Fast Multichannel Nonnegative Matrix Factorization with Directivity-Aware Jointly-Diagonalizable Spatial Covariance Matrices for Blind Source Separation

Kouhei Sekiguchi, *Member, IEEE*, Yoshiaki Bando, *Member, IEEE*, Aditya Arie Nugraha, *Member, IEEE*, Kazuyoshi Yoshii, *Member, IEEE*, and Tatsuya Kawahara, *Fellow, IEEE*

***Abstract*—This paper describes a computationally-efficient blind source separation (BSS) method based on the independence, low-rankness, and directivity of the sources. A typical approach to BSS is unsupervised learning of a probabilistic model that consists of a source model representing the time-frequency structure of source images and a spatial model representing their inter-channel covariance structure. Building upon the low-rank source model based on nonnegative matrix factorization (NMF), which has been considered to be effective for inter-frequency source alignment, multichannel NMF (MNMF) assumes source images to follow multivariate complex Gaussian distributions with unconstrained full-rank spatial covariance matrices (SCMs). An effective way of reducing the computational cost and initialization sensitivity of MNMF is to restrict the degree of freedom of SCMs. While a variant of MNMF called independent low-rank matrix analysis (ILRMA) severely restricts SCMs to rank-1 matrices under an idealized condition that only directional and less-echoic sources exist, we restrict SCMs to jointly-diagonalizable yet full-rank matrices in a frequency-wise manner, resulting in FastMNMF1. To help inter-frequency source alignment, we then propose FastMNMF2 that shares the directional feature of each source over all frequency bins. To explicitly consider the directivity or diffuseness of each source, we also propose rank-constrained FastMNMF that enables us to individually specify the ranks of SCMs. Our experiments showed the superiority of FastMNMF over MNMF and ILRMA in speech separation and the effectiveness of the rank constraint in speech enhancement.**

***Index Terms*—Blind source separation, multichannel nonnegative matrix factorization, joint diagonalization, full-rank spatial covariance matrix.**

I. Introduction

Multichannel source separation is one of the most fundamental techniques for computational auditory scene analysis including automatic speech recognition and acoustic event detection [1], [2]. As a front end of these tasks, deep neural networks (DNNs) are often trained by using pairs of mixture and isolated signals [3]–[7]. Although such a supervised approach works well in a known environment, it often fails to generalize to an unseen environment [8], [9].

Another approach is blind source separation (BSS) based on unsupervised learning of a probabilistic model that represents a multichannel mixture spectrogram as the sum of multichannel source spectrograms called *images*. Such a probabilistic model typically consists of a source model representing the time-frequency (TF) structure of source images and a spatial model representing their inter-channel covariance structure. In particular, the low-rank source model based on nonnegative matrix factorization (NMF) [10] has widely been used for mitigating the permutation problem, *i.e.*, source component alignment over all frequency bins. In a typical spatial model, the TF bins of each source image are assumed to independently follow multivariate complex Gaussian distributions with spatial covariance matrices (SCMs) [11].

In an *underdetermined* condition where the number of microphones M is less than that of sources N ($M \leq N$), multichannel nonnegative matrix factorization (MNMF) consisting of a low-rank source model and a full-rank spatial model has been investigated actively [12]–[17] (Fig. 1)[1]. Although the full-rank spatial model is capable of representing a wide variety of source directivity (diffuseness) under an echoic condition, MNMF tends to get stuck at bad local optima because a large number of unconstrained SCMs are estimated iteratively. MNMF thus suffers from the strong sensitivity to parameter initialization

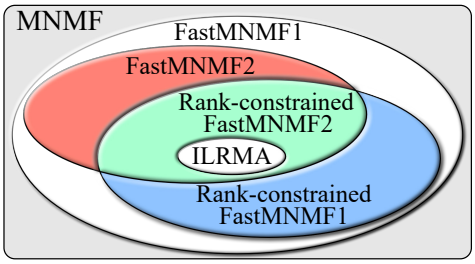


Fig. 1: The relations between MNMF and its variants.

Manuscript received XXXX XX, 2020; revised XXXX XX, 2020; accepted XXXX XX, 2020. Date of publication XXXX XX, 2020; date of current version XXXX XX, 2020. This work was supported by JSPS KAKENHI No. 19H04137, JST ERATO No. JPMJER1401, and NII CRIS collaborative Research Program. The associate editor coordinating the review of this manuscript and approving it for publication is XXX XXX. *(Corresponding author: Kouhei Sekiguchi.)*

K. Sekiguchi, A. A. Nugraha, and K. Yoshii are with the Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo 103-0027, Japan (email: {kouhei.sekiguchi, adityaarie.nugraha, kazuyoshi.yoshii}@riken.jp).

K. Sekiguchi, K. Yoshii, T. Kawahara are with the Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (email: {sekiguch, yoshii, kawahara}@sap.ist.i.kyoto-u.ac.jp).

Y. Bando is with the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, 135-0064, Japan (email: y.bando@aist.go.jp).

[1]In general, a full-rank SCM is used for representing each source image except for [12] using a full-rank noise SCM and rank-1 source SCMs, resulting in a full-rank spatial model for the mixture.

and a high computational cost. An effective solution to these problems is to restrict the degree of freedom of the SCMs. Under a less-echoic condition, the SCMs become close to rank-1 matrices and a *determined* ($M = N$) variant of MNMF with rank-1 SCMs called independent low-rank matrix analysis (ILRMA) was proposed [18]. Although ILRMA has been empirically known to work well for directional sources in practice, its performance is severely limited because the SCMs are not exactly rank-1 matrices in a real environment.

As an intermediate *(over)determined* ($M \geq N$) BSS method between MNMF and ILRMA, we proposed a computationally-efficient variant of MNMF called FastMNMF1 that restricts all source SCMs of each frequency bin to jointly-diagonalizable (JD) yet full-rank matrices [19]. Note that another FastMNMF1 based on the same formulation had been developed independently and concurrently [20]. To estimate the SCMs, we used a convergence-guaranteed iterative projection (IP) algorithm [21], while [20] used a fixed-point iteration (FPI) algorithm without convergence guarantee. FastMNMF1 is as fast as ILRMA and superior to ILRMA and MNMF in separation performance.

To reduce the initialization sensitivity of FastMNMF1, in this paper we propose a well-behaved constrained version of FastMNMF1 called FastMNMF2 that shares the directional feature of each source over all frequency bins. The JD spatial model of FastMNMF1 assumes that in each frequency bin, the SCM of each source is represented as the weighted sum of M common rank-1 SCMs expected to correspond to M different directions. While the direction weights of each source vary over frequency bins in FastMNMF1, they are shared over all bins in FastMNMF2. This idea has connection to *non-blind* direction-aware MNMF [15]–[17], which represents the SCM of each source as the weighted sum of SCMs precomputed for all possible directions. Our directivity-aware spatial model would mitigate the permutation problem, which has mainly been tackled by improving the source model.

To explicitly consider the directivity or diffuseness of each source, we further propose a rank-constrained version of FastMNMF1 or FastMNMF2 (collectively called FastMNMF) that enables us to individually specify the ranks of SCMs. When one or more people are talking in a noisy environment, for example, our goal is to separate an observed mixture into directional speech sources and diffuse noise sources. Such speech enhancement or separation can be achieved by initializing the direction weights of speech and noise sources to one-hot and all-one vectors, respectively, because the number of non-zero weights indicates the rank of an SCM. Through the iterative optimization, the speech SCMs are kept to rank-1 matrices and the noise SCMs to full-rank matrices thanks to the nature of the multiplicative update algorithm. If the SCMs of all M sources are restricted to rank-1 matrices in a determined case, rank-constrained FastMNMF reduces to ILRMA.

The main contribution of this paper is to propose *directivity-aware* full-rank spatial models and derive FastMNMF2 and its rank-constrained version. Moreover, we report comprehensive comparative evaluation of the conventional and proposed variants of MNMF. In a speech separation experiment, we show that FastMNMF2 works better than FastMNMF1, especially when a larger number of microphones are used or short-time Fourier transform (STFT) with a longer window is used, and the source model contains a fewer number of bases. We also investigate the computational cost and convergence speed of each variant and compare four methods of initializing the spatial models. In a speech enhancement experiment, we show the superiority of rank-constrained FastMNMF over a conventional ILRMA-based method that sequentially estimates the rank-1 SCMs of directional speech and the full-rank SCMs of diffuse noise in different steps [22].

The rest of the paper is organized as follows. Section II reviews related work on BSS methods. Section III explains the conventional methods based on the full-rank and rank-1 spatial models and Section IV describes the proposed methods based on the JD spatial models. Section V reports comparative experiments. Finally, Section VI concludes this paper.

II. Related Work

This section reviews existing BSS methods based on rank-1 and full-rank spatial covariance matrices for instantaneous mixing processes in which sources are propagated and superimposed instantaneously in the frequency domain. Let N and M be the number of sources and that of microphones, respectively.

A. BSS Methods Based on Rank-1 Spatial Models

BSS methods based on rank-1 spatial models can be used under a determined condition ($N = M$). The rank-1 spatial model can represent a *less-echoic* sound propagation process[2], where the reverberation time is shorter than the window length of STFT. In theory, the rank-1 spatial model is able to represent a uni-, multi-, or non-directional sound propagation process, *i.e.*, the same sound is allowed to arrive from multiple or all directions. In reality, however, such cases are unlikely to occur under a less-echoic condition and sound propagation processes are usually assumed to be uni-directional.

Frequency-domain independent component analysis (ICA) [23] is the most basic BSS method based on the independence of sources. A determined mixing process enables us to consider its inverse process called a demixing process. The goal of ICA is thus to estimate frequency-wise demixing matrices such that source spectrograms determined by multiplying the demixing matrices to a multichannel mixture spectrogram in a TF-wise manner are independent. Note that ICA is considered to have a simple source model that assumes the TF bins of each source to follow *independent univariate* complex non-Gaussian distributions. This causes the permutation problem because all frequency bins are processed independently.

To solve this problem, Kim *et al.* [24] proposed independent vector analysis (IVA) based on a modified source model that assumes the time frames of each source to follow *independent multivariate* complex non-Gaussian distributions. To accelerate and stabilize IVA, Ono [21] proposed a convergence-guaranteed parameter estimation method called iterative projection (IP). IVA has recently been extended for dealing with an overdetermined condition ($N < M$) [25], [26], where $M - N$ sources of

[2]Literally speaking, the term "rank-1" means that a multivariate complex Gaussian distribution has a rank-1 covariance matrix. We use the term "rank-1" in a wider sense to include ICA and IVA based on linear demixing matrices.

no interest in addition to N sources are internally considered to recover a determined condition in exchange for the acceptable increase of the computational cost.

To further mitigate the permutation problem left in IVA, Kitamura *et al.* [18] proposed independent low-rank matrix analysis (ILRMA) based on a low-rank source model that assumes the TF bins of each source to follow *conditionally-independent univariate* complex Gaussian distributions with power spectral densities (variances) factorized by NMF. Under a less-echoic condition, each TF bin of each source image can be said to follow a *degenerate multivariate* complex Gaussian distribution with a rank-1 SCM. For parameter estimation, an efficient convergence-guaranteed optimization algorithm iterating NMF and IP was derived. In spite of the severely restricted ability of the rank-1 spatial model, ILRMA is empirically known to work stably in real environments.

Several attempts have been made to enable ILRMA to deal with an overdetermined condition ($N < M$). Kitamura *et al.* [27] used ILRMA with M microphones for estimating M *components* clustered into N sources. Note that the component-source association should be specified in advance. In practice, $M = NP$ should be required for stable estimation, where P is the number of components associated to each source and represents the rank of the SCMs of the source. If $P = 2$, for example, each source would be represented by two components corresponding to direct and reflective propagation paths (multimodal directivity). Kubo *et al.* [22] used ILRMA for speech enhancement. Specifically, the rank-1 SCMs of directional speech and the rank-$(M-1)$ SCMs of diffuse noise are estimated with ILRMA and the missing rank-1 SCMs and power spectral densities (PSDs) of speech and noise are then estimated in an independent step. This method is similar to our rank-constrained FastMNMF in a sense that the rank of the SCMs of each source is specified explicitly according to its directivity. A key difference is that rank-constrained FastMNMF jointly estimate the SCMs of N sources with different ranks.

B. BSS Methods Based on Full-Rank Spatial Models

In theory, BSS methods based on full-rank spatial models can be used under either of determined ($N = M$), overdetermined ($M > N$), and underdetermined ($M < N$) conditions. In particular, the overdetermined condition is considered as the most important because it is often the case that at most two or three sources of interest are overlapped. From a practical point of view, more sources are considered to be hard to separate reasonably. The full-rank spatial model can represent an *echoic* sound propagation process, where each bin of each source image is assumed to follow a *multivariate* complex Gaussian distribution with a full-rank SCM.

Duong *et al.* [11] pioneered a BSS method based on the full-rank spatial model, which was called full-rank spatial covariance analysis (FCA) in [28], [29]. Because FCA has no specific source model, it suffers from the permutation problem like ICA. To alleviate this problem, multichannel NMF (MNMF) based on the low-rank source model has been developed [12]–[14]. ILRMA was originally derived by integrating the low-rank source model into IVA and was shown to be a special case of MNMF obtained by restricting the SCMs of sources to rank-1 matrices. Because in multichannel speech enhancement and separation, the low-rank assumption does not hold for speech spectrograms, semi-supervised BSS methods that use as a source model a deep generative model of speech trained from clean speech data [30]–[33] have been developed, inspired by monaural speech enhancement methods [34], [35]. Although a convergence-guaranteed closed-form iterative optimization algorithm has been developed for MNMF [20], it tends to easily get stuck at bad local optima because of the strong initialization sensitivity and suffers from the high computational cost because of the repeated heavy matrix operations.

In this paper we focus on the joint diagonalization of covariance matrices for accelerated computation. This idea was originally proposed for reducing the huge computational cost of a covariance-aware ultimate extension of NMF [10] called correlated tensor factorization (CTF) [36], resulting in independent low-rank tensor analysis (ILRTA a.k.a. FastCTF) [37] with an ILRMA-like convergence-guaranteed optimization algorithm. In application of CTF for monaural source separation, one needs to estimate the full-rank frequency and temporal covariance matrices of sources whose diagonal elements correspond to the basis spectra and temporal activations of NMF. In FastCTF, the frequency and temporal covariance matrices are assumed to be JD, respectively. Inspired by FastCTF and ILRMA, Ikeshita and Nakatani [38], [39] proposed multichannel BSS methods based on the JD full-rank frequency and temporal covariance matrices and the rank-1 SCMs.

Several studies have been proposed to restrict the SCMs of sources to JD yet full-rank matrices for multichannel BSS. Ito and Nakatani proposed a fast version of FCA called FastFCA [28], [29] and then proposed a fast version of MNMF called FastMNMF1 [20] independently and concurrently with our work [19]. To estimate a non-singular matrix called *diagonalizer* used for jointly diagonalizing the SCMs of sources at each frequency bin, we used a convergence-guaranteed IP method as in FastCTF [37], while a fixed point iteration (FPI) method without convergence guarantee was used in [20]. The diagonalizer can be estimated with a regularization technique that assumes the diagonalizer to be distributed around a demixing matrix estimated by ILRMA [40].

Several studies use fixed diagonalizers for efficient source separation in a transformed space [41]–[44]. Lee *et al.* [41], for example, used as the diagonalizer a beamspace transform matrix calculated from premeasured steering vectors. Mitsufuji *et al.* [42] proposed a variant of FastMNMF fixing the diagonalizer to the discrete Fourier transform (DFT) matrix for a straight-shape array of a large number (e.g., 32) of equally-spanned microphones. To relax this condition, the steering vectors of all possible directions were used in [44]. Taniguchi *et al.* [43] proposed another variant of FastMNMF and found that a demixing matrix estimated by IVA works better as the diagonalizer than the beamspace transform matrix.

III. Conventional BSS Methods

This section explains conventional BSS methods that integrate the low-rank source model with the full-rank or rank-1 spatial model.

A. Problem Specification

Suppose that a mixture of N sources are recorded by M microphones. Let $\mathbf{X} = \{\mathbf{x}_{ft}\}_{f=1,t=1}^{F,T} \in \mathbb{C}^{F\times T\times M}$ be the observed multichannel mixture spectrogram, where F and T represent the number of frequency bins and that of time frames, respectively. Let $\mathbf{S}_n = \{s_{ftn}\}_{f=1,t=1}^{F,T} \in \mathbb{C}^{F\times T}$ be the single-channel spectrogram of source n and $\mathbf{X}_n = \{\mathbf{x}_{ftn}\}_{f=1,t=1}^{F,T} \in \mathbb{C}^{F\times T\times M}$ be the *image* of source n. Assuming the additivity of complex spectra, $\mathbf{x}_{ft} \in \mathbb{C}^M$ is given by

$$\mathbf{x}_{ft} = \sum_{n=1}^{N} \mathbf{x}_{ftn}. \tag{1}$$

Given $\mathbf{X}$ as observed data, the goal of BSS is to estimate the latent source images $\{\mathbf{X}_n\}_{n=1}^N$ (not $\{\mathbf{S}_n\}_{n=1}^N$ in this paper).

B. Low-Rank Source Model

The source model represents a probabilistic generative process of the source spectrogram $\mathbf{S}_n$. We assume s_{ftn} to follow a univariate complex Gaussian distribution as follows:

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

where λ_{ftn} represents the PSD of source n at frequency f and time t. In the low-rank source model based on NMF, the PSDs $\{\lambda_{ftn}\}_{f,t=1}^{F,T}$ of each source n is assumed to have low-rank structure as follows:

$$\lambda_{ftn} = \sum_{k=1}^{K} w_{nkf} h_{nkt}, \tag{3}$$

where K is the number of bases, $w_{nkf} \geq 0$ is the magnitude of basis k of source n at frequency f, and $h_{nkt} \geq 0$ is the activation of basis k of source n at time t.

C. Spatial Models

The spatial model represents a probabilistic generative model of the source image $\mathbf{X}_n$.

1) Rank-1 Spatial Model: If the sound propagation process (room acoustics) is time-invariant, we have

$$\mathbf{x}_{ftn} = \mathbf{a}_{nf} s_{ftn}, \tag{4}$$

where $\mathbf{a}_{nf} \in \mathbb{C}^M$ is the steering vector of source n at frequency f. Using Eqs. (2) and (4), we have

$$\mathbf{x}_{ftn} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \lambda_{ftn}\mathbf{G}_{nf}) \triangleq \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{Y}_{ftn}), \tag{5}$$

where $\mathbf{G}_{nf} = \mathbf{a}_{nf}\mathbf{a}_{nf}^{\mathrm{H}} \in \mathbb{S}_+^M$ is the rank-1 SCM of source n at frequency f and $\mathbb{S}_+^M$ indicates the set of positive semidefinite matrices of size M. Using Eqs. (1) and (5) and the additive property of the Gaussian distribution, we have

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n=1}^{N} \lambda_{ftn}\mathbf{G}_{nf}\right) \triangleq \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{Y}_{ft}), \tag{6}$$

where $N \geq M$ is a necessary condition to keep Eq. (6) from being a degenerate distribution.

We then narrow our focus on a determined condition ($N = M$). Substituting Eq. (4) into Eq. (1), we get

$$\mathbf{x}_{ft} = \mathbf{A}_f \mathbf{s}_{ft}, \tag{7}$$

where $\mathbf{A}_f = [\mathbf{a}_{1f}, \ldots, \mathbf{a}_{Nf}] \in \mathbb{C}^{M\times N}$ is a non-singular square matrix called a mixing matrix. The source spectrum $\mathbf{s}_{ft} = [s_{ft1}, \cdots, s_{ftN}]^{\mathrm{T}}$ can be estimated directly as follows:

$$\mathbf{s}_{ft} = \mathbf{D}_f \mathbf{x}_{ft}, \tag{8}$$

where $\mathbf{D}_f = \mathbf{A}_f^{-1} \in \mathbb{C}^{N\times M}$ is a demixing matrix.

2) Full-Rank Spatial Model: The full-rank spatial model has the same formulation as Eq. (6) except that $\mathbf{G}_{nf}$ is a full-rank matrix. The source image $\mathbf{x}_{ftn}$ can be estimated with a Wiener filtering as follows:

$$\mathbf{x}_{ftn}|\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\Big(\mathbf{Y}_{ftn}\mathbf{Y}_{ft}^{-1}\mathbf{x}_{ft}, \mathbf{Y}_{ft} - \mathbf{Y}_{ftn}\mathbf{Y}_{ft}^{-1}\mathbf{Y}_{ftn}\Big), \tag{9}$$

$$\textit{i.e.,}\quad \mathbb{E}[\mathbf{x}_{ftn}|\mathbf{x}_{ft}] = \mathbf{Y}_{ftn}\mathbf{Y}_{ft}^{-1}\mathbf{x}_{ft}. \tag{10}$$

The full-rank spatial model can be used even under an underdetermined condition ($M < N$) unlike the rank-1 model, but its parameter estimation is harder because of the considerably larger number of parameters ($NFM(M+1)/2 \gg NFM$).

D. Integration of Source and Spatial Models

Two types of BSS methods, ILRMA [18] and MNMF [14] can be formulated by integrating the low-rank source model with the rank-1 and full-rank spatial models, respectively.

1) Independent Low-Rank Matrix Analysis: Using Eqs. (2), (3), and (8), the log-likelihood of the observed multichannel mixture spectrogram $\mathbf{X}$ is given by

$$\begin{aligned}\log p(\mathbf{X}|\mathbf{W},\mathbf{H},\mathbf{D}) &= \log p(\mathbf{S}|\mathbf{W},\mathbf{H},\mathbf{D}) + \sum_{f,t=1}^{F,T} \log\left|\frac{\partial \mathbf{s}_{ft}}{\partial \mathbf{x}_{ft}}\right| \\ &\overset{\mathrm{c}}{=} -\sum_{f,t,n=1}^{F,T,N}\left(\frac{|s_{ftn}|^2}{\lambda_{ftn}} + \log\lambda_{ftn}\right) + T\sum_{f=1}^{F}\log|\mathbf{D}_f\mathbf{D}_f^{\mathrm{H}}|,\end{aligned} \tag{11}$$

where $\mathbf{W} = \{w_{nkf}\}_{n,k,f=1}^{N,K,F}$, $\mathbf{H} = \{h_{nkt}\}_{n,k,t=1}^{N,K,T}$, and $\lambda_{ftn} = \sum_k w_{nkf}h_{nkt}$. The parameters $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{D}$ are estimated such that the log-likelihood function given by Eq. (11) is maximized.

2) Multichannel NMF: Using Eqs. (3) and (6), the log-likelihood of the observed spectrogram $\mathbf{X}$ is given by

$$\log p(\mathbf{X}|\mathbf{W},\mathbf{H},\mathbf{G}) \overset{\mathrm{c}}{=} -\sum_{f,t=1}^{F,T}\left(\mathrm{tr}\left(\mathbf{Y}_{ft}^{-1}\mathbf{X}_{ft}\right) + \log|\mathbf{Y}_{ft}|\right), \tag{12}$$

where $\mathbf{X}_{ft} = \mathbf{x}_{ft}\mathbf{x}_{ft}^{\mathrm{H}}$, $\mathbf{Y}_{ft} = \sum_{n,k} w_{nkf}h_{nkt}\mathbf{G}_{nf}$, and $\mathbf{G}_{nf}$ is a full-rank matrix. The parameters $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{G}$ are estimated such that the log-likelihood function given by Eq. (12) is maximized.

IV. Proposed Methods

This section explains the proposed BSS methods based on the low-rank source model and the jointly-diagonalizable (JD) full-rank spatial models. First, FastMNMF1 is derived from MNMF by restricting the SCMs to JD matrices at each frequency bin. Second, FastMNMF2 is derived from FastMNMF1 by sharing the directional feature of each source over all frequency bins. Finally, rank-constrained FastMNMF is derived such that the rank of the SCMs corresponding to each source can be specified explicitly according to the directivity of the source.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.3019181, IEEE/ACM Transactions on Audio, Speech, and Language Processing

A. FastMNMF1

We formulate the probabilistic model of FastMNMF1 and then derive an efficient parameter estimation algorithm based on iterations of nonnegative tensor factorization (NTF) and IP.

1) Model Formulation: The SCMs of N sources $\{\mathbf{G}_{nf}\}_{n=1}^{N}$ are assumed to be JD as follows:

$$\forall n\ \mathbf{G}_{nf} = \mathbf{Q}_f^{-1}\mathrm{Diag}(\tilde{\mathbf{g}}_{nf})\mathbf{Q}_f^{-\mathrm{H}}, \tag{13}$$

where $\tilde{\mathbf{g}}_{nf} = [\tilde{g}_{nf1}, \ldots, \tilde{g}_{nfM}] \in \mathbb{R}_+^M$ is a nonnegative vector, and $\mathbf{Q}_f = [\mathbf{q}_{f1}, \ldots, \mathbf{q}_{fM}]^{\mathrm{H}} \in \mathbb{C}^{M\times M}$ is a non-singular matrix called a *diagonalizer*, which is not limited to a unitary matrix. Substituting Eq. (13) into Eq. (6), we have

$$\mathbf{Q}_f\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n=1}^{N}\lambda_{ftn}\mathrm{Diag}(\tilde{\mathbf{g}}_{nf})\right). \tag{14}$$

This means that the elements of $\mathbf{Q}_f\mathbf{x}_{ft}$ are all independent. Regarding $\mathbf{Q}_f\mathbf{x}_{ft}$ as observed data, MNMF for $\mathbf{Q}_f\mathbf{x}_{ft}$ reduces to nonnegative tensor factorization (NTF) for the PSDs of $\mathbf{Q}_f\mathbf{x}_{ft}$, which can be performed efficiently (Fig. 2). The log-likelihood function of the parameters $\mathbf{W}$, $\mathbf{H}$, $\tilde{\mathbf{G}} = \{\tilde{\mathbf{g}}_{nf}\}_{n,f=1}^{N,F}$, and $\mathbf{Q}$ is then given by

$$\begin{aligned}
&\log p(\mathbf{X}|\mathbf{W},\mathbf{H},\tilde{\mathbf{G}},\mathbf{Q}) \\
&= \sum_{f,t=1}^{F,T}\log\mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{ft}\,\middle|\,\mathbf{0}, \sum_{n=1}^{N}\lambda_{ftn}\mathbf{Q}_f^{-1}\mathrm{Diag}(\tilde{\mathbf{g}}_{nf})\mathbf{Q}_f^{-\mathrm{H}}\right) \\
&\overset{\mathrm{c}}{=} -\sum_{f,t,m=1}^{F,T,M}\left(\frac{\tilde{x}_{ftm}}{\tilde{y}_{ftm}} + \log\tilde{y}_{ftm}\right) + T\sum_{f=1}^{F}\log\left|\mathbf{Q}_f\mathbf{Q}_f^{\mathrm{H}}\right|,
\end{aligned} \tag{15}$$

where $\tilde{x}_{ftm} = |\mathbf{q}_{fm}^{\mathrm{H}}\mathbf{x}_{ft}|^2$ and $\tilde{y}_{ftm} = \sum_{n,k} w_{nkf}h_{nkt}\tilde{g}_{nfm}$.

To avoid the scale ambiguity, we put normalization constraints on $\mathbf{Q}$, $\tilde{\mathbf{G}}$, and $\mathbf{W}$ as follows:

$$\mathbf{q}_{fm}^{\mathrm{H}}\mathbf{q}_{fm} = 1, \tag{16}$$

$$\sum_{m=1}^{M}\tilde{g}_{nfm} = 1, \tag{17}$$

$$\sum_{f=1}^{F}w_{nkf} = 1. \tag{18}$$

2) Source Separation: To estimate the source images $\mathbf{x}_{ftn} \in \mathbb{C}^M$, we use a Wiener filtering given by Eq. (10), which can be rewritten using Eq. (13) as follows:

$$\begin{aligned}
\mathbb{E}[\mathbf{x}_{ftn}|\mathbf{x}_{ft}] &= \mathbf{Y}_{ftn}\mathbf{Y}_{ft}^{-1}\mathbf{x}_{ft} \\
&= \mathbf{Q}_f^{-1}\mathrm{Diag}\left(\frac{\lambda_{ftn}\tilde{\mathbf{g}}_{nf}}{\sum_n\lambda_{ftn}\tilde{\mathbf{g}}_{nf}}\right)\mathbf{Q}_f\mathbf{x}_{ft}.
\end{aligned} \tag{19}$$

3) Parameter Estimation: Our goal is to jointly estimate $\mathbf{W}$, $\mathbf{H}$, $\tilde{\mathbf{G}}$, and $\mathbf{Q}$ such that the log-likelihood function given by Eq. (15) is maximized. Because Eq. (15) has the same form as the log-likelihood function of ILRMA given by Eq. (11), we can derive a convergence-guaranteed optimization algorithm based on iterations of NTF and IP in the same way as ILRMA based on iterations of NMF and IP.

Because the first term of Eq. (15) is the negative Itakura-Saito (IS) divergence between $\tilde{x}_{ftm}$ and $\tilde{y}_{ftm}$, the maximization of the log-likelihood with respect to $\mathbf{W}$, $\mathbf{H}$, and $\tilde{\mathbf{G}}$ is equivalent to NTF. The multiplicative update (MU) rules for $\mathbf{W}$, $\mathbf{H}$, and $\tilde{\mathbf{G}}$ are given by

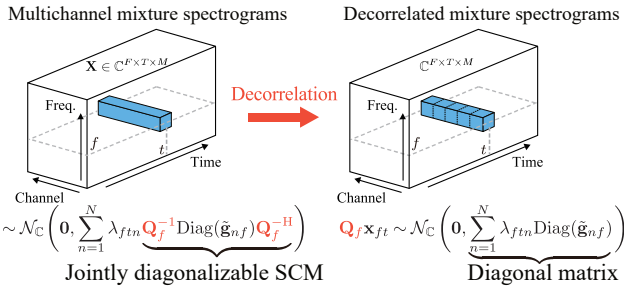


Fig. 2: The jointly-diagonalizable full-rank spatial model.

$$w_{nkf} \leftarrow w_{nkf}\sqrt{\frac{\sum_{t,m=1}^{T,M}h_{nkt}\tilde{g}_{nfm}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{t,m=1}^{T,M}h_{nkt}\tilde{g}_{nfm}\tilde{y}_{ftm}^{-1}}}, \tag{20}$$

$$h_{nkt} \leftarrow h_{nkt}\sqrt{\frac{\sum_{f,m=1}^{F,M}w_{nkf}\tilde{g}_{nfm}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{f,m=1}^{F,M}w_{nkf}\tilde{g}_{nfm}\tilde{y}_{ftm}^{-1}}}, \tag{21}$$

$$\tilde{g}_{nfm} \leftarrow \tilde{g}_{nfm}\sqrt{\frac{\sum_{t,k=1}^{T,K}w_{nkf}h_{nkt}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{t,k=1}^{T,K}w_{nkf}h_{nkt}\tilde{y}_{ftm}^{-1}}}. \tag{22}$$

As in IVA [21] and ILRMA [18], the IP rules of $\mathbf{Q}_f$ are given by

$$\mathbf{V}_{fm} \overset{\mathrm{def}}{=} \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_{ft}\tilde{y}_{ftm}^{-1}, \tag{23}$$

$$\mathbf{q}_{fm} \leftarrow (\mathbf{Q}_f\mathbf{V}_{fm})^{-1}\mathbf{e}_m, \tag{24}$$

$$\mathbf{q}_{fm} \leftarrow (\mathbf{q}_{fm}^{\mathrm{H}}\mathbf{V}_{fm}\mathbf{q}_{fm})^{-\frac{1}{2}}\mathbf{q}_{fm}, \tag{25}$$

where $\mathbf{e}_m$ is a one-hot vector whose m-th element is 1. The diagonalizer $\mathbf{Q}_f$ is estimated such that the M *components* of $\{\mathbf{Q}_f\mathbf{x}_{ft}\}_{f,t=1}^{F,T}$ become independent. In ILRMA under a determined condition ($N = M$), a demixing matrix $\mathbf{D}_f$ is estimated such that the M *sources* of $\{\mathbf{D}_f\mathbf{x}_{ft}\}_{f,t=1}^{F,T}$ become independent. Therefore, $\mathbf{Q}_f$ and $\mathbf{D}_f$ are estimated in the almost same way, and expected to play a similar role.

To satisfy the normalization constraints given by Eqs. (16), (17), and (18), we adjust the scales of $\mathbf{Q}$, $\tilde{\mathbf{G}}$, and $\mathbf{W}$ in this order in each iteration as follows:

$$\mu_{fm} = \mathbf{q}_{fm}^{\mathrm{H}}\mathbf{q}_{fm}, \quad \begin{cases}\mathbf{q}_{fm} \leftarrow \mu_{fm}^{-\frac{1}{2}}\mathbf{q}_{fm}, \\ \tilde{\mathbf{g}}_{nfm} \leftarrow \mu_{fm}^{-1}\tilde{\mathbf{g}}_{nfm},\end{cases} \tag{26}$$

$$\phi_{nf} = \sum_{m=1}^{M}\tilde{g}_{nfm}, \quad \begin{cases}\tilde{\mathbf{g}}_{nfm} \leftarrow \phi_{nf}^{-1}\tilde{\mathbf{g}}_{nfm}, \\ w_{nkf} \leftarrow \phi_{nf}w_{nkf},\end{cases} \tag{27}$$

$$\nu_{nk} = \sum_{f=1}^{F}w_{nkf}, \quad \begin{cases}w_{nkf} \leftarrow \nu_{nk}^{-1}w_{nkf}, \\ h_{nkt} \leftarrow \nu_{nk}h_{nkt}.\end{cases} \tag{28}$$

4) Physical Interpretation: We here discuss the physical interpretation of the joint-diagonalization constraint given by Eq. (13), which can be rewritten as

$$\mathbf{G}_{nf} = \mathbf{U}_f\mathrm{Diag}(\tilde{\mathbf{g}}_{nf})\mathbf{U}_f^{\mathrm{H}} = \sum_{m=1}^{M}\tilde{g}_{nfm}\mathbf{u}_{fm}\mathbf{u}_{fm}^{\mathrm{H}}, \tag{29}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.3019181, IEEE/ACM Transactions on Audio, Speech, and Language Processing

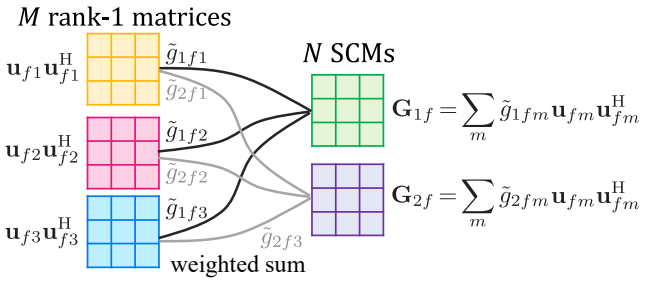


Fig. 3: Interpretation of the jointly-diagonalizable full-rank spatial model with $M = 3$ and $N = 2$.

where $\mathbf{U}_f = \mathbf{Q}_f^{-1} = [\mathbf{u}_{f1}, \cdots, \mathbf{u}_{fM}] \in \mathbb{C}^{M\times M}$. Eq. (29) means that $\mathbf{G}_{nf}$ is given as the weighted sum of M rank-1 matrices $\{\mathbf{U}_{fm} \triangleq \mathbf{u}_{fm}\mathbf{u}_{fm}^{\mathrm{H}}\}_{m=1}^{M}$, where the weights are given by $\tilde{\mathbf{g}}_{nf} = \{\tilde{g}_{nfm}\}_{m=1}^{M}$ (Fig. 3). The number of non-zero elements of $\tilde{\mathbf{g}}_{nf}$ thus indicates the rank of $\mathbf{G}_{nf}$.

We clarify the relation between FastMNMF1 and ILRMA under a determined condition ($N = M$). If $\tilde{\mathbf{g}}_{nf} = \mathbf{e}_n$, $\mathbf{G}_{nf}$ is a rank-1 matrix, and FastMNMF1 given by Eq. (15) reduces to ILRMA given by Eq. (11), where $\mathbf{Q}_f = \mathbf{D}_f$ ($\mathbf{U}_f = \mathbf{A}_f$) and $\tilde{y}_{ftm} = \sum_n \lambda_{ftn}\tilde{g}_{nfm} = \lambda_{ftm}$. This means the JD full-rank spatial model includes the rank-1 spatial model as its special case. Note that if $\tilde{\mathbf{g}}_{nf} = \mathbf{e}_n$, $\mathbf{u}_{fm}$ is equal to the steering vector of a certain direction. If $\tilde{\mathbf{g}}_{nf} \neq \mathbf{e}_n$, $\tilde{\mathbf{g}}_{nf}$ is considered to indicate the weights of M directions for source n. This hypothesis is experimentally validated in Section V-A.

We then discuss FastMNMF1 and ILRMA under an overdetermined condition ($N < M$). Thanks to the *soft* clustering of M directions into N sources, FastMNMF1 can directly deal with the overdetermined condition. Note that FastMNMF1 can be formulated mathematically unlike ILRMA even under an underdetermined condition ($N > M$), but does not work in practice because at most M directions can be covered by the JD full-rank spatial model. A possible way of using ILRMA under the overdetermined condition is to perform *hard* clustering of M components estimated by ILRMA into N sources [27]. This two-step method, however, is considered to be sub-optimal and the SCMs of N sources are rank-deficient. As a *soft*-clustering version of [27], one can formulate two-step FastMNMF1 that fixes $\mathbf{Q}_f$ to $\mathbf{D}_f$ estimated with ILRMA and estimates only $\mathbf{W}$, $\mathbf{H}$, and $\tilde{\mathbf{G}}$. This method can be considered as an extension of [43] using IVA for estimating $\mathbf{D}_f$. The superiority of FastMNMF over these two-step methods is experimentally validated in Sections V-E and V-G.

B. FastMNMF2

We formulate the probabilistic model of FastMNMF2 based on the weight-shared version of the JD full-rank spatial model and then derive a modified parameter estimation algorithm.

1) Model Formulation: In light of discussions described in Section IV-A4, we propose to make the direction weights $\tilde{\mathbf{g}}_{nf}$ of FastMNMF1 consistent over all frequency bins. More specifically, Eqs. (13) and (29) are replaced with

$$\forall n\ \mathbf{G}_{nf} = \mathbf{Q}_f^{-1}\mathrm{Diag}(\tilde{\mathbf{g}}_n)\mathbf{Q}_f^{-\mathrm{H}} = \sum_{m=1}^{M}\tilde{g}_{nm}\mathbf{u}_{fm}\mathbf{u}_{fm}^{\mathrm{H}}, \quad (30)$$

where $\tilde{\mathbf{g}}_n = [\tilde{g}_{n1}, \ldots, \tilde{g}_{nM}] \in \mathbb{R}_+^M$ is a *frequency-invariant* nonnegative vector. We refer to this model as the weight-shared jointly-diagonalizable (WJD) full-rank spatial model. Because $\tilde{\mathbf{g}}_n$ is estimated by taking all frequency bins into account, the permutation problem is expected to be mitigated, resulting in performance improvement from FastMNMF1.

Substituting Eq. (30) into Eq. (6), the log-likelihood function of the parameters $\mathbf{W}$, $\mathbf{H}$, $\tilde{\mathbf{G}}$, and $\mathbf{Q}$ is given by

$$\begin{aligned}&\log p(\mathbf{X}|\mathbf{W},\mathbf{H},\tilde{\mathbf{G}},\mathbf{Q})\\ &= \sum_{f,t=1}^{F,T}\log\mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{ft}\middle|\mathbf{0},\sum_{n=1}^{N}\lambda_{ftn}\mathbf{Q}_f^{-1}\mathrm{Diag}(\tilde{\mathbf{g}}_n)\mathbf{Q}_f^{-\mathrm{H}}\right)\\ &\overset{\mathrm{c}}{=} -\sum_{f,t,m=1}^{F,T,M}\left(\frac{\tilde{x}_{ftm}}{\tilde{y}_{ftm}}+\log\tilde{y}_{ftm}\right)+T\sum_{f=1}^{F}\log\left|\mathbf{Q}_f\mathbf{Q}_f^{\mathrm{H}}\right|, \quad (31)\end{aligned}$$

where $\tilde{x}_{ftm} = |\mathbf{q}_{fm}^{\mathrm{H}}\mathbf{x}_{ft}|^2$ and $\tilde{y}_{ftm} = \sum_{n,k} w_{nkf}h_{nkt}\tilde{g}_{nm}$.

To avoid the scale ambiguity, we put the normalization constraints given by Eq. (18) and

$$\sum_{m=1}^{M}\tilde{g}_{nm} = 1, \quad (32)$$

$$\mathrm{tr}\left(\mathbf{Q}_f\mathbf{Q}_f^{\mathrm{H}}\right) = M. \quad (33)$$

FastMNMF2 is a special case of FastMNMF1, and ILRMA is a special case of FastMNMF2. The numbers of parameters of MNMF, FastMNMF1, FastMNMF2, and ILRMA for SCMs are $FNM(M+1)/2$, FM^2+FNM, FM^2+NM, and FM^2, respectively. The computational times, convergence speeds, and performances of these methods with different values of N, M, K, and F are evaluated in Sections V-C and V-D.

2) Parameter Estimation: The parameters $\mathbf{W}$, $\mathbf{H}$, $\tilde{\mathbf{G}}$, and $\mathbf{Q}$ are updated in the same way as FastMNMF1. The MU update rules for $\mathbf{W}$, $\mathbf{H}$, and $\tilde{\mathbf{G}}$ are given by

$$w_{nkf} \leftarrow w_{nkf}\sqrt{\frac{\sum_{t,m=1}^{T,M}h_{nkt}\tilde{g}_{nm}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{t,m=1}^{T,M}h_{nkt}\tilde{g}_{nm}\tilde{y}_{ftm}^{-1}}}, \quad (34)$$

$$h_{nkt} \leftarrow h_{nkt}\sqrt{\frac{\sum_{f,m=1}^{F,M}w_{nkf}\tilde{g}_{nm}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{f,m=1}^{F,M}w_{nkf}\tilde{g}_{nm}\tilde{y}_{ftm}^{-1}}}, \quad (35)$$

$$\tilde{g}_{nm} \leftarrow \tilde{g}_{nm}\sqrt{\frac{\sum_{f,t,k=1}^{F,T,K}w_{nkf}h_{nkt}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{f,t,k=1}^{F,T,K}w_{nkf}h_{nkt}\tilde{y}_{ftm}^{-1}}}. \quad (36)$$

$\mathbf{Q}_f$ is updated in the same way as FastMNMF1 using Eq. (24). To satisfy the normalization constraints given by Eqs. (18), (32), and (33), we adjust the scales of $\mathbf{Q}$, $\tilde{\mathbf{G}}$, and $\mathbf{W}$ in this order in each iteration by using

$$\mu_f = \frac{1}{M}\mathrm{tr}(\mathbf{Q}_f\mathbf{Q}_f^{\mathrm{H}}), \quad \begin{cases}\mathbf{Q}_f \leftarrow \mu_f^{-\frac{1}{2}}\mathbf{Q}_f,\\ w_{nkf} \leftarrow \mu_f^{-1}w_{nkf},\end{cases} \quad (37)$$

$$\phi_n = \sum_{m=1}^{M}\tilde{g}_{nm}, \quad \begin{cases}\tilde{g}_{nm} \leftarrow \phi_n^{-1}\tilde{g}_{nm},\\ w_{nkf} \leftarrow \phi_n w_{nkf},\end{cases} \quad (38)$$

and Eq. (28).

3) Connection to Direction-Aware MNMF: FastMNMF2 has a connection to direction-aware MNMF [15]–[17] based on a factorizable full-rank spatial model given by

$$\mathbf{G}_{nf} = \sum_{d=1}^{D} z_{nd} \mathbf{R}_{fd}, \tag{39}$$

where D is the number of possible directions taken into account, $\mathbf{R}_{fd}$ is the SCM of direction d at frequency f, and z_{nd} is the weight of direction d for source n. Similarly to Eq. (30), Eq. (39) represents $\mathbf{G}_{nf}$ as the weighted sum of basis SCMs, and the weights are shared over all frequency bins. A difference is that direction-aware MNMF can be used under a non-blind condition; only the magnitude part of $\mathbf{R}_{fd}$ is estimated, while the phase part of $\mathbf{R}_{fd}$ is fixed to that of the geometrically-computed SCM of direction d at frequency f. This method thus tends to fail in an unseen acoustic environment.

C. Parameter Initialization

We explain four parameter initialization methods for FastMNMF, *i.e.*, random, diagonal, circular, and gradual initialization methods. The parameters $\mathbf{W}$ and $\mathbf{H}$ of the low-rank source model are initialized randomly and the parameters $\tilde{\mathbf{G}}$ and $\mathbf{Q}$ of the JD full-rank spatial model are initialized with one of the four methods. As experimentally shown in Section V-E, the gradual initialization method works best in practice. We here discuss FastMNMF1 under an (over)determined condition ($N \leq M$), which is considered to be practically important, because FastMNMF2 can be initialized in the same way.

1) Random Initialization: The diagonalizer $\mathbf{Q}_f$ is initialized to an identity matrix and $\tilde{\mathbf{g}}_{nf}$ is initialized randomly. Although FastMNMF is considered to be less sensitive to the initialization than MNMF because of the restricted model complexity, FastMNMF is still more likely to get stuck in bad local optima than ILRMA (constrained version of FastMNMF), when the random initialization method is used.

2) Diagonal Initialization: Inspired by the relation between FastMNMF1 and ILRMA discussed in Section IV-A4, $\mathbf{Q}_f$ is initialized to an identity matrix and $\tilde{\mathbf{g}}_{:f} \in \mathbb{R}^{N \times M}$ is initialized to a pseudo-diagonal matrix as follows:

$$\tilde{\mathbf{g}}_{:f} = \begin{pmatrix} 1 & \epsilon & \dots & \epsilon & \epsilon & \epsilon & \dots \\ \epsilon & 1 & \dots & \epsilon & \epsilon & \epsilon & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \\ \epsilon & \epsilon & \dots & 1 & \epsilon & \epsilon & \dots \end{pmatrix}, \tag{40}$$

where ':' indicates a set of all indices and ϵ is a small number (*e.g.*, $\epsilon = 10^{-2}$). Under a determined condition ($N = M$), $\tilde{\mathbf{g}}_{:f}$ is a square matrix close to an identity matrix. Although $\tilde{\mathbf{g}}_{nf}$ is updated iteratively, FastMNMF1 starting with $\tilde{\mathbf{g}}_{nf} \approx \mathbf{e}_n$ is expected to work as stably as ILRMA with $\tilde{\mathbf{g}}_{nf} = \mathbf{e}_n$. Under an overdetermined condition ($N < M$), however, the pseudo-demixing filters $\{\mathbf{q}_{fm}\}_{m=N+1}^{M}$ work ineffectively in the early iterations because the direction weights $\{\tilde{g}_{nfm}\}_{m=N+1}^{M}$ of each source n are small, *i.e.*, at most only N possible directions can be considered for N sources. In fact, we found that overdetermined FastMNMF1 with M microphones is comparable with determined FastMNMF1 using only the first N microphones, when the diagonal initialization method is used.

3) Circular Initialization: To solve the potential problem of the diagonal initialization under an overdetermined condition, $\mathbf{Q}_f$ is set to an identity matrix and $\tilde{\mathbf{g}}_{:f} \in \mathbb{R}^{N \times M}$ is set to a pseudo-circulant matrix as follows:

$$\tilde{\mathbf{g}}_{:f} = \begin{pmatrix} 1 & \epsilon & \dots & \epsilon & 1 & \epsilon & \dots \\ \epsilon & 1 & \dots & \epsilon & \epsilon & 1 & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \\ \epsilon & \epsilon & \dots & 1 & \epsilon & \epsilon & \dots \end{pmatrix}. \tag{41}$$

In this case, M possible directions are considered for N sources even in the early iterations and $M - N$ extra directions are gradually removed during the iterations.

4) Gradual Initialization: Inspired by the stable behavior of ILRMA with $K = 2$, FastMNMF1 with $K = 2$ is initialized by the circular initialization method. After updating $\mathbf{W}$, $\mathbf{H}$, $\mathbf{Q}$, and $\tilde{\mathbf{G}}$ 50 times, K is increased to a larger number and only $\mathbf{W}$ and $\mathbf{H}$ are reset to random values. This method was found to work stably among several possible implementations of gradual initialization.

In an overdetermined case ($N_{\text{true}} < M$), for example, another option is to first use determined FastMNMF1 ($N = M$) with $K = 2$ assuming the presence of $N - N_{\text{true}}$ extra sources (noise or reverberation), where N_{true} represents the number of true sources. After updating $\mathbf{W}$, $\mathbf{H}$, $\mathbf{Q}$, and $\tilde{\mathbf{G}}$ 50 times, K is set to a larger number and $\mathbf{W}$ and $\mathbf{H}$ are reset to random values. To obtain $\tilde{\mathbf{g}}_{:f} \in \mathbb{R}^{N_{\text{true}} \times M}$, one needs to select N_{true} rows from the estimated $\tilde{\mathbf{g}}_{:f} \in \mathbb{R}^{N \times M}$. Although this could be done as in rank-constrained FastMNMF (Section IV-D1), wrong selection degrades the separation performance. This option was thus not used in our experiment.

D. Rank-Constrained FastMNMF

We propose rank-constrained FastMNMF, a special case of FastMNMF that enables us to explicitly specify the rank of the SCMs $\{\mathbf{G}_{nf}\}_{f=1}^{F}$ of each source n according to its directivity. We here discuss rank-constrained FastMNMF1 because rank-constrained FastMNMF2 can be derived straightforwardly. As discussed in Section IV-A4, the number of non-zero elements of $\tilde{\mathbf{g}}_{nf}$ indicates the rank of $\mathbf{G}_{nf}$ in the JD full-rank spatial model. In the MU rule given by Eq. (22), once some elements of $\tilde{\mathbf{g}}_{nf}$ are set to zero, they are kept to zero. Rank-constrained FastMNMF1 can thus be obtained by initializing a specified number of elements of $\tilde{\mathbf{g}}_{nf}$ to zero, where the dimensions of those elements should be selected carefully for each source n according to the surrounding acoustic environment. Typically, $\tilde{\mathbf{g}}_{nf}$ is initialized to a one-hot or all-one vector for a directional or diffuse sound, respectively. We explain how to initialize rank-constrained FastMNMF1 for source enhancement or separation, where one or more directional sources (*e.g.*, speakers) exist with diffuse noise, respectively.

1) Source Separation: Suppose that there are L directional sources (target) and $N - L$ diffuse sources (noise). First, $\mathbf{Q}$ is initialized with the gradual initialization method described in Section IV-C4. Since $\mathbf{Q}_f$ is a pseudo-demixing matrix at frequency f, the spectrogram of component m is given by $\{\hat{x}_{ftm} = \mathbf{q}_{fm}^{\mathrm{H}} \mathbf{x}_{ft}\}_{f,t=1}^{F,T}$. Using the projection-back method [45] for solving the scale ambiguity of each component,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.3019181, IEEE/ACM Transactions on Audio, Speech, and Language Processing

the image of component m is given by $\{u_{fmm'}\hat{x}_{ftm}\}_{f,t,m'=1}^{F,T,M}$, where $\mathbf{u}_{fm}$ (column vectors of $\mathbf{Q}_f^{-1} = \mathbf{U}_f$) is a pseudo-steering vector of component m at frequency f. Let v_m be the maximum frame-wise power of component m, *i.e.*,

$$v_m = \max_t \sum_{f,m'=1}^{F,M} |u_{fmm'}\mathbf{q}_{fm}^{\mathrm{H}}\mathbf{x}_{ft}|^2. \tag{42}$$

The component indices $(1 \leq m \leq M)$ are then sorted in a descending order with respect to the significance $\{v_m\}_{m=1}^M$ and the M rows of $\mathbf{Q}_f$ are permuted accordingly. Assuming that the top L components correspond to the target $(L < N \leq M)$, $\{\tilde{\mathbf{g}}_{nf}\}_{n=1}^L$ and $\{\tilde{\mathbf{g}}_{nf}\}_{n=L+1}^N$ are initialized to one-hot and all-one vectors, respectively, as follows:

$$\tilde{\mathbf{g}}_{:f} = \begin{pmatrix} 1 & 0 & \dots & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & \dots & 0 \\ 1 & 1 & \dots & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \dots & 1 & \dots & 1 \end{pmatrix}. \tag{43}$$

2) Source Enhancement: Source enhancement is equivalent to source separation with $L = 1$. Assuming that the target source is predominant in the mixture, the pseudo-steering vectors $\mathbf{u}_{f1}$ and $\{\mathbf{u}_{fm}\}_{m=2}^M$ are initialized to the most principal and the remaining eigenvectors of the empirical SCM given by $\sum_t \mathbf{X}_{ft}$, respectively. Then, $\mathbf{Q}$ is initialized with the gradual initialization method described in Section IV-C4. The subsequent part is the same as source separation.

The key feature of rank-constrained FastMNMF is that the rank-1 SCMs of L directional target sources and the full-rank SCMs of $N - L$ diffuse noise sources are estimated jointly, where N is an arbitrary number and the noise sources are assumed to exist on M directions including the target directions. When ILRMA is used, in contrast, M directional sources are assumed to *exclusively* exist on M directions, resulting in L rank-1 target SCMs and a rank-$(M - L)$ noise SCM. An additional step is thus required for recovering the full-rank noise SCM [22]. The superiority of rank-constrained FastMNMF in speech enhancement and source separation is experimentally validated in Sections V-G and V-H.

V. Evaluation

This section reports comparative experiments conducted for evaluating the effectiveness of FastMNMF. First, we investigated the physical interpretation of the joint diagonalization constraint described in Section IV-A4. Second, we compared the separation performances and computational efficiencies of FastMNMF, ILRMA [18], and MNMF [14] for speech separation while the theoretical complexities of these methods are given in Section IV-B1. To draw the the full potential of FastMNMF, we comprehensively investigated the configuration of N, M, K, and F and compared the four initialization methods described in Section IV-C. Finally, we tested rank-constrained FastMNMF for speech enhancement and separation as described in Section IV-D. Through all experiments, audio signals were sampled at 16 kHz and processed by STFT with a shifting interval of 512 points and a Hann window of 2048 points ($F = 1025$), unless otherwise noted.

A. Validation of Directivity Awareness

We validated our hypothesis that $\tilde{\mathbf{g}}_{nf}$ indicates the weights of M directions for source n. If this is true, some column vectors of $\mathbf{U}_f = \mathbf{Q}_f^{-1}$ estimated by FastMNMF would coincide with the steering vectors of source directions because $\mathbf{U}_f$ can be regarded as a pseudo-mixing matrix.

1) Experimental Conditions: We investigated a determined case ($N = M = 2$) and an overdetermined case ($N = 2, M = 4$), where the sources and microphones were located as depicted in Fig. 4(a) and only the upper two microphones were used in the determined case. Using *Pyroomacoustics* library [46], we simulated the steering vectors $\{\mathbf{a}_{fd} \in \mathbb{C}^M\}_{d=1}^D$ of equally-spanned directions (azimuths; $D = 72$) with the reverberation time of $\mathrm{RT}_{60} = 100$ ms. We made an M-channel mixture signal of 6.9 seconds by spatially mixing two speech signals ($N = 2$) randomly selected from the WSJ-0 corpus [47] with the steering vectors $\mathbf{a}_{f1}$ and $\mathbf{a}_{f13}$ corresponding to the source directions (0 and 60 degrees).

FastMNMF1 with the circular initialization method and $K = 16$ was used for estimating the pseudo-demixing matrix $\mathbf{Q}_f = \mathbf{U}_f^{-1}$. The Euclidean distances between each pseudo-steering vector $\mathbf{u}_{fm}$ (the m-th column vector of $\mathbf{U}_f$) and the true steering vectors $\{\mathbf{a}_{fd}\}_{d=1}^D$ were computed at each frequency f and the average distance was computed over all frequency bins. Note that all vectors were normalized in advance such that the L2 norm of each vector is equal to 1 and the phase of the first channel was equal to zero.

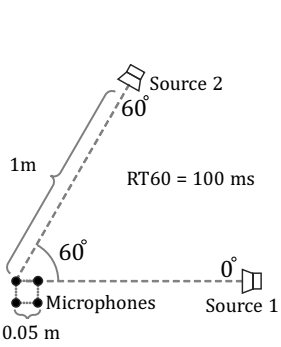
2) Experimental Results: As shown in Fig. 4(b), when $N = M = 2$, we observed the expected correspondence, *i.e.*, $\mathbf{u}_{f1} \approx \mathbf{a}_{f13}$ (60 degree) and $\mathbf{u}_{f2} \approx \mathbf{a}_{f1}$ (0 degrees). Note that $\mathbf{u}_{f1} \approx \mathbf{a}_{f61}$ (300 degrees) was also observed because of the front-back ambiguity of the straight-shape microphone array. We also confirmed that $\tilde{\mathbf{g}}_{1f} \approx [1, \epsilon]^{\mathrm{T}}$ and $\tilde{\mathbf{g}}_{2f} \approx [\epsilon, 1]^{\mathrm{T}}$ indicate the weights of the two directions. As shown in Fig. 4(c), when $N = 2$ and $M = 4$, we found $\mathbf{u}_{f3} \approx \mathbf{a}_{f13}$ (60 degrees) and $\mathbf{u}_{f4} \approx \mathbf{a}_{f1}$ (0 degree), and $\mathbf{u}_{f1}$ and $\mathbf{u}_{f2}$, which did not correspond to any of $\{\mathbf{a}_{fd}\}_{d=1}^D$, were considered to represent the non-directional reverberation of the two sources. In fact, $\tilde{\mathbf{g}}_{1f} \approx [\epsilon, \epsilon, 1, \epsilon]^{\mathrm{T}}$ and $\tilde{\mathbf{g}}_{2f} \approx [\epsilon, \epsilon, \epsilon, 1]^{\mathrm{T}}$ indicate the large weights of the two clear directions (the third and fourth dimensions corresponding to 60 and 0 degrees) and the small weights of the two vague directions (the first and second dimensions corresponding to the reverberation). This result clearly supports our hypothesis on the directivity awareness of FastMNMF1 and justify the inter-frequency weight sharing of FastMNMF2.

B. Basic Configurations for Speech Separation

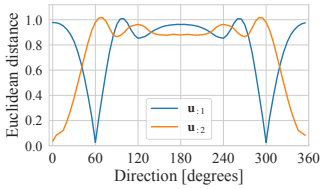
We compared the separation performances and computational efficiencies of FastMNMF, ILRMA [18], and MNMF [14] in a speech separation task. We randomly selected 100 simulated echoic three-speaker mixture signals ($N_{\text{true}} = 3, M = 8$) from the evaluation dataset of spatialized WSJ0-mix [7], where the positions of sources and microphones had been randomly

TABLE I: Elapsed times [sec] per iteration for processing 8ch signals of 10 [sec] on CPU (Intel Xeon W-2145 3.70 GHz).

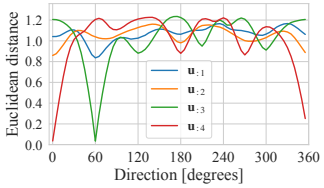
Method		ILRMA					FastMNMF1					FastMNMF2					MNMF				
# of bases K		2	4	16	64	256	2	4	16	64	256	2	4	16	64	256	2	4	16	64	256
# of sources N	3	-	-	-	-	-	1.61	1.60	1.69	2.02	3.47	1.58	1.60	1.65	1.99	3.39	21.7	21.7	21.8	22.1	23.7
	8	1.41	1.43	1.56	2.22	4.87	2.09	2.13	2.35	3.23	6.89	2.04	2.11	2.31	3.18	6.90	30.0	30.1	30.3	31.1	35.0



(a) Recording condition.



(b) $N = M = 2$.



(c) $M = 4$ and $N = 2$.

Fig. 4: The Euclidean distances between the estimated pseudo-steering vectors $\{\mathbf{u}_{fm}\}_{m=1}^{M}$ of M directions and the ground-truth steering vectors $\{\mathbf{a}_{fd}\}_{d=1}^{D}$ of all possible D directions accumulated over all frequency bins.

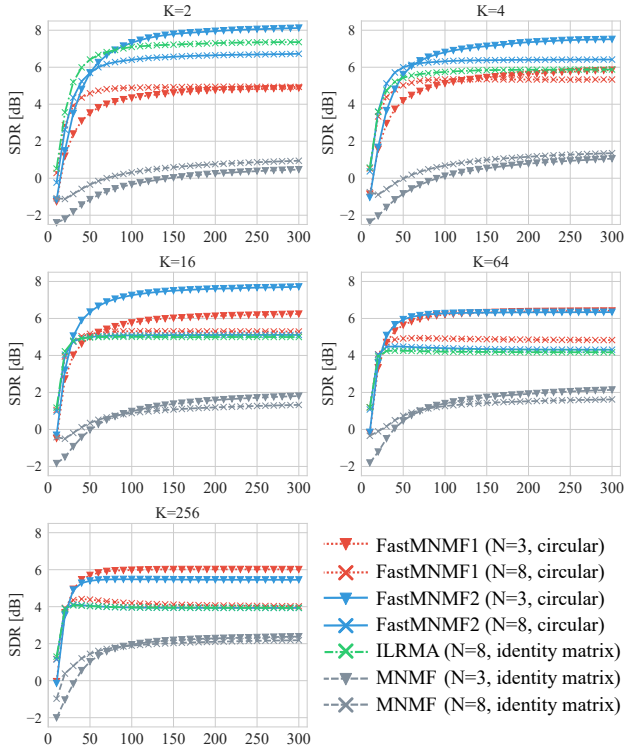


Fig. 5: The evolutions of average SDRs [dB]. Crosses and triangles indicate the determined and overdertermined settings, respectively.

determined for each mixture. The average SDR of the input mixture signals (the first channel) was -3.1 dB. FastMNMF and MNMF were directly tested with the overdetermined setting ($N = 3, M = 8$). In addition, all methods were tested with the determined setting ($N = M = 8$). For evaluation, N_{true} sources were selected from N estimated sources in a retrospective manner such that the SDR was maximized. Although this strategy was advantageous for the determined setting, we aimed to eliminate the impact of an arbitrary selection method and show the maximum potential of determined BSS methods including ILRMA. For the low-rank source model (Section III-B), $K \in \{2, 4, 16, 64, 256\}$ was used, and $\mathbf{W}$ and $\mathbf{H}$ were initialized randomly.

FastMNMF (Section IV) was initialized with the circular or gradual initialization methods (Sections IV-C3 and IV-C4). For ILRMA (Section III-D1), the demixing matrices $\mathbf{D}$ were initialized to identity matrices. Alternatively, ILRMA was initialized with the gradual initialization method because it was a special case of FastMNMF. For MNMF (Section III-D2), the SCMs $\mathbf{G}$ were initialized to identity matrices. Alternatively, MNMF was initialized with ILRMA, *i.e.*, the SCM of each dominant source n ($1 \leq n \leq N_{\text{true}}$) was given by $\mathbf{G}_{nf} = \mathbf{a}_{nf}\mathbf{a}_{nf}^{\mathrm{H}} + \epsilon\mathbf{I}$, where $\mathbf{a}_{nf}$ is a steering vector such that $\mathbf{A}_f = [\mathbf{a}_{1f}, \ldots, \mathbf{a}_{Nf}] = \mathbf{D}_f^{-1}$ and $\epsilon = 10^{-2}$ is a small number to make $\mathbf{G}_f$ a full-rank matrix. In all methods, the total number of iterations (including 50 iterations for initial FastMNMF with $K = 2$ in the gradual initialization method or 50 iterations for ILRMA in the initialization of MNMF) was set to 200.

The signal-to-distortion ratio (SDR) [48], [49] was used as a standard criterion for evaluating the separation performance.

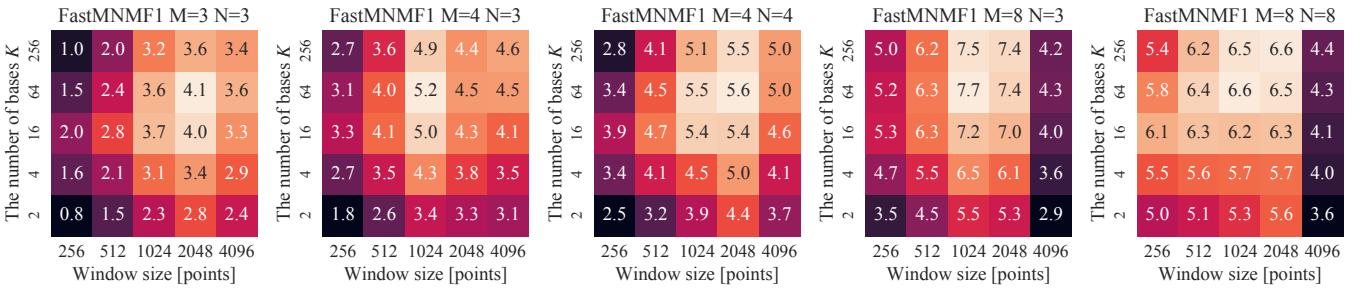
To investigate the convergence speed of each method, the SDR evolution was monitored along iterations. The elapsed time per iteration for processing a 10-s mixture signal was measured on Intel Xeon W-2145 (3.70 GHz).

C. Comparison of FastMNMF with ILRMA and MNMF

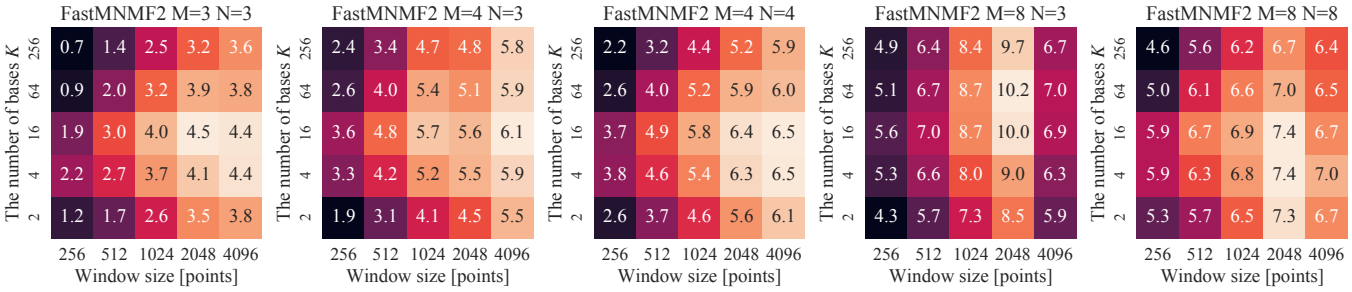
Table I lists the elapsed times per iteration. There was no significant difference between FastMNMF1 and FastMNMF2. FastMNMF was more than 10 or 5 times faster than MNMF for $K = 2$ or $K = 256$, respectively. An interesting finding was that ILRMA with $N = 8$ was 1.5 times faster than FastMNMF with $N = 8$ for any K, but almost as fast as FastMNMF with $N = 3$. Especially, FastMNMF with $N = 3$ with larger K tended to be faster than ILRMA. This indicates the effectiveness of considering only N_{true} sources in saving the computational cost under an overdetermined condition.

Fig. 5 shows the SDR evolutions of FastMNMF, ILRMA, and MNMF averaged over the 100 mixtures. In each method, the determined version converged faster than the overdetermined version, especially for $K \in \{2, 4, 16\}$. In FastMNMF1, $\mathbf{Q}$ was updated with IP and $\mathbf{W}$, $\mathbf{H}$, and $\tilde{\mathbf{G}}$ were updated with NTF, where at each frequency f and time t, M components

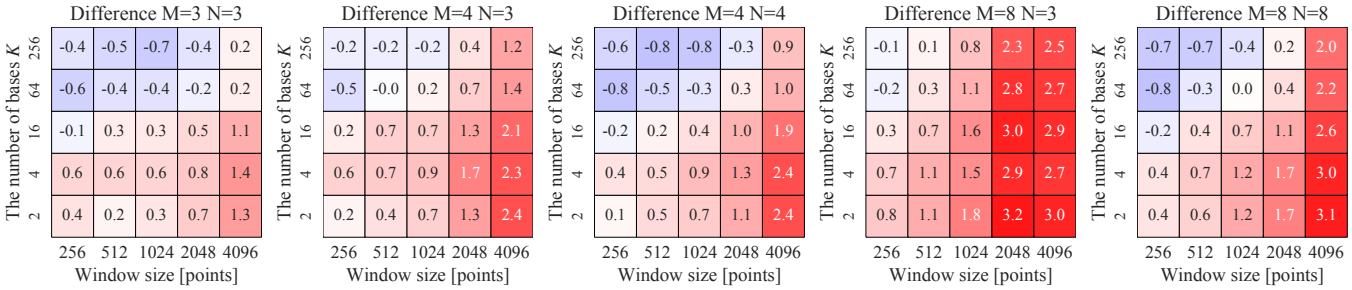
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.3019181, IEEE/ACM Transactions on Audio, Speech, and Language Processing



(a) SDRs obtained by FastMNMF1.



(b) SDRs obtained by FastMNMF2.



(c) SDR gaps between FastMNMF1 and FastMNMF2 (positive values indicate the superiority of FastMNMF2).

Fig. 6: The average SDRs obtained by FastMNMF1 and FastMNMF2 with gradual initialization.

$\{\tilde{x}_{ftm} = |\mathbf{q}_{fm}^{\mathrm{H}}\mathbf{x}_{ft}|^2\}_{m=1}^{M}$ were approximated by the weighted sums of N sources $\{\lambda_{ftn} = \sum_{k=1}^{K} w_{nfk}h_{nkt}\}_{n=1}^{N}$ with the weights $\{\tilde{g}_{nfm}\}_{n=1,m=1}^{N,M}$. When $N = M$, $\tilde{\mathbf{g}}_{nf}$ was initialized to a one-hot-like vector ($\tilde{\mathbf{g}}_{nf} \approx \mathbf{e}_n$) and M components were almost exclusively associated with N sources. $\mathbf{W}$ and $\mathbf{H}$ were thus estimated in an almost component- or source-wise manner, leading to faster convergence.

We found that FastMNMF with larger K converged faster, but did not necessarily work better. For larger K, the factorized model $\tilde{y}_{ftm} = \sum_{n,k} w_{nkf}h_{nkt}\tilde{g}_{nm}$ tended to overfit $\tilde{x}_{ftm} = |\mathbf{q}_{fm}^{\mathrm{H}}\mathbf{x}_{ft}|^2$ before $\mathbf{Q}$ was sufficiently estimated because of the rich expressive power of NMF with larger K. FastMNMF with larger K was thus likely to get stuck at local optima within several tens of iterations.

For any K, FastMNMF2 with $N = 3$ outperformed FastMNMF2 with $N = 8$, ILRMA, and MNMF. FastMNMF2 outperformed FastMNMF1 for $K \in \{2, 4, 16\}$, but slightly underperformed FastMNMF1 for $K \in \{64, 256\}$. Note that FastMNMF1 and FastMNMF2 are intermediate methods between MNMF and ILRMA and relatively closer to MNMF and ILRMA, respectively (Fig. 1). Because the PSDs of each source n should be estimated as finely as possible for estimating the SCMs of source n with a high degree of freedom, FastMNMF1 and MNMF tended to work better for larger K. The separation performance, however, was limited because of the insufficient optimization of $\mathbf{Q}$, as discussed above. Because the excessively-simplified low-rank PSDs of source n were sufficient for estimating the strictly-constrained SCMs of source n, FastMNMF2 and ILRMA worked better for smaller K.

D. Comparison of Model Complexities for FastMNMF

In the speech separation task, we comprehensively investigated the SDRs obtained by FastMNMF with different complexities. More specifically, we tested FastMNMF with $N \in \{3, 4, 8\}$ sources, $M \in \{3, 4, 8\}$ microphones ($N = M$ or $N = N_{\text{true}} = 3$), $K \in \{2, 4, 16, 64, 256\}$ bases, and $F \in \{129, 257, 513, 1025, 2049\}$ frequency bins, where STFT with a Hann window of $2(F-1)$ points and a shifting interval of $(F-1)/2$ points was used. In a determined setting, N_{true} sources were selected from N estimated sources for evaluation, as noted in Section V-B. To draw the full potential of FastMNMF, it was initialized with the gradual initialization method described in Section IV-C4.

Figs. 6(a) and 6(b) show the SDRs of FastMNMF1 and FastMNMF2 averaged over the 100 mixtures, respectively, and Fig. 6(c) shows the SDR gaps. FastMNMF2 with $N = 3$, $M = 8$, $K = 64$, and $F = 1025$ attained the highest SDR (10.2 dB).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.3019181, IEEE/ACM Transactions on Audio, Speech, and Language Processing

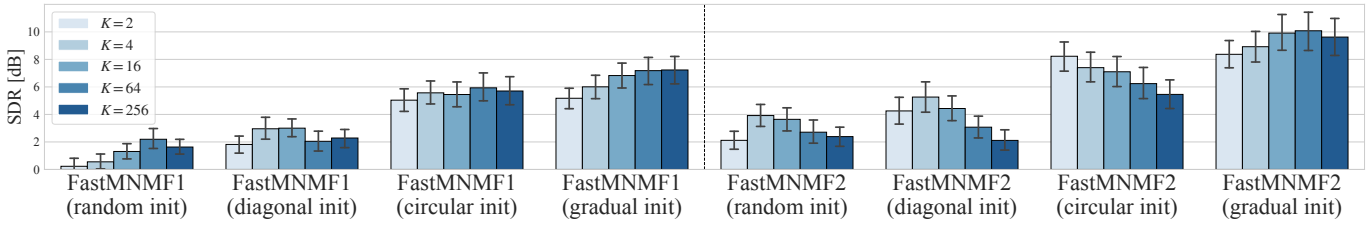


Fig. 7: SDRs obtained by FastMNMF with the four initialization methods.

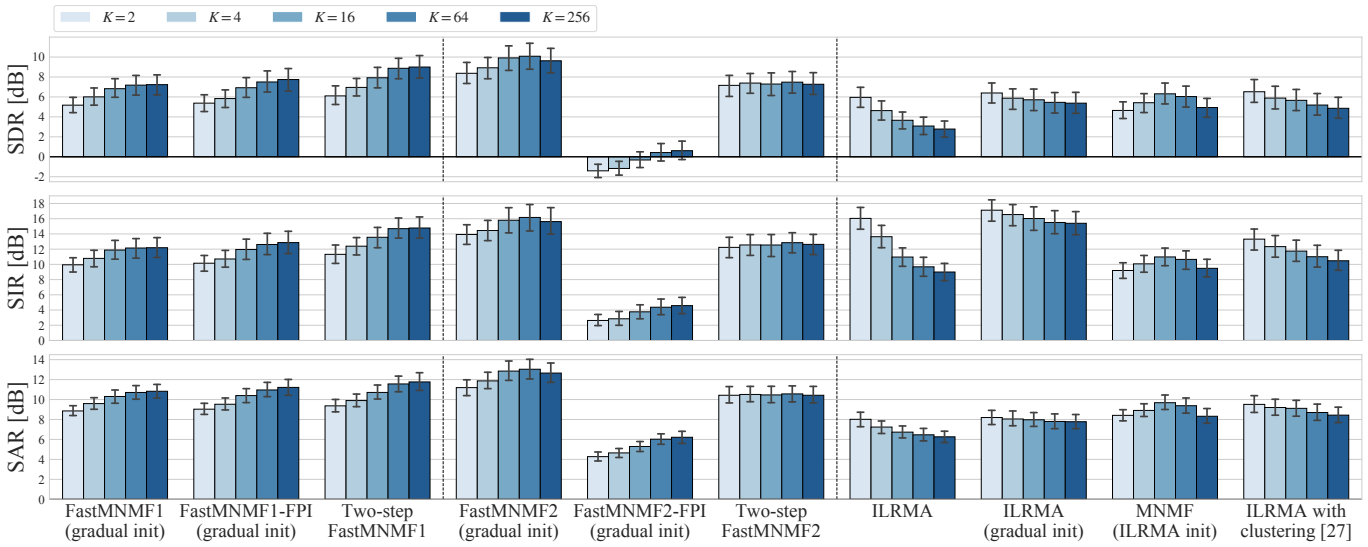


Fig. 8: SDRs, SIRs, and SARs obtained by FastMNMF, FastMNMF-FPI, Two-step FastMNMF, ILRMA, and MNMF.

For $K \in \{2, 4, 16\}$, FastMNMF2 outperformed FastMNMF1 in almost all settings. For larger M and F, FastMNMF2 noticeably outperformed FastMNMF1. Since the spatial models of FastMNMF1 and FastMNMF2 have $FM^2 + FNM$ and $FM^2 + NM$ parameters, respectively, FastMNMF1 with larger M and F tended to get stuck at bad local optima.

E. Comparison of Initialization Methods for FastMNMF

In the speech separation task, we further investigated the SDRs obtained by FastMNMF with the random, diagonal, circular, and gradual initialization methods described in Section IV-C. The model complexities were set to $N = 3$, $M = 8$, $K \in \{2, 4, 16, 64, 256\}$, and $F = 1025$.

Fig. 7 shows the average SDRs over the 100 mixtures. FastMNMF2 with $K = 64$ and the gradual initialization method achieved the highest SDR (10.2 dB). For any K, FastMNMF1 with the circular initialization method significantly outperformed FastMNMF1 with the random or diagonal initialization method. The same can be said for FastMNMF2. While FastMNMF2 with the circular initialization method worked better for smaller K, FastMNMF2 with the gradual initialization method worked better for larger K. This indicates that FastMNMF2 with $K = 2$ was effectively used for mitigating the initialization sensitivity of FastMNMF2 with larger K in the gradual initialization method.

F. Comparison with the State-of-the-Art BSS Methods

We compared the proposed FastMNMF with the IP method and another FastMNMF with the FPI method (FastMNMF1-FPI [20] and FastMNMF2-FPI). We also tested ILRMA with the component clustering mechanism [27], where ILRMA with $M = 8$ was used for estimating M components that were hardly clustered to $N = 4$ sources in advance ($P = 2$ components each) as described in Section IV-A4. Moreover, we tested the soft-clustering version of [27] called two-step FastMNMF described in Section IV-A4.

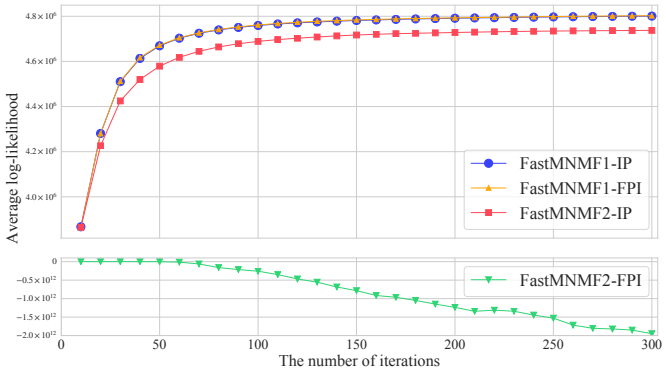


Fig. 9: The evolutions of the average log-likelihoods obtained by FastMNMF1, FastMNMF2, FastMNMF1-FPI, and FastMNMF2-FPI with the circular initialization.

Fig. 8 shows the average SDRs, SIRs, and SARs over the 100 mixtures. Because these three measures were consistent, we henceforth focus on the SDRs only. FastMNMF2 with the gradual initialization method (10.2 dB) outperformed two-step FastMNMF1 (9.1 dB), two-step FastMNMF2 (7.8 dB), and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.3019181, IEEE/ACM Transactions on Audio, Speech, and Language Processing

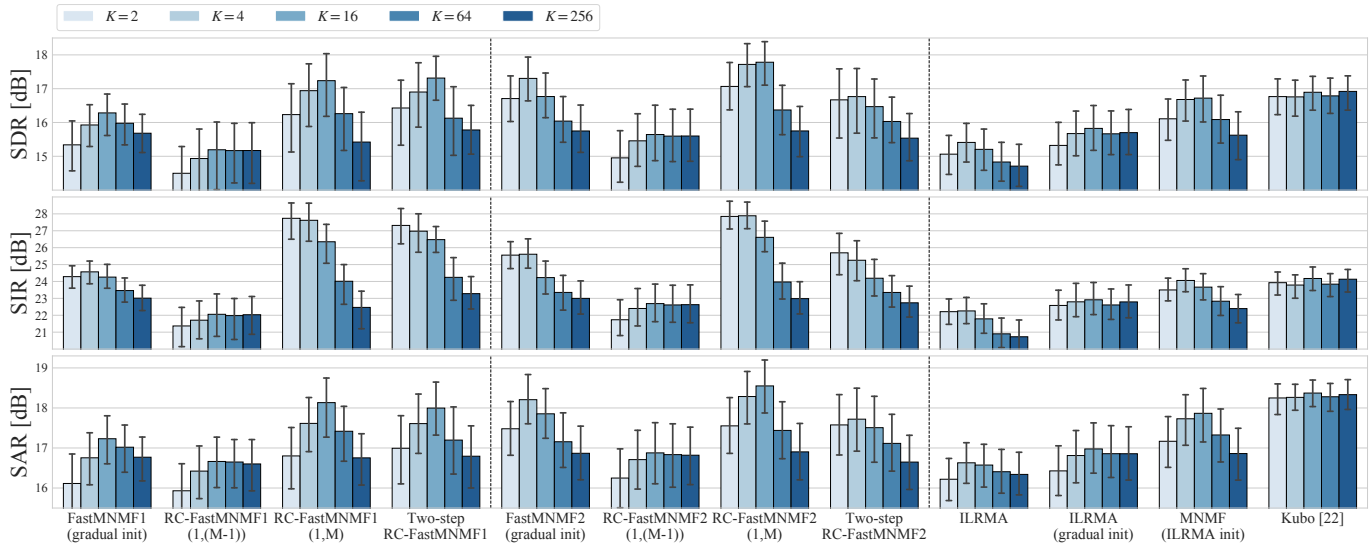


Fig. 10: SDRs, SIRs, and SARs obtained by rank-constrained FastMNMF, FastMNMF, ILRMA, and MNMF in speech enhancement.

ILRMA with the clustering mechanism (6.8 dB). This indicates that $\mathbf{Q}$ estimated by ILRMA was not optimal under the overdetermined condition and that the joint component separation and clustering was effective for improving the performance. A reason why two-step FastMNMF1 outperformed two-step FastMNMF2 was that when $\mathbf{Q}$ was fixed to one estimated by ILRMA, there was more room for performance improvement in FastMNMF1 with a higher degree of freedom in the subsequent step.

Comparing FastMNMF with FastMNMF-FPI, FastMNMF1-FPI (7.9 dB) performed slightly better than FastMNMF1 (7.4 dB), while FastMNMF2-FPI failed in all cases (1.1 dB). Fig. 9 shows the evolutions of the average log-likelihoods obtained by FastMNMF and FastMNMF-FPI with the circular initialization. The log-likelihoods of FastMNMF1 were almost same as those of FastMNMF1-FPI and higher than those of FastMNMF2 because FastMNMF1 has a higher degree of freedom than FastMNMF2. While the log-likelihoods of FastMNMF1, FastMNMF1-FPI, and FastMNMF2 increased monotonically, those of FastMNMF2-FPI tended to decrease because the FPI method has no guarantee to increase the likelihood.

G. Rank-Constrained FastMNMF for Speech Enhancement

We evaluated the effectiveness of rank-constrained FastMNMF in speech enhancement.

1) Experimental Conditions: We used the evaluation dataset of CHiME3 [50], which contains 1320 noisy speech signals simulated for a tablet with six microphones in four types of noisy environments: bus (BUS), cafe (CAF), pedestrian area (PED), and street junction (STR). 25 utterances were selected randomly for each environment (100 utterances in total). The average SDR of the input noisy signals (the fifth channel) was 7.5 dB. $M = 5$ channels excluding the second channel behind the tablet were used and $N = 5$ sources (one speech source and four noise sources) were assumed to exist. In this experiment, STFT with a shifting interval of 256 points and a Hann window of 1024 points was used ($F = 513$).

We evaluated the proposed rank-constrained FastMNMF, named RC-FastMNMF$_{(1,M)}$ (Section IV-D2), where the SCMs of source 1 (directional speech) were restricted to rank-1 matrices and those of source $n \in \{2, \cdots, N\}$ (diffuse noise) were left as full-rank matrices. For comparison, we tested RC-FastMNMF$_{(1,M-1)}$ with the rank-1 SCMs of source 1 and the rank-$(M-1)$ SCMs of source $n \in \{2, \cdots, N\}$ obtained by initializing $\tilde{\mathbf{g}}_{n(\geq 2)f}$ with $[0, 1, \ldots, 1]$. We also tested a speech enhancement method based on ILRMA [22]. Specifically, the rank-1 SCMs of speech and the rank-$(M-1)$ SCMs of noise were estimated with ILRMA and the missing rank-1 SCMs of noise and the PSDs of speech and noise were then estimated in an independent step. In addition, we tested two-step RC-FastMNMF, where $\mathbf{Q}$ was estimated with ILRMA and $\mathbf{W}$, $\mathbf{H}$, and $\tilde{\mathbf{G}}$ were then estimated (Section IV-A4). Note that [22] is similar to two-step RC-FastMNMF1. A main difference is that in [22] an inverse gamma prior distribution with a shape of 0.7 and a scale of 10^{-16} was put on the speech PSDs. These methods have the same advantage that the rank of the SCMs of each source can be specified explicitly according to its directivity.

As general-purpose BSS methods, we tested vanilla FastMNMF with the gradual initialization method, MNMF with ILRMA-based initialization, and ILRMA with the diagonal or gradual initialization method. For evaluation, the most dominant source in terms of the average power was selected as target speech from N estimated sources.

2) Experimental Results: Fig. 10 shows the SDRs, SIRs, and SARs of the compared methods averaged over the 100 utterances. In almost all versions of rank-constrained FastMNMF, the SDRs and SARs were maximized when $K = 16$, and the SIRs were maximized when $K = 2$. RC-FastMNMF2$_{(1,M)}$ achieved the highest SDR (17.8 dB) and outperformed RC-FastMNMF2$_{(1,M-1)}$ (15.6 dB) and ILRMA with $K = 16$ and the gradual initialization method (15.8 dB). Similarly, RC-FastMNMF1$_{(1,M)}$ (17.2 dB) outperformed RC-FastMNMF1$_{(1,M-1)}$ (15.2 dB). This indicates that the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.3019181, IEEE/ACM Transactions on Audio, Speech, and Language Processing

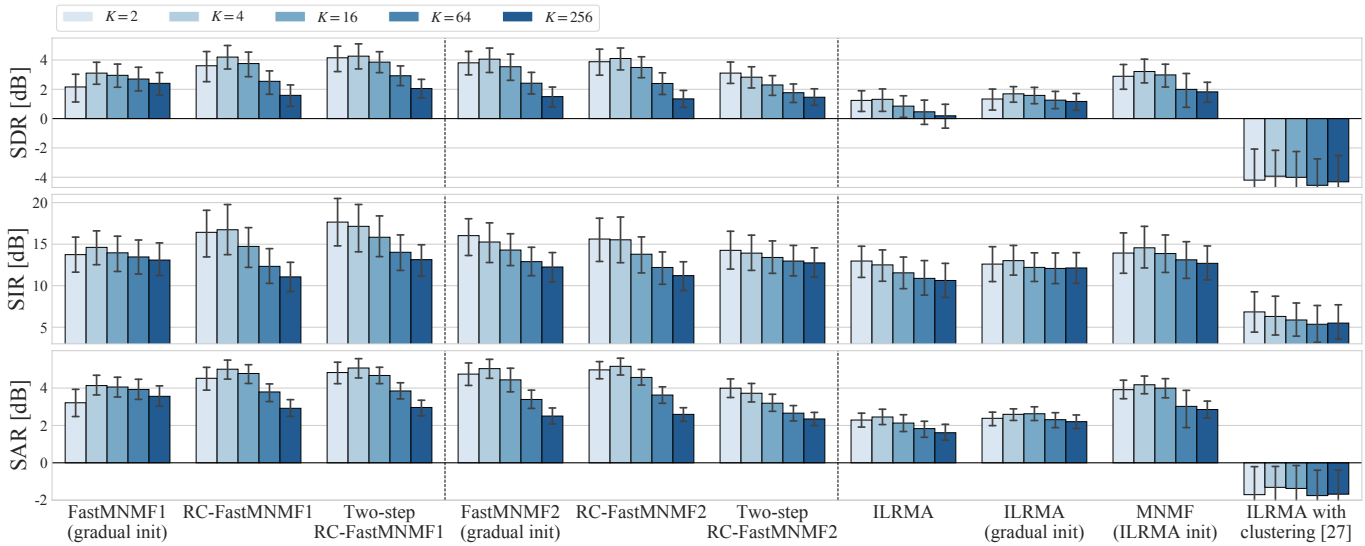


Fig. 11: SDRs, SIRs, and SARs obtained by rank-constrained FastMNMF, FastMNMF, ILRMA, and MNMF in speech separation.

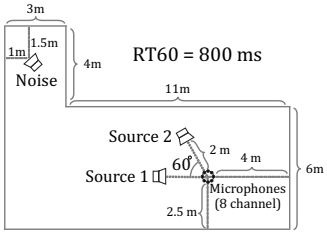


Fig. 12: Recording condition of the real data in Section V-H.

full-rankness of the noise SCMs was important for speech enhancement. RC-FastMNMF2$_{(1,M)}$ outperformed vanilla FastMNMF2 with $K = 4$ and the gradual initialization method (17.3 dB). When $\tilde{\mathbf{g}}_n$ of each source n was initialized to a one-hot-like vector, the noise SCMs estimated by FastMNMF2 were often close to rank-deficient matrices. RC-FastMNMF2$_{(1,M)}$ outperformed two-step RC-FastMNMF1 with $K = 16$ (17.3 dB) and two-step RC-FastMNMF2 with $K = 4$ (16.8 dB). This indicates the importance of jointly estimating $\mathbf{Q}$, $\mathbf{W}$, $\mathbf{H}$, and $\tilde{\mathbf{G}}$, as reported in Section V-F.

H. Rank-Constrained FastMNMF for Speech Separation

We evaluated the effectiveness of rank-constrained FastMNMF in speech separation using a dataset recorded in a real environment.

1) Experimental Conditions: An eight-channel microphone array ($M = 8$) and three loudspeakers corresponding to two speech sources and one noise source were put in a spacious, heavily-echoic room with RT$_{60}$ = 800 ms. (Fig. 12). The loudspeaker placed far away from the microphones was used for emitting a noise signal to the wall to simulate diffuse noise. We randomly selected 20 clean speech signals from the WSJ-0 corpus and four noise signals from the CHiME3 evaluation dataset. To obtain ground-truth images, these signals were recorded individually and 20 mixtures were synthesized by superimposing randomly-selected speech and noise signals, where the signal-to-noise ratio (SNR) was set to 0 dB. The average SDR of the input mixture signals (the first channel) was -4.1 dB.

We tested RC-FastMNMF, FastMNMF with the gradual initialization method, and two-step RC-FastMNMF, where $N = 8$ sources were assumed to exist in order to deal with heavy reverberation (a number of virtual sources were considered). In RC-FastMNMF and two-step RC-FastMNMF, the SCMs of two speech sources were restricted to rank-1 matrices and those of six noise sources were full-rank matrices. For comparison, we tested ILRMA with the diagonal or gradual initialization method, MNMF with ILRMA-based initialization, and ILRMA with the clustering mechanism [27]. For evaluation, two dominant sources in terms of the average power were selected from N sources as target speech sources.

2) Experimental Results: Fig. 11 shows the SDRs, SIRs, and SARs averaged over the 20 mixtures. In FastMNMF and RC-FastMNMF, the SDRs were maximized when $K = 4$. In this experiment, we found no significant difference between RC-FastMNMF2 and FastMNMF2 (4.1 dB) because the rank-1 assumption on the SCMs of speech was violated by the heavy reverberation, which was much longer than the STFT window size. In contrast, RC-FastMNMF1 (4.2 dB) outperformed FastMNMF1 (3.1 dB) because the inter-frequency direction weight sharing in two speech sources ($\tilde{\mathbf{g}}_{1f} = \mathbf{e}_n$ and $\tilde{\mathbf{g}}_{2f} = \mathbf{e}_2$ for any f) helped parameter estimation as in FastMNMF2. Two-step RC-FastMNMF1 with $K = 4$ (4.3 dB) outperformed RC-FastMNMF2 (4.1 dB), although $\mathbf{Q}$ estimated by ILRMA was considered to be sub-optimal, as discussed in Sections V-E and V-G. This indicates that RC-FastMNMF1 with $\tilde{\mathbf{g}}_{nf}$ might be more suitable than RC-FastMNMF2 with $\tilde{\mathbf{g}}_n$ for representing strong diffuse noise in a highly reverberant environment.

VI. Conclusion

In this paper, we proposed a versatile and computationally-efficient BSS method called FastMNMF based on directivity-aware jointly-diagonalizable full-rank SCMs. FastMNMF is a special case of MNMF based on unconstrained full-rank

SCMs [14]. More specifically, at each frequency bin, we represent the full-rank SCMs of sources as the weighted sums of common rank-1 matrices corresponding to different directions, resulting in FastMNMF1. Given that the directional feature of each source should be consistent over frequency bins, we make the direction weights of FastMNMF1 shared over frequency bins, resulting in FastMNMF2. To avoid bad local optima in iterative parameter optimization, we proposed and experimentally compared four initialization methods. To explicitly consider the directivity or diffuseness of each source, we further derived rank-constrained FastMNMF that enables us to individually specify the ranks of SCMs.

In a speech separation experiment, we confirmed that FastMNMF2 outperformed FastMNMF1, especially for larger numbers of microphones and frequency bins. We found that the circular and gradual initialization methods worked well. In a speech enhancement experiment, RC-FastMNMF2 with rank-1 speech SCMs and full-rank noise SCMs achieved the best performance. In the future, we plan to extend FastMNMF for joint separation and dereverberation.

REFERENCES

[1] S. Makino, Ed., *Audio Source Separation*. Springer, 2018.

[2] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*. Wiley, 2018.

[3] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7092–7096.

[4] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[5] H. Erdogan, J. Hershey, S. Watanabe, M. Mandel, and J. L. Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Interspeech*, 2016, pp. 1981–1985.

[6] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 196–200.

[7] Z.-Q. Wang, J. L. Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 1–5.

[8] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised speech enhancement based on multichannel NMF-informed beamforming for noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 5, pp. 960–971, 2019.

[9] A. Pandey and D. Wang, "On cross-corpus generalization of deep learning based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020, accepted for publication.

[10] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. with application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

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

[12] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[13] S. Arberet, A. Ozerov, N. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *ISSPA*, 2010, pp. 1–4.

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

[15] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 727–739, 2014.

[16] ——, "Multichannel audio separation by direction of arrival based spatial covariance model and non-negative matrix factorization," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 6677–6681.

[17] 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 Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1512–1527, 2018.

[18] 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 Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[19] K. Sekiguchi, A. A. Nugraha, Y. Bando, and K. Yoshii, "Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices," in *European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[20] N. Ito and T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 371–375.

[21] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 189–192.

[22] Y. Kubo, N. Takamune, D. Kitamura, and H. Saruwatari, "Efficient full-rank spatial covariance estimation using independent low-rank matrix analysis for blind source separation," in *European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[23] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, 2004.

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

[25] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 180–184.

[26] R. Ikeshita, T. Nakatani, and S. Araki, "Overdetermined independent vector analysis," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 591–595.

[27] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Relaxation of rank-1 spatial constraint in overdetermined blind source separation," in *European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1271–1275.

[28] N. Ito, S. Araki, and T. Nakatani, "FastFCA: A joint diagonalization based fast algorithm for audio source separation using a full-rank spatial covariance model," in *European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1667–1671.

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

[30] K. Sekiguchi, Y. Bando, K. Yoshii, and T. Kawahara, "Bayesian multichannel speech enhancement with a deep speech prior," in *APSIPA*, 2018, pp. 1233–1239.

[31] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 101–105.

[32] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Semi-supervised multichannel speech enhancement with a deep speech prior," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2197–2212, 2019.

[33] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neuro Computation*, vol. 31, no. 9, pp. 1–24, 2019.

[34] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 716–720.

[35] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *IEEE MLSP*, 2018, pp. 1–6.

[36] K. Yoshii, "Correlated tensor factorization for audio source separation," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 731–735.
[37] K. Yoshii, K. Kitamura, Y. Bando, E. Nakamura, and T. Kawahara, "Independent low-rank tensor analysis for audio source separation," in *European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1671–1675.
[38] R. Ikeshita, N. Ito, T. Nakatani, and H. Sawada, "A unifying framework for blind source separation based on a joint diagonalizability constraint," in *European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.
[39] ——, "Independent low-rank matrix analysis with decorrelation learning," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 288–292.
[40] K. Kamo, Y. Kubo, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Regularized fast multichannel nonnegative matrix factorization with ILRMA-based prior distribution of joint-diagonalization process," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 606–610.
[41] S. Lee, S. H. Park, and K. M. Sung, "Beamspace-domain multichannel nonnegative matrix factorization for audio source separation," *IEEE Signal Processing Letters*, vol. 19, no. 1, pp. 43–46, 2012.
[42] Y. Mitsufuji, S. Koyama, and H. Saruwatari, "Multichannel blind source separation based on non-negative tensor factorization in wavenumber domain," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2016, pp. 56–60.
[43] T. Taniguchi and T. Masuda, "Linear demixed domain multichannel nonnegative matrix factorization for speech enhancement," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 476–480.
[44] Y. Mitsufuji, S. Uhlich, N. Takamune, D. Kitamura, S. Koyama, and H. Saruwatari, "Multichannel non-negative matrix factorization using banded spatial covariance matrices in wavenumber domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 49–60, 2019.
[45] 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.
[46] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 351–355.
[47] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.
[48] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
[49] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "mir_eval: A transparent implementation of common MIR metrics," in *ISMIR*, 2014, pp. 367–372.
[50] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes," *Computer Speech & Language*, vol. 46, pp. 605–626, 2017.