Multichannel Non-Negative Matrix Factorization Using Banded Spatial Covariance Matrices in Wavenumber Domain

Yuki Mitsufuji, *Member, IEEE*, Stefan Uhlich, *Member, IEEE*, Norihiro Takamune, Daichi Kitamura, *Member, IEEE*, Shoichi Koyama, *Member, IEEE*, and Hiroshi Saruwatari, *Member, IEEE*

***Abstract*—Blind source separation exploiting multichannel information has long been a popular topic, and recently proposed methods based on the local Gaussian model have shown promising results despite its high computational cost for the case of many microphone signals. The low updating speed for such a model is mainly due to the inversion of a spatial covariance matrix, for which the complexity increases with the number of microphones, M, and is generally of order $O(M^3)$. Several projection-based approaches that attempt to concentrate energy on the diagonal part of the spatial covariance matrix have been introduced to circumvent the matrix inversion, which can reduce the complexity to $O(M)$. In this article, we focus on the fast Fourier transform as a projection method because the energy concentration on the diagonal can be efficiently achieved compared with other projection-based methods. For the case where the diagonalization is imperfect, for example, owing to discontinuities at the edge of a linear array, we also developed a more robust algorithm approximating the tri-diagonal part of the spatial covariance matrix, which requires a complexity of $O(M^2)$ for the inversion by applying the Thomas algorithm. To remove the ad-hoc integration of post clustering after the decomposition, we also examine a self-clustering algorithm. Our evaluation shows better results than other previously proposed methods in terms of the separation quality under reverberant conditions as well as higher efficiency than multichannel non-negative matrix factorization.**

***Index Terms*—Multichannel source separation, non-negative factorization, spatial covariance model, wavenumber domain, local Gaussian model.**

Manuscript received January 8, 2019; revised July 30, 2019 and October 7, 2019; accepted October 7, 2019. Date of publication October 21, 2019; date of current version December 24, 2019. This work was supported by the SECOM Science and Technology Foundation and JSPS KAKENHI under Grants JP19H01116 and JP19K20306. The associate editor coordinating the review of this manuscript and approving it for publication was J. Du. This article was presented in part at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Shanghai, China, March 2016. *(Corresponding author: Yuki Mitsufuji.)*

Y. Mitsufuji, N. Takamune, D. Kitamura, S. Koyama, and H. Saruwatari are with the Graduate School of Information Science and Technology, University of Tokyo, Bunkyo City, Tokyo 113-8654, Japan (e-mail: yuki_mitsufuji@ipc.i.u-tokyo.ac.jp; norihiro_takamune@ipc.i.u-tokyo.ac.jp; daichi_kitamura@ipc.i.u-tokyo.ac.jp; koyama.shoichi@ieee.org; hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp).

S. Uhlich was with the University of Stuttgart 70174, Stuttgart, Germany. He is now with Sony Europe Ltd, ZNL Deutschland, 70327 Stuttgart, Germany (e-mail: stefan.uhlich@sony.com).

Digital Object Identifier 10.1109/TASLP.2019.2948770

I. Introduction

MULTICHANNEL music source separation is one of the most actively studied topics in the audio signal processing field and various approaches have been proposed to tackle this difficult problem. In recent years, owing to the advent of deep learning, supervised methods based on training spectro-temporal information of audio signals have been proven to yield notable results [1]–[6]. In contrast, unsupervised source separation, where training data are not available, still remains a challenging and open problem.

In the last decade, the local Gaussian model has gained much attention as one of the most promising unsupervised approaches exploiting multichannel coherence. In 2005 the local Gaussian model was first applied to multichannel source separation [7], [8], in which the spectrum of each time-frequency bin is modeled as an instantaneous mixture of complex multivariate Gaussians. To cope with more complex mixing conditions such as convolutive environments, the model was further extended to incorporate a full-rank model, and a generalized expectation-maximization (GEM) algorithm was employed to derive the update rules to obtain the model parameters [9], [10]. Ozerov and Févotte applied a low-rank factorization in this framework for modeling source amplitudes of time-frequency bins [11]. Their approach can be regarded as the multichannel extension of the well-known non-negative matrix factorization (NMF) [12]. This original approach was limited to a rank-1 matrix, which was later generalized to the full-rank case so that the algorithm can also be used under reverberant conditions [13]–[15]. The separation quality of multichannel NMF was highlighted in the literature as outperforming other existing methods, such as l_1-norm minimization [16], l_p-norm minimization [17], and binary clustering [18]. Its huge computational cost and slow convergence, were addressed as two major problems to be tackled in the future. In detail, the computational cost explodes with increasing the number of microphones, M, as $O(M^3)$ owing to the multiple matrix inversions during the parameter updates [19]. For the convergence of multichannel NMF, GEM-based parameter updates were shown to be much slower than multiplicative updates by comparison with non-negative tensor factorization (NTF) [20].

The convergence speed of multichannel NMF was increased by incorporating multiplicative updates into the M-step of the source parameter updates [21]. A different update method was

proposed in [22] consisting solely of multiplicative updates with the majorization–minimization algorithm [23]. While the convergence problem was mitigated by such means, the authors reported that the separation performance was prone to local minima resulting from the initialization of the model parameters.

Nikunen and Virtanen developed a direction-of-arrival (DOA)-based method to overcome the initialization problem [24]. The algorithm provides a series of DOA kernels enabling the spatial properties observed in the multichannel signals to be encoded. A DOA kernel is composed of outer products of steering vectors. In [24], it was reported that the method yielded robust results even in the case of initializing parameters with random values. To reduce the number of parameters to be updated, the DOA-based method was further enhanced by splitting a DOA kernel into two parts: a fixed kernel consisting of phase covariances and an updatable kernel consisting of amplitude covariances [25].

In parallel with the advancement of multichannel NMF, other authors devoted their time to improving NTF [26]–[29]. Their methods treat multichannel spectrograms as three-way tensors and apply non-negative factorization inspired by NMF. NTF can also be regarded as a simplified approach to multichannel NMF because the model extracts the diagonal part of a spatial covariance matrix (SCM) while discarding the off-diagonal part that contains information of interchannel phase differences. Since NTF assumes that the original sources are mixed instantaneously, exploiting only the diagonal part is often not sufficient to model more realistic mixing conditions. In addition to NTF modeling of the diagonal part, an NMF-based treatment of off-diagonal elements has recently been proposed [30].

To overcome the weakness of the high computational cost in multichannel NMF, several authors have attempted to apply different types of orthogonal transforms to an SCM, enabling the energy of the SCM to be concentrated in the diagonal part. As a result, a matrix inversion in the update can be replaced with element-wise diagonal divisions, thus reducing the high computational cost. The authors in [31] leveraged a steering matrix to convert the SCM into the so-called *beamspace* domain. The method was further generalized in the framework called $PROJET$ [32]. There have been several works on the iterative estimation of the projection matrix by independent vector analysis (IVA) [33], either jointly with NMF updates [34] or independently followed by the NMF approach [35].

In this paper, we first propose the use of a fast Fourier transform (FFT) to project signals into the wavenumber domain and to model only the band elements of an SCM. The conversion can be achieved more efficiently than by other projection-based methods while making use of the property that plane waves can be sparsely represented in the wavenumber domain. The diagonal approximation of an SCM in the wavenumber domain was first introduced in an international conference paper [36] by the authors, in which only the scenario of using a uniform linear array with a large number of microphones was evaluated. However, it can be assumed that the performance gradually degrades as the number of microphones is reduced because the SCM of projected signals cannot contain sufficient information in the diagonal part when the number of microphones is small. Evaluating the robustness to such a scenario is one of the focuses of this paper.

To further increase the robustness, secondly, we also devised a tri-diagonal approximation approach, where the algorithm not only takes into account the diagonal part but also exploits the adjacent lower and upper bands of the matrix. It relies on the assumption that the tri-diagonal part contains more spatial information than the diagonal part. A difficulty exists in the extraction step of the tri-diagonal elements because a simple truncation of off-diagonal elements cannot maintain the positive semi-definitiveness of the matrix, which is required to ensure the convergence of the iterative update rules.

The inversions of tri-diagonal matrices, which appear a few times in the multiplicative updates, can be efficiently computed by means of the Thomas algorithm [37], which can achieve a matrix inversion of order $O(M^2)$.

Finally, we examine the incorporation of self-clustering to remove the dependency of clustering methods after the decomposition, which was reported in [38] to yield a quality improvement.

The novelties and the contributions of this work are as follows:

- an in-depth evaluation of the FFT-based projection method [36] in the case of a small number of microphones,
- an extension to tri-diagonal SCM approximation, where we ensure the positive semi-definitiveness of the matrix, followed by an efficient inverse calculation based on the Thomas algorithm, and,
- the incorporation of self-clustering to iteratively group NMF components into several source types.

The paper is organized as follows. First, we introduce the spatial covariance model in Section II. We then describe in detail our proposed method in Section III, which is then evaluated in Section IV. Finally, Section V gives the conclusions.

The following notations are used throughout this paper: $\mathbf{x}$ denotes a column vector and $\mathbf{X}$ a matrix, where $\mathbf{I}$ is the identity matrix. The trace operator, determinant, matrix transpose, conjugate matrix transpose, Euclidean vector norm, and Frobenius matrix norm are denoted by $\mathrm{tr}\{.\}$, $\det\{.\}$, $(.)^T$, $(.)^H$, $\|.\|$, and $\|.\|_F$, respectively. $\mathbf{X} \succ \mathbf{0}$, $\mathbf{X} \succeq \mathbf{0}$ means that $\mathbf{X}$ is symmetric and positive definite / semi-definite. Furthermore, $\mathrm{tridiag}\{\mathbf{X}\}$ returns a matrix of the same size as $\mathbf{X}$ that contains the tri-diagonal part of $\mathbf{X}$.

II. Models and Related Methods

A. Local Gaussian Model

This section provides an introduction to the model underlying multichannel NMF and its variant, DOA-based multichannel NMF [24]. It assumes that an M-channel vector of a short-time Fourier transform (STFT) bin can be modeled as a multivariate complex Gaussian, i.e.,

$$\mathbf{s}_{fn}^{i} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{R}_{fn}^{i}\right), \tag{1}$$

where M denotes the number of microphones, $\mathbf{s}_{fn}^{i} \in \mathbb{C}^M$ denotes the spatial image of the ith source in the STFT domain, $\mathbf{R}_{fn}^{i} = \mathrm{E}[\mathbf{s}_{fn}^{i}{\mathbf{s}_{fn}^{i}}^{H}] \in \mathbb{C}^{M\times M}$ denotes the SCM of the complex

Gaussian distribution $\mathcal{N}_{\mathbb{C}}$, f is the frequency bin index, and n is the time frame index.

The spatial image of a mixture of multiple sources $\mathbf{x}_{fn} \in \mathbb{C}^M$ is represented as a sum of complex Gaussians, i.e.,

$$\mathbf{x}_{fn} = \sum_i \mathbf{s}_{fn}^i \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{R}_{fn}\right), \tag{2}$$

where $\mathbf{R}_{fn} \in \mathbb{C}^{M\times M}$ denotes the SCM. Assuming that the sources are mutually independent, the SCM of the mixture $\mathbf{R}_{fn}$ is given by the sum of the SCMs of all sources, i.e.,

$$\mathbf{R}_{fn} = \mathrm{E}\left[\mathbf{x}_{fn}\mathbf{x}_{fn}^H\right] = \sum_i \mathbf{R}_{fn}^i. \tag{3}$$

The log-likelihood of the spatial image $\mathbf{x}_{fn}$ for the model parameters $\boldsymbol{\theta}$ under the assumption of the local Gaussian model (2) is given by

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{fn} \log \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{fn}|\mathbf{0}, \hat{\mathbf{R}}_{fn}(\boldsymbol{\theta})\right), \tag{4}$$

where the model parameter θ will be defined later in Sections II-B and II-C. This likelihood can be interpreted as the log-determinant divergence [39] between the empirical SCM, $\tilde{\mathbf{R}}_{fn} = \mathbf{x}_{fn}\mathbf{x}_{fn}^H$, and the estimated SCM, $\hat{\mathbf{R}}_{fn}(\boldsymbol{\theta}) \in \mathbb{C}^{M\times M}$:

$$\begin{aligned} C(\boldsymbol{\theta}) &= \sum_{fn} D_{\mathrm{LD}}\left(\tilde{\mathbf{R}}_{fn}|\hat{\mathbf{R}}_{fn}(\boldsymbol{\theta})\right) \\ &\equiv \sum_{fn} \mathrm{tr}\left(\tilde{\mathbf{R}}_{fn}\hat{\mathbf{R}}_{fn}(\boldsymbol{\theta})^{-1}\right) + \log\det\left(\hat{\mathbf{R}}_{fn}(\boldsymbol{\theta})\right), \end{aligned} \tag{5}$$

where $C(\boldsymbol{\theta})$ can be seen as a cost function which we want to minimize with respect to the model parameters $\boldsymbol{\theta}$. We denote the log-determinant divergence by D_{LD}.

B. Multichannel NMF

In the framework of multichannel NMF proposed in [22] where $\boldsymbol{\theta} = \{\mathbf{A}_{fk}, w_{fk}, h_{kn}\}$, the SCM $\hat{\mathbf{R}}_{fn}$ is assumed to be a superposition of time-invariant normalized SCMs $\mathbf{A}_{fk} \in \mathbb{C}^{M\times M}$ coupled with a scale value that represents the power spectral density. The scale value is decomposed into a non-negative frequency weight w_{fk} and a non-negative activation h_{kn},

$$\hat{\mathbf{R}}_{fn}(\boldsymbol{\theta}) = \sum_k \mathbf{A}_{fk} w_{fk} h_{kn}, \tag{6}$$

where the NMF component index is denoted by k. Even though the performance of multichannel NMF exceeds that of other multichannel based approaches, such as IVA [33], it is known to be sensitive to the initialization of the parameters. The drawback can be mitigated by employing other multichannel methods as an initializer of multichannel NMF [34]. Another major problem of multichannel NMF is its computational cost because the complexity increases with the number of microphones, M. In particular, in the case of multichannel NMF based on the log-determinant divergence (5), numerous matrix inversions in the update rules of w_{fk} and h_{kn} and eigendecompositions in the update of $\mathbf{A}_{fk}$ result in a very high computational cost of order $O(M^3)$ [19].

C. Multichannel NMF With Fixed DOA Kernels

In the extended approach of multichannel NMF proposed in [24] where $\boldsymbol{\theta} = \{z_{ko}, w_{fk}, h_{kn}\}$, the time-invariant normalized SCM is further decomposed into a set of fixed DOA kernels $\mathbf{J}_{fo} \in \mathbb{C}^{M\times M}$ and the corresponding directional weights $z_{ko} \in \mathbb{R}_+^{K\times O}$,

$$\hat{\mathbf{R}}_{fn}(\boldsymbol{\theta}) = \sum_k \sum_o \mathbf{J}_{fo} z_{ko} w_{fk} h_{kn}, \tag{7}$$

where $\mathbf{A}_{fk} = \sum_o \mathbf{J}_{fo} z_{ko}$ results in the same equation as (6) and o denotes the index of steering directions for the DOA kernels. By using a fixed basis of DOA kernels throughout the separation process, the stability of the performance is improved even when the parameters are initialized with random values. Furthermore, there is also a major advantage in terms of computational cost owing to the elimination of the update of the normalized SCM $\mathbf{A}_{fk}$.

Opposed to multichannel NMF, which blindly estimates the normalized SCM, the DOA-based approach requires that the geometry of the array is known in order to compute the DOA kernels. In the case of a uniform linear array, the DOA kernel is composed of an outer product of steering vectors $\mathbf{q}_{fo} \in \mathbb{C}^M$, where each steering vector is represented as a function of the time delay determined by the steering direction and the microphone distance, i.e.,

$$\mathbf{J}_{fo} = \mathbf{q}_{fo}\mathbf{q}_{fo}^H, \tag{8}$$

with

$$\mathbf{q}_{fo}^T = \begin{bmatrix} 1 & e^{j\omega_f\gamma_o} & \cdots & e^{j\omega_f(M-1)\gamma_o} \end{bmatrix}, \tag{9}$$

where ω_f denotes the frequency and γ_o denotes the time delay between each microphone and the array center. Although the overall computational cost is greatly reduced by fixing the DOA kernels, the presence of the matrix inversions in the updates of z_{ko}, w_{fk}, and h_{kn} prevents us from applying the algorithm to the scenario where a large number of microphones are required, i.e., where M is large.

III. Proposed Method

A. Motivation

To further reduce the computational complexity owing to the inverse operations in the update rules, several authors have attempted to diagonalize the SCM in different ways such that inverse operations can be replaced with element-wise divisions [31], [32], [36]. In this paper, we employ a wavenumber domain transform because it can be achieved by applying an FFT operation, which only requires $O(M\log M)$ operations. The wavenumber domain, also referred to as spatial Fourier domain, can be applied to the multichannel observations. If the microphones are uniformly distributed in space, the SCM can be diagonalized by the wavenumber transform [40]. This approach makes use of the property that plane waves can be sparsely represented in the wavenumber domain, as can be seen in Fig. 1. The spectrograms in the wavenumber domain are characterized by several sparsely located peaks that represent plane waves.

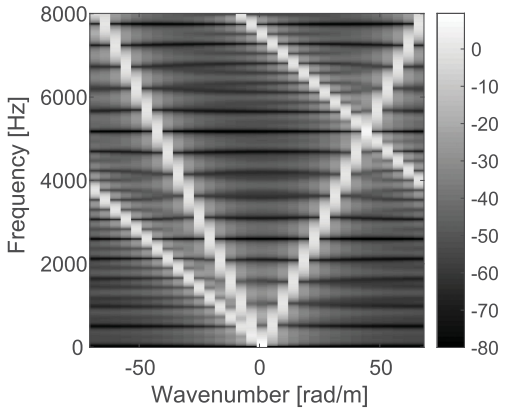

Fig. 1. Virtual plane waves represented in the wavenumber domain, originating from $\pi/20$, $8\pi/20$, and $14\pi/20$.

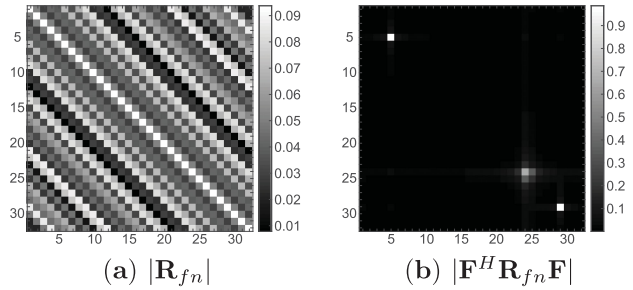

Fig. 2. SCM of mixed plane waves (2 kHz) originating from $\pi/20$, $8\pi/20$, and $14\pi/20$. The number of microphones, M, is 32. $|.|$ denotes the element-wise absolute value of the matrix.

B. Spatial Transform of SCMs

The STFT multichannel signals are converted into the wavenumber domain, where the underlying probability model is based on the zero-mean complex Gaussian distribution,

$$\mathbf{F}^H\mathbf{x}_{fn} = \mathbf{F}^H \sum_i \mathbf{s}^i_{fn} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{R}^{\mathrm{SP}}_{fn}\right), \tag{10}$$

with

$$\mathbf{R}^{\mathrm{SP}}_{fn} = \mathbf{F}^H\tilde{\mathbf{R}}_{fn}\mathbf{F}, \tag{11}$$

where $\mathbf{R}^{\mathrm{SP}}_{fn} \in \mathbb{C}^{M\times M}$ denotes the SCM in the wavenumber domain and $\mathbf{F} \in \mathbb{C}^{M\times M}$ denotes the discrete Fourier transform (DFT) matrix. Fig. 2 shows an example of a comparison between an SCM for $M = 32$ in the time-frequency domain and the converted matrix in the wavenumber domain. It is clear from the figure that the SCM in the wavenumber domain has strong peaks in the band elements, whereas the one in the time-frequency domain has quasi-uniformly distributed values. A comparison for $M = 4$ is shown in Fig. 3. Again, the SCM in wavenumber domain exhibits a small number of strong peaks. To numerically assess the sparseness of the SCMs, we computed the percentage of elements that are smaller in magnitude than 20% of the maximum element. For the case of $M = 32$ (Fig. 2), we obtain 6.1% in (a) and 99.5% in (b). Also, for the case of $M = 4$ (Fig. 3), we obtain 0% in (a) and 81.3% in (b). Thus, even for $M = 4$, more than 80% of all SCM elements in wavenumber domain are smaller than the threshold whereas it was none in the

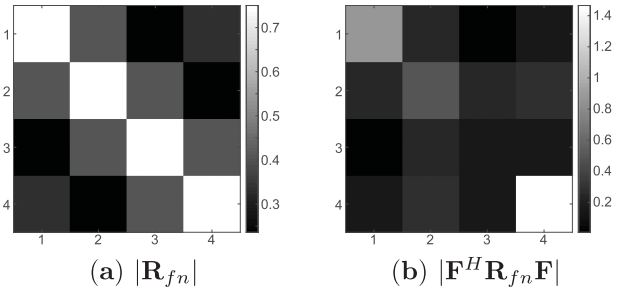

Fig. 3. SCM of mixed plane waves (2 kHz) originating from $\pi/20$, $8\pi/20$, and $14\pi/20$. The number of microphones, M, is 4. $|.|$ denotes the element-wise absolute value of the matrix.

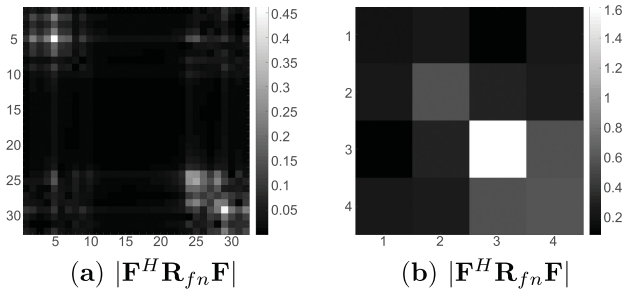

Fig. 4. SCM of three point sources (2 kHz) originating from $\pi/20$, $8\pi/20$, and $14\pi/20$, simulated in the reverberant scenario with the image method (7.0 m $\times$ 12.0 m $\times$ 3.0 m, T60 $=$ 400 ms). The number of microphones, M, is 32 for (a), and 4 for (b). $|.|$ denotes the element-wise absolute value of the matrix.

time-frequency domain. Finally, to observe the effect of room reverberation on the sparseness of the SCM, we conducted a simulation based on the image method [41]. The simulated room size is 7.0 m $\times$ 12.0 m $\times$ 3.0 m and the reverberation time is 400 ms. The setup is the same as for Figs. 2 and 3, i.e., the three sources are placed at the corresponding directions. Fig. 4 shows the resultant wavenumber-domain SCMs obtained in this simulation. In the reverberant scenario, the sparseness of the SCMs are 91.8% for $M = 32$ and 68.8% for $M = 4$. Although the values decrease gradually compared with the anechoic case, the sparse nature in the wavenumber domain still remains. Furthermore, the concentration on the diagonal elements can still be seen in Fig. 4, as in Figs. 2 and 3.

Given $\hat{\mathbf{R}}^{\mathrm{SP}}_{fn} = \mathbf{F}^H\hat{\mathbf{R}}_{fn}\mathbf{F}$, the cost function for the local Gaussian model in (5) can then be modified by replacing the SCM in the time-frequency domain with the matrix in the wavenumber domain,

$$C_{\mathrm{sp}}(\boldsymbol{\theta}) = \sum_{fn} D_{\mathrm{LD}}\left(\mathbf{R}^{\mathrm{SP}}_{fn}|\hat{\mathbf{R}}^{\mathrm{SP}}_{fn}\right). \tag{12}$$

Note that $C_{\mathrm{sp}}(\boldsymbol{\theta})$ is equivalent to $C(\boldsymbol{\theta})$ in (5) as can be seen from

$$\begin{aligned} C_{\mathrm{sp}}(\boldsymbol{\theta}) &= \sum_{fn} D_{\mathrm{LD}}\left(\mathbf{F}^H\tilde{\mathbf{R}}_{fn}\mathbf{F}|\mathbf{F}^H\hat{\mathbf{R}}_{fn}\mathbf{F}\right) \\ &= \sum_{fn} \mathrm{tr}\left(\mathbf{F}^H\tilde{\mathbf{R}}_{fn}\mathbf{F}\mathbf{F}^{-1}\hat{\mathbf{R}}^{-1}_{fn}\mathbf{F}^{-H}\right) \\ &\quad + \log\det\left(\mathbf{F}^H\hat{\mathbf{R}}_{fn}\mathbf{F}\right) \end{aligned}$$

$$= \sum_{fn} \mathrm{tr}\left(\mathbf{F}^H\mathbf{F}\mathbf{F}^{-1}\mathbf{F}^{-H}\tilde{\mathbf{R}}_{fn}\hat{\mathbf{R}}_{fn}^{-1}\right) + \log\det\left(\hat{\mathbf{R}}_{fn}\right) - \log\det\left(\mathbf{F}\right) + \log\det\left(\mathbf{F}\right) = C(\boldsymbol{\theta}). \quad (13)$$

Since w_{fk} and h_{kn} are not dependent on the channel dimension, the spatial transform of the estimated SCMs, supposed to be performed in every iteration, can be replaced with a single spatial transform of the fixed DOA kernels $\mathbf{J}_{fo}$ at the initialization stage,

$$\hat{\mathbf{R}}_{fn}^{\mathrm{SP}} = \mathbf{F}^H\hat{\mathbf{R}}_{fn}\mathbf{F} = \sum_k\sum_o \mathbf{J}_{fo}^{\mathrm{SP}} z_{ko}w_{fk}h_{kn}, \quad (14)$$

with

$$\mathbf{J}_{fo}^{\mathrm{SP}} = \mathbf{F}^H\mathbf{J}_{fo}\mathbf{F}. \quad (15)$$

C. Wiener Filtering in Wavenumber Domain

Given the estimated model parameters, the STFT coefficients of each source can be recovered by a multichannel Wiener filter, i.e., a minimum mean squared error (MMSE) estimator [42]. Since $\hat{\mathbf{R}}_{fn} = \mathbf{F}\hat{\mathbf{R}}_{fn}^{\mathrm{SP}}\mathbf{F}^H$ and $\mathbf{J}_{fo} = \mathbf{F}\mathbf{J}_{fo}^{\mathrm{SP}}\mathbf{F}^H$ also hold owing to $\mathbf{F}\mathbf{F}^H = \mathbf{I}$ where $\mathbf{I}$ denotes an identity matrix, the MMSE estimator can be given in the wavenumber domain by

$$\begin{aligned}\hat{\mathbf{s}}_{fn}^i &= \left(\sum_{k\in K_i}\sum_o \mathbf{J}_{fo}z_{ko}w_{fk}h_{kn}\right)\left(\hat{\mathbf{R}}_{fn}\right)^{-1}\mathbf{x}_{fn}\\ &= \left(\sum_{k\in K_i}\sum_o \mathbf{F}\mathbf{J}_{fo}^{\mathrm{SP}}\mathbf{F}^H z_{ko}w_{fk}h_{kn}\right)\left(\mathbf{F}\hat{\mathbf{R}}_{fn}^{\mathrm{SP}}\mathbf{F}^H\right)^{-1}\mathbf{x}_{fn}\\ &= \mathbf{F}\left(\sum_{k\in K_i}\sum_o \mathbf{J}_{fo}^{\mathrm{SP}} z_{ko}w_{fk}h_{kn}\right)\left(\hat{\mathbf{R}}_{fn}^{\mathrm{SP}}\right)^{-1}\mathbf{F}^H\mathbf{x}_{fn},\end{aligned} \quad (16)$$

where K_i denotes the set of components that belong to the ith source. The set of NMF components K_i can be determined by a clustering method, such as LPC-based [43] or Mel-spectrum-based clustering [44].

D. Diagonal Approximation

If we assume that $\mathbf{F}^H\hat{\mathbf{R}}_{fn}\mathbf{F}$ and $\mathbf{F}^H\mathbf{J}_{fo}\mathbf{F}$ are diagonal matrices (see Fig. 2), then $\mathbf{F}^H\hat{\mathbf{R}}_{fn}\mathbf{F}$ and $\mathbf{F}^H\mathbf{J}_{fn}\mathbf{F}$ can be well approximated by considering only their diagonal elements,

$$\mathbf{F}^H\hat{\mathbf{R}}_{fn}\mathbf{F} \approx \hat{\mathbf{R}}_{fn}^{\mathrm{Diag}} = \begin{pmatrix}\hat{a}_{1fn} & 0 & \dots & 0\\ 0 & \hat{a}_{2fn} & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & \hat{a}_{Mfn}\end{pmatrix}, \quad (17)$$

$$\mathbf{F}^H\mathbf{J}_{fo}\mathbf{F} \approx \mathbf{J}_{fo}^{\mathrm{Diag}} = \begin{pmatrix}b_{1fo} & 0 & \dots & 0\\ 0 & b_{2fo} & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & b_{Mfo}\end{pmatrix}. \quad (18)$$

Thus, the cost function (13) can also be approximated by focusing on the diagonal elements,

$$C_{\mathrm{sp}}(\boldsymbol{\theta}) \approx \sum_{fn} D_{\mathrm{LD}}\left(\mathbf{R}_{fn}^{\mathrm{SP}}|\hat{\mathbf{R}}_{fn}^{\mathrm{Diag}}\right) = \sum_{fn} D_{\mathrm{LD}}\left(\mathbf{R}_{fn}^{\mathrm{SP}}|\sum_k\sum_o \mathbf{J}_{fo}^{\mathrm{Diag}} z_{ko}w_{fk}h_{kn}\right). \quad (19)$$

The update rules reflecting the approximation can be written such that they only contain the inversions of diagonal matrices, allowing the algorithm to run at a computational cost of order $O(M)$ in each iteration,

$$z_{ko} \leftarrow z_{ko}\sqrt{\frac{\sum_{fn}\mathrm{tr}\left(\left(\hat{\mathbf{R}}_{fn}^{\mathrm{Diag}}\right)^{-1}\mathbf{R}_{fn}^{\mathrm{SP}}\left(\hat{\mathbf{R}}_{fn}^{\mathrm{Diag}}\right)^{-1}\mathbf{J}_{fo}^{\mathrm{Diag}}\right)w_{fk}h_{kn}}{\sum_{fn}\mathrm{tr}\left(\left(\hat{\mathbf{R}}_{fn}^{\mathrm{Diag}}\right)^{-1}\mathbf{J}_{fo}^{\mathrm{Diag}}\right)w_{fk}h_{kn}}}, \quad (20a)$$

$$w_{fk} \leftarrow w_{fk}\sqrt{\frac{\sum_{on}\mathrm{tr}\left(\left(\hat{\mathbf{R}}_{fn}^{\mathrm{Diag}}\right)^{-1}\mathbf{R}_{fn}^{\mathrm{SP}}\left(\hat{\mathbf{R}}_{fn}^{\mathrm{Diag}}\right)^{-1}\mathbf{J}_{fo}^{\mathrm{Diag}}\right)z_{ko}h_{kn}}{\sum_{on}\mathrm{tr}\left(\left(\hat{\mathbf{R}}_{fn}^{\mathrm{Diag}}\right)^{-1}\mathbf{J}_{fo}^{\mathrm{Diag}}\right)z_{ko}h_{kn}}}, \quad (20b)$$

$$h_{kn} \leftarrow h_{kn}\sqrt{\frac{\sum_{fo}\mathrm{tr}\left(\left(\hat{\mathbf{R}}_{fn}^{\mathrm{Diag}}\right)^{-1}\mathbf{R}_{fn}^{\mathrm{SP}}\left(\hat{\mathbf{R}}_{fn}^{\mathrm{Diag}}\right)^{-1}\mathbf{J}_{fo}^{\mathrm{Diag}}\right)z_{ko}w_{fk}}{\sum_{fo}\mathrm{tr}\left(\left(\hat{\mathbf{R}}_{fn}^{\mathrm{Diag}}\right)^{-1}\mathbf{J}_{fo}^{\mathrm{Diag}}\right)z_{ko}w_{fk}}}. \quad (20c)$$

Since a matrix inversion can be written in a more compact way by using element-wise divisions, the above update rules can be simplified, i.e.,

$$z_{ko} \leftarrow z_{ko}\sqrt{\frac{\sum_{fn}\sum_m \frac{a_{mfn}}{\hat{a}_{mfn}^2}b_{mfo}w_{fk}h_{kn}}{\sum_{fn}\sum_m \frac{1}{\hat{a}_{mfn}}b_{mfo}w_{fk}h_{kn}}}, \quad (21a)$$

$$w_{fk} \leftarrow w_{fk}\sqrt{\frac{\sum_{on}\sum_m \frac{a_{mfn}}{\hat{a}_{mfn}^2}b_{mfo}z_{ko}h_{kn}}{\sum_{on}\sum_m \frac{1}{\hat{a}_{mfn}}b_{mfo}z_{ko}h_{kn}}}, \quad (21b)$$

$$h_{kn} \leftarrow h_{kn}\sqrt{\frac{\sum_{fo}\sum_m \frac{a_{mfn}}{\hat{a}_{mfn}^2}b_{mfo}z_{ko}w_{fk}}{\sum_{fo}\sum_m \frac{1}{\hat{a}_{mfn}}b_{mfo}z_{ko}w_{fk}}}. \quad (21c)$$

where a_{mfn} denotes the diagonal elements of $\mathbf{R}_{fn}^{\mathrm{SP}}$. The detailed derivation can be found in Appendix. The proposed diagonal algorithm is given in Algorithm 1.

Algorithm 1: Diagonal Approximation Approach.

Input: Mixture $\mathbf{x}_{fn}$
Compute $\tilde{\mathbf{R}}_{fn} = \mathbf{x}_{fn}\mathbf{x}_{fn}^H$
Apply spatial transform to $\tilde{\mathbf{R}}_{fn}$ with (11)
Apply spatial transform to $\mathbf{J}_{fo}$ with (15)
Extract diagonal part of $\mathbf{F}^H\mathbf{J}_{fo}\mathbf{F}$ with (18)
Initialize z_{ko}, w_{fk}, h_{kn} with randomized values
Compute $\hat{\mathbf{R}}_{fn}^{\mathrm{SP}}$ with (14)
for $i = 0$ to MM iteration **do**
  $z_{ko} \leftarrow (21a)$
  Compute $\hat{\mathbf{R}}_{fn}^{\mathrm{SP}}$ with (14)
  $w_{fk} \leftarrow (21b)$
  Compute $\hat{\mathbf{R}}_{fn}^{\mathrm{SP}}$ with (14)
  $h_{kn} \leftarrow (21c)$
  Compute $\hat{\mathbf{R}}_{fn}^{\mathrm{SP}}$ with (14)
  Normalize z_{ko}, w_{fk}
end for
Cluster NMF components based on [43], [44]
Apply Wiener filtering with (16)
Output: Estimates $\hat{\mathbf{s}}_{fn}^{i}$

E. Extension to Tri-Diagonal Approximation

Owing to the finite length of the DFT matrix, the SCM cannot be perfectly diagonalized and off-diagonal elements cannot be avoided. This "smear" becomes more dominant in the case of a uniform linear array because the matrix cannot be regarded as a circulant matrix for which diagonalization can be perfectly achieved. Different from the previous subsection, where we exploited only the diagonal part, we also make use of the upper and lower adjacent bands, i.e., the tri-diagonal part. The matrix entries outside the tri-diagonal part are not taken into account in the optimization process.

It should be emphasized that simply discarding the off-tri-diagonal part does not ensure the convergence of the multiplicative update rules because setting these entries to zero does not guarantee the positive semi-definitiveness of the resulting tri-diagonal SCM, which is an essential assumption throughout the update of model parameters in multichannel NMF. To maintain the positive semi-definitiveness of the matrix during the tri-diagonalization process, we must solve the optimization problem

$$\begin{aligned}\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}} &= \arg\min_{\mathbf{V}} \left\| \hat{\mathbf{R}}_{fn}^{\mathrm{SP}} - \mathrm{tridiag}\{\mathbf{V}\} \right\|_F^2 \\ &\text{subject to} \quad \mathrm{tridiag}\{\mathbf{V}\} \succeq \mathbf{0}. \end{aligned} \tag{22}$$

The operation $\mathrm{tridiag}\{\mathbf{V}\}$ returns a tri-diagonal matrix in which elements that are not on the main diagonal and on the diagonal above/below are set to zero. The optimization problem (22) is a *nearest-matrix* problem where we solve for the positive semi-definite, tri-diagonal matrix $\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}} \in \mathbb{C}^{M\times M}$ that is nearest to $\mathbf{R}_{fn}^{\mathrm{SP}}$ using the Frobenius norm. We solve the semi-definite programming problem (22) by using YALMIP [45] with SeDuMi [46].

Although the optimization (22) in order to obtain $\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}}$ is costly, it can be replaced with solving the same problem for the DOA kernel $\mathbf{J}_{fo}$, which only must be carried out once in the initialization process as the DOA kernels are fixed throughout the update iterations.

$$\begin{aligned}\hat{\mathbf{J}}_{fo}^{\mathrm{Tri}} &= \arg\min_{\mathbf{V}} \left\| \mathbf{J}_{fo}^{\mathrm{SP}} - \mathrm{tridiag}\{\mathbf{V}\} \right\|_F^2 \\ &\text{subject to} \quad \mathrm{tridiag}\{\mathbf{V}\} \succeq \mathbf{0}. \end{aligned} \tag{23}$$

$$\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}} = \sum_k \sum_o \hat{\mathbf{J}}_{fo}^{\mathrm{Tri}} z_{ko} w_{fk} h_{kn}. \tag{24}$$

Using the approximated tri-diagonal SCM, the cost function in (13) can be modified to

$$\begin{aligned} C_{\mathrm{sp}}(\boldsymbol{\theta}) &\approx \sum_{fn} D_{\mathrm{LD}}\left(\mathbf{R}_{fn}^{\mathrm{SP}} | \hat{\mathbf{R}}_{fn}^{\mathrm{Tri}}\right) \\ &= \sum_{fn} D_{\mathrm{LD}}\left(\mathbf{R}_{fn}^{\mathrm{SP}} | \sum_k \sum_o \mathbf{J}_{fo}^{\mathrm{Tri}} z_{ko} w_{fk} h_{kn}\right). \end{aligned} \tag{25}$$

To minimize the log-determinant divergence between two SCMs, we employ the majorization–minimization algorithm to reduce the cost monotonically. The detailed derivation can be found in Appendix.

The multiplicative update rules for the model parameters z_{ko}, w_{fk}, and h_{kn} are given by

$$z_{ko} \leftarrow z_{ko} \sqrt{\frac{\sum_{fn} \mathrm{tr}\left(\left(\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}}\right)^{-1} \mathbf{R}_{fn}^{\mathrm{SP}} \left(\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}}\right)^{-1} \hat{\mathbf{J}}_{fo}^{\mathrm{Tri}}\right) w_{fk} h_{kn}}{\sum_{fn} \mathrm{tr}\left(\left(\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}}\right)^{-1} \hat{\mathbf{J}}_{fo}^{\mathrm{Tri}}\right) w_{fk} h_{kn}}}, \tag{26a}$$

$$w_{fk} \leftarrow w_{fk} \sqrt{\frac{\sum_{on} \mathrm{tr}\left(\left(\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}}\right)^{-1} \mathbf{R}_{fn}^{\mathrm{SP}} \left(\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}}\right)^{-1} \hat{\mathbf{J}}_{fo}^{\mathrm{Tri}}\right) z_{ko} h_{kn}}{\sum_{on} \mathrm{tr}\left(\left(\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}}\right)^{-1} \hat{\mathbf{J}}_{fo}^{\mathrm{Tri}}\right) z_{ko} h_{kn}}}, \tag{26b}$$

$$h_{kn} \leftarrow h_{kn} \sqrt{\frac{\sum_{fo} \mathrm{tr}\left(\left(\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}}\right)^{-1} \mathbf{R}_{fn}^{\mathrm{SP}} \left(\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}}\right)^{-1} \hat{\mathbf{J}}_{fo}^{\mathrm{Tri}}\right) z_{ko} w_{fk}}{\sum_{fo} \mathrm{tr}\left(\left(\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}}\right)^{-1} \hat{\mathbf{J}}_{fo}^{\mathrm{Tri}}\right) z_{ko} w_{fk}}}. \tag{26c}$$

Even in the case of a huge number of microphones, M, the inversion of the estimated SCM is not costly when employing the Thomas algorithm [37], which has a computational complexity of $O(M^2)$. The proposed tri-diagonal algorithm is given in Algorithm 2.

Algorithm 2: Extension to Tri-Diagonal Approximation.

Input: Mixture $\mathbf{x}_{fn}$
Compute $\tilde{\mathbf{R}}_{fn} = \mathbf{x}_{fn}\mathbf{x}_{fn}^H$
Apply spatial transform to $\tilde{\mathbf{R}}_{fn}$ with (11)
Apply spatial transform to $\mathbf{J}_{fo}$ with (15)
Obtain tri-diagonal approximation $\hat{\mathbf{J}}_{fo}^{\mathrm{Tri}}$ with (23)
Initialize z_{ko}, w_{fk}, h_{kn} with randomized values
Compute $\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}}$ with (24)
for $i = 0$ to MM iteration **do**
  $z_{ko} \leftarrow (26a)$
  Compute $\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}}$ with (24)
  $w_{fk} \leftarrow (26b)$
  Compute $\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}}$ with (24)
  $h_{kn} \leftarrow (26c)$
  Compute $\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}}$ with (24)
  Normalize z_{ko}, w_{fk}
end for
Cluster NMF components based on [43], [44]
Apply Wiener filtering with (16)
Output: Estimates $\hat{\mathbf{s}}_{fn}^i$

F. Extension to Self-Clustering

Inspired by the prior works [22], [26] in which the models have clustering capability, we extended (19) as follows to possess a grouping factor g_{ik}:

$$\hat{a}_{mfn} = \sum_i \sum_k \sum_o b_{mfo} z_{io} g_{ik} w_{fk} h_{kn}. \tag{27}$$

The update rules are derived in a similar way to in Section III-D and are given by

$$z_{io} \leftarrow z_{io} \sqrt{\frac{\sum_{kfn} \sum_m \frac{a_{mfn}}{\hat{a}_{mfn}^2} b_{mfo} g_{ik} w_{fk} h_{kn}}{\sum_{kfn} \sum_m \frac{1}{\hat{a}_{mfn}} b_{mfo} g_{ik} w_{fk} h_{kn}}}, \tag{28a}$$

$$g_{ik} \leftarrow g_{ik} \sqrt{\frac{\sum_{ofn} \sum_m \frac{a_{mfn}}{\hat{a}_{mfn}^2} b_{mfo} z_{io} w_{fk} h_{kn}}{\sum_{ofn} \sum_m \frac{1}{\hat{a}_{mfn}} b_{mfo} z_{io} w_{fk} h_{kn}}}, \tag{28b}$$

$$w_{fk} \leftarrow w_{fk} \sqrt{\frac{\sum_{ion} \sum_m \frac{a_{mfn}}{\hat{a}_{mfn}^2} b_{mfo} z_{io} g_{ik} h_{kn}}{\sum_{ion} \sum_m \frac{1}{\hat{a}_{mfn}} b_{mfo} z_{io} g_{ik} h_{kn}}}, \tag{28c}$$

$$h_{kn} \leftarrow h_{kn} \sqrt{\frac{\sum_{ifo} \sum_m \frac{a_{mfn}}{\hat{a}_{mfn}^2} b_{mfo} z_{io} g_{ik} w_{fk}}{\sum_{ifo} \sum_m \frac{1}{\hat{a}_{mfn}} b_{mfo} z_{io} g_{ik} w_{fk}}}. \tag{28d}$$

In [22], it was shown that incorporating self-clustering into the iterations improved their performance. We compare self-clustering with various other post-clustering approaches in Section IV-D and observe similar behavior for the average performance over all instruments. Note that the self-clustering extension can also be applied to the tri-diagonal case. The proposed self-clustering extension can be found in Algorithm 3.

TABLE I
EXPERIMENTAL SETUP

Sampling rate	16 kHz
STFT frame size	1024
STFT hop size	512

Algorithm 3: Extension to Self-Clustering.

Input: Mixture $\mathbf{x}_{fn}$
Compute $\tilde{\mathbf{R}}_{fn} = \mathbf{x}_{fn}\mathbf{x}_{fn}^H$
Apply spatial transform to $\tilde{\mathbf{R}}_{fn}$ with (11)
Apply spatial transform to $\mathbf{J}_{fo}$ with (15)
Extract diagonal part of $\mathbf{F}^H\mathbf{J}_{fo}\mathbf{F}$ with (18)
Initialize $z_{io}, b_{ik}, w_{fk}, h_{kn}$ with randomized values
Compute $\hat{\mathbf{R}}_{fn}^{\mathrm{SP}}$ with (14)
for $i = 0$ to MM iteration **do**
  $z_{io} \leftarrow (28a)$
  Compute $\hat{\mathbf{R}}_{fn}^{\mathrm{SP}}$ with (14)
  $g_{ik} \leftarrow (28b)$
  Compute $\hat{\mathbf{R}}_{fn}^{\mathrm{SP}}$ with (14)
  $w_{fk} \leftarrow (28c)$
  Compute $\hat{\mathbf{R}}_{fn}^{\mathrm{SP}}$ with (14)
  $h_{kn} \leftarrow (28d)$
  Compute $\hat{\mathbf{R}}_{fn}^{\mathrm{SP}}$ with (14)
  Normalize z_{io}, g_{ik}, w_{fk}
end for
Apply Wiener filtering with (16)
Output: Estimates $\hat{\mathbf{s}}_{fn}^i$

IV. EVALUATION

A. Experimental Conditions

To evaluate the proposed algorithms, Algorithms 1, 2, and 3, we conducted various experiments in a music separation scenario. The experimental conditions are listed in Table I. As an evaluation metric, we employed the SDR improvement, which can be computed by subtracting the outputs of BSS Eval Toolbox [47] for the original mixture from the BSS Eval values of the separations. Furthermore, the SIR improvement and the SAR are also given for the experiment in Section IV-B. Please note that we did not compute SAR improvements as the input SAR is infinity and, therefore, an improvement is not computable. Moreover, we omit the SIR improvement and the SAR for the other experiments because the tendencies of these metrics are the same as that of the corresponding SDR improvement, as can be seen in Section IV-B.

B. Comparison With Other Methods Under Reverberant Conditions

Reverberant room conditions were simulated to compare the proposed method with various other methods under realistic conditions. To answer the issue addressed in Section I regarding the case of imperfect diagonalization of SCM for small M, the number of microphones was set to $M = 4$ to observe our proposed method in the case of a small number of microphones.

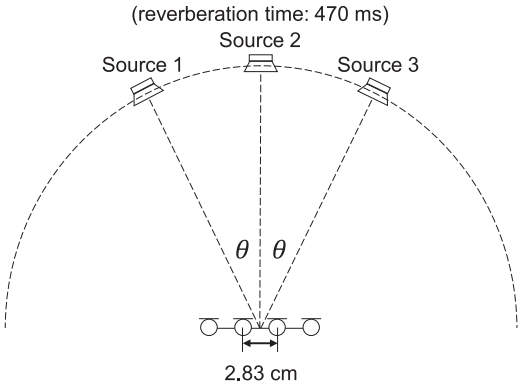

Fig. 5. Relationship between source positions with respect to the center of the microphone array.

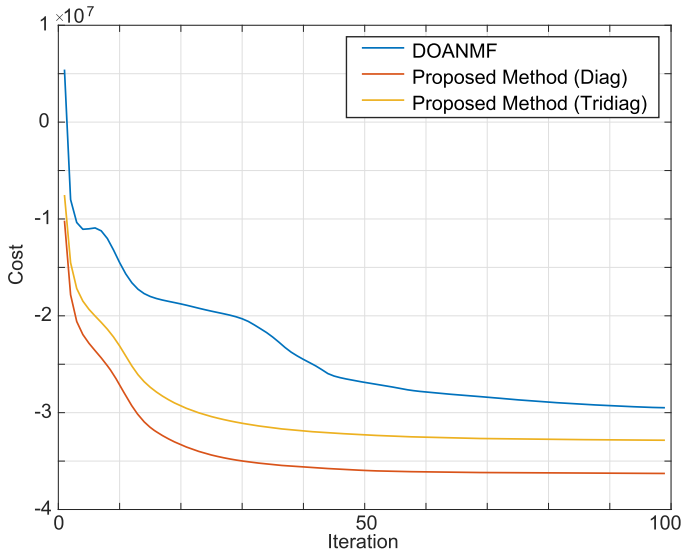

Fig. 6. Convergence curves for DOANMF, the proposed Algorithm 1 (Diag), and Algorithm 2 (Tridiag). The costs for DOANMF, Algorithm 1 (Diag), and Algorithm 2 (Tridiag) are computed by (5), (19), and (25), respectively.

Note that the case of many microphones was investigated in [36], where we studied an example with $M = 32$ microphones and showed the effectiveness of our diagonal approximation in this scenario. Three instrument signals of 10 seconds each were taken from the SiSEC (Signal Separation Evaluation Campaign) database 2008[1] for the task of under-determined speech and music mixtures. The angle between adjacent sources and the array center is denoted as θ, as shown in Fig. 5. To create reverberant signals, the RWCP impulse response [48] was used and convolved with the source signals. The reverberation time of the selected room response was 470 ms, which corresponds to the reverberation of a standard conference room. The angle θ was varied from 10 to 30° to verify the robustness to angle differences. To exclude the dependence of random initializations, the average over 10 trials per angle was computed and the result is labeled as "mean" in Fig. 7. The maximum value over 10 trials is also shown to observe the potential of each algorithm and the corresponding bar is labeled as "max" in Fig. 7. The locations of the three sources were rotated three times to avoid the bias of spatial locations. In addition to the proposed Algorithm 1 (Diag) and Algorithm 2 (Tridiag) with $O = 5$ kernels, five different algorithms were evaluated as baselines: minimum variance distortionless response (MVDR) beamformer, IVA [33], independent low-rank matrix analysis (ILRMA) [34], full-rank multichannel NMF (MNMF) [22], and multichannel NMF with fixed DOA kernels ($O = 5, 8$), described in Section II-C (DOANMF). For MVDR beamformer, the DOAs of all three sources as well as the oracle SCM of interferences were fed to the algorithm as prior knowledge. The cost function for ILRMA, MNMF, and DOANMF was based on the log-determinant divergence. The number of iterations for all the iterative methods was set to 100, empirically determined based on the convergence plot shown in Fig. 6. The NMF clustering method in [44] was carried out for MNMF, DOANMF, and the proposed algorithms. The SDR improvements, SIR improvements, and raw SARs for the three different angles between the sources are shown in Fig. 7. For all three angles, both the diagonal approach and the tri-diagonal approach consistently outperformed the other methods. A more rigorous comparison between the proposed algorithms is described in Section IV-C. For the conventional methods, DOANMF exhibits inferior performance to MNMF, in contrast to our expectation that a DOA-based method should be robust against random initializations as reported in [24]. The performance of DOANMF was improved as the number of DOA kernels was increased from $O = 5$ to $O = 8$ (the best-performance case), but it cannot reach the performance of the proposed methods. Although ILRMA exhibits superior performance to IVA owing to the NMF-based source model on top of the IVA spatial model, it did not achieve as good results as our approaches. This is due to the fact that one of the algorithmic assumptions of ILRMA, i.e., the spatially rank-1 property for each source, does not hold under our simulated reverberant conditions. In this simulation, the STFT frame size was set to 1024 points, corresponding to a length of 64 ms, which is much shorter than the reverberation time of 470 ms and thus, no valid time-invariant demixing matrices exist in ILRMA and IVA. For MVDR beamformer, although the above mentioned prior knowledge was given to the system, it does not perform as good as ILRMA. It is natural to observe such results because MVDR beamformer's prior information on the target source is only the direct-wave direction (steering vector) without taking the reverberant components into account. Thus, in the reverberant condition, the target source component has much leakage. On the other hand, ILRMA can estimate the optimal separation matrix, which consists of multiple beamformers' weights to cancel each of interferences with their reverberant components, resulting in less leakage (the detailed mechanism has been reported in [49]).

[1][Online]. Available: http://sisec2008.wiki.irisa.fr/tiki-index.html

C. Robustness to Different Angles

To evaluate the effectiveness of adding upper and lower bands to the diagonal elements in the tri-diagonal approach, we further

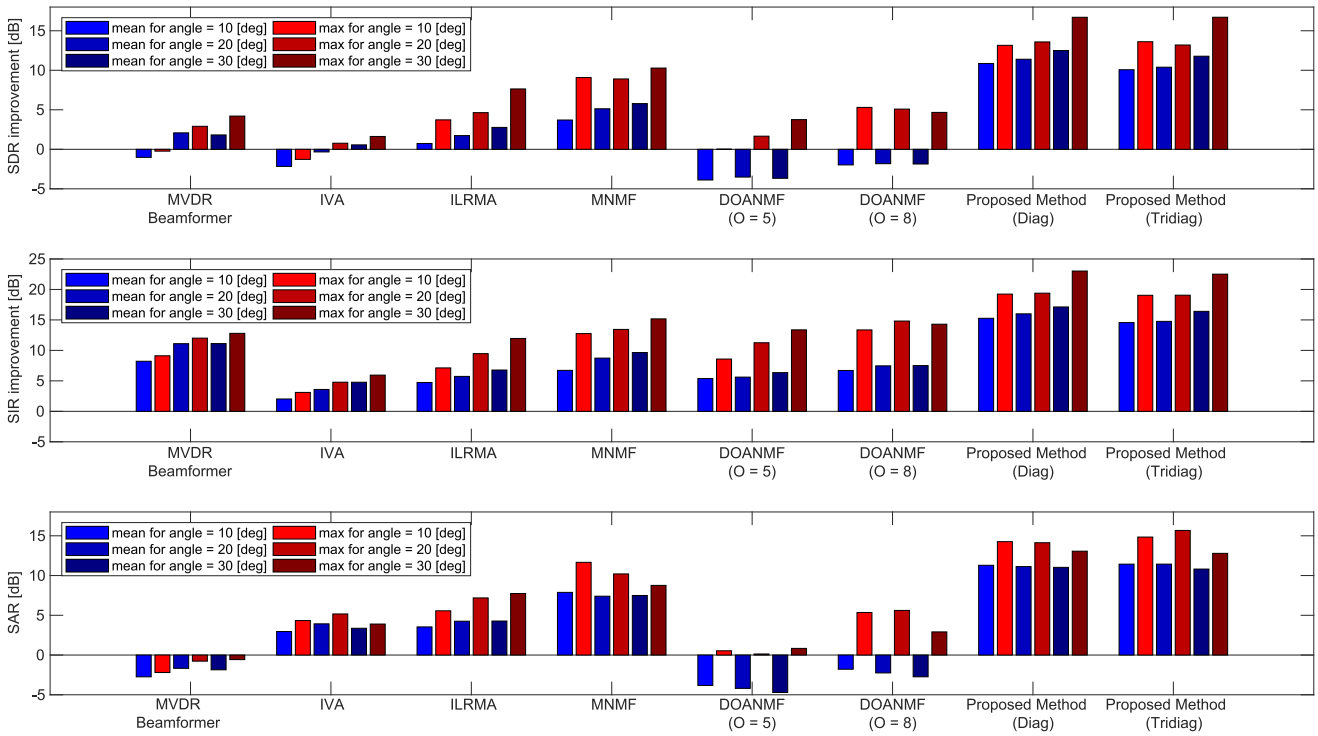

Fig. 7. SDR improvements, SIR improvements, and SARs under the reverberant conditions shown in Fig. 5, where the reverberation time is 470 ms.

TABLE II
MAXIMUM SDR IMPROVEMENTS AMONG 10 TRIALS FOR DIFFERENT SOURCE ANGLES UNDER ANECHOIC CONDITION

Approach	10 [deg]	20 [deg]	30 [deg]	40 [deg]	50 [deg]	60 [deg]	70 [deg]	80 [deg]	average over angles
Tridiag	**16.50**	**17.49**	**17.35**	**17.65**	**17.79**	**18.09**	**18.67**	**17.51**	**17.63**
Diag	16.15	16.78	16.29	16.32	17.43	17.45	17.43	16.80	16.83

TABLE III
AVERAGED SDR IMPROVEMENTS OVER 10 TRIALS FOR DIFFERENT SOURCE ANGLES UNDER ANECHOIC CONDITION

Approach	10 [deg]	20 [deg]	30 [deg]	40 [deg]	50 [deg]	60 [deg]	70 [deg]	80 [deg]	average over angles
Tridiag	**14.23**	13.45	**13.58**	**13.52**	**10.52**	12.87	**12.13**	8.91	**12.40**
Diag	13.88	**14.16**	12.64	7.58	10.41	**12.95**	11.55	**11.71**	11.86

conducted an experiment comparing Algorithm 1 (Diag) and Algorithm 2 (Tridiag) by changing the angle between the sources. We assume that the tri-diagonal approach is more robust against changes to the angle because it contains additional information in the lower and upper bands. The source angle was varied from 10 to 80°. To minimize the overlap with the previous experiment, this experiment was performed under anechoic conditions and the number of DOA kernels was set to $O = 8$. Other than the angle variation and the room conditions, the experimental settings listed in Table I were retained in this experiment. The SDR improvements are shown in Tables II and III. The better result for each angle is highlighted in bold. In Table II, the maximum SDR improvement among 10 trials with different initializations is given for each source angle θ. Regardless of the source angle, the results show consistent improvements upon adding band elements to the diagonal matrix. In contrast, Table III did not show the clear superiority of the tri-diagonal approach over the diagonal approach. We assume that the reasons for this are twofold. First, the Thomas algorithm used in the tri-diagonal approach is probably not as stable as diagonal division [37]. Second, since the tri-diagonal approach involves the approximation of the SCM, the error resulting from the optimization of (22) is not negligible.

D. Different Clustering Algorithms

To evaluate the self-clustering algorithm for the diagonal approach, we incorporated the state-of-the-art LPC-based clustering algorithm [43] into Algorithm 1 (Diag). Two other clustering methods [44] were also compared as baselines. The three clustering algorithms are denoted as, LPC k-medoid, MFCC k-means, and Mel-NMF. The SDR improvements for the different sources are shown in Fig. 8. On average, we can see that self-clustering has the best performance among the four clustering algorithms. Note that the SDR improvements for each instrument do not show a clear trend in the performance for different algorithms.

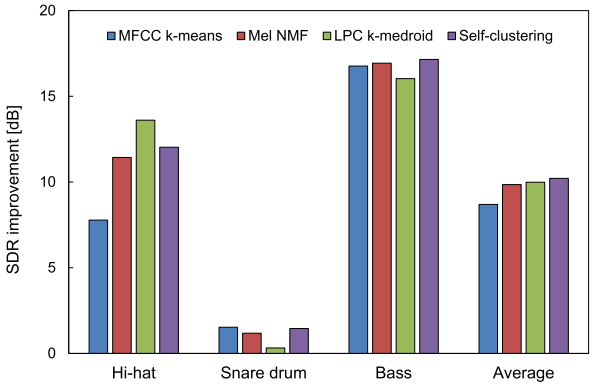

Fig. 8. Comparison of SDR improvement for various clustering methods. Three instrument signals (Hi-hat, Snare drum, Bass) obtained from SiSEC database were used to evaluate clustering methods.

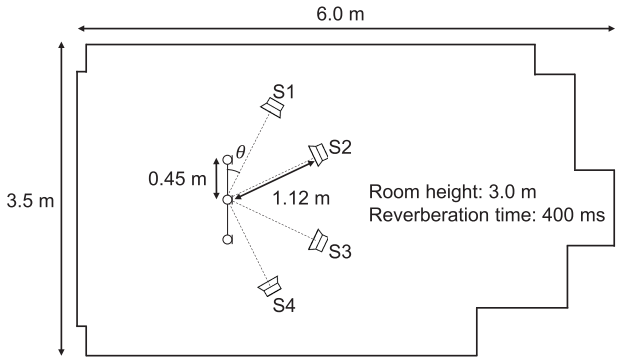

Fig. 9. The bird-view geometry of the common room used for the real data recordings in Section IV-E. Three microphones were linearly placed with a distance of 0.45 m while four loudspeakers were placed in a circle configuration with a radius of 1.12 m.

TABLE IV
RELATIONSHIP BETWEEN MUSICAL INSTRUMENTS AND LOUDSPEAKER POSITIONS

Musical instrument	Oboe	Flute	Piano	Trombone
Loudspeaker position	S1	S2	S3	S4
Angle θ [deg]	26.6	63.4	116.6	153.4

TABLE V
SDR IMPROVEMENTS UNDER THE CONDITIONS SHOWN IN FIG. 9

Approach	mean	max
ILRMA [34]	-1.44	0.47
MNMF [22]	2.65	4.83
Proposed method (Diag)	3.63	5.17
Proposed method (Tridiag)	**3.90**	**5.21**

TABLE VI
COMPUTATIONAL COSTS

Approach	time [s]
MVDR beamformer	2.9
IVA [33]	8.7
MNMF [22]	785.2
ILRMA [34]	22.9
DOANMF	2223.3
Proposed method (Diag)	68.0
Proposed method (Tridiag)	1447.7

E. Under-Determined Case With Real Recordings

To take into account data obtained from more realistic acoustic conditions, we recorded music sources consisting of four musical instruments emitted by four loudspeakers with box enclosures. Three omni-directional microphones are placed with a distance of 0.45 m. The four-second long sources of musical instruments were obtained from the songKitamura dataset.[2] Garritan Personal Orchestra 4 was chosen as MIDI source as it is considered more realistic than the other provided MIDI sources. More details about the dataset can be found in [50]. Four loudspeakers are placed in a circle with an angle θ clockwise with respect to the microphone array. The relationship between the musical instruments and the positions of the loudspeakers are listed in Table IV. The rough size of the room was 3.5 m $\times$ 6.0 m $\times$ 3.0 m and the reverberation time was 400 ms. The shape of the room can be found in Fig. 9. We chose four best performing methods in terms of SDR improvements shown in Section IV-B, i.e., ILRMA, MNMF, Algorithm 1 (Diag), and Algorithm 2 (Tridiag). The number of iterations was set to 500 to ensure convergence of the methods. Table V shows mean and maximum values of the SDR improvements if all four methods are run ten times. Algorithm 2 (Tridiag) outperformed the other three methods in both mean and maximum values. ILRMA does not perform as good as in Section IV-B as it is not designed for such an under-determined case.

F. Computation Time

The computation times for all the methods compared in Section IV-B were measured to investigate the efficiencies of the proposed algorithms. The measurement was carried out by inserting MATLAB time commands in the space before inputting the STFT signals and after outputting the separated time-domain signals. We ran the programs on a Xeon E5-2620 v4 CPU where each core has 2.1 GHz CPU capability. The number of DOA kernels was set to $O = 5$ as in Section IV-B. The results are listed in Table VI. MVDR beamformer, the non-iterative method, has the shortest computation time. By comparing Algorithm 1 (Diag) of order $O(M)$, Algorithm 2 (Tridiag) of order $O(M^2)$, and DOANMF of order $O(M^3)$, the efficiency of the proposed algorithms can be clearly observed. The speed of Algorithm 2 (Tridiag) can be further improved if a specialized matrix product that assumes that one of the two matrices is sparse is used instead of a standard matrix product. We can expect that the difference in efficiency between the algorithms will be even more significant when the number of microphones, M, is large. The diagonal approach and ILRMA have similar results because both algorithms are based on diagonal approximations of SCMs computed in the projected space. ILRMA is faster than the diagonal approach because the signals are projected to a more compact space, i.e., source space. In this experiment, the dimension of the projected space for ILRMA is equal to the number of sources, 3, whereas that for the diagonal approach is equal to the size of spatial DFT, 4.

[2][Online]. Available: http://d-kitamura.net/dataset.htm

V. Conclusion

To reduce the computational cost of multichannel NMF, mainly owing to the inverse calculation of an SCM, in this paper we proposed the conversion of STFT signals to the wavenumber domain, where the power of the SCM can be concentrated on diagonal elements, thus enabling us to apply element-wise divisions to obtain the inverse matrix. The approximation can reduce the computational complexity from order $O(M^3)$ to $O(M)$. The diagonal algorithm can be further extended by incorporating the self-clustering framework. To increase the accuracy of the matrix approximation, we also devised another extension that makes use of the upper and lower bands in addition to the diagonal elements. The inverse of the tri-diagonal matrix can be computed with a computational complexity of $O(M^2)$ by applying the Thomas algorithm. To ensure that the tri-diagonal matrix is positive semi-definite while truncating the off-diagonal elements, a semi-definite problem was solved. The experimental results show that our two approximations consistently gave better results in terms of SDR improvement than various unsupervised methods. The computation time measured by MATLAB time commands showed the greater efficiency of the tri-diagonal approach than multichannel NMF in the case of four channels, and we expect that this tendency will become more significant when the number of microphones is large.

Appendix

The upper bounds of the cost functions (19) and (25) are constructed by applying two inequalities to the convex part and the concave part, respectively [51]. This yields

$$C_{\mathrm{sp}}^{+}\left(\boldsymbol{\theta}, \mathbf{T}_{fnko}, \mathbf{U}_{fn}\right) = \sum_{fn}\left(\sum_{ko}\frac{\mathrm{tr}\left(\mathbf{R}_{fn}^{\mathrm{SP}}\mathbf{T}_{fnko}^{H}\left(\mathbf{J}_{fo}^{*}\right)^{-1}\mathbf{T}_{fnko}\right)}{z_{ko}w_{fk}h_{kn}} + \log\det\mathbf{U}_{fn} + \mathrm{tr}\left(\mathbf{U}_{fn}^{-1}\hat{\mathbf{R}}_{fn}^{*}\right) - M\right), \quad (29)$$

where $\mathbf{T}_{fnko}$ and $\mathbf{U}_{fn}$ are hidden variable matrices that satisfy $\sum_{ko}\mathbf{T}_{fnko} = \mathbf{I}, \mathbf{T}_{fnko} \succeq 0$, and $\mathbf{U}_{fn} \succeq 0$. The banded SCM and DOA kernel are denoted as $\hat{\mathbf{R}}_{fn}^{*}$ and $\mathbf{J}_{fo}^{*}$, respectively.

The partial derivatives with respect to z_{ko}, w_{fk}, and h_{kn} are derived by minimizing the upper bound function $C_{\mathrm{sp}}^{+}\left(\boldsymbol{\theta}, \mathbf{T}_{fnko}, \mathbf{U}_{fn}\right)$. The derivatives of $C_{\mathrm{sp}}^{+}\left(\boldsymbol{\theta}, \mathbf{T}_{fnko}, \mathbf{U}_{fn}\right)$ with respect to the model parameters are

$$\frac{\partial C_{\mathrm{sp}}^{+}}{\partial z_{ko}} = \sum_{fn}\left(-\frac{\mathrm{tr}\left(\mathbf{R}_{fn}^{\mathrm{SP}}\mathbf{T}_{fnko}^{H}\left(\mathbf{J}_{fo}^{*}\right)^{-1}\mathbf{T}_{fnko}\right)}{z_{ko}^{2}w_{fk}h_{kn}} + \mathrm{tr}\left(\mathbf{U}_{fn}^{-1}\mathbf{J}_{fo}^{*}\right)w_{fk}h_{kn}\right), \quad (30a)$$

$$\frac{\partial C_{\mathrm{sp}}^{+}}{\partial w_{fk}} = \sum_{no}\left(-\frac{\mathrm{tr}\left(\mathbf{R}_{fn}^{\mathrm{SP}}\mathbf{T}_{fnko}^{H}\left(\mathbf{J}_{fo}^{*}\right)^{-1}\mathbf{T}_{fnko}\right)}{z_{ko}w_{fk}^{2}h_{kn}} + \mathrm{tr}\left(\mathbf{U}_{fn}^{-1}\mathbf{J}_{fo}^{*}\right)z_{ko}h_{kn}\right), \quad (30b)$$

$$\frac{\partial C_{\mathrm{sp}}^{+}}{\partial h_{kn}} = \sum_{fo}\left(-\frac{\mathrm{tr}\left(\mathbf{R}_{fn}^{\mathrm{SP}}\mathbf{T}_{fnko}^{H}\left(\mathbf{J}_{fo}^{*}\right)^{-1}\mathbf{T}_{fnko}\right)}{z_{ko}w_{fk}h_{kn}^{2}} + \mathrm{tr}\left(\mathbf{U}_{fn}^{-1}\mathbf{J}_{fo}^{*}\right)z_{ko}w_{fk}\right). \quad (30c)$$

The equality of the auxiliary function and the cost function holds only when the hidden variables satisfy

$$\mathbf{T}_{fnko} = \left(\mathbf{J}_{fo}^{*}z_{ko}w_{fk}h_{kn}\right)\left(\hat{\mathbf{R}}_{fn}^{*}\right)^{-1}, \quad (31a)$$

$$\mathbf{U}_{fn} = \hat{\mathbf{R}}_{fn}^{*}. \quad (31b)$$

The multiplicative update rules for the model parameters z_{ko}, w_{fk}, and h_{kn} are obtained by equating the partial derivatives to zero. Equations. (21a)-(21c) for the diagonal case are the result of substituting $\hat{\mathbf{R}}_{fn}^{*}$ and $\mathbf{J}_{fo}^{*}$ by $\hat{\mathbf{R}}_{fn}^{\mathrm{Diag}}$ and $\mathbf{J}_{fo}^{\mathrm{Diag}}$, respectively, while (26a)-(26c) for the tri-diagonal case are obtained by substituting $\hat{\mathbf{R}}_{fn}^{*}$ and $\mathbf{J}_{fo}^{*}$ by $\hat{\mathbf{R}}_{fn}^{\mathrm{Tri}}$ and $\hat{\mathbf{J}}_{fo}^{\mathrm{Tri}}$, respectively.

Acknowledgment

The authors thank Mr. Yuki Kubo for collecting the multichannel recordings that are used in Section IV-E.

References

[1] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 2135–2139.

[2] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1652–1664, Sep. 2016.

[3] S. Uhlich *et al.*, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 261–265.

[4] S. I. Mimilakis, K. Drossos, T. Virtanen, and G. Schuller, "A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.

[5] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band DenseNets for audio source separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 21–25.

[6] N. Takahashi, N. Goswami, and Y. Mitsufuji, "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. Int. Workshop Acoust. Signal Enhanc.*, 2018, pp. 106–110.

[7] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2005, pp. 78–81.

[8] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local Gaussian modeling," in *Proc. 8th Int. Conf. Independent Compon. Anal. Signal Separ.*, 2009, pp. 775–782.

[9] 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.

[10] N. Q. 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.

[11] 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.

[12] D. Lee, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[13] S. Arberet *et al.*, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. 10th Int. Conf. Inf. Sci., Signal Process. Appl.*, 2010, pp. 1–4.

[14] A. Ozerov, E. Vincent, and F. Bimbot, "A general modular framework for audio source separation," in *Proc. Latent Variable Anal. Signal Separ.*, 2010, pp. 33–40.

[15] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1118–1133, May 2012.

[16] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Process.*, vol. 81, no. 11, pp. 2353–2362, 2001.

[17] E. Vincent, "Complex nonconvex l_p norm minimization for underdetermined source separation," in *Proc. 7th Int. Conf. Independent Compon. Anal. Signal Separ.*, 2007, pp. 430–437.

[18] A. Jourjine, S. Rickard, and Ö. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, pp. 2985–2988.

[19] S. Boyd and L. Vandenberghe, *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[20] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations – Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. New York, NY, USA: Wiley, 2009.

[21] A. Ozerov, C. Févotte, R. Blouet, and J. 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.

[22] 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.

[23] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Am. Stat.*, vol. 58, no. 1, pp. 30–37, 2004.

[24] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 22, no. 3, pp. 727–739, Mar. 2014.

[25] J. J. Carabias-Orti, J. Nikunen, T. Virtanen, and P. Vera-Candeas, "Multichannel blind sound source separation using spatial covariance model with level and time differences and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1512–1527, Sep. 2018.

[26] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation: Statistical insights and towards self-clustering of the spatial cues," in *Proc. Int. Symp. Comput. Music Model. Retriev.*, 2010, pp. 102–115.

[27] D. Fitzgerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Comput. Intell. Neurosci.*, vol. 2008, 2008, Art. no. 872425.

[28] Y. Mitsufuji and A. Roebel, "On the use of a spatial cue as prior information for stereo sound source separation based on spatially weighted non-negative tensor factorization," *EURASIP J. Adv. Signal Process.*, vol. 2014, pp. 1–9, 2014.

[29] Y. Mitsufuji, M. Liuni, A. Baker, and A. Roebel, "Online non-negative tensor deconvolution for source detection in 3DTV audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 3082–3086.

[30] M. Fakhry, P. Svaizer, and M. Omologo, "Audio source separation in reverberant environments using β-divergence-based nonnegative factorization," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 7, pp. 1462–1476, Jul. 2017.

[31] S. Lee, S. H. Park, and K. Sung, "Beamspace-domain multichannel nonnegative matrix factorization for audio source separation," *IEEE Signal Process. Lett.*, vol. 19, no. 1, pp. 43–46, Jan. 2012.

[32] D. Fitzgerald, A. Liutkus, and R. Badeau, "Projection-based demixing of spatial audio," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1560–1572, Sep. 2016.

[33] 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.

[34] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, Sep. 2016.

[35] T. Taniguchi and T. Masuda, "Linear demixed domain multichannel nonnegative matrix factorization for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 476–480.

[36] Y. Mitsufuji, S. Koyama, and H. Saruwatari, "Multichannel blind source separation based on non-negative tensor factorization in wavenumber domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 56–60.

[37] L. H. Thomas, "Elliptic problems in linear differential equations over a network," Tech. Rep., Watson Sci. Comput. Lab Report Columbia University New York, NY, 1949.

[38] J. Nikunen and T. Virtanen, "Multichannel audio separation by direction of arrival based spatial covariance model and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 6677–6681.

[39] B. Kulis, M. A. Sustik, and I. S. Dhillon, "Low-rank kernel learning with Bregman matrix divergences," *J. Mach. Learn. Res.*, vol. 10, pp. 341–376, 2009.

[40] J. Ahrens, *Analytic Methods of Sound Field Synthesis*. Berlin, Germany: Springer, 2012.

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

[42] S. M. Kay, *Fundamentals of Statistical Signal Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[43] X. Guo, S. Uhlich, and Y. Mitsufuji, "NMF-based blind source separation using a linear predictive coding error clustering criterion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 261–265.

[44] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *Proc. Int. Conf. Digit. Audio Effects*, 2009. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8336997

[45] J. Löfberg, "YALMIP: A toolbox for modeling and optimization in MATLAB," in *Proc. CACSD Conf.*, 2004, pp. 284–289.

[46] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optim. Methods Softw.*, vol. 11, no. 1-4, pp. 625–653, 1999.

[47] 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.

[48] 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 Eval.*, 2000, pp. 965–968.

[49] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 650–664, May 2009.

[50] 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.

[51] K. Kitamura, Y. Bando, K. Itoyama, and K. Yoshii, "Student's t multichannel nonnegative matrix factorization for blind source separation," in *Proc. IEEE Int. Workshop Acoust. Signal Enhanc.*, 2016, pp. 1–5.