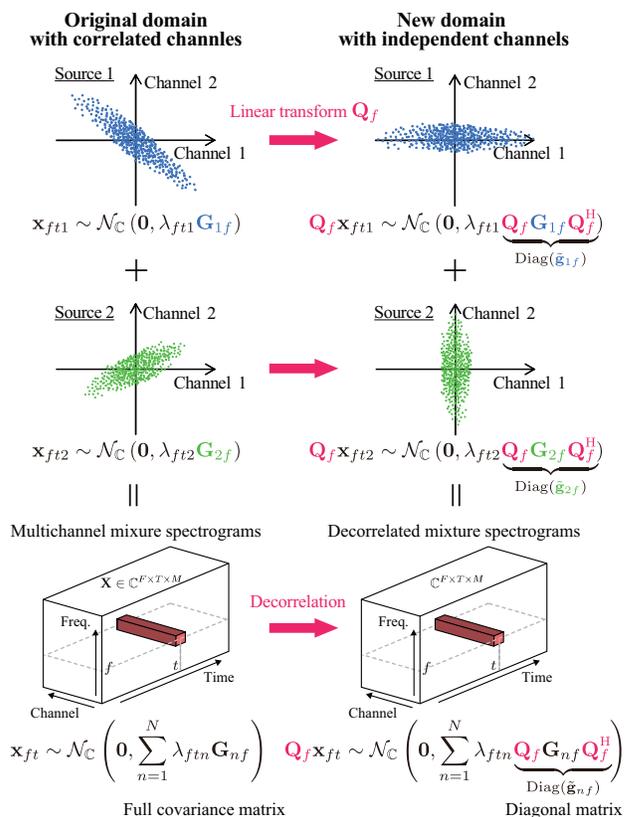
Fig. 1: The full-rank spatial model for the *correlated* channels of the original data is equivalent to the diagonal spatial model for the *independent* channels of the decorrelated data.

posed problem, one can take a statistical approach based on a *spatial model* representing a sound propagation process and a *source model* representing the power spectral densities (PSDs) of each source. Duong *et al.* [4] pioneered this approach by integrating a full-rank spatial model using the frequency-wise full-rank SCMs of sources with a source model assuming the source spectra to be complex Gaussian distributed, which we call full-rank spatial covariance analysis (FCA) as in [5]–[7]. To alleviate the frequency permutation problem of FCA, multichannel nonnegative matrix factorization (MNMF) that uses an NMF-based source model based on the co-occurrence and low-rankness of frequency components has been studied [8]–[10].

Such a low-rank source model, however, does not fit speech spectra. In speech enhancement, a semi-supervised approach that uses as source models a DNN-based speech model (deep prior, DP) trained from clean speech data and an NMF-based noise model learned on the fly has thus recently been investigated (called MNMF-DP) [11]–[13].

The major drawbacks common to these methods based on the full-rank SCMs are the high computational cost due to the repeated heavy operations (*e.g.*, inversion) of the SCMs and the difficulty of parameter optimization due to the large degree of freedom (DOF) of the spatial model. Kitamura *et al.* [14] thus proposed a constrained version of MNMF called independent low-rank matrix analysis (ILRMA) that restricts the SCMs to *rank-1* matrices. While ILRMA is an order of magnitude faster and practically performed better than MNMF, it suffers from the severe loss of the spatial modeling ability. Ito and Nakatani recently proposed a fast version of FCA [5], [6] and that of MNMF [7] by restricting the SCMs of each frequency bin to *jointly-diagonalizable* matrices. For parameter estimation, an expectation-maximization (EM) algorithm [5] or a majorization-minimization (MM) algorithm [6], [7] was used with a fixed-point iteration (FPI) method, but its convergence was not theoretically guaranteed.

In this paper we propose a versatile convergence-guaranteed method for estimating the jointly-diagonalizable SCMs of the full-rank spatial model and its application to FCA, MNMF, and MNMF-DP called FastFCA, FastMNMF, and FastMNMF-DP, respectively. Note that FastMNMF (restricted full-rank SCMs) has an intermediate spatial modeling ability between MNMF (unconstrained full-rank SCMs) and ILRMA (rank-1 SCMs). While all channels are correlated in the original domain, they are independent in the transformed domain obtained by multiplying a diagonalizer (linear transform) to each frequency bin (Fig. 1). MNMF for the original *complex* spectrograms is thus equivalent to computationally-efficient nonnegative tensor factorization (NTF) for the *nonnegative* power spectrograms in the transformed domain. To estimate such a diagonalizer, we use a convergence-guaranteed iterative projection (IP) method instead of the FPI method in a way similar to the estimation of a demixing filter in independent vector analysis (IVA) [15]. Our algorithm based on the iteration of NTF and IP is similar in form to ILRMA [14] based on that of NMF and IP.

One of the important contributions of this paper is to show the effectiveness of the joint diagonalization of covariance matrices in a general context of multichannel source separation. This idea was first discussed for an ultimate, computationally-prohibitive extension of NTF called correlated tensor factorization (CTF) [16] based on multi-way full-rank covariance matrices, resulting in a fast version of CTF called independent low-rank tensor analysis (ILRTA) [17]. While ILRTA was used for single-channel BSS based on jointly-diagonalizable *frequency* covariance matrices, in this paper we focus on multi-channel BSS based on jointly-diagonalizable *spatial* covariance matrices. In fact, our FastMNMF can be regarded as a special case of ILRTA because in both methods NTF and IP are iterated for convergence-guaranteed parameter optimization.

## II. MULTICHANNEL SOURCE SEPARATION

This section reviews existing multichannel source separation methods based on a full-rank spatial model, *i.e.*, full-rank spatial covariance analysis (FCA) [4] based on an unconstrained source model, MNMF [10] based on an NMF-based source model, and its adaptation to speech enhancement called MNMF-DP [12] based on a DNN-based speech model and an NMF-based noise model.

### A. Full-Rank Spatial Model

*1) Model Formulation:* Suppose that $N$ sources are observed by $M$ microphones. Let $\mathbf{X} = \{\mathbf{x}_{ft}\}_{f,t=1}^{F,T} \in \mathbb{C}^{F \times T \times M}$ be the observed multichannel complex spectra, where $F$ and $T$ are the number of frequency bins and that of frames, respectively. Let $\mathbf{x}_{ftn} = [x_{ftn1}, \cdots, x_{ftnM}]^{\mathrm{T}} \in \mathbb{C}^M$ be the image of source $n$ assumed to be circularly-symmetric complex Gaussian distributed as follows:

$$\mathbf{x}_{ftn} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \lambda_{ftn}\mathbf{G}_{nf}\right), \tag{1}$$

where $\lambda_{ftn}$ is the PSD of source $n$ at frequency $f$ and time $t$, $\mathbf{G}_{nf}$ is the $M \times M$ positive definite full-rank SCM of source $n$ at frequency $f$. Using the reproductive property of the Gaussian distribution, the observed spectrum $\mathbf{x}_{ft} = \sum_{n=1}^{N} \mathbf{x}_{ftn}$ is given by

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n=1}^{N} \lambda_{ftn}\mathbf{G}_{nf}\right). \tag{2}$$

Given the mixture spectrum $\mathbf{x}_{ft}$ and the model parameters $\mathbf{G}_{nf}$ and $\lambda_{ftn}$, the posterior expectation of the source image $\mathbf{x}_{ftn}$ is obtained by multichannel Wiener filtering (MWF):

$$\mathbf{x}_{ftn} = \mathbb{E}[\mathbf{x}_{ftn}|\mathbf{x}_{ft}] = \mathbf{Y}_{ftn}\mathbf{Y}_{ft}^{-1}\mathbf{x}_{ft}, \tag{3}$$

where $\mathbf{Y}_{ftn} \stackrel{\text{def}}{=} \lambda_{ftn}\mathbf{G}_{nf}$ and $\mathbf{Y}_{ft} \stackrel{\text{def}}{=} \sum_{n=1}^{N}\mathbf{Y}_{ftn}$.

*2) Parameter Estimation:* Our goal is to estimate the parameters $\mathbf{G} = \{\mathbf{G}_{nf}\}_{n,f=1}^{N,F}$ and $\mathbf{\Lambda} = \{\lambda_{ftn}\}_{f,t,n=1}^{F,T,N}$ that maximize the log-likelihood given by Eq. (2):

$$\log p(\mathbf{X}|\mathbf{G}, \mathbf{\Lambda}) \stackrel{c}{=} -\sum_{f,t=1}^{F,T}\left(\mathrm{tr}\left(\mathbf{X}_{ft}\mathbf{Y}_{ft}^{-1}\right) + \log|\mathbf{Y}_{ft}|\right), \tag{4}$$

where $\mathbf{X}_{ft} \stackrel{\text{def}}{=} \mathbf{x}_{ft}\mathbf{x}_{ft}^{\mathrm{H}}$. In this paper we use a majorization-minimization (MM) algorithm [10] that iteratively maximizes a lower bound of Eq. (4). As in [16], [17], the closed-form update rule of $\mathbf{G}$ was recently found to be given by

$$\mathbf{A}_{nf} \stackrel{\text{def}}{=} \sum_{t=1}^{T}\lambda_{ftn}\mathbf{Y}_{ft}^{-1}\mathbf{X}_{ft}\mathbf{Y}_{ft}^{-1}, \tag{5}$$

$$\mathbf{B}_{nf} \stackrel{\text{def}}{=} \sum_{t=1}^{T}\lambda_{ftn}\mathbf{Y}_{ft}^{-1}, \tag{6}$$

$$\mathbf{G}_{nf} \leftarrow \mathbf{B}_{nf}^{-1}\left(\mathbf{B}_{nf}\mathbf{G}_{nf}\mathbf{A}_{nf}\mathbf{G}_{nf}\right)^{\frac{1}{2}}. \tag{7}$$

### B. Source Models

*1) Unconstrained Source Model:* The unconstrained model directly uses $\mathbf{\Lambda}$ as free parameters. Using the MM algorithm, the multiplicative update (MU) rule of $\mathbf{\Lambda}$ is given by

$$\lambda_{ftn} \leftarrow \lambda_{ftn}\sqrt{\frac{\mathrm{tr}\left(\mathbf{G}_{nf}\mathbf{Y}_{ft}^{-1}\mathbf{X}_{ft}\mathbf{Y}_{ft}^{-1}\right)}{\mathrm{tr}\left(\mathbf{G}_{nf}\mathbf{Y}_{ft}^{-1}\right)}}. \tag{8}$$

*2) NMF-Based Source Model:* If the PSDs $\{\lambda_{ftn}\}_{f,t=1}^{F,T}$ of a source $n$ (*e.g.*, noise and music) have low-rank structure, the PSDs can be factorized as follows [10]:

$$\lambda_{ftn} = \sum_{k=1}^{K} w_{nkf} h_{nkt}, \tag{9}$$

where $K$ is the number of bases, $w_{nkf} \geq 0$ is the magnitude of basis $k$ of source $n$ at frequency $f$, and $h_{nkt} \geq 0$ is the activation of basis $k$ of source $n$ at time $t$. Using the MM algorithm [18], the MU rules of $\mathbf{W}$ and $\mathbf{H}$ are given by

$$w_{nkf} \leftarrow w_{nkf} \sqrt{\frac{\sum_{t=1}^{T} h_{nkt} \operatorname{tr}\left(\mathbf{G}_{nf} \mathbf{Y}_{ft}^{-1} \mathbf{X}_{ft} \mathbf{Y}_{ft}^{-1}\right)}{\sum_{t=1}^{T} h_{nkt} \operatorname{tr}\left(\mathbf{G}_{nf} \mathbf{Y}_{ft}^{-1}\right)}}, \tag{10}$$

$$h_{nkt} \leftarrow h_{nkt} \sqrt{\frac{\sum_{f=1}^{F} w_{nkf} \operatorname{tr}\left(\mathbf{G}_{nf} \mathbf{Y}_{ft}^{-1} \mathbf{X}_{ft} \mathbf{Y}_{ft}^{-1}\right)}{\sum_{f=1}^{F} w_{nkf} \operatorname{tr}\left(\mathbf{G}_{nf} \mathbf{Y}_{ft}^{-1}\right)}}. \tag{11}$$

*3) DNN-Based Source Model:* To represent the complicated characteristics of the PSDs $\{\lambda_{ftn}\}_{f,t=1}^{F,T}$ of a source $n$ (*e.g.*, speech), a deep generative model can be used as follows [11]:

$$\lambda_{ftn} = u_{nf} v_{nt} [\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\mathbf{z}_{nt})]_f \tag{12}$$

where $\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\cdot)$ is a nonlinear function (DNN) with parameters $\boldsymbol{\theta}$ that maps a latent variable $\mathbf{z}_{nt} \in \mathbb{R}^D$ to a nonnegative spectrum $\mathbf{r}_{nt} \stackrel{\text{def}}{=} \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\mathbf{z}_{nt}) \in \mathbb{R}_+^F$ at each time $t$, $[\cdot]_f$ indicates the $f$-th element of a vector, $u_{nf} \geq 0$ is a scaling factor at frequency $f$, and $v_{nt} \geq 0$ is an activation at time $t$.

To update the latent variables $\mathbf{Z}_n = \{\mathbf{z}_{nt}\}_{t=1}^T$, we use Metropolis sampling. A proposal $\mathbf{z}_{nt}^{\text{new}} \sim \mathcal{N}(\mathbf{z}_{nt}^{\text{old}}, \epsilon \mathbf{I})$ is accepted with probability $\min(1, \gamma_{nt})$, where $\gamma_{nt}$ is given by

$$\log \gamma_{nt} = -\sum_{f=1}^{F} \left(\frac{1}{\lambda_{ftn}^{\text{new}}} - \frac{1}{\lambda_{ftn}^{\text{old}}}\right) \operatorname{tr}\left(\mathbf{G}_{nf} \mathbf{Y}_{ft}^{-1} \mathbf{X}_{ft} \mathbf{Y}_{ft}^{-1}\right)$$
$$- \sum_{f=1}^{F} \left(\lambda_{ftn}^{\text{new}} - \lambda_{ftn}^{\text{old}}\right) \operatorname{tr}\left(\mathbf{G}_{nf} \mathbf{Y}_{ft}^{-1}\right), \tag{13}$$

where $\lambda_{ftn}^{\text{new}} = u_{nf} v_{nt} [\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\mathbf{z}_{nt}^{\text{new}})]_f$, $\lambda_{ftn}^{\text{old}} = u_{nf} v_{nt} [\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\mathbf{z}_{nt}^{\text{old}})]_f$. In practice, we update $\mathbf{Z}_n$ several times without updating $\mathbf{Y}_{ft}$ to reduce the computational cost of calculating $\mathbf{Y}_{ft}^{-1}$.

In the same way as the NMF-based source model, the MU rules of $\mathbf{U}$ and $\mathbf{V}$ are given by

$$u_{nf} \leftarrow u_{nf} \sqrt{\frac{\sum_{t=1}^{T} v_{nt} r_{ntf} \operatorname{tr}\left(\mathbf{G}_{nf} \mathbf{Y}_{ft}^{-1} \mathbf{X}_{ft} \mathbf{Y}_{ft}^{-1}\right)}{\sum_{t=1}^{T} v_{nt} r_{ntf} \operatorname{tr}\left(\mathbf{G}_{nf} \mathbf{Y}_{ft}^{-1}\right)}}, \tag{14}$$

$$v_{nt} \leftarrow v_{nt} \sqrt{\frac{\sum_{f=1}^{F} u_{nf} r_{ntf} \operatorname{tr}\left(\mathbf{G}_{nf} \mathbf{Y}_{ft}^{-1} \mathbf{X}_{ft} \mathbf{Y}_{ft}^{-1}\right)}{\sum_{f=1}^{F} u_{nf} r_{ntf} \operatorname{tr}\left(\mathbf{G}_{nf} \mathbf{Y}_{ft}^{-1}\right)}}. \tag{15}$$

*C. Integration of Spatial and Source Models*

*1) Full-Rank Spatial Covariance Analysis:* FCA [4] is obtained by integrating the full-rank spatial model and the unconstrained source model. While the EM algorithm was originally used in [4], in this paper we use the MM algorithm expected to converge faster as proposed in [6].

*2) Multichannel NMF:* MNMF [10] is obtained by integrating the NMF-based source model into FCA.

*3) MNMF with a Deep Prior:* MNMF-DP [12] specialized for speech enhancement is obtained by integrating the full-rank spatial model and the DNN- and NMF-based source models representing speech and noise sources, respectively. Assuming a source indexed by $n = 1$ corresponds to the speech, $\lambda_{ft1}$ and $\lambda_{ft(n \geq 2)}$ are given by Eq. (12) and Eq. (9), respectively.

## III. FAST MULTICHANNEL SOURCE SEPARATION

This section proposes the fast versions of FCA, MNMF, and MNMF-DP based on the jointly diagonalizable SCMs.

*A. Jointly Diagonalizable Full-Rank Spatial Model*

*1) Model Formulation:* To reduce the computational cost of the full-rank spatial model, we put a constraint that the SCMs $\{\mathbf{G}_{nf}\}_{n=1}^{N}$ can be jointly diagonalized as follows:

$$\mathbf{Q}_f \mathbf{G}_{nf} \mathbf{Q}_f^{\text{H}} = \operatorname{Diag}(\tilde{\mathbf{g}}_{nf}), \tag{16}$$

where $\mathbf{Q}_f = [\mathbf{q}_{f1}, \cdots, \mathbf{q}_{fM}]^{\text{H}} \in \mathbb{C}^{M \times M}$ is a non-singular matrix called a *diagonalizer* and $\tilde{\mathbf{g}}_{nf} = [\tilde{g}_{nf1}, \cdots, \tilde{g}_{nfM}] \in \mathbb{R}_+^M$ is a nonnegative vector. The observed spectrum $\mathbf{x}_{ft}$ is projected into a new space where the elements of the projected spectrum $\mathbf{Q}_f \mathbf{x}_{ft}$ are all independent (Fig. 1).

*2) Parameter Estimation:* Our goal is to jointly estimate $\mathbf{Q}$, $\tilde{\mathbf{G}}$, and $\boldsymbol{\Lambda}$ that maximize the log-likelihood given by substituting Eq. (16) into Eq. (2) as follows:

$$\log p(\mathbf{X} | \mathbf{Q}, \tilde{\mathbf{G}}, \boldsymbol{\Lambda})$$
$$= \sum_{f,t=1}^{F,T} \log \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{ft} \middle| \mathbf{0}, \sum_{n=1}^{N} \lambda_{ftn} \mathbf{Q}_f^{-1} \operatorname{Diag}(\tilde{\mathbf{g}}_{nf}) \mathbf{Q}_f^{-\text{H}}\right)$$
$$\stackrel{\text{c}}{=} \sum_{f,t,m=1}^{F,T,M} \left(-\frac{\tilde{x}_{ftm}}{\tilde{y}_{ftm}} - \log \tilde{y}_{ftm}\right) + T \sum_{f=1}^{F} \log |\mathbf{Q}_f \mathbf{Q}_f^{\text{H}}|, \tag{17}$$

where $\tilde{\mathbf{x}}_{ft} = \operatorname{Diag}(\mathbf{Q}_f \mathbf{X}_{ft} \mathbf{Q}_f^{\text{H}}) = |\mathbf{Q}_f \mathbf{x}_{ft}|^{\circ 2}$, $|\cdot|^{\circ 2}$ indicates the element-wise absolute square, and $\tilde{\mathbf{y}}_{ft} = \sum_{n=1}^{N} \lambda_{ftn} \tilde{\mathbf{g}}_{nf}$.

Since Eq. (17) has the same form as the log-likelihood function of IVA [15], $\mathbf{Q}_f$ can be updated by using the convergence-guaranteed iterative projection (IP) method as follows:

$$\mathbf{V}_{fm} \stackrel{\text{def}}{=} \frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_{ft} \tilde{y}_{ftm}^{-1}, \tag{18}$$

$$\mathbf{q}_{fm} \leftarrow (\mathbf{Q}_f \mathbf{V}_{fm})^{-1} \mathbf{e}_m, \tag{19}$$

$$\mathbf{q}_{fm} \leftarrow (\mathbf{q}_{fm}^{\text{H}} \mathbf{V}_{fm} \mathbf{q}_{fm})^{-\frac{1}{2}} \mathbf{q}_{fm}, \tag{20}$$

where $\mathbf{e}_m$ is a one-hot vector whose $m$-th element is 1. A diagonalizer $\mathbf{Q}_f$ is estimated so that the $M$ components (*channels*) of $\{\mathbf{Q}_f \mathbf{x}_{ft}\}_{f,t=1}^{F,T}$ become independent. In IVA [15] and ILRMA [14] under a determined condition ($M = N$), a demixing matrix $\mathbf{D}_f$ is estimated so that the $M$ components (*sources*) of $\{\mathbf{D}_f \mathbf{x}_{ft}\}_{f,t=1}^{F,T}$ become independent. In any case, the characteristics of the components (*e.g.*, low-rankness in the NMF-based source model) represented by $\{\tilde{\mathbf{y}}_{ft}\}_{f,t=1}^{F,T}$ are considered. This implies that our method could work as fast as ILRMA even in an underdetermined condition ($M < N$) while keeping the full-rank spatial modeling ability.

Since the first term of Eq. (17) is the negative Itakura-Saito (IS) divergence between $\tilde{x}_{ftm}$ and $\tilde{y}_{ftm}$, the MU rule of $\tilde{\mathbf{G}}$ is given by using the MM algorithm for IS-NMF [18] as follows:

$$\tilde{g}_{nfm} \leftarrow \tilde{g}_{nfm}\sqrt{\frac{\sum_{t=1}^{T} \lambda_{ftn}\tilde{y}_{ftm}^{-1}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-1}}{\sum_{t=1}^{T} \lambda_{ftn}\tilde{y}_{ftm}^{-1}}}. \quad (21)$$

### B. Source Models

*1) Unconstrained Source Model:* Using the MM algorithm for IS-NMF [18], the MU rule of $\mathbf{\Lambda}$ is given by

$$\lambda_{ftn} \leftarrow \lambda_{ftn}\sqrt{\frac{\sum_{m=1}^{M} \tilde{g}_{nfm}\tilde{y}_{ftm}^{-1}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-1}}{\sum_{m=1}^{M} \tilde{g}_{nfm}\tilde{y}_{ftm}^{-1}}}. \quad (22)$$

*2) NMF-Based Source Model:* Similarly, the MU rules of $\mathbf{W}$ and $\mathbf{H}$ included in Eq. (9) are given by

$$w_{nkf} \leftarrow w_{nkf}\sqrt{\frac{\sum_{t,m=1}^{T,M} h_{nkt}\tilde{g}_{nfm}\tilde{y}_{ftm}^{-1}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-1}}{\sum_{t,m=1}^{T,M} h_{nkt}\tilde{g}_{nfm}\tilde{y}_{ftm}^{-1}}}, \quad (23)$$

$$h_{nkt} \leftarrow h_{nkt}\sqrt{\frac{\sum_{f,m=1}^{F,M} w_{nkf}\tilde{g}_{nfm}\tilde{y}_{ftm}^{-1}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-1}}{\sum_{f,m=1}^{F,M} w_{nkf}\tilde{g}_{nfm}\tilde{y}_{ftm}^{-1}}}. \quad (24)$$

*3) DNN-Based Source Model:* To update the latent variables $\mathbf{Z}_n$ included in Eq. (12), we use Metropolis sampling. A proposal $\mathbf{z}_{nt}^{\text{new}} \sim \mathcal{N}(\mathbf{z}_{nt}^{\text{old}}, \epsilon\mathbf{I})$ is accepted with probability $\min(1, \gamma_{nt})$, where $\gamma_{nt}$ is given by

$$\log\gamma_{nt} = -\sum_{f,m=1}^{F,M}\left(\frac{\tilde{x}_{ftm}}{\lambda_{ftn}^{\text{new}}\tilde{g}_{nfm} + \tilde{y}_{ftm}^{\neg n}} - \frac{\tilde{x}_{ftm}}{\lambda_{ftn}^{\text{old}}\tilde{g}_{nfm} + \tilde{y}_{ftm}^{\neg n}}\right)$$
$$- \sum_{f,m=1}^{F,M} \log\frac{\lambda_{ftn}^{\text{new}}\tilde{g}_{nfm} + \tilde{y}_{ftm}^{\neg n}}{\lambda_{ftn}^{\text{old}}\tilde{g}_{nfm} + \tilde{y}_{ftm}^{\neg n}}, \quad (25)$$

where $\tilde{y}_{ftm}^{\neg n} \stackrel{\text{def}}{=} \sum_{n'\neq n}^{N} \lambda_{ftn'}\tilde{g}_{n'fm}$ is a reconstruction without the component of source $n$. As in the NMF-based source model, the MU rules of $\mathbf{U}$ and $\mathbf{V}$ are given by

$$u_{nf} \leftarrow u_{nf}\sqrt{\frac{\sum_{t,m=1}^{T,M} v_{nt}[\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{2}(\mathbf{z}_{nt})]_{f}\tilde{g}_{nfm}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{t,m=1}^{T,M} v_{nt}[\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{2}(\mathbf{z}_{nt})]_{f}\tilde{g}_{nfm}\tilde{y}_{ftm}^{-1}}}, \quad (26)$$

$$v_{nt} \leftarrow v_{nt}\sqrt{\frac{\sum_{f,m=1}^{F,M} u_{nf}[\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{2}(\mathbf{z}_{nt})]_{f}\tilde{g}_{nfm}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{f,m=1}^{F,M} u_{nf}[\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{2}(\mathbf{z}_{nt})]_{f}\tilde{g}_{nfm}\tilde{y}_{ftm}^{-1}}}. \quad (27)$$

### C. Integration of Spatial and Source Models

*1) FastFCA:* The fast version of FCA is obtained by integrating the jointly diagonalizable full-rank spatial model and the unconstrained source model. While the MM algorithm that uses the FPI method for estimating $\mathbf{Q}$ was proposed in [6], in this paper we use the MM algorithm with the convergence-guaranteed IP method.

*2) FastMNMF:* The fast version of MNMF is obtained by integrating the NMF-based source model into FastFCA.

*3) FastMNMF-DP:* The fast version of MNMF-DP is obtained by integrating the jointly diagonalizable full-rank spatial model, the DNN- and NMF-based source models representing speech and noise sources, respectively.

## IV. Evaluation

This section reports speech enhancement experiments.

### A. Experimental Conditions

100 simulated noisy speech signals sampled at 16 kHz were randomly selected from the evaluation dataset of CHiME3 [19]. These data were supposed to be recorded by six microphones attached to a tablet device. Five channels ($M = 5$) excluding the second channel behind the tablet were used. The short-time Fourier transform with a window length of 1024 points ($F = 513$) and a shifting interval of 256 points was used. To evaluate the performance of speech enhancement, the signal-to-distortion ratio (SDR) was measured [20], [21]. To evaluate the computational efficiency, the elapsed time per iteration for processing 8 sec data was measured on Intel Xeon W-2145 (3.70 GHz) or NVIDIA TITAN RTX.

FastFCA (Section III-C1), FastMNMF (Section III-C2), and FastMNMF-DP (Section III-C3) based on the *jointly diagonalizable* SCMs were compared with FCA [6] (Section II-C1), MNMF [10] (Section II-C2), MNMF-DP [12] (Section II-C3) based on the *unconstrained* SCMs. All methods used the MM algorithms (and the IP method in the fast versions) described in this paper. The original FastFCA-FPI [6] that uses the FPI method and ILRMA [14] were also tested.

The number of sources $N$ was set as $2 \leq N \leq M$ except for ILRMA used only in a determined condition $N = M = 5$. The number of iterations was 100. For the NMF-based source model, the number of bases $K$ was set to 4, 16, or 64. For the DNN-based source model, the latent variables $\mathbf{Z}_1$ with $D = 16$ were updated 30 times per iteration and the proposal variance $\epsilon$ was set to $10^{-4}$. The parameters $\boldsymbol{\theta}$ were trained in advance from clean speech data of about 15 hours included in WSJ-0 corpus [22] as described in [23]. More specifically, a DNN-based decoder $\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{2}$ that generates $\mathbf{X}$ from $\mathbf{Z}$ and a DNN-based encoder that infers $\mathbf{Z}$ from $\mathbf{X}$ were trained jointly in a variational autoencoding manner [24]. The SCM of speech $\mathbf{G}_{1f}$ was initialized as the average of the observed SCMs and the SCMs of noise $\mathbf{G}_{(n\geq 2)f}$ were initialized as the identity matrices. $\tilde{\mathbf{G}}$ and $\mathbf{Q}$ were initialized with spectral decomposition of $\mathbf{G}$. $\mathbf{Z}_1$ was initialized by feeding $\mathbf{X}$ to the encoder.

### B. Experimental Results

Tables I-(a) and I-(b) list the elapsed times per iteration and Table II lists the average SDRs. The proposed FastFCA was around 10% faster than FastFCA-FPI [6] on the CPU because the IP method and the FPI method execute the matrix inversion only once and twice, respectively, for updating $\mathbf{Q}$. FastFCA, FastMNMF, and FastMNMF-DP were an order of magnitude faster and performed even better than their original versions. It is known that any two positive definite (covariance) matrices can be exactly jointly diagonalized and this does not apply to more than two matrices. Although the fast versions with $N \geq 3$ are thus inferior to the original versions in terms of the DOF, the restriction of the DOF of the spatial model was proved to be practically effective for avoiding bad local optima. The fast versions with $N = 2$ are comparable with the original versions

TABLE I: The elapsed times per iteration for processing 5ch noisy speech signals of 8 [sec].

(a) Elapsed times [sec] on CPU (Intel Xeon W-2145 3.70 GHz)

| Method | FastFCA-FPI [6] | FCA [6] / FastFCA | ILRMA [14] | | | MNMF [10] / FastMNMF | | | MNMF-DP [12] / FastMNMF-DP | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| # of bases $K$ | – | – | 4 | 16 | 64 | 4 | 16 | 64 | 4 | 16 | 64 |
| # of sources $N$   2 | 0.47 | 3.3 / 0.43 | – | – | – | 5.1 / 0.67 | 5.2 / 0.74 | 5.5 / 1.2 | 12 / 1.9 | 12 / 1.9 | 12 / 2.1 |
| 3 | 0.50 | 4.0 / 0.47 | – | – | – | 6.1 / 0.71 | 6.2 / 0.84 | 6.8 / 1.7 | 14 / 1.9 | 14 / 2.1 | 14 / 2.6 |
| 4 | 0.61 | 4.7 / 0.56 | – | – | – | 7.0 / 0.81 | 7.3 / 0.99 | 8.0 / 2.2 | 16 / 2.1 | 16 / 2.2 | 16 / 3.1 |
| 5 | 0.69 | 5.3 / 0.63 | 0.51 | 0.61 | 1.0 | 8.1 / 0.87 | 8.3 / 1.1 | 9.2 / 2.7 | 18 / 2.2 | 18 / 2.4 | 18 / 3.7 |

(b) Elapsed times [decisec] on GPU (NVIDIA TITAN RTX)

| Method | FastFCA-FPI [6] | FCA [6] / FastFCA | ILRMA [14] | | | MNMF [10] / FastMNMF | | | MNMF-DP [12] / FastMNMF-DP | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| # of bases $K$ | – | – | 4 | 16 | 64 | 4 | 16 | 64 | 4 | 16 | 64 |
| # of sources $N$   2 | 0.082 | 1.5 / 0.088 | – | – | – | 3.0 / 0.12 | 3.0 / 0.21 | 3.2 / 0.54 | 5.3 / 0.60 | 5.3 / 0.64 | 5.3 / 0.81 |
| 3 | 0.088 | 2.1 / 0.095 | – | – | – | 4.2 / 0.15 | 4.2 / 0.27 | 4.5 / 0.77 | 7.0 / 0.63 | 7.1 / 0.71 | 7.1 / 1.1 |
| 4 | 0.093 | 2.7 / 0.099 | – | – | – | 5.3 / 0.17 | 5.4 / 0.33 | 5.7 / 1.0 | 8.8 / 0.65 | 8.8 / 0.78 | 8.9 / 1.3 |
| 5 | 0.10 | 3.4 / 0.11 | 0.10 | 0.16 | 0.39 | 6.6 / 0.19 | 6.7 / 0.40 | 7.1 / 1.2 | 11 / 0.68 | 11 / 0.86 | 11 / 1.6 |

TABLE II: The average SDRs [dB] for 100 noisy speech signals.

| Method | FastFCA-FPI [6] | FCA [6] / FastFCA | ILRMA [14] | | | MNMF [10] / FastMNMF | | | MNMF-DP [12] / FastMNMF-DP | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| # of bases $K$ | – | – | 4 | 16 | 64 | 4 | 16 | 64 | 4 | 16 | 64 |
| # of sources $N$   2 | 10.6 | 8.6 / 10.5 | – | – | – | 11.4 / 15.5 | 11.1 / 15.5 | 10.5 / 15.1 | 17.5 / 17.5 | 18.1 / 18.2 | 18.5 / 18.6 |
| 3 | 11.3 | 8.8 / 11.1 | – | – | – | 12.3 / 16.1 | 12.0 / 16.4 | 11.3 / 15.8 | 18.0 / 18.3 | 18.4 / 18.6 | 18.6 / 18.8 |
| 4 | 11.8 | 8.8 / 11.6 | – | – | – | 13.0 / 16.2 | 12.7 / 16.7 | 11.9 / 16.1 | 18.0 / 18.4 | 18.4 / **18.9** | 18.4 / **18.9** |
| 5 | 12.0 | 8.9 / 11.9 | 15.1 | 15.1 | 14.9 | 13.2 / 16.4 | 13.1 / 16.8 | 12.4 / 16.3 | 18.2 / 18.6 | 18.2 / 18.8 | 18.1 / 18.8 |

in terms of the DOF, but the fast versions were still better. One reason would be that while only the SCM of speech $\mathbf{G}_{1f}$ was initialized to a reasonable value in the original versions, the initialization of $\mathbf{Q}_f$ based on $\mathbf{G}_{1f}$ contributed to initializing the SCM of noise in the fast versions. We also observed that when $N = 5$ and $K = 4$ (the best condition for ILRMA), FastMNMF was as fast as and outperformed ILRMA.

## V. CONCLUSION

This paper presented a full-rank spatial model based on the jointly diagonalizable SCMs of sound sources and its application to existing methods such as FCA, MNMF, and MNMF-DP. For such fast versions, we proposed a general convergence-guaranteed MM algorithm that uses the IP method for estimating the SCMs. We experimentally showed that our approach is effective for improving both the separation performance and computational efficiency. One important direction is to develop online FastMNMF-DP for real-time noisy speech recognition because the real-time factor of FastMNMF-DP could be less than 1. We also plan to simultaneously consider the jointly diagonalizable *full-rank* spatial and frequency covariance matrices of sound sources as suggested in [17].

## ACKNOWLEDGMENT

## REFERENCES

[1] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *ICASSP*, 2013, pp. 7092–7096.

[2] H. Erdogan *et al.*, "Improved MVDR beamforming using single-channel mask prediction networks," in *Interspeech*, 2016, pp. 1981–1985.

[3] J. Heymann *et al.*, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*, 2016, pp. 196–200.

[4] N. Q. K. Duong *et al.*, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE TASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[5] N. Ito and T. Nakatani, "FastFCA-AS: Joint diagonalization based acceleration of full-rank spatial covariance analysis for separating any number of sources," in *IWAENC*, 2018, pp. 151–155.

[6] ——, "Multiplicative updates and joint diagonalization based acceleration for under-determined BSS using a full-rank spatial covariance model," in *GlobalSIP*, 2018, pp. 231–235.

[7] ——, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *ICASSP*, 2019, pp. 371–375.

[8] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE TASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[9] S. Arberet *et al.*, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *ISSPA*, 2010, pp. 1–4.

[10] H. Sawada *et al.*, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE TASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[11] Y. Bando *et al.*, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *ICASSP*, 2018, pp. 716–720.

[12] K. Sekiguchi *et al.*, "Bayesian multichannel speech enhancement with a deep speech prior," in *APSIPA*, 2018, pp. 1233–1239.

[13] S. Leglaive *et al.*, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *ICASSP*, 2019, pp. 101–105.

[14] D. Kitamura *et al.*, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE TASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

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

[16] K. Yoshii, "Correlated tensor factorization for audio source separation," in *ICASSP*, 2018, pp. 731–735.

[17] K. Yoshii *et al.*, "Independent low-rank tensor analysis for audio source separation," in *EUSIPCO*, 2018, pp. 1671–1675.

[18] M. Nakano *et al.*, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with $\beta$-divergence," in *MLSP*, 2010, pp. 283–288.

[19] J. Barker *et al.*, "The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes," *Computer Speech & Language*, vol. 46, pp. 605–626, 2017.

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

[21] C. Raffel *et al.*, "mir_eval: A transparent implementation of common MIR metrics," in *ISMIR*, 2014, pp. 367–372.

[22] J. Garofalo *et al.*, "CSR-I (WSJ0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[23] S. Leglaive *et al.*, "A variance modeling framework based on variational autoencoders for speech enhancement," in *IEEE MLSP*, 2018.

[24] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *ICLR*, 2014.