Determined Blind Source Separation Unifying Independent Vector Analysis and Nonnegative Matrix Factorization

Daichi Kitamura, *Member, IEEE*, Nobutaka Ono, *Senior Member, IEEE*, Hiroshi Sawada, *Senior Member, IEEE*, Hirokazu Kameoka, *Member, IEEE*, and Hiroshi Saruwatari, *Member, IEEE*

***Abstract*—This paper addresses the determined blind source separation problem and proposes a new effective method unifying independent vector analysis (IVA) and nonnegative matrix factorization (NMF). IVA is a state-of-the-art technique that utilizes the statistical independence between sources in a mixture signal, and an efficient optimization scheme has been proposed for IVA. However, since the source model in IVA is based on a spherical multivariate distribution, IVA cannot utilize specific spectral structures such as the harmonic structures of pitched instrumental sounds. To solve this problem, we introduce NMF decomposition as the source model in IVA to capture the spectral structures. The formulation of the proposed method is derived from conventional multichannel NMF (MNMF), which reveals the relationship between MNMF and IVA. The proposed method can be optimized by the update rules of IVA and single-channel NMF. Experimental results show the efficacy of the proposed method compared with IVA and MNMF in terms of separation accuracy and convergence speed.**

***Index Terms*—Blind source separation, determined, independent vector analysis, nonnegative matrix factorization.**

I. Introduction

BLIND source separation (BSS) is a technique for separating specific sources from a recorded sound without any information about the recording environment, mixing system, or source locations. In a determined or overdetermined situation (number of microphones $\geq$ number of sources), independent component analysis (ICA) [1] is the method most commonly used to solve the BSS problem, and many ICA-based techniques have been proposed [2]–[6]. On the other hand, for an underdetermined situation (number of microphones $<$ number of sources) including monaural recording, nonnegative matrix factorization (NMF) [7], [8] has received much attention [9], [10]. BSS is generally used to solve speech separation problems, but recently the use of BSS for music signals has also become an active research area [11].

As a means of solving the permutation problem [12]–[15] (grouping the components of the same source signal over frequency bins) in frequency domain ICA (FDICA), independent vector analysis (IVA) [16]–[18] has been proposed. Recently, a fast and stable optimization scheme based on the auxiliary function technique has been proposed for FDICA [19] and IVA [20]. Such ICA-based methods assume independence between the sources to estimate a demixing matrix. In addition, IVA generally assumes a spherical multivariate distribution (e.g., spherical Laplace distribution) as the source model to ensure higher-order correlations between frequency bins in each source. This source model does not include any specific information on the spectral structures of sources, meaning that it can be generally used for various types of sound. However, some sources have specific spectral structures such as the harmonic structure of instrumental sounds or music tones. Therefore, the introduction of a better source model has the potential to improve the source separation performance.

In NMF-based methods, NMF decomposes the given spectrogram into several spectral bases and temporal activations, which must be clustered in every source to achieve source separation. One effective way of achieving this is to utilize a sample sound of the target signal [21]–[23]. However, such supervision cannot be utilized in BSS. To solve this problem, multichannel NMF (MNMF) has been proposed [24]–[30]. In particular, MNMF methods [27]–[30] treat convolutive mixtures similarly to FDICA and IVA and estimate a mixing system for the sources, which is utilized for the clustering of bases. In these MNMFs, the spatial covariance [31], [32], which is the covariance matrix of a zero-mean multivariate Gaussian distribution, has been utilized to model the mixing conditions of the recording environment. In [27], an MNMF method with rank-1 spatial covariance for each source was first proposed (hereafter referred to as *Ozerov's MNMF*), where the rank-1 constraint of the covariance matrix corresponds to the assumption of instantaneous mixtures in the frequency domain. Ozerov's MNMF also includes a full-rank spatial covariance that models an additive noise component. The mixing matrix, NMF variables, and noise component are simultaneously estimated by update rules based on expectation-maximization (EM) and multiplicative update (MU) algorithms. As a new method of modeling for more reverberant mixture signals, MNMF with a full-rank spatial covariance [32] for each source has been proposed [28], [30]. In particular, the MNMF in [30] (hereafter referred to as *Sawada's MNMF*) can be considered as a natural extension of simple NMF because Hermitian

Manuscript received September 29, 2015; revised February 21, 2016 and May 11, 2016; accepted May 30, 2016. Date of publication June 07, 2016; date of current version July 04, 2016. This work was supported in part by Grant-in-Aid for JSPS Fellows Number 26.10796. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. DeLiang Wang.

D. Kitamura is with the Department of Informatics, School of Multidisciplinary Sciences, SOKENDAI (The Graduate University for Advanced Studies), Kanagawa 240-0193 Japan (e-mail: d-kitamura@nii.ac.jp).

N. Ono is with the National Institute of Informatics, Tokyo 101-8430, Japan (e-mail: onono@nii.ac.jp).

H. Sawada and H. Kameoka are with the NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Kyoto 619-0237, Japan (e-mail: sawada.hiroshi@lab.ntt.co.jp; kameoka.hirokazu@lab.ntt.co.jp).

H. Saruwatari is with the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan (e-mail: hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp).

Digital Object Identifier 10.1109/TASLP.2016.2577880

positive semi-definiteness is utilized as a multichannel counterpart of nonnegativity, and new MU update rules are derived in a generic form. However, it was reported that the algorithms [27]–[30] are sensitive to the initial values in source separation tasks.

In this paper, we only focus on the determined BSS problem, and propose a new effective method unifying IVA and NMF. The proposed method exploits NMF decomposition to capture the spectral structures of each source as the source model in IVA. Intriguingly, the formulation of the proposed method can be derived from Sawada's MNMF by introducing a rank-1 spatial covariance for each source. This fact reveals the relationship between MNMF and IVA. Also, the proposed method can be optimized by the fast and stable update rules of IVA and conventional single-channel NMF. Experimental results show the efficacy of the proposed method compared with IVA and conventional MNMFs in terms of separation performance and convergence speed.

The rest of this paper is organized as follows. In Section II, conventional IVA and MNMFs are described. In Section III, we propose a new method unifying IVA and NMF and derive its update rules based on the auxiliary function technique. In Section IV, we discuss the inherent difference between the conventional and proposed methods on the basis of a simple experiment using artificial sources. In Section V, the efficacy of the proposed method is confirmed from the results of BSS experiments using speech and music signals. Following a discussion on the results of the experiments, we present our conclusions in Section VI. Note that this paper is partially based on an international conference paper [33] written by the authors. The contribution of this paper is that we provide a new analysis of the proposed method and report enhanced experiments carried out under various conditions.

II. Conventional Methods

A. Formulation

Let the numbers of sources and microphones (channels) be N and M, respectively. The multichannel source and the observed and separated signals in each time-frequency slot are described as

$$\boldsymbol{s}_{ij} = (s_{ij,1} \cdots s_{ij,N})^{\mathrm{t}}, \tag{1}$$

$$\boldsymbol{x}_{ij} = (x_{ij,1} \cdots x_{ij,M})^{\mathrm{t}}, \tag{2}$$

$$\boldsymbol{y}_{ij} = (y_{ij,1} \cdots y_{ij,N})^{\mathrm{t}}, \tag{3}$$

where $i=1,\ldots,I$; $j=1,\ldots,J$; $n=1,\ldots,N$; and $m=1,\ldots,M$ are the integral indexes of the frequency bins, time frames, sources, and channels, respectively, $^{\mathrm{t}}$ denotes the vector transpose, and all the entries of these vectors are complex values. When the window size in a short-time Fourier transform (STFT) is sufficiently long compared with the impulse responses between sources and microphones, we can approximately represent the observed signal as

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij}, \tag{4}$$

where $\boldsymbol{A}_i = (\boldsymbol{a}_{i,1} \ldots \boldsymbol{a}_{i,N})$ is an $M \times N$ mixing matrix and $\boldsymbol{a}_{i,n}$ is the steering vector for each source. For the case of an overdetermined signal ($M > N$), a standard approach is to apply principal component analysis (PCA) in advance to reduce the dimension of $\boldsymbol{x}_{ij}$ so that $M=N$. If the mixing matrix $\boldsymbol{A}_i$ is invertible and $M=N$, we can define the demixing matrix $\boldsymbol{W}_i = (\boldsymbol{w}_{i,1} \ldots \boldsymbol{w}_{i,M})^{\mathrm{h}}$ as the inverse of the mixing matrix, and the separated signal can be represented as

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}, \tag{5}$$

where $^{\mathrm{h}}$ denotes the Hermitian transpose.

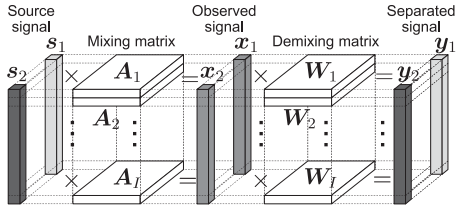

Fig. 1. Conceptual model of IVA ($N = M = 2$).

B. Independent Vector Analysis

IVA [16]–[18] is a multivariate extension of FDICA and can solve the permutation problem [12]–[14]. These ICA-based methods including IVA can only be applied to the determined situation ($M=N$) with rank-1 spatial model because they estimate demixing matrix $\boldsymbol{W}_i$. For simplicity, in this section, let M be equal to N. In IVA, we assume multivariate vector variable $\boldsymbol{y}_{j,m}$ and source $\boldsymbol{s}_{j,m}$ that consists of all frequency bins for time frame j and source m as

$$\boldsymbol{y}_{j,m} = (y_{1j,m} \cdots y_{Ij,m})^{\mathrm{t}}, \tag{6}$$

$$\boldsymbol{s}_{j,m} = (s_{1j,m} \cdots s_{Ij,m})^{\mathrm{t}}. \tag{7}$$

Also, a super-Gaussian spherical multivariate distribution is assumed as the source prior $p(\boldsymbol{s}_{j,m})$. In the literature [16]–[18], the following spherical Laplace distribution is often used:

$$p(\boldsymbol{s}_{j,m}) = \rho \exp\left(-\sqrt{\sum_i \left| \frac{s_{ij,m}}{r_{j,m}} \right|^2} \right), \tag{8}$$

where ρ is a normalization term and $r_{j,m}$ is the uniform variance over the frequency bins, which corresponds to the power spectrum of each source (source model). Such spherical symmetry of the source prior ensures a higher-order correlation between frequency bins. Fig. 1 shows the conceptual model of IVA. IVA can be used to estimate the demixing matrix $\boldsymbol{W}_i$ by assuming both independence between the vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and a higher-order correlation between the frequency bins in each vector.

The cost function of IVA is defined as follows:

$$Q_{\mathrm{IVA}}(\boldsymbol{W}_i) = \sum_m \frac{1}{J} \sum_j G(\boldsymbol{y}_{j,m}) - \sum_i \log |\det \boldsymbol{W}_i|, \tag{9}$$

where J is the number of time frames and $G(\boldsymbol{y}_{j,m})$ is a contrast function. When $\boldsymbol{y}_{j,m}$ has the distribution $p(\boldsymbol{y}_{j,m})$, the contrast function $G(\boldsymbol{y}_{j,m})$ is given as $-\log p(\boldsymbol{y}_{j,m})$. If we assume that the source components have the spherical Laplace distribution (8), $G(\boldsymbol{y}_{j,m}) = \|\boldsymbol{y}_{j,m}\|_2$ can be used, where $\|\cdot\|_2$ denotes the L_2 norm. For the minimization of (9), fast and stable update rules (hereafter referred to as iterative projection: IP) based on the auxiliary function technique have been proposed [20].

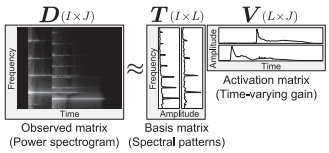

Fig. 2. Decomposition model of simple NMF ($L=2$).

For speech signal separation, IVA can achieve better separation performance than FDICA [16]. However, since only the higher-order correlation defined in (8) is utilized as the source model, IVA cannot treat the specific harmonic structures of each source and lacks flexibility. For this reason, IVA is not suitable for sources that have characteristic (specific) spectral structures, such as instrumental sounds or music signals.

C. Nonnegative Matrix Factorization

NMF is a type of sparse representation algorithm that decomposes a nonnegative matrix into two nonnegative matrices as

$$\boldsymbol{D} \approx \boldsymbol{T}\boldsymbol{V}, \tag{10}$$

where $\boldsymbol{D}(\in \mathbb{R}_{\geq 0}^{I\times J})$ is a nonnegative data matrix, which is the power spectrogram used when applying NMF to an acoustic signal, $\boldsymbol{T}(\in \mathbb{R}_{\geq 0}^{I\times L})$ is a basis matrix, which includes bases (frequently appearing spectral patterns in $\boldsymbol{D}$) as column vectors, and $\boldsymbol{V}(\in \mathbb{R}_{\geq 0}^{L\times J})$ is an activation matrix, which involves time-varying gains of each basis in $\boldsymbol{T}$ as row vectors. Also, L is the number of bases. Fig. 2 depicts the decomposition model of NMF, where the number of bases L equals two. In this figure, the basis matrix includes two types of spectral pattern as the bases to represent the observed matrix using time-varying gains in the activation matrix. In the decomposition of NMF, the variables $\boldsymbol{T}$ and $\boldsymbol{V}$ are optimized by minimizing the cost function. In this paper, we only focus on the following Itakura-Saito-divergence-based cost function [34]:

$$Q_{\mathrm{NMF}} = \sum_{i,j}\left(\frac{d_{ij}}{\sum_l t_{il}v_{lj}} + \log\sum_l t_{il}v_{lj}\right), \tag{11}$$

where constant terms are omitted, $l=1,\ldots,L$ is the integral index of the spectral bases, and d_{ij}, t_{il}, and v_{lj} are the nonnegative entries of $\boldsymbol{D}$, $\boldsymbol{T}$, and $\boldsymbol{V}$, respectively. The MU rules for $\boldsymbol{T}$ and $\boldsymbol{V}$ that minimize (11) are given by [34]

$$t_{il} \leftarrow t_{il}\sqrt{\frac{\sum_j d_{ij}v_{lj}\left(\sum_{l'} t_{il'}v_{l'j}\right)^{-2}}{\sum_j v_{lj}\left(\sum_{l'} t_{il'}v_{l'j}\right)^{-1}}}, \tag{12}$$

$$v_{lj} \leftarrow v_{lj}\sqrt{\frac{\sum_i d_{ij}t_{il}\left(\sum_{l'} t_{il'}v_{l'j}\right)^{-2}}{\sum_i t_{il}\left(\sum_{l'} t_{il'}v_{l'j}\right)^{-1}}}. \tag{13}$$

The source separation problem using NMF can be considered as how to cluster the decomposed bases into specific sources, and in the blind situation, it is still a very difficult problem. However, if the observed signal is obtained in a multichannel format, we can use information between channels (differences between gains and phases) for the clustering of the spectral bases. For this reason, the multichannel extension of NMF has been proposed as described in the next section.

D. Existing Works Related to Multichannel Extensions of NMF

There have been several multichannel extensions of NMF [27]–[30] and related methods [31], [32]. In these methods, the probability distribution of multichannel STFT coefficients $\boldsymbol{x}_{ij}$ is modeled by a phase-invariant multivariate zero-mean Gaussian distribution with a time-frequency variant covariance matrix as follows:

$$p(\boldsymbol{x}_{ij}) = \frac{1}{|\pi \boldsymbol{R}_{ij}^{(\boldsymbol{x})}|}\exp\left(-\boldsymbol{x}_{ij}^{\mathrm{h}}{\boldsymbol{R}_{ij}^{(\boldsymbol{x})}}^{-1}\boldsymbol{x}_{ij}\right), \tag{14}$$

where $\boldsymbol{R}_{ij}^{(\boldsymbol{x})}$ is called the *spatial covariance* of the observed multichannel signal. Existing MNMFs and their related works can be characterized in terms of two features: the spatial model $\boldsymbol{R}_{ij}^{(\boldsymbol{x})}$ and the modeling of source spectrograms.

Table I summarizes the models of the spatial covariance and the power spectrograms in the existing methods. The models proposed in [31], [32] have the most general representations. These methods decompose the spatial covariance $\boldsymbol{R}_{ij}^{(\boldsymbol{x})}$ into the power spectrogram $r_{ij,n}$ and the time-invariant spatial covariance $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ for each source, where $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ represents the spatial position and the spatial spread of the nth source. Several types of $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ have been investigated including rank-1 and full-rank matrices. Ozerov's MNMF [27] was the first method to model the power spectrogram $r_{ij,n}$ using NMF decomposition, namely, $r_{ij,n}=\sum_l t_{il,n}v_{lj,n}$. In this method, the sourcewise spatial covariance $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ is constrained by a rank-1 matrix, and an additive noise component $\boldsymbol{b}_{ij}$ is also assumed. The update rules of the variables based on both EM and MU algorithms have been derived. In the EM algorithm, an annealing technique for the noise component $\boldsymbol{b}_{ij}$ has been proposed because the spatial covariance for noise $\boldsymbol{R}_{i}^{(\boldsymbol{b})}$ is necessary for the stable optimization of $\boldsymbol{A}_i$. Ozerov's MNMF was extended to a full-rank spatial model in [28]. Also, a more flexible source model with a partitioning function z_{nk}, which clusters the bases into each source, was introduced in [29]. As another optimization scheme, an MU algorithm based on an auxiliary function technique has been proposed as Sawada's MNMF [30]. It also employs the full-rank $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ and the flexible source model with z_{nk} and NMF variables.

Since our proposed method, which will be described in the next section, is based on Sawada's MNMF, we here explain its formulation in detail. In Sawada's MNMF, the observed signal is represented as

$$\mathsf{X}_{ij} = \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{h}}, \tag{15}$$

where X_{ij} is a Hermitian positive definite matrix of size $M\times M$, which indicates the instantaneous covariance of the observed signal at the ij time-frequency slot. Therefore, the entire multichannel signal X can be considered as a fourth-order tensor, which has an $M\times M$ matrix as each element of the $I\times J$ matrix (see right-hand side of Fig. 3). The diagonal elements of X_{ij} represent real-valued nonnegative powers observed using each microphone, and the nondiagonal elements represent the complex-valued correlations between the microphones. The de-

TABLE I
MODELS OF MIXING SYSTEM, SPATIAL COVARIANCE, POWER SPECTROGRAM, AND THEIR OPTIMIZATION IN EACH METHOD

Literature	Mixing system	Spatial covariance	Power spectrogram	Optimization
Ozerov and Févotte [27]	$\boldsymbol{R}_{ij}^{(\boldsymbol{x})} = \sum_{n,l} t_{il,n}\, v_{lj,n}\, \boldsymbol{R}_{i,n}^{(\boldsymbol{s})} + \boldsymbol{R}_i^{(\boldsymbol{b})}$ $(\boldsymbol{x}_{ij} = \boldsymbol{A}_i\, \boldsymbol{s}_{ij} + \boldsymbol{b}_{ij})$	Rank-1 matrix $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ and diagonal matrix $\boldsymbol{R}_i^{(\boldsymbol{b})}$	NMF w/o partitioning function	EM and MU for $\boldsymbol{A}_i$, $\boldsymbol{R}_i^{(\boldsymbol{b})}$, $\boldsymbol{T}_n$, and $\boldsymbol{V}_n$
Arberet *et al.* [28]	$\boldsymbol{R}_{ij}^{(\boldsymbol{x})} = \sum_{n,l} t_{il,n}\, v_{lj,n}\, \boldsymbol{R}_{i,n}^{(\boldsymbol{s})} + \boldsymbol{R}_i^{(\boldsymbol{b})}$	Full-rank matrix $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ and diagonal matrix $\boldsymbol{R}_i^{(\boldsymbol{b})}$	NMF w/o partitioning function	EM for $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$, $\boldsymbol{R}_i^{(\boldsymbol{b})}$, $\boldsymbol{T}_n$, and $\boldsymbol{V}_n$
Duong *et al.* [32]	$\boldsymbol{R}_{ij}^{(\boldsymbol{x})} = \sum_n r_{ij,n}\, \boldsymbol{R}_{i,n}$	Several types of $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ including rank-1 and full-rank matrices	$r_{ij,n}$ (w/o NMF)	EM for $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$
Ozerov *et al.* [29]	$\boldsymbol{R}_{ij}^{(\boldsymbol{x})} = \sum_k (\sum_n \boldsymbol{R}_{i,n}^{(\boldsymbol{s})} z_{nk}) t_{ik}\, v_{kj} + \boldsymbol{R}_i^{(\boldsymbol{b})}$ $(\boldsymbol{x}_{ij} = \boldsymbol{A}_i\, \boldsymbol{s}_{ij} + \boldsymbol{b}_{ij})$	Rank-1 matrix $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ and diagonal matrix $\boldsymbol{R}_i^{(\boldsymbol{b})}$	NMF with partitioning function	EM and MU for $\boldsymbol{A}_i$, $\boldsymbol{R}_i^{(\boldsymbol{b})}$, $\boldsymbol{Z}$, $\boldsymbol{T}$, and $\boldsymbol{V}$
Sawada *et al.* [30]	$\boldsymbol{R}_{ij}^{(\boldsymbol{x})} = \sum_k (\sum_n \boldsymbol{R}_{i,n}^{(\boldsymbol{s})} z_{nk}) t_{ik}\, v_{kj}$	Full-rank matrix $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$	NMF with partitioning function	MU for $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$, $\boldsymbol{Z}$, $\boldsymbol{T}$, and $\boldsymbol{V}$
Kitamura *et al.* [33]	$\boldsymbol{R}_{ij}^{(\boldsymbol{x})} = \sum_k (\sum_n \boldsymbol{R}_{i,n}^{(\boldsymbol{s})} z_{nk}) t_{ik}\, v_{kj}$ $(\boldsymbol{x}_{ij} = \boldsymbol{A}_i\, \boldsymbol{s}_{ij})$	Rank-1 matrix $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$	NMF with partitioning function	IP for $\boldsymbol{W} = \boldsymbol{A}^{-1}$ and MU for $\boldsymbol{Z}$, $\boldsymbol{T}$, and $\boldsymbol{V}$

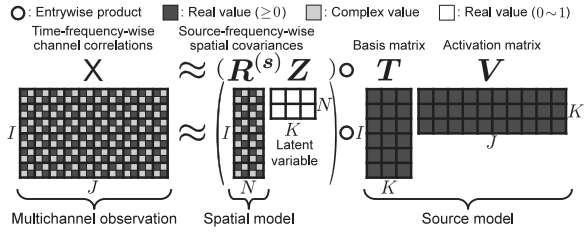

Fig. 3. Decomposition model of Sawada's MNMF ($I=6$, $J=10$, $M=N=2$, and $K=3$).

composition model of Sawada's MNMF is expressed as

$$\mathsf{X}_{ij} \approx \hat{\mathsf{X}}_{ij} = \sum_k \left(\sum_n \boldsymbol{R}_{i,n}^{(\boldsymbol{s})} z_{nk} \right) t_{ik} v_{kj}, \tag{16}$$

where $k=1,\ldots,K$ is the integral index of the bases and $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ is an $M \times M$ Hermitian positive definite matrix, which comprises the spatial covariance for each frequency i and source n. In addition, $z_{nk} (\in \mathbb{R}_{[0,1]})$ is a latent variable (partitioning function) that indicates whether the kth basis belongs to the nth source ($z_{nk}=1$) or not ($z_{nk}=0$) and satisfies $\sum_n z_{nk}=1$; t_{ik} and v_{kj} are the nonnegative elements of the basis matrix $\boldsymbol{T} (\in \mathbb{R}_{\geq 0}^{I \times K})$ and activation matrix $\boldsymbol{V} (\in \mathbb{R}_{\geq 0}^{K \times J})$, respectively.

Fig. 3 shows the decomposition model of Sawada's MNMF. This method decomposes the observed signal into $\boldsymbol{T}$ and $\boldsymbol{V}$ and simultaneously optimizes the spatial covariances for each source $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$. Then, the sources are separated by associating these variables $\boldsymbol{T}$ and $\boldsymbol{V}$ with $\boldsymbol{R}^{(\boldsymbol{s})}$ by using a partitioning function $\boldsymbol{Z} (\in \mathbb{R}_{\geq 0}^{N \times K})$.

The cost function based on Itakura–Saito divergence is defined as [30]

$$Q_{\mathrm{MNMF}} = \sum_{i,j} \left[\mathrm{tr} \left(\mathsf{X}_{ij} \hat{\mathsf{X}}_{ij}^{-1} \right) + \log \det \hat{\mathsf{X}}_{ij} \right], \tag{17}$$

where constant terms are omitted. Note that this cost function coincides with that of Ozerov's MNMF, which is a criterion of maximum log-likelihood. MU update rules based on the auxiliary function technique to minimize (17) were derived in [30].

In Sawada's MNMF, the spatial covariance matrix $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ is estimated as a full-rank matrix. This means that Sawada's MNMF has the capability to separate sources even if the mixing system cannot be represented as a rank-1 spatial model given by (4). However, Sawada's MNMF requires many iterations for optimization, and the separation performance strongly depends on the initial values of $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ and the NMF variables. This is because a large number of variables should be optimized and because there is no constraint for optimizing the full-rank spatial covariances.

III. PROPOSED METHOD

A. Motivation and Strategy

The separation performance of IVA is degraded for music signals because the source model defined from the multivariate distribution is not flexible. When we use a spherical multivariate distribution, IVA assumes that all the frequency bins have the same activations (time-varying gains). In contrast, MNMF can capture specific harmonic structures of the sources because it utilizes NMF decomposition to represent a power spectrogram $r_{ij,n}$. However, the optimization of the full-rank spatial covariance $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ is a difficult problem, and Sawada's MNMF lacks robustness. Ozerov's MNMF also suffers from sensitivity to the initial values even though it employs the rank-1 spatial covariance for each source.

In this paper, we derive a new and efficient algorithm by considering the determined situation ($M=N$). and the linear time-invariant mixing system (rank-1 spatial model) described by (4), which is similar to IVA. This is a special case of existing models. For example, our model is identical to Ozerov's MNMF in the determined situation when the noise component is zero and is identical to Sawada's MNMF in the determined situation when all the source covariance matrices are rank-1. However, the derived optimization algorithm is considerably different. In Ozerov's MNMF, the noise component $\boldsymbol{b}_{ij}$ is inherently necessary for optimization because the EM algorithm does not work (the update rule becomes $\boldsymbol{A}_i \leftarrow \boldsymbol{A}_i$) when $\boldsymbol{b}_{ij}=\boldsymbol{0}$, which was clearly addressed by the authors [27]. In contrast, in our approach, thanks to the assumption of the invertibility of $\boldsymbol{A}_i$, we can transform the spatial optimization in MNMF into an estimation of the demixing matrix $\boldsymbol{W}_i$. Therefore, the proposed method employs a flexible source model as in MNMF and

rapidly estimates the demixing matrix $\boldsymbol{W}_i$ in a stable manner as in auxiliary-function-based IVA (AuxIVA) [20]. Hereafter, the proposed method is referred to as *determined rank-1 MNMF*. Note that the proposed method cannot be applied to the underdetermined BSS problem because the mixing matrix $\boldsymbol{A}_i$ must be invertible in this approach.

Similarly to standard ICA or IVA, the proposed method is applicable for the overdetermined case ($M > N$) with dimensionality reduction using PCA. The authors have also proposed another method [35] for the overdetermined case when M is sufficiently larger than N such as when $M = 2N$ or $3N$.

B. Derivation of Cost Function

If we assume that the mixing system is represented by the mixing matrix $\boldsymbol{A}_i = (\boldsymbol{a}_{i,1} \dots \boldsymbol{a}_{i,N})$ appearing in (4), the spatial covariance for each source $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ can be modeled by a rank-1 matrix that is a product of the steering vector $\boldsymbol{a}_{i,n}$ as follows [27], [32]:

$$\boldsymbol{R}_{i,n}^{(\boldsymbol{s})} = \boldsymbol{a}_{i,n} \boldsymbol{a}_{i,n}^{\mathrm{h}}. \tag{18}$$

To introduce the rank-1 spatial model into Sawada's MNMF, we substitute (18) into (16) and reformulate $\hat{\mathsf{X}}_{ij}$ using the mixing matrix $\boldsymbol{A}_i$ as follows:

$$\begin{aligned} \hat{\mathsf{X}}_{ij} &= \sum_k \left(\sum_n \boldsymbol{a}_{i,n} \boldsymbol{a}_{i,n}^{\mathrm{h}} z_{nk} \right) t_{ik} v_{kj} \\ &= \sum_n \boldsymbol{a}_{i,n} \boldsymbol{a}_{i,n}^{\mathrm{h}} \sum_k z_{nk} t_{ik} v_{kj} \\ &= \boldsymbol{A}_i \mathsf{D}_{ij} \boldsymbol{A}_i^{\mathrm{h}}, \end{aligned} \tag{19}$$

where

$$\mathsf{D}_{ij} = \begin{pmatrix} d_{ij,1} & 0 & \cdots & 0 \\ 0 & d_{ij,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & d_{ij,N} \end{pmatrix}, \tag{20}$$

$$d_{ij,n} = \sum_k z_{nk} t_{ik} v_{kj}. \tag{21}$$

By substituting (19) into the cost function of Sawada's MNMF (17), we obtain

$$Q = \sum_{i,j} \left[\mathrm{tr} \left(\boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathrm{h}} \left(\boldsymbol{A}_i^{\mathrm{h}} \right)^{-1} \mathsf{D}_{ij}^{-1} \boldsymbol{A}_i^{-1} \right) + \log \det \boldsymbol{A}_i \mathsf{D}_{ij} \boldsymbol{A}_i^{\mathrm{h}} \right]. \tag{22}$$

In the determined situation (we let M equal N for simplicity), the demixing matrix $\boldsymbol{W}_i$ exists and we can transform the variables, i.e., the observed signal $\boldsymbol{x}_{ij}$ and the mixing matrix $\boldsymbol{A}_i$, to the separated signal $\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}$ and the demixing matrix $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1}$, respectively, as follows:

$$\begin{aligned} Q = & \sum_{i,j} \Big[\mathrm{tr} \left(\boldsymbol{W}_i^{-1} \boldsymbol{y}_{ij} \boldsymbol{y}_{ij}^{\mathrm{h}} \left(\boldsymbol{W}_i^{\mathrm{h}} \right)^{-1} \boldsymbol{W}_i^{\mathrm{h}} \mathsf{D}_{ij}^{-1} \boldsymbol{W}_i \right) \\ & + \log \left(\det \boldsymbol{A}_i \right) \left(\det \mathsf{D}_{ij} \right) \left(\det \boldsymbol{A}_i^{\mathrm{h}} \right) \Big] \\ = & \sum_{i,j} \Big[\mathrm{tr} \left(\boldsymbol{W}_i \boldsymbol{W}_i^{-1} \boldsymbol{y}_{ij} \boldsymbol{y}_{ij}^{\mathrm{h}} \left(\boldsymbol{W}_i^{\mathrm{h}} \right)^{-1} \boldsymbol{W}_i^{\mathrm{h}} \mathsf{D}_{ij}^{-1} \right) \\ & + 2 \log |\det \boldsymbol{A}_i| + \log \det \mathsf{D}_{ij} \Big] \\ = & \sum_{i,j} \left[\mathrm{tr} \left(\boldsymbol{y}_{ij} \boldsymbol{y}_{ij}^{\mathrm{h}} \mathsf{D}_{ij}^{-1} \right) - 2 \log |\det \boldsymbol{W}_i| + \sum_m \log d_{ij,m} \right] \\ = & \sum_{i,j} \left[\sum_m \frac{|y_{ij,m}|^2}{\sum_k z_{mk} t_{ik} v_{kj}} - 2 \log |\det \boldsymbol{W}_i| \right. \\ & \left. + \sum_m \log \sum_k z_{mk} t_{ik} v_{kj} \right], \end{aligned} \tag{23}$$

where $y_{ij,m} = \boldsymbol{w}_{i,m}^{\mathrm{h}} \boldsymbol{x}_{ij}$. In conventional MNMFs, the separated signal is obtained by clustering the source model $\boldsymbol{TV}$ into specific sources using the spatial covariance $\boldsymbol{R}^{(\boldsymbol{s})}$ and partitioning function $\boldsymbol{Z}$. The proposed method estimates the demixing matrix $\boldsymbol{W}_i$ to obtain the separated signal $\boldsymbol{y}_{ij}$, where we approximately decompose $|y_{ij,m}|^2$ into z_{mk}, t_{ik}, and v_{kj} as model spectrograms $r_{ij,m}$ for the sources in each iteration.

C. Relationship Between IVA and MNMF

The first and second terms in the cost function (23) are equivalent to the IVA cost function (9) except for the contrast function $G(\cdot)$, and the first and third terms in (23) are equivalent to a single-channel NMF cost function (11). Therefore, the proposed cost function is a superposition of those of IVA and NMF. This fact reveals the relationship between IVA and MNMF, namely, MNMF with a rank-1 spatial model, which assumes a linear time-invariant mixing system in the time-frequency domain, is essentially equivalent to IVA with a flexible source model using NMF decomposition. Therefore, the proposed determined rank-1 MNMF given by (23) can be considered as an intermediate model between IVA and MNMF with the full-rank spatial model in terms of the model flexibility. From the IVA side, we introduced the source model using NMF with K bases to capture the specific spectral patterns, and from the MNMF side, a rank-1 spatial model was introduced to transform the variable $\boldsymbol{A}_i$ into $\boldsymbol{W}_i$ and to make the optimization more efficient. However, the source priors of IVA and the proposed method are different. IVA generally assumes the spherical Laplace distribution, which has the same variance for all frequency bins, as the source prior by setting $G(\boldsymbol{y}_{j,m}) = \|\boldsymbol{y}_{j,m}\|_2$ in (9). The proposed method using (23) assumes independent complex Gaussian distributions in each time-frequency slot [34], similarly to conventional MNMFs. This issue will be discussed in Section IV.

The original cost function (23) is defined as the Itakura-Saito divergence between observation X_{ij} and estimation $\hat{\mathsf{X}}_{ij}$. Intriguingly, the proposed method utilizes the independence between sources to separate them because the IVA cost function appears. This is because minimizing the Itakura-Saito divergence im-

plicitly leads to the independence between sources, namely, we implicitly assume independent complex Gaussian distributions for each time and frequency slot, which model the mutually independent sources [34].

D. Update Rules

1) Spatial Model: For the optimization of ICA or IVA, update rules based on the auxiliary function technique have been proposed [19], [20], and it has been reported that these update rules are faster and more stable than those for a conventional update scheme (e.g., natural gradient method) and that the step size parameter can be omitted in each iteration. In particular, when the contrast function is $G=|y_{ij,m}|^2/r_{ij,m}$ ($r_{ij,m}$ is the variance of the complex Gaussian distribution), the optimization of $\boldsymbol{W}_i$ in (23) becomes equivalent to that in auxiliary-function-based IVA [20]. For this reason, the update rules of $\boldsymbol{W}_i$ can easily be derived as follows:

$$V_{i,m} = \frac{1}{J}\sum_j \frac{1}{r_{ij,m}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{h}}, \tag{24}$$

$$\boldsymbol{w}_{i,m} \leftarrow (\boldsymbol{W}_i V_{i,m})^{-1}\boldsymbol{e}_m, \tag{25}$$

$$\boldsymbol{w}_{i,m} \leftarrow \boldsymbol{w}_{i,m}\left(\boldsymbol{w}_{i,m}^{\mathrm{h}} V_{i,m}\boldsymbol{w}_{i,m}\right)^{-\frac{1}{2}}, \tag{26}$$

where $r_{ij,m}$ is the estimated variance of each source (model spectrogram for mth source [34]), and $\boldsymbol{e}_m$ denotes the unit vector with the mth element equal to unity. After the update of $\boldsymbol{W}_i$, the separated signal $\boldsymbol{y}_{ij}$ should be updated as

$$y_{ij,m} \leftarrow \boldsymbol{w}_{i,m}^{\mathrm{h}}\boldsymbol{x}_{ij}. \tag{27}$$

Note that the derivation of the update rules for $\boldsymbol{W}_i$ based on the auxiliary function technique is out of the scope in the contribution of this paper.

2) Source Model: Here, we propose two types of update rule for the source models, which are related to the presence of z_{mk}. If we eliminate the partitioning function z_{mk} in (23), the cost function (23) can be rewritten as follows:

$$Q_{\mathrm{mod}} = \sum_{i,j}\left[\sum_m \frac{|y_{ij,m}|^2}{\sum_l t_{il,m}v_{lj,m}} - 2\log|\det\boldsymbol{W}_i| + \sum_m \log\sum_l t_{il,m}v_{lj,m}\right], \tag{28}$$

where $t_{il,m}$ and $v_{lj,m}$ are the *sourcewise* bases and activations, respectively, and we assume that $z_{mk}\in\{0,1\}$ and all the sources are modeled by the same number of bases L (namely, $L\times M=K$). In this formulation, the differentials $\partial Q_{\mathrm{mod}}/\partial t_{il,m}$ and $\partial Q_{\mathrm{mod}}/\partial v_{lj,m}$ become identical to $\partial Q_{\mathrm{NMF}}/\partial t_{il}$ and $\partial Q_{\mathrm{NMF}}/\partial v_{lj}$, respectively. Therefore, for the modified cost function (28), the update rules of $t_{il,m}$ and $v_{lj,m}$ are the same as those of simple NMF (12) and (13), i.e.,

$$t_{il,m} \leftarrow t_{il,m}\sqrt{\frac{\sum_j |y_{ij,m}|^2 v_{lj,m}\left(\sum_{l'} t_{il',m}v_{l'j,m}\right)^{-2}}{\sum_j v_{lj,m}\left(\sum_{l'} t_{il',m}v_{l'j,m}\right)^{-1}}}, \tag{29}$$

$$v_{lj,m} \leftarrow v_{lj,m}\sqrt{\frac{\sum_i |y_{ij,m}|^2 t_{il,m}\left(\sum_{l'} t_{il',m}v_{l'j,m}\right)^{-2}}{\sum_i t_{il,m}\left(\sum_{l'} t_{il',m}v_{l'j,m}\right)^{-1}}}. \tag{30}$$

Also, the estimated source model (the variance of the complex Gaussian distribution) is represented as

$$r_{ij,m} = \sum_l t_{il,m}v_{lj,m}. \tag{31}$$

Alternatively, if we employ a continuous-valued z_{mk} to cluster K bases into M specific sources, we can derive the update rules of z_{mk}, t_{ik}, and v_{kj} by minimizing (23) by the auxiliary function technique. Here, we design an upper bound function of (23) as the auxiliary function. The first term in (23) is a convex function for the variables. Applying Jensen's inequality to this term with an auxiliary variable $\alpha_{ijk}\geq 0$ that satisfies $\sum_k \alpha_{ijk}=1$, we have

$$\frac{1}{\sum_k z_{mk}t_{ik}v_{kj}} \leq \sum_k \frac{\alpha_{ijk}^2}{z_{mk}t_{ik}v_{kj}}. \tag{32}$$

Also, the third term in (23) is a concave function, and we can apply the tangent line inequality to this term with an auxiliary variable $\beta_{ij}\geq 0$ as

$$\log\sum_k z_{mk}t_{ik}v_{kj} \leq \frac{1}{\beta_{ij}}\left(\sum_k z_{mk}t_{ik}v_{kj} - \beta_{ij}\right) + \log\beta_{ij}. \tag{33}$$

The equality of (32) and (33) holds if and only if the auxiliary variables are set as follows:

$$\alpha_{ijk} = \frac{z_{mk}t_{ik}v_{kj}}{\sum_{k'} z_{mk'}t_{ik'}v_{k'j}}, \tag{34}$$

$$\beta_{ij} = \sum_k z_{mk}t_{ik}v_{kj}. \tag{35}$$

Using these upper bounds, we can design the auxiliary function of (23) as

$$Q \leq Q^+ = \sum_{i,j}\left[\sum_{m,k}\frac{|y_{ij,m}|^2\alpha_{ijk}^2}{z_{mk}t_{ik}v_{kj}} - 2\log|\det\boldsymbol{W}_i| + \frac{1}{\beta_{ij}}\left(\sum_k z_{mk}t_{ik}v_{kj} - \beta_{ij}\right) + \log\beta_{ij}\right]. \tag{36}$$

The update rules for Q^+ with respect to each variable are determined by setting the gradient to zero. From $\partial Q^+/\partial z_{mk}=0$, we obtain

$$\sum_{i,j}\left[-\frac{|y_{ij,m}|^2\alpha_{ijk}^2}{z_{mk}^2 t_{ik}v_{kj}} + \frac{1}{\beta_{ij}}t_{ik}v_{kj}\right] = 0. \tag{37}$$

By transposing the first term in (37) to the right-hand side and multiplying both sides by z_{mk}^2, we have

$$z_{mk}^2 \sum_{i,j} \frac{1}{\beta_{ij}} t_{ik} v_{kj} = \sum_{i,j} \frac{|y_{ij,m}|^2 \alpha_{ijk}^2}{t_{ik} v_{kj}}. \tag{38}$$

Finally, the MU rule of z_{mk} can be derived by substituting (34) and (35) into (38) as follows:

$$z_{mk} \leftarrow z_{mk} \sqrt{\frac{\sum_{i,j} |y_{ij,m}|^2 t_{ik} v_{kj} \left(\sum_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-2}}{\sum_{i,j} t_{ik} v_{kj} \left(\sum_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-1}}}, \tag{39}$$

where we calculate $z_{mk} \leftarrow z_{mk} / \sum_{m'} z_{m'k}$ to ensure $\sum_m z_{mk} = 1$ at each iteration. Similarly to (39), the update rules of t_{ik} and v_{kj} are obtained as

$$t_{ik} \leftarrow t_{ik} \sqrt{\frac{\sum_{j,m} |y_{ij,m}|^2 z_{mk} v_{kj} \left(\sum_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-2}}{\sum_{j,m} z_{mk} v_{kj} \left(\sum_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-1}}}, \tag{40}$$

$$v_{kj} \leftarrow v_{kj} \sqrt{\frac{\sum_{i,m} |y_{ij,m}|^2 z_{mk} t_{ik} \left(\sum_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-2}}{\sum_{i,m} z_{mk} t_{ik} \left(\sum_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-1}}}, \tag{41}$$

and the estimated source model is represented as

$$r_{ij,m} = \sum_k z_{mk} t_{ik} v_{kj}. \tag{42}$$

3) Normalization: We can estimate all the variables that minimize (23) by iterating (24)–(27) and (29)–(31) or (39)–(42) alternately. Note that a scale ambiguity exists between $\boldsymbol{W}_i$ and $r_{ij,m}$ because both of them can determine the scale. Therefore, the estimated variance $r_{ij,m}$ has a risk of diverging. To avoid this problem, we normalize $\boldsymbol{W}_i$ and $r_{ij,m}$ at each iteration as follows:

$$\lambda_m = \sqrt{\frac{1}{IJ} \sum_{i,j} |y_{ij,m}|^2}, \tag{43}$$

$$\boldsymbol{w}_{i,m} \leftarrow \boldsymbol{w}_{i,m} \lambda_m^{-1}, \tag{44}$$

$$y_{ij,m} \leftarrow y_{ij,m} \lambda_m^{-1}, \tag{45}$$

$$r_{ij,m} \leftarrow r_{ij,m} \lambda_m^{-2}. \tag{46}$$

Note that these normalizations never change the value of the cost function (23). The signal scale can be restored by applying a back-projection technique [13] after the optimization.

IV. Experimental Analysis of Determined Rank-1 MNMF

In this section, we discuss the inherent difference between conventional ICA-based methods (IVA and FDICA) and determined rank-1 MNMF. In addition, we evaluate them via BSS experiments using artificial sources and show that determined rank-1 MNMF possesses better flexibility than IVA and FDICA for both the source and spatial models.

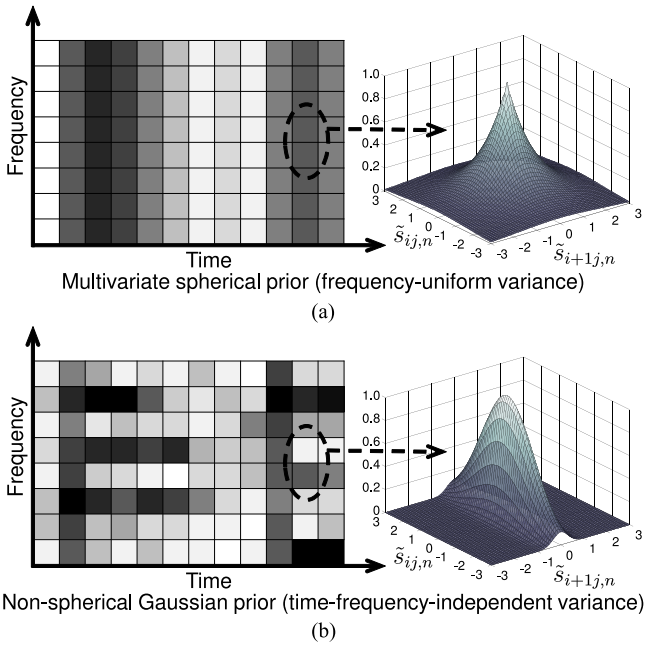

Fig. 4. Illustration of source models (model spectrograms) for one source in (a) IVA and (b) determined rank-1 MNMF, where grayscale of each time-frequency slot indicates value of variance and $\tilde{s}$ denotes only real or imaginary part of complex-valued component s. In IVA, multivariate spherical prior is assumed so that components with higher-order dependence are modeled as one source, and this method can be interpreted as NMF with only one spectral basis. In contrast, determined rank-1 MNMF can express more complex spectrogram using NMF with arbitrary number of spectral bases.

A. Difference Between Assumption in Source Model

In IVA, we generally introduce a spherical multivariate distribution to ensure the higher-order correlation between frequency bins, namely, it is assumed that all the frequency bins have the same activations (time-varying gains) as shown in Fig. 4(a). This is because the variances of all frequency bins are the same. This simple and nonflexible source model in IVA can be interpreted as a specific NMF that has only one frequency-uniform (flat) basis for each source. Therefore, the number of bases in the IVA model spectrogram always becomes one (see the spectrogram in Fig. 4(a)).

On the other hand, conventional MNMF and proposed determined rank-1 MNMF assume independent complex Gaussian distributions for each time-frequency slot [34], and their cost functions (17) and (23) are based on a log-determinant divergence, which is a matrix version of Itakura-Saito divergence. Therefore, the estimated variances $r_{ij,m}$ can explicitly express a model spectrogram via NMF decomposition with an arbitrary number of bases (see the spectrogram in Fig. 4(b)). For this reason, the source model in determined rank-1 MNMF is more flexible than that in IVA. In addition, IVA can be thought of as a special case of determined rank-1 MNMF. If we set the number of bases for each source to one, both methods are essentially equivalent except for the prior distributions.

For conventional FDICA, its source model depends on how the permutation problem is solved, and two well-known techniques for solving the permutation problem have been proposed. One technique is to utilize the correlation between frequency bins [13], which is an essentially equivalent approach to IVA.

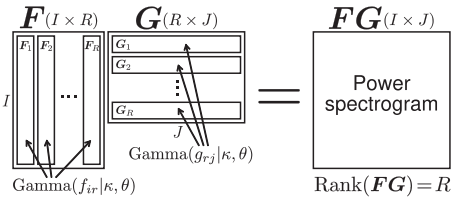

Fig. 5. Artificial source that consists of R bases.

The other well-known permutation solver is to estimate and utilize the direction of arrival (D.O.A.) of each source [12]. Hereafter, we refer to the combined method of FDICA and this permutation solver as *FDICA+DOA*. FDICA+DOA uses the estimated steering vectors (estimated spatial model), and there is no explicit source model except for non-Gaussianity in the time series for each frequency bin.

B. Difference Between Assumption in Spatial Model

In IVA and determined rank-1 MNMF, there is no explicit assumption in the spatial model (mixing system). Both methods only use the statistical independence between source models (model spectrograms) and the observed multichannel mixtures to estimate the demixing matrix.

In contrast, FDICA+DOA directly uses the difference between the estimated spatial conditions for each source to solve the permutation problem. Therefore, the separation performance of FDICA+DOA is sensitive to the spatial setup of the sources; if the positions of the sources become close or the reverberation becomes strong, the error of the permutation solver may increase. In summary, FDICA+DOA is severely affected by the spatial conditions rather than source modeling, whereas IVA and determined rank-1 MNMF are not.

C. Experimental Validation

We validate the difference between both the source and spatial models among IVA, FDICA+DOA, and determined rank-1 MNMF. In this validation, for simplicity, the numbers of sources and microphones are set to two, namely, $N = M = 2$.

1) Design of Artificial Spectrograms With R Bases: From the difference between the source models of IVA and determined rank-1 MNMF, we can expect that the number of bases in the power spectrogram will affect the separation performance for IVA. If the power spectrogram of each source consists of only one basis, both IVA and determined rank-1 MNMF can separate the sources with high accuracy. However, if the sources have more complicated power spectrograms, the source model in IVA cannot represent them in principle, and the separation performance may decrease.

To investigate this issue, in this experiment, we produce artificial sources whose power spectrograms can be represented by R bases. Fig. 5 shows the power spectrogram that we produced. To simulate a nonnegative sparse spectrogram, we generate nonnegative random values f_{ir} and g_{rj} that obey independent and identically distributed (i.i.d.) gamma distributions, where $r = 1, \ldots, R$ is the integral index of the basis in matrices $\boldsymbol{F}$ and $\boldsymbol{G}$. The power spectrogram is a product of $\boldsymbol{F}$ and $\boldsymbol{G}$ and its size is $I \times J$. The gamma distribution can be represented as

$$\mathrm{Gamma}(\chi|\kappa,\theta) = \chi^{\kappa-1}\frac{e^{-\chi/\theta}}{\Gamma(\kappa)\theta^{\kappa}}, \tag{47}$$

where χ is a random variable, and κ and θ are shape and scale parameters, respectively. After producing the power spectrogram $\boldsymbol{FG}$, we add random phases that obey a uniform distribution in the range $[0, 2\pi]$ to $\boldsymbol{FG}$, and the produced complex spectrogram ($\boldsymbol{FG}$ with random phases) is used as an artificial source whose power spectrogram has R bases. Therefore, in this procedure, we simulate the variances of complex Gaussian distributions [34] with an outer product of variables that obey i.i.d. gamma distributions and their linear combination.

In this artificial source, it is important to set κ to an appropriate value. For example, when we set κ to a constant value regardless of R, the random values in the power spectrogram $\boldsymbol{FG}$ become close to a Gaussian distribution. This is because the kurtosis of the element $\sum_{r=1}^{R} f_{ir}g_{rj}$ in $\boldsymbol{FG}$ converges to three by the central limit theorem when R increases. For this reason, the separation accuracy of ICA-based methods decreases as R increases. To avoid this influence, we adjust the shape parameter κ for each value of R so that the kurtosis of $\boldsymbol{FG}$ is always the same value regardless of R. Such a κ can be derived using the moment-cumulant transform [36] (see Appendix). The following equation gives the shape parameter κ used to adjust the kurtosis of $\boldsymbol{FG}$:

$$\frac{\zeta(\kappa,R)}{\xi(\kappa,R)} - \mathrm{kurt} = 0, \tag{48}$$

where kurt is the intended value for the kurtosis of $\boldsymbol{FG}$ and

$$\begin{aligned}\zeta(\kappa,R) = {} & 84\kappa^3 + 174\kappa^2 + 132\kappa + 36 \\ & + R\left(52\kappa^4 + 60\kappa^3 + 19\kappa^2\right) \\ & + R^2\left(12\kappa^5 + 6\kappa^4\right) + R^3\kappa^6,\end{aligned} \tag{49}$$

$$\begin{aligned}\xi(\kappa,R) = {} & R\left(4\kappa^4 + 4\kappa^3 + \kappa^2\right) + R^2\left(4\kappa^5 + 2\kappa^4\right) \\ & + R^3\kappa^6.\end{aligned} \tag{50}$$

Since no closed-form solution exists that satisfies (48), we calculate the optimal κ by a greedy search. Table II shows the estimated shape parameter values when $\mathrm{kurt} = 50$. We experimentally confirmed that the kurtosis of the produced power

TABLE II
ESTIMATED VALUES OF SHAPE PARAMETER κ SO THAT KURTOSIS OF $\boldsymbol{FG}$ IS ADJUSTED TO 50 FOR EACH R

Number of bases R	Shape parameter κ
1	0.83809
2	0.54962
3	0.43450
4	0.36929
5	0.32617
6	0.29504
7	0.27124
8	0.25231

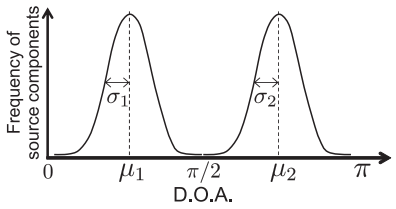

Fig. 6. Artificial DOA with Gaussian distributions.

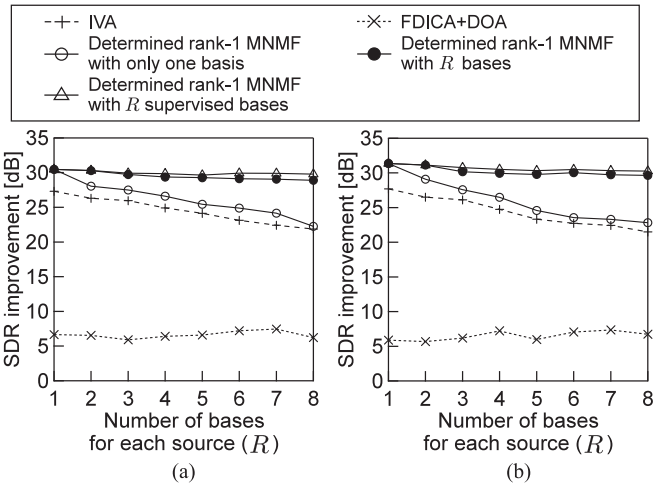

Fig. 7. Separation results of (a) source 1 and (b) source 2 for various numbers of bases.

spectrogram $\boldsymbol{FG}$ is always controlled to be approximately 50 using these shape parameters.

2) Design of Artificial Mixing Systems: For a mixing system, we designed an artificial D.O.A. that consists of $N=2$ Gaussian distributions, as shown in Fig. 6, where μ_n and σ_n are the mean value (position of the source) and standard deviation of the nth source, respectively. This modeling mimics an actual acoustical phenomenon in which the D.O.A.s of wavefronts in different frequencies i are randomly distributed owing to the room reverberation effect. We produced steering vectors $\boldsymbol{a}_{i,n}$ that obey the Gaussian distributions in Fig. 6 and prepared an artificial mixing matrix $\boldsymbol{A}_i$. Finally, we produced an artificial observed signal $\boldsymbol{x}_{ij}$ with artificial sources and an artificial mixing system using (4).

3) Experiment on Variational Artificial Spectrogram: In this experiment, we assume the following conditions: $I=J=257$, $\mathrm{kurt}=50$, $\theta=1$, $\mu_1=5\pi/12$, $\mu_2=7\pi/12$, $\sigma_1^2=\sigma_2^2=0.05$, and the interelement spacing of microphones is set to 5.66 cm. Fig. 7 shows the improvement of the signal-to-distortion ratio (SDR) [37] for various numbers of bases R, where SDR indicates the total separation performance and the improvement in the SDR is the increment from the SDR value of the observed signal. For determined rank-1 MNMF, we use a simple formulation without the partitioning function (the cost function (28) optimized by (29) and (30)). Also, we evaluate three patterns, namely, the case of $L=1$ (determined rank-1 MNMF with only one basis), the case of a suitable number of spectral bases $L=R$ (determined rank-1 MNMF with R bases), and the case of a supervised source model by setting $\boldsymbol{T}=\boldsymbol{F}$ and $\boldsymbol{V}=\boldsymbol{G}$ (determined rank-1 MNMF with R supervised bases). The difference between IVA and determined rank-1 MNMF with only one basis is simply the type of assumed distribution; IVA assumes the spherical Laplace distribution and determined rank-1 MNMF assumes time-frequency-independent Gaussian distributions. From Fig. 7, the separation scores of IVA and determined rank-1 MNMF with only one basis decrease when the number of bases of each source, R, increases because they cannot capture the exact power spectrograms. In contrast, determined rank-1 MNMF with R bases can maintain high SDR values because the power spectrogram of each source can be represented by a model spectrogram using R spectral bases in $\boldsymbol{T}$. This clearly demonstrates the flexibility of the source model in determined rank-1 MNMF.

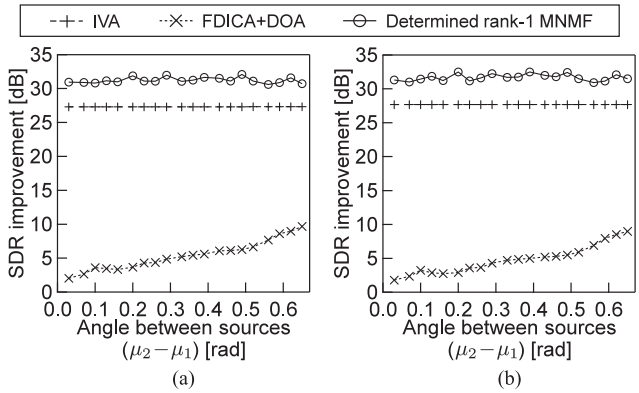

Fig. 8. Separation results of (a) source 1 and (b) source 2 for various angles.

4) Experiment on Variational Artificial Mixing Systems: From the difference between the spatial models in FDICA+DOA and determined rank-1 MNMF, we can expect that the mixing system (spatial conditions of each source) will affect the separation performance for FDICA+DOA. If the source positions are close or the variance of the D.O.A.s is large, a large error of D.O.A. clustering occurs in FDICA+DOA, resulting in marked degradation of the separation. However, since IVA and determined rank-1 MNMF do not use the explicit properties of the mixing condition (spatial model), we can expect that their separation performance will not strongly depend on the source positions or the variance of the D.O.A.s. To investigate this issue, in this experiment, we produce observed signals with various mixing conditions and evaluate the separation performance. We use the artificial sources described in Section IV-C1, where the power spectrograms of these sources are generated with $\mathrm{kurt}=50$ and $R=1$. The mixing system is produced by the artificial D.O.A. shown in Fig. 6 with various μ_1, μ_2, σ_1^2, and σ_2^2. Note that the experiment in which σ_1^2 and σ_2^2 are changed does not simulate a change in the reverberation time. It only controls the variance of the D.O.A.s over the frequencies, and the length of the impulse response does not change. Therefore, even when using larger σ_1^2 and σ_2^2, the rank-1 spatial model is always valid in this simulation. For determined rank-1 MNMF, the number of bases L is set to one, which is equal to R. The other conditions are the same as those in Section IV-C1.

Fig. 8 shows the SDR results for various positions of the sources (μ_1 and μ_2), where the horizontal axis indicates the angle between the two sources, $\mu_2-\mu_1$, and the variances are fixed to $\sigma_1^2=\sigma_2^2=0.05$. Also, Fig. 9 shows the SDR results for various variances (σ_1^2 and σ_2^2), where σ_1^2 and σ_2^2 are always set to the same value and the positions of the sources are fixed to

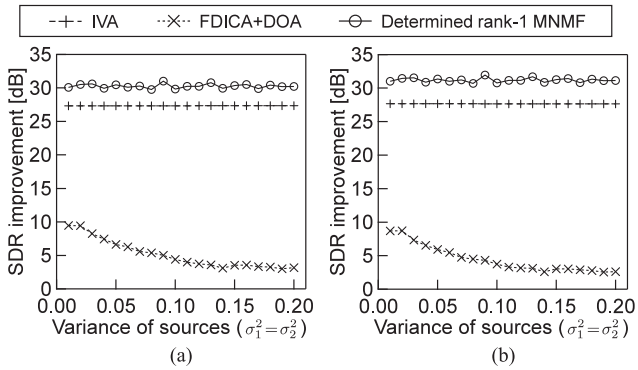

Fig. 9. Separation results of (a) source 1 and (b) source 2 for various variances.

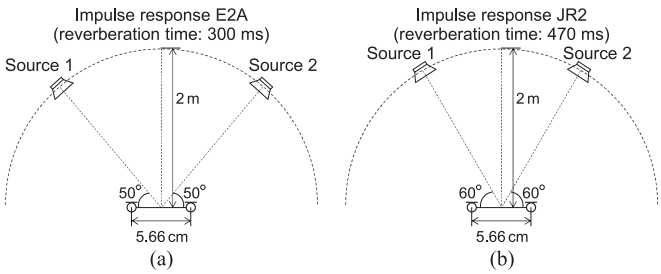

Fig. 10. Recording condition of impulse responses (a) E2A and (b) JR2 for two-source case.

$\mu_1 = 5\pi/12$ and $\mu_2 = 7\pi/12$. From these results, we can confirm that the separation performance of FDICA+DOA is sensitive to the mixing system. In particular, when the source positions become close (around 0.0 on the horizontal axis in Fig. 8) or the variance of the D.O.A.s. is large (around 0.20 on the horizontal axis in Fig. 9), the permutation solver using the D.O.A. cannot cluster the sources correctly, resulting in large permutation errors. In contrast, IVA and determined rank-1 MNMF achieve good performance regardless of the mixing system because these methods do not have explicit spatial constraints. This shows the flexibility of the spatial model in determined rank-1 MNMF.

V. Experiments on Speech and Music Separation

A. Datasets

In this section, we confirm the efficacy of the proposed method for separating mixture signals. In this experiment, we investigated two cases: speech signal and music signal cases. In the speech signal case, we used live recorded mixture signals obtained from an underdetermined BSS task in SiSEC2011 [38]. This dataset includes 12 mixture signals (*dev1* and *dev2* datasets) with female and male speech, where the reverberation time is 130 ms/250 ms and the microphone spacing is 1 m/5 cm. Details of the other conditions for this dataset can be found in [38]. Note that since this dataset is for underdetermined BSS, three sources ($M=3$) are provided as stereo recordings ($N=2$). In this experiment, we used only the first and second speech sources to make the task determined ($M=N=2$). In the music signal case, the observed signals were produced by convoluting the impulse response *E2A* or *JR2*, which was obtained from the RWCP database [39], with each source. Fig. 10 shows the recording conditions of impulse responses E2A and JR2. As the music sources, we used professionally produced music obtained from a music separation task in SiSEC2011. The titles of the music and the instruments used are shown in Table III.

TABLE III
Music Sources for Two-Source Case

ID	Song name	Source (1/2)
1	bearlin-roads	acoustic_guit_main/vocals
2	another_dreamer-the_ones_we_love	guitar/vocals
3	fort_minor-remember_the_name	violins_synth/vocals
4	ultimate_nz_tour	guitar/synth

TABLE IV
Experimental Conditions

Sampling frequency	16 kHz
FFT length	256 ms in speech signal case and 512 ms in music signal case
Window shift length	128 ms in both speech and music signal cases
Initialization	$\boldsymbol{W}_i$: identity matrix NMF variables: uniform random values $[0, 1]$
Number of iterations	200

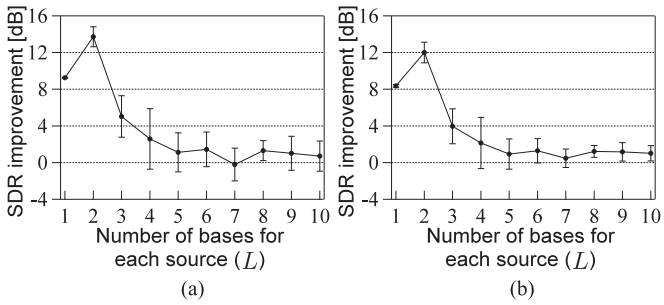

Fig. 11. Average SDR improvements for female speech (dev1) with 1 m microphone spacing and 130 ms reverberation time: (a) first speaker and (b) second speaker.

B. Experimental Analysis of Optimal Number of Bases

In this section, we give an experimental analysis of the optimal number of bases in determined rank-1 MNMF. Since NMF decomposition is more suitable for music than speech because of the stable pitch of instruments, we expect that the optimal number of bases will be different between them. For this reason, we evaluated the separation performance of determined rank-1 MNMF named *Proposed method w/o partitioning function* (updated using (24)–(27) and (29)–(31)) with various numbers of bases for each source, where this method models all the sources with the same fixed number of bases L. The experimental conditions used are shown in Table IV. As the evaluation score, we used the SDR improvement.

Figs. 11 and 12 show the average SDR improvements and their deviations in 10 trials with different various pseudorandom seeds, where the speech signal (Fig. 11) is a female speech from the dev1 dataset with 130 ms reverberation time and 1 m microphone spacing, and the music signal (Fig. 12) is song ID4 with impulse response E2A. From these results, we confirm that determined rank-1 MNMF cannot achieve a good

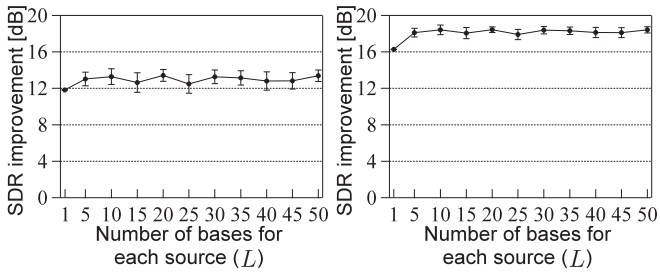

Fig. 12. Average SDR improvements for song ID4 with impulse response E2A: (a) guitar and (b) synth.

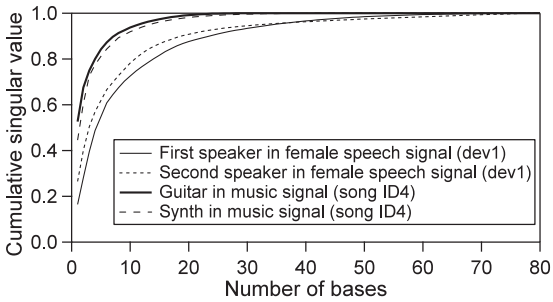

Fig. 13. Cumulative singular values of each source spectrogram in female speech (dev1) and music (song ID4), where all sources have the same length.

separation performance for speech signals when the number of bases is large. This is due to the structural complexity of the speech spectrogram. Fig. 13 shows cumulative singular values of each source spectrogram in the speech and music signals. The speech sources require more than 50 bases to represent the spectrogram while the music sources are saturated with 25 bases. Because of the time-varying pitch, it is difficult to capture speech spectrograms using NMF decomposition. If determined rank-1 MNMF fails to capture the correct spectrogram of each speech in the optimization, the demixing matrix will be trapped into a poor solution (local minimum). On the other hand, owing to the low-rankness of music spectrograms, determined rank-1 MNMF gives a better performance for music separation even if the number of bases increases.

C. Comparison of Separation Performance

1) Experimental Conditions: We compare the separation performance of eight methods, namely, *directional clustering* [40], *IVA*, *Ozerov's MNMF*, *Ozerov's MNMF with random initialization*, *Sawada's MNMF*, *Proposed method w/o partitioning function*, *Proposed method with partitioning function* (updated using (24)–(27) and (39)–(42)), and *Sawada's MNMF initialized by proposed method*. Directional clustering is a simple separation technique, which clusters all the STFT coefficients into specific sources using both powers and phases. In this experiment, we use k-means clustering in directional clustering, which corresponds to a double-disjoint assumption, namely, we assume that each time-frequency slot has only one source component. In Ozerov's MNMF, we used the experimental conditions described in [27], as shown in Table V, where the mixing matrices and the source models are initialized by estimation using Soft-LOST [41] with the permutation solver [15]. Also,

TABLE V
EXPERIMENTAL CONDITIONS USED IN OZEROV'S MNMF

Sampling frequency	16 kHz
FFT length	128 ms
Window shift length	64 ms
Number of bases	10 bases for each speech source and 4 bases for each music source
Initialization of mixing matrices	Mixing matrices estimated by Soft-LOST [41] and permutation solver [15]
Initialization of source models (NMF variables)	Pretrained bases and activations using simple NMF based on Kullback-Leibler divergence with sources estimated by Soft-LOST and [15]
Annealing for EM algorithm	Annealing with noise injection proposed in [27]
Number of iterations	500

Ozerov's MNMF with random initialization has the same conditions as Ozerov's MNMF except for the initialization, namely, the mixing matrices and the source models are initialized by the identity matrix and the uniform random values $[0, 1]$, respectively. In the other methods, the experimental conditions shown in Table IV were used. In Proposed method with partitioning function, we only set the total number of bases, K, and the sources are flexibly modeled with the optimal number of bases using the partitioning function $\boldsymbol{Z}$. Sawada's MNMF initialized by proposed method has the same algorithm as Sawada's MNMF, but the initial values of the spatial covariance matrix $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ are given by (18), where the steering vector $\boldsymbol{a}_{i,n}$ is calculated from the inverse of the demixing matrix $\boldsymbol{W}_i$ estimated by Proposed method w/o partitioning function.

On the basis of the results in Section V-B, we set the number of bases of each source to $L=2$ for the speech signals and $L=30$ for the music signals in Proposed method w/o partitioning function. In Proposed method with partitioning function and Sawada's MNMF, we set the total number of bases to $K=2\times N$ for the speech signals and $K=30\times N$ for the music signals. The number of bases used in Ozerov's MNMF is shown in Table V.

2) Results: Figs. 14 and 15 show examples of results for speech signals given by the average SDR improvements and their deviations in 10 trials with different pseudorandom seeds. Also, Figs. 16 and 17 show examples of results for music signals. The total average scores are shown in Tables VI and VII. From these results, we confirm that directional clustering cannot separate the sources because of the imperfect double-disjoint assumption and the deviation of the D.O.A.s in reverberant environments. Also, IVA cannot achieve satisfactory separation because the source model in IVA is not flexible as described in Section IV-A. Ozerov's MNMF outperforms IVA for the music signals, but the separation performance for speech signals is inferior to that of IVA. In addition, Ozerov's MNMF with random initialization cannot solve the BSS problem. This method must be initialized by other methods to find a good solution. The results of Sawada's MNMF have large error bars, namely, this method is also sensitive to initial values. However, for the music signals, Sawada's MNMF gives better performance than IVA and Ozerov's MNMF. The proposed methods achieve a high and stable performance. For the speech signals, Proposed method w/o partitioning function is preferable to Proposed method with partitioning function. This might be due to the sensitiv-

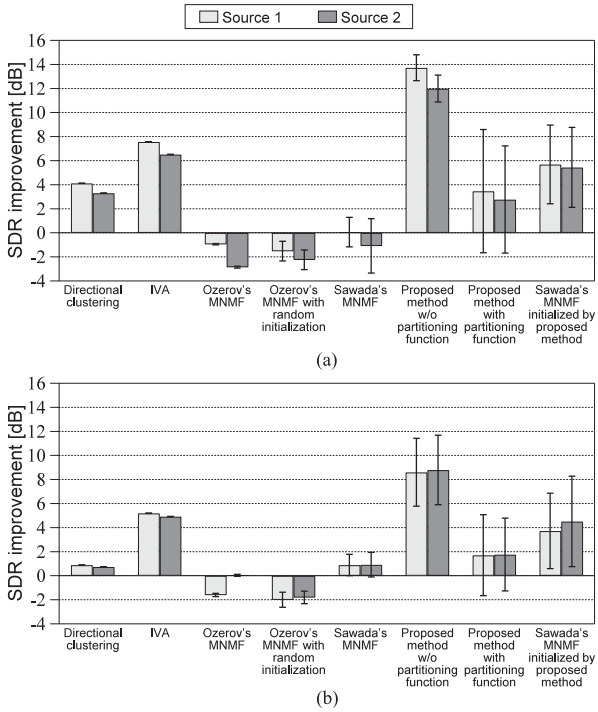

Fig. 14. Average SDR improvements for female speech (dev1) with 1 m microphone spacing, where reverberation time is (a) 130 ms and (b) 250 ms.

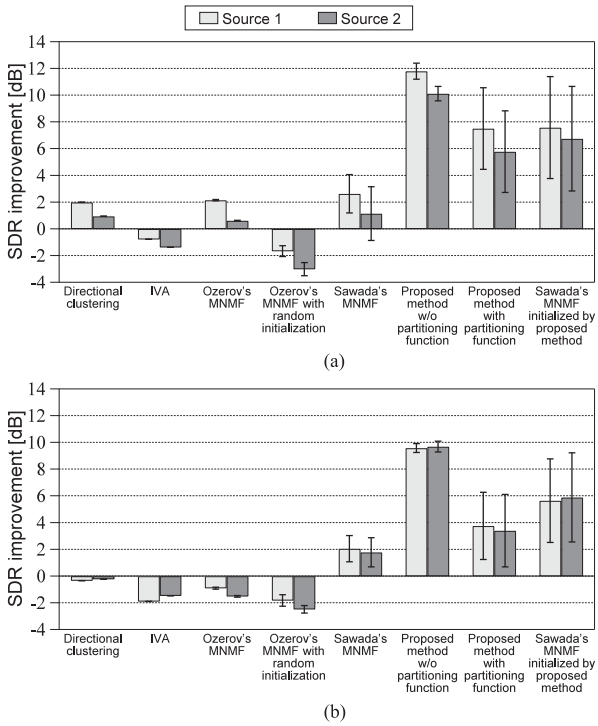

Fig. 15. Average SDR improvements for male speech (dev1) with 1 m microphone spacing, where reverberation time is (a) 130 ms and (b) 250 ms.

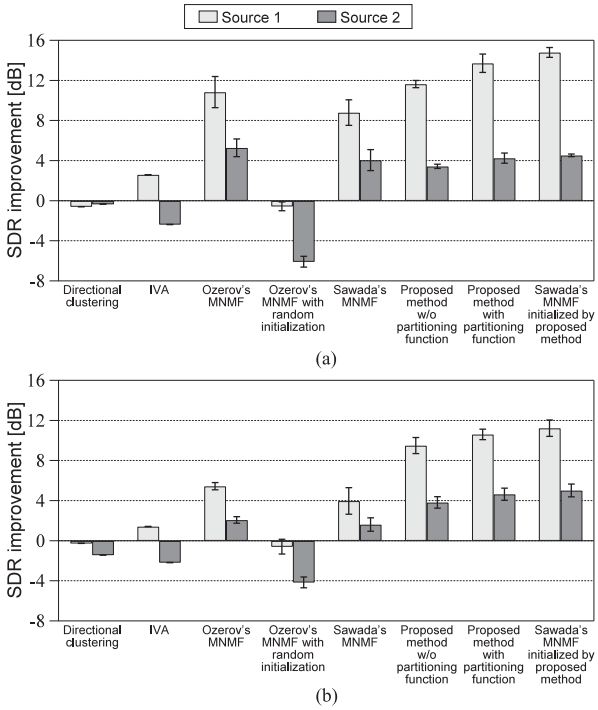

Fig. 16. Average SDR improvements for music signal song ID3 with impulse response (a) E2A and (b) JR2.

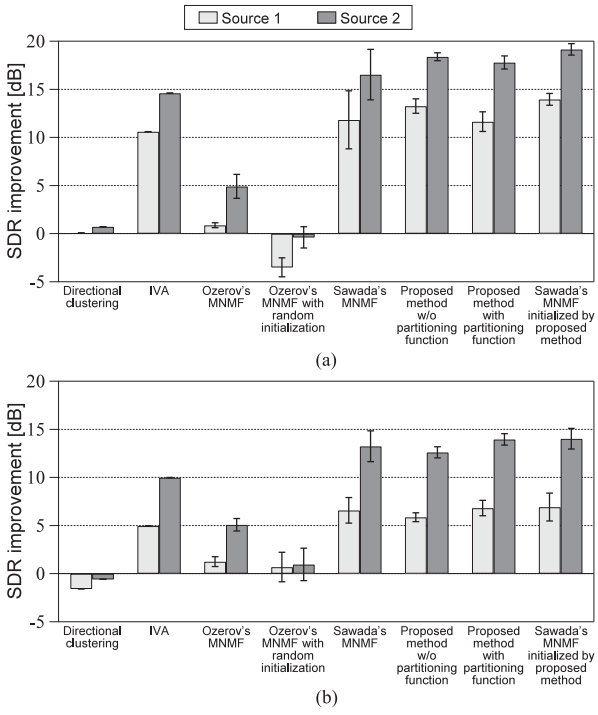

Fig. 17. Average SDR improvements for music signal song ID4 with impulse response (a) E2A and (b) JR2.

ity of the performance to the number of bases, as discussed in Section V-B. In contrast, for the music signals, Proposed method with partitioning function exhibits slightly higher performance than Proposed method w/o partitioning function. This improvement is achieved by modeling the sources with the optimal number of bases using the partitioning function z_{mk}. Fig. 18 shows an example of the convergence of the partitioning function z_{1k} from $k=1$ to $k=K$ in the music signal case. These values indicate whether the kth basis contributes to only source one ($z_{1k}=1$) or only source two ($z_{1k}=0$). We can confirm that almost all the partitioning functions converge to one or zero and that all the sources are effectively modeled with the optimal number of bases.

TABLE VI
AVERAGED SDR IMPROVEMENTS OVER VARIOUS SPEECH SIGNALS AND SOURCES WITH SAME RECORDING CONDITIONS IN TWO-SOURCE CASE

Recording conditions (rev. time and mic. spacing)	Directional clustering	IVA	Ozerov's MNMF	Ozerov's MNMF with random initialization	Sawada's MNMF	Proposed method w/o partitioning function	Proposed method with partitioning function	Sawada's MNMF initialized by proposed method
130 ms and 1 m	2.59	2.98	1.35	−2.11	0.68	**11.91**	4.88	6.36
130 ms and 5 cm	−1.51	2.86	2.13	−0.22	1.13	**8.97**	3.48	5.60
250 ms and 1 m	0.14	2.03	0.49	−2.02	0.48	**7.34**	2.09	4.19
250 ms and 5 cm	−1.56	2.43	0.91	−1.06	0.47	**6.43**	1.91	3.95

TABLE VII
AVERAGED SDR IMPROVEMENTS OVER VARIOUS MUSIC SIGNALS AND SOURCES WITH SAME IMPULSE RESPONSE IN TWO-SOURCE CASE

Impulse response	Directional clustering	IVA	Ozerov's MNMF	Ozerov's MNMF with random initialization	Sawada's MNMF	Proposed method w/o partitioning function	Proposed method with partitioning function	Sawada's MNMF initialized by proposed method
E2A	−0.73	5.72	5.73	−2.70	10.32	12.29	12.29	**14.41**
JR2	−1.18	1.77	2.37	0.75	6.11	6.62	7.40	**9.06**

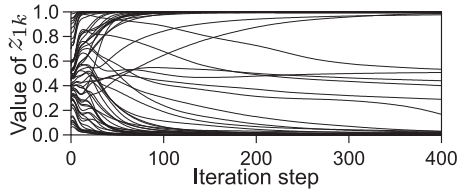

Fig. 18. Convergence of z_{1k} from $k=1$ to $k=K$ in music signal case.

The deviations of the proposed methods are smaller than those of Ozerov's and Sawada's MNMFs, which is particularly evident in Proposed method w/o partitioning function. This is because the optimization of the demixing matrix using the IVA update rules results in a stable separation performance. In fact, we experimentally confirmed that the initialization using Soft-LOST [41] and the permutation solver [15], which was employed in Ozerov's MNMF, did not improve the separation performance of Proposed method w/o partitioning function. This fact means that the proposed method is robust against the initial values. For music signals with impulse response JR2 (Figs. 16(b) and 17(b)), the SDRs of the proposed methods are markedly degraded compared with those with impulse response E2A because the reverberation time is longer than impulse response E2A and is close to the length of the window function in the STFT. Even if Sawada's MNMF has the potential to model such a mixing system by employing a full-rank spatial model, it is a very difficult problem to find the optimal $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$. However, Sawada's MNMF initialized by proposed method can achieve high and very stable separation performance even with impulse response JR2. This means that the demixing matrix estimated by the proposed methods can be a good initial value of the spatial model $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ in order to find the full-rank spatial covariance.

Fig. 19 shows an example of the SDR convergence for each method in music signal case. Both IVA and the proposed methods show much faster convergence than Sawada's MNMF. Also, the numbers of required iterations in Sawada's MNMF is greatly reduced by the initialization of the rank-1 spatial covariance.

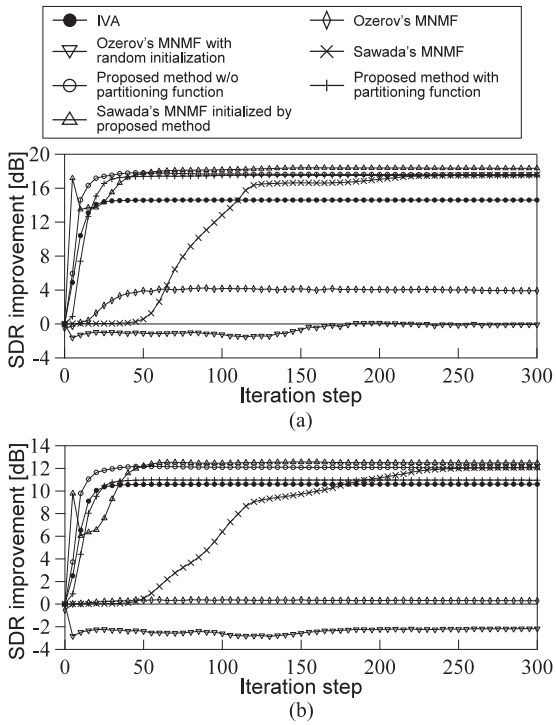

Fig. 19. SDR convergence for music signal song ID4 with impulse response E2A: (a) guitar and (b) synth.

This result shows the difficulty of optimizing the full-rank spatial covariance $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$.

D. Experiments on Three-Source Case With Music Signals

We also conducted an experiment involving three sources and three microphones ($M=N=3$) with music signals. Similarly to the music dataset described in Section V-A, we produced the observed signals using the same songs and the three instruments shown in Table VIII with the impulse responses shown in Fig. 20. The experimental conditions are those in Table IV, where we

TABLE VIII
MUSIC SOURCES FOR THREE-SOURCE CASE

ID	Song name	Source (1/2/3)
1	bearlin-roads	acoustic_guit_main/bass/vocals
2	another_dreamer-the_ones_we_love	drums/guitar/vocals
3	fort_minor-remember_the_name	drums/violins_synth/vocals
4	ultimate_nz_tour	guitar/synth/vocals

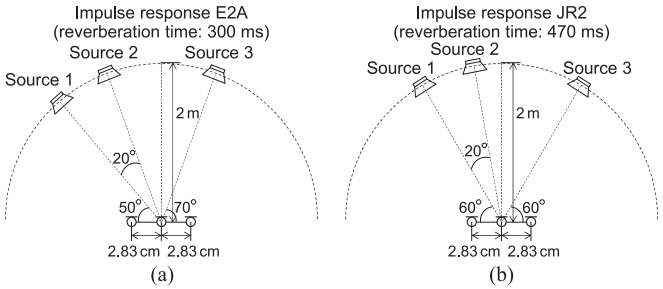

Fig. 20. Recording condition of impulse responses (a) E2A and (b) JR2 for three-source case.

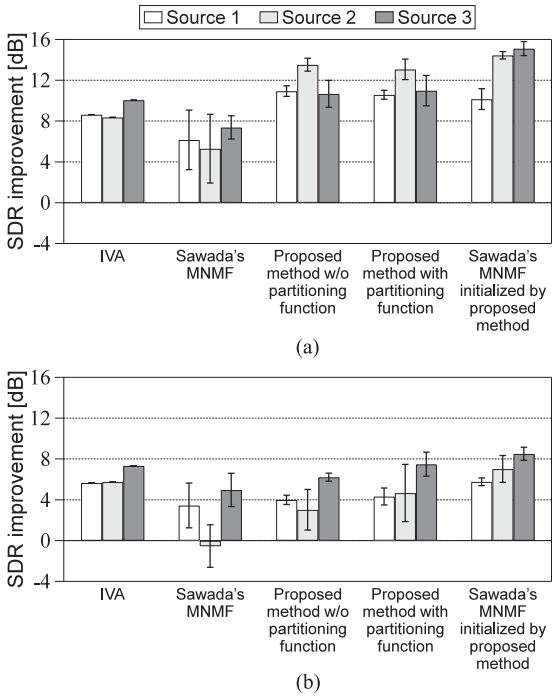

Fig. 21. Average SDR improvements for music signal song ID4 in three-source case with impulse response (a) E2A and (b) JR2.

here omit the results of directional clustering, Ozerov's MNMF, and Ozerov's MNMF with random initialization.

Fig. 21 shows examples of results, and Table IX shows the total average scores in the three-source case. Similarly to the previous results, the proposed method achieves better and more stable performance than Sawada's MNMF, and the spatial model estimated by the proposed method provides an efficient initialization for Sawada's MNMF. Table X shows the actual computational time for each method in the three-source case, where the calculations were performed using MATLAB 8.3 (64-bit)

TABLE IX
AVERAGED SDR IMPROVEMENTS OVER VARIOUS MUSIC SIGNALS AND SOURCES WITH SAME IMPULSE RESPONSE IN THREE-SOURCE CASE

Impulse response	IVA	Sawada's MNMF	Proposed method w/o partitioning function	Proposed method with partitioning function	Sawada's MNMF initialized by proposed method
E2A	3.86	7.77	8.03	6.18	**9.44**
JR2	2.81	4.44	5.03	4.11	**7.00**

TABLE X
COMPUTATIONAL TIMES (S) FOR SEPARATION OF SONG ID1 WITH IMPULSE RESPONSE E2A IN THREE-SOURCE CASE

IVA	Sawada's MNMF	Proposed method w/o partitioning function	Proposed method with partitioning function
91.6	4498.4	121.0	173.4

with an Intel Core i7-4790 (3.60 GHz) CPU. The computational times of the proposed methods are less than twice that of IVA. Sawada's MNMF requires a longer computational time because the eigenvalue decomposition of a $2M \times 2M$ matrix is required for each update iteration of $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$. From these results, the proposed methods are advantageous in terms of the convergence speed and computational cost while maintaining comparable separation performance with Sawada's MNMF.

VI. CONCLUSION

This paper proposes a new determined BSS technique that estimates a spatial model using IVA and a source model by low-rank decomposition using NMF. Also, the relationship between conventional MNMF and IVA is revealed: the proposed method is equivalent to Sawada's MNMF employing rank-1 modeling of the spatial covariance matrix, and IVA can be thought of as a special case of the proposed method, namely, the proposed method can be thought of as IVA with increased flexibility of the model. The proposed method can be optimized using the fast update rules of IVA and single-channel NMF based on the auxiliary function technique. The experimental results show that the proposed method achieves faster convergence and better results than the conventional BSS techniques.

APPENDIX
DERIVATION OF SHAPE PARAMETER FOR ARTIFICIAL RANDOM SPECTROGRAM WITH CONSTANT KURTOSIS

To produce an artificial random spectrogram $\boldsymbol{FG}$ with constant kurtosis, we derive the optimal shape parameter κ for each value of R. Hereafter, we denote an $I \times J$ matrix whose elements are $f_{ir}g_{rj}$ as $\boldsymbol{F}_r\boldsymbol{G}_r$, namely, $\boldsymbol{FG} = \sum_r \boldsymbol{F}_r\boldsymbol{G}_r$. Also, we denote a pth-order moment and pth-order cumulant of $\boldsymbol{F}_r\boldsymbol{G}_r$ as μ_{pr} and c_{pr} and those of $\boldsymbol{FG}$ as μ'_p and c'_p, respectively. When R increases beyond one, the matrix $\boldsymbol{FG}$ becomes a linear combination expressed as $\sum_{r=1}^{R} \boldsymbol{F}_r\boldsymbol{G}_r$. Therefore, the kurtosis of $\boldsymbol{FG}$ can be derived via the moment-cumulant transform [36]. Since f_{ir} and g_{rj} are generated from i.i.d. gamma distributions, μ_{pr} is equal to the product of pth-order moments of $\boldsymbol{F}_r$ and $\boldsymbol{G}_r$

as follows:

$$\mu_{pr} = \theta^{2p} \prod_{q=0}^{p-1} (\kappa + q)^2. \tag{51}$$

By the moment-cumulant transform, c_{pr} from $p=1$ to $p=4$ can be represented as

$$c_{1r} = \mu_{1r}, \tag{52}$$

$$c_{2r} = \mu_{2r} - \mu_{1r}^2, \tag{53}$$

$$c_{3r} = \mu_{3r} - 3\mu_{1r}\mu_{2r} + 2\mu_{1r}^3, \tag{54}$$

$$c_{4r} = \mu_{4r} - 4\mu_{1r}\mu_{3r} - 3\mu_{2r}^2 + 12\mu_{1r}^2\mu_{2r} - 6\mu_{1r}^4. \tag{55}$$

Since a cumulant satisfies additivity for the variables, the cumulant of $\boldsymbol{FG}$ is easily derived as follows:

$$c'_p = \sum_{r=1}^{R} c_{pr} = R c_{pr}. \tag{56}$$

The moments of $\boldsymbol{FG}$ for $p=2$ and $p=4$ are given by the moment-cumulant transform as

$$\mu'_2 = c'_2 + c'^2_1, \tag{57}$$

$$\mu'_4 = c'_4 + 3c'^2_2 + 4c'_1c'_3 + 6c'^2_1c'_2 + c'^4_1. \tag{58}$$

Finally, the kurtosis of $\boldsymbol{FG}$ can be derived as

$$\mathrm{kurtosis}(\boldsymbol{FG}) = \mu'_4/\mu'^2_2 = \zeta(\kappa, R)/\xi(\kappa, R). \tag{59}$$

Therefore, by solving (48), we can obtain the shape parameter κ so that the kurtosis of $\boldsymbol{FG}$ has the same value (kurt) for any value of R.

REFERENCES

[1] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[3] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 2, pp. 109–116, Mar. 2003.

[4] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Convolutive blind source separation for more than two sources in the frequency domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. III-885–III-888.

[5] H. Buchner, R. Aichner, and W. Kellerman, "A generalization of blind source separation algorithms for convolutive mixtures based on second order statistics," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 120–134, Jan. 2005.

[6] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 666–678, Mar. 2006.

[7] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inform. Process. Syst.*, 2000, vol. 13, pp. 556–562.

[9] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[10] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 121–124.

[11] H. Kameoka, M. Nakano, K. Ochiai, Y. Imoto, K. Kashino and S. Sagayama, "Constrained and regularized variants of non-negative matrix factorization incorporating music-specific constraints," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 5365–5368.

[12] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, vol. 5, pp. 3140–3143.

[13] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, 2001.

[14] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 530–538, Sep. 2004.

[15] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2007, pp. 3247–3250.

[16] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation*, 2006, pp. 165–172.

[17] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation*, 2006, pp. 601–608.

[18] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70–79, 2007.

[19] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2010, pp. 165–172.

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

[21] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. Int. Conf. Independent Compon. Anal. Signal Separation*, 2007, pp. 414–421.

[22] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi and K. Kondo, "Music signal separation based on supervised nonnegative matrix factorization with orthogonality and maximum-divergence penalties," *IEICE Trans. Fundamentals Electron. Commun. Comput. Sci.*, vol. E97-A, no. 5, pp. 1113–1118, 2014.

[23] D. Kitamura, H. Saruwatari, H. Kameoka, Y. Takahashi, K. Kondo and S. Nakamura, "Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 654–669, Apr. 2015.

[24] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," in *Proc. Irish Signals Syst. Conf.*, 2005, pp. 8–12.

[25] R. M. Parry and I. A. Essa, "Estimating the spatial position of spectral components in audio," in *Proc. Int. Conf. Independent Component Anal. Blind Source Separation*, Berlin, Germany: Springer, 2006, pp. 666–673.

[26] Y. Mitsufuji and A. Roebel, "Sound source separation based on non-negative tensor factorization incorporating spatial cue as prior knowledge," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 71–75.

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

[28] S. Arberet *et al.*, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. Inform. Sci. Signal Process. Appl.*, 2010, pp. 1–4.

[29] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 257–260.

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

[31] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial covariance models for under-determined reverberant audio source separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 129–132.

[32] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.
[33] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 276–280.
[34] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, 2009.
[35] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Relaxation of rank-1 spatial constraint in overdetermined blind source separation," in *Proc. Eur. Signal Process. Conf.*, 2015, pp. 1271–1275.
[36] Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Musical-noise analysis in methods of integrating microphone array and spectral subtraction based on higher-order statistics," *EURASIP J. Adv. Signal Process.*, vol. 2010, 2010, Art. no. 431347.
[37] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[38] S. Araki *et al.*, "The 2011 signal separation evaluation campaign (SiSEC2011):-audio source separation," in *Proc. Latent Variable Anal. Signal Separation*, 2012, pp. 414–422.
[39] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. Int. Conf. Lang. Resources Evaluation*, 2000, pp. 965–968.
[40] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833–1847, 2007.
[41] P. D. O'Grady and B. A. Pearlmutter, "Soft-LOST: EM on a mixture of oriented lines," in *Proc. Independent Compon. Anal. Blind Signal Separation*, 2004, pp. 430–436.

Daichi Kitamura received the M.E. degree in engineering from the Nara Institute of Science and Technology, Nara, Japan, in 2014. He is currently working toward the Ph.D. degree in informatics at the SOKENDAI (The Graduate University for Advanced Studies), Kanagawa, Japan.

His research interests include audio signal processing, statistical signal processing, source separation, and machine learning. He is a Member of the Institute of Electronics, Information and Communication Engineers and the Acoustical Society of Japan.

Nobutaka Ono received the B.E., M.S., and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 1996, 1998, 2001, respectively. He joined the University of Tokyo in 2001 as a Research Associate and became a Lecturer in 2005. He moved to National Institute of Informatics, Japan, as an Associate Professor in 2011. His research interests include source localization, blind source separation, and optimization algorithms for them.

He is a Senior Member of IEEE Signal Processing Society and has been a Member of IEEE Audio and Acoustic Signal Processing Technical Committee since 2014. He received the Best Paper Award from IEEE ISIE in 2008, the Best Paper Award from the IEEE IS3C in 2014, the Excellent Paper Award from IIHMSP in 2014, and the Unsupervised Learning ICA Pioneer Award from SPIE.DSS in 2015.

Hiroshi Sawada received the B.E., M.E., and Ph.D. degrees in information science from the Kyoto University, Kyoto, Japan, in 1991, 1993 and 2001, respectively. He joined NTT Corporation in 1993. He is now an Executive Manager at the NTT Communication Science Laboratories, Kyoto, Japan. His research interests include statistical signal processing, audio source separation, array signal processing, machine learning, latent variable model, graph-based data structure, and computer architecture.

From 2006 to 2009, he served as an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. He is an Associate Member of the Audio and Acoustic Signal Processing Technical Committee of the IEEE SP Society. He received the Best Paper Award of the IEEE Circuit and System Society in 2000, the SPIE ICA Unsupervised Learning Pioneer Award in 2013, and the Best Paper Award of the IEEE Signal Processing Society in 2014. He is a Member of the IEICE and the ASJ.

Hirokazu Kameoka received B.E., M.S., and Ph.D. degrees all from the University of Tokyo, Japan, in 2002, 2004, and 2007, respectively. He is currently a Distinguished Researcher and a Senior Research Scientist at NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and an Adjunct Associate Professor at the National Institute of Informatics.

From 2011 to 2016, he was an Adjunct Associate Professor at the University of Tokyo. His research interests include audio, speech, and music signal processing and machine learning. Since 2015, he has been working as an Associate Editor of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. He received 13 awards over the past 10 years, including the IEEE Signal Processing Society 2008 SPS Young Author Best Paper Award. He is the author or co-author of about 100 articles in journal papers and peer-reviewed conference proceedings.

Hiroshi Saruwatari received the B.E., M.E., and Ph.D. degrees from the Nagoya University, Nagoya, Japan, in 1991, 1993, and 2000, respectively. He joined the Intelligent System Laboratory, SECOM Co., Ltd., Tokyo, Japan, in 1993, where he was engaged in the research on the ultrasonic array system for the acoustic imaging. He is currently a Professor of Graduate School of Information Science and Technology, the University of Tokyo, Tokyo. His research interests include noise reduction, array signal processing, blind source separation, and sound field reproduction. He received paper awards from the IEICE in 2001 and 2006, from the Telecommunications Advancement Foundation in 2004 and 2009, and at the IEEE-IROS2005 in 2006. He won the first prize at the IEEE MLSP2007 Data Analysis Competition for BSS. He is a Member of the IEICE, Japan VR Society, and the ASJ.