Fast Multichannel Correlated Tensor Factorization for Blind Source Separation

Kazuyoshi Yoshii*† Kouhei Sekiguchi†* Yoshiaki Bando‡ Mathieu Fontaine† Aditya Arie Nugraha†
*Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
Email: yoshii@kuis.kyoto-u.ac.jp
†Center for Advanced Intelligence Project (AIP), RIKEN, Chuo-ku, Tokyo 103-0027, Japan
Email: {kouhei.sekiguchi,mathieu.fontaine,adityaarie.nugraha}@riken.jp
‡AIRC, National Institute of Advanced Industrial Science and Technology (AIST), Koto-ku, Tokyo 135-0064, Japan
Email: y.bando@aist.go.jp

***Abstract*—This paper describes an ultimate covariance-aware multichannel extension of nonnegative matrix factorization (NMF) for blind source separation (BSS). A typical approach to BSS is to integrate a low-rank source model with a full-rank spatial model as multichannel NMF (MNMF) based on full-rank spatial covariance matrices (CMs) or its efficient version named FastMNMF based on jointly-diagonalizable spatial CMs do. The NMF-based phase-unaware source model, however, can deal with only the *positive cooccurrence* relations between time-frequency bins. To overcome this limitation, we propose an efficient multichannel extension of correlated tensor factorization (CTF) named FastMCTF based on jointly-diagonalizable temporal, frequency, and spatial CMs. Integration of the jointly-diagonalizable full-rank source model proposed by FastCTF with the jointly-diagonalizable full-rank spatial model proposed by FastMNMF enables us to completely consider the *positive and negative covariance* relations between frequency bins, time frames, and channels. We derive a convergence-guaranteed parameter estimation algorithm based on the multiplicative update and iterative projection and experimentally show the potential of the proposed method.**

I. Introduction

Blind source separation (BSS) of multichannel mixture spectrograms (complex-valued three-mode tensor having the time, frequency, and channel axes) plays a vital role for audio event detection and noisy speech recognition. In general, the inter-channel (spatial) covariance structure has mainly been used for BSS. In independent component analysis (ICA) [1] assuming that the time-frequency (TF) bins of source spectrograms to follow independent non-Gaussian distributions, for example, a linear demixing filter (transform matrix) is estimated at each frequency such that the channels are made independent after applying the demixing filter to the mixture spectrograms. ICA, however, suffers from the permutation problem; one needs to align the orders of sources between frequency bins by focusing on the temporal, frequency, and spatial features of sources.

The permutation problem has recently been tackled concurrently with spatial filtering by integrating a *source model* representing the TF structure of sources with a *spatial model* representing the inter-channel structure of source images. To represent the higher-order inter-frequency correlations of source spectra, an extension of ICA called independent vector analysis (IVA) [2], [3] assumes source spectra to follow multivariate complex non-Gaussian distributions. If particular spectral patterns (*e.g.*, musical instrument sounds) appear repeatedly in a source spectrogram, its power spectral densities (PSDs) have low-rank structure and can thus be approximated as the sum of products between a fewer number of basis PSD patterns and that of the corresponding temporal envelopes in the framework of nonnegative matrix factorization (NMF) [4].

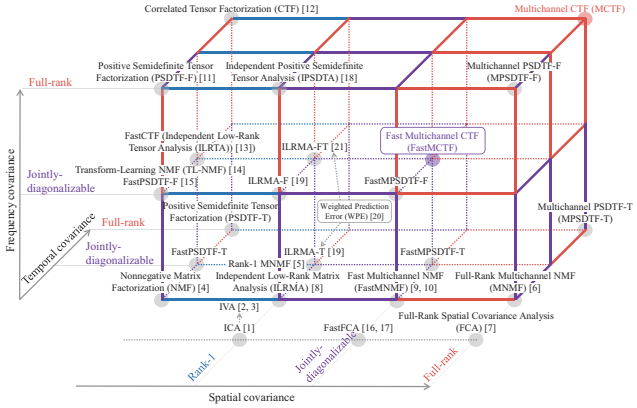

Fig. 1. An overview of BSS methods based on time, frequency, and spatial covariance matrices.

A versatile BSS method called Multichannel NMF (MNMF) integrates an NMF-based source model with a spatial model using rank-1 or full-rank spatial covariance matrices (CMs) [5], [6] Although the full-rank spatial model can represent reverberant and diffuse sounds [7], MNMF tends to get stuck at bad local optima because of the large degree of freedom. To mitigate this problem, spatial CMs are restricted to rank-1 and jointly-diagonalizable matrices in independent low-rank matrix analysis (ILRMA) [8] and an efficient version of MNMF called FastMNMF [9], [10], respectively. A common feature of these methods is that spatial filters are estimated such that source spectrograms are made independent *and* low-rank.

The fundamental limitation of the NMF-based source model based on the *positive coocurrence* relations between TF bins is that the *positive and negative covariance* relations, which could help to solve the permutation problem, are not considered. To overcome this limitation, one can use a covariance-aware extension of NMF such as positive semidefinite tensor factorization (PSDTF) [11] and correlated tensor factorization (CTF) [12]. In CTF, the empirical CM over all TF bins is

approximated as the sum of Kronecker products between a fewer number of temporal CMs and that of the corresponding frequency CMs. Although CTF is an ultimate low-rank decomposition technique, its over-parametrization nature and huge computational complexity prevent stable parameter estimation. To solve this problem, an efficient version of CTF named independent low-rank tensor analysis (ILRTA) [13] (a.k.a. FastCTF) based on jointly-diagonalizable temporal and frequency CMs was proposed for single-channel BSS.

In this paper we propose an efficient BSS method named fast multichannel CTF (FastMCTF) that integrates a FastCTF-based source model using *jointly-diagonalizable temporal and frequency* CMs with a full-rank spatial model using *jointly-diagonalizable spatial* CMs. FastMCTF can thus be viewed as a covariance-aware extension of FastMNMF and as a multichannel extension of FastCTF (Fig. 1). Using the minorization-maximization principle, we derive a convergence-guaranteed covariance estimation algorithm consisting of multiplicative update and iterative projection [3] as in [13]. Finally, the time-frequency-channel (TFC) elements of each source image are inferred at once by decomposing a multichannel mixture spectrogram with a big Wiener filter based on the huge TFC CMs of source images. This process is equivalent to *element-wise* Wiener filters for a linearly-transformed mixture spectrogram, in which all elements are made independent and low-rank thanks to the joint diagonalizability of temporal, frequency, and spatial CMs.

II. Related Work

This section reviews single- and multi-channel BSS methods in terms of temporal, frequency, spatial covariance modeling (Fig. 1). Versatile BSS methods have been developed based on the independence, low-rankness, and/or nonnegativity of audio signals. Let M, F, and T be the number of channels, that of frequency bins, and that of frames.

Most studies assume the independence of TF bins. In theory, discrete Fourier transform (DFT) of infinitely-long stationary signals yields independent frequency components. In reality, when short-time Fourier transform (STFT) is applied to non-stationary signals, the inter-frequency covariance is induced by the non-stationary characteristics of a target signal, which could be a useful clue for BSS. In addition, adjacent frames are strongly correlated with each other because STFT yields a redundant representation of a target signal.

A. Single-channel BSS

A standard approach to single-channel BSS is to use NMF based on the independence and low-rankness of sources. Under an assumption that source spectrograms follow complex Gaussian distributions, NMF based on the Itakura-Saito divergence (IS-NMF) [4] is theoretically justified. Using TF-wise Wiener filters based on the parameters of NMF, a mixture spectrogram is decomposed into the sum of source spectrograms while the phase information is not changed. The quality of separated source signals is thus limited.

To solve this problem, we have developed covariance-aware BSS methods based on the independence and positive semidefiniteness of sources. While NMF [4] assumes that each *time-frequency bin* of a source spectrogram follows a complex Gaussian distribution, PSDTF [11] assumes that each *time or frequency axis-aligned slice* follows a multivariate complex Gaussian distribution with a temporal or frequency CM (PSDTF-T or PSDTF-F). CTF [12] assumes that the *entire* source spectrogram follows a multivariate complex Gaussian distribution. Using a Wiener filter with the huge TF CMs of source spectrograms, complex source spectrograms with appropriate phase information can be estimated. Although CTF is a mathematically ultimate extension of NMF, it is computationally prohibitive and extremely sensitive to initialization because of its over-parametrization nature. The computational complexities of NMF, PSDTF-T, PSDTF-F, CTF are given by $O(TF)$, $O(T^3F)$, $O(TF^3)$, and $O(T^3F^3)$, respectively.

To solve this problem, efficient covariance-aware BSS methods that restrict temporal and/or frequency CMs to jointly-diagonalizable ones have been proposed. An efficient version of CTF called FastCTF [13], for example, was proposed with a convergence-guaranteed parameter estimation algorithm [3]. Concurrently and independently, transform-learning NMF (TL-NMF) [14] was proposed for low-rank decomposition in an optimized domain. While NMF is generally performed in the frequency domain, TL-NMF finds an optimal transform better than DFT such that a spectrogram-like representation in the new domain fits NMF. An efficient version of PSDTF-F called FastPSDTF-F [15] was also proposed with an optimization algorithm having no convergence guarantee. Interestingly, a matrix optimized for jointly diagonalizing frequency CMs in FastPSDTF-F can be interpreted as a transform matrix of TL-NMF and both methods are special cases of FastCTF based on jointly-diagonalizable temporal and frequency CMs. The computational complexity of FastPSDTF-T, FastPSDTF-F, FastCTF is $O(TF(T+F))$.

B. Multichannel BSS

Integration of source and spatial models has been considered to be a promising approach to multichannel BSS. MNMF, for example, originally integrates an NMF-based source model with a rank-1 spatial model [5]. Although no specific source model was considered, a full-rank spatial model was proposed for representing the inter-channel characteristics of source images [7] (called full-rank spatial covariance analysis (FCA) in [9], [16]). Another MNMF was then proposed for integrating an NMF-based source model with a full-rank spatial model [6]. MNMF, however, suffers from the initialization sensitivity and the large computational complexity.

A remedy for this problem is to restrict full-rank spatial CMs to *jointly-diagonalizable* ones. FastFCA [16], [17] and FastMNMF [9], [10], for example, were derived by incorporating the joint diagonalizability constraint into FCA [7] and MNMF [6], respectively. Another approach is to use *rank-1* spatial CMs in exchange for the severe loss of the spatial modeling ability as in IVA [3] and ILRMA [8].

Covariance-aware extensions of ILRMA have recently been proposed for handling the temporal and frequency CMs of sources. Integrating a PSDTF-based source model with the rank-1 spatial model yielded independent positive semidefinite tensor analysis (IPSDTA) [18], which was further extended to ILRMA-F [19] based on jointly-diagonalizable frequency CMs and ILRMA-T [19] based on jointly-diagonalizable temporal CMs inspired by a dereverberation method called the weighted prediction error (WPE) [20]. Integrating these two methods finally yielded ILRMA-FT [21], which, however, still uses the restrictive rank-1 spatial model and considers the TF covariance structure only partially for deriving an optimization algorithm.

Our FastMCTF has the same motivation as ILRMA-FT, but is the first method that can consider the full covariance structure of three-mode tensors spanned by the time, frequency, and channel axes. The computational complexity of FastMCTF is $O(TFM(T+FM+M))$ at a manageable level.

III. Existing Methods

This section introduces CTF [12] and FastCTF [13] used for single-channel BSS and MNMF [6] and FastMNMF [10] used for multichannel BSS as the basis of the proposed method.

A. Single-channel BSS

Our goal is to decompose a mixture spectrogram $\mathbf{S} \in \mathbb{C}^{T\times F}$ to the sum of K *basis* spectrograms $\{\mathbf{S}_k \in \mathbb{C}^{T\times F}\}_{k=1}^K$. Let $\mathbf{s} \in \mathbb{C}^{TF}$ and $\mathbf{s}_k \in \mathbb{C}^{TF}$ be vectors obtained by serializing $\mathbf{S}$ and $\mathbf{S}_k$ in the frequency-major manner, respectively. Here $\mathbf{s}_k$ is assumed to follow a multivariate complex Gaussian distribution with a CM $\mathbf{Y}_k \in \mathbb{S}_+^{TF}$ as follows:

$$\mathbf{s}_k \mid \mathbf{Y}_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{s}_k \mid \mathbf{0}, \mathbf{Y}_k). \tag{1}$$

Using the additivity $\mathbf{s} = \sum_k \mathbf{s}_k$, we have

$$\mathbf{s} \mid \mathbf{Y} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{s} \mid \mathbf{0}, \mathbf{Y}), \tag{2}$$

where $\mathbf{Y} = \sum_k \mathbf{Y}_k$. The log-likelihood of $\mathbf{Y}$ to be optimized for the observed mixture $\mathbf{s}$ is given by

$$\log p(\mathbf{s} \mid \mathbf{Y}) \stackrel{c}{=} -\log|\mathbf{Y}| - \mathrm{tr}(\mathbf{X}\mathbf{Y}^{-1}), \tag{3}$$

where $\mathbf{X} = \mathbf{s}\mathbf{s}^{\mathrm{H}}$. The maximization of Eq. (3) is equivalent to the maximization of the log-det (LD) divergence given by

$$\mathcal{D}_{\mathrm{LD}}(\mathbf{X}|\mathbf{Y}) = -\log\left|\mathbf{X}\mathbf{Y}^{-1}\right| + \mathrm{tr}\left(\mathbf{X}\mathbf{Y}^{-1}\right) - TF, \tag{4}$$

Once $\mathbf{Y}$ is estimated, $\mathbf{s}_k$ can be inferred from $\mathbf{s}$ by using a multivariate Wiener filter as follows:

$$p(\mathbf{s}_k \mid \mathbf{s}, \mathbf{Y}) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{s}_k \mid \mathbf{Y}_k\mathbf{Y}^{-1}\mathbf{s}, \mathbf{Y} - \mathbf{Y}_k\mathbf{Y}^{-1}\mathbf{Y}_k\right). \tag{5}$$

Finally, inverse STFT is applied to $\mathbb{E}[\mathbf{s}_k] = \mathbf{Y}_k\mathbf{Y}^{-1}\mathbf{s}$.

1) Correlated Tensor Factorization (CTF): CTF aims to approximate observed $\mathbf{X} \in \mathbb{S}_+^{TF}$ as the sum of Kronecker products between a set of temporal CMs $\mathbf{H} = \{\mathbf{H}_k \in \mathbb{S}_+^{\mathrm{T}}\}_{k=1}^K$ and a set of the corresponding frequency CMs $\mathbf{W} = \{\mathbf{W}_k \in \mathbb{S}_+^F\}_{k=1}^K$ as follows (Fig. 2):

$$\mathbf{X} \approx \mathbf{Y} = \sum_{k=1}^K \mathbf{H}_k \otimes \mathbf{W}_k, \tag{6}$$

where $\mathbf{Y}_k = \mathbf{H}_k \otimes \mathbf{W}_k$. Let $[\mathbf{z}]$ denote a diagonal matrix whose diagonal elements are given by a nonnegative vector

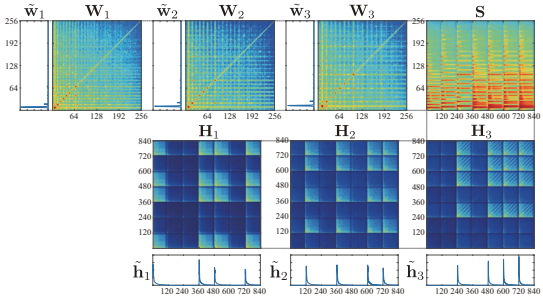

Fig. 2. Covariance-aware low-rank spectrogram modeling based on CTF.

$\mathbf{z}$. If $\mathbf{H}_k = [\tilde{\mathbf{h}}_k]$ *and* $\mathbf{W}_k = [\tilde{\mathbf{w}}_k]$, CTF reduces to NMF [4]. If $\mathbf{H}_k = [\tilde{\mathbf{h}}_k]$ *or* $\mathbf{W}_k = [\tilde{\mathbf{w}}_k]$, CTF reduces to PSDTF-F or PSDTF-T [11], respectively. Eq. (4) can be minimized with respect to $\mathbf{W}$ and $\mathbf{H}$ by using a convergence-guaranteed iterative optimization algorithm [12] with a complexity of $O(T^3F^3)$.

2) FastCTF: FastCTF tries to find an optimal domain other than the TF domain such that all elements are made independent and low-rank for justifying NMF. Specifically, $\mathbf{H}$ and $\mathbf{W}$ are assumed to be jointly diagonalizable as follows:

$$\mathbf{H}_k = \mathbf{R}^{-1}[\tilde{\mathbf{h}}_k]\mathbf{R}^{-\mathrm{H}}, \quad \mathbf{W}_k = \mathbf{P}^{-1}[\tilde{\mathbf{w}}_k]\mathbf{P}^{-\mathrm{H}}, \tag{7}$$

where $\mathbf{R} = [\mathbf{r}_1, \cdots, \mathbf{r}_T]^{\mathrm{H}} \in \mathbb{C}^{T\times T}$ and $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_F]^{\mathrm{H}} \in \mathbb{C}^{F\times F}$ are non-singular matrices called *diagonalizers* (not limited to unitary matrices unlike [14]) and $\tilde{\mathbf{h}}_k \in \mathbb{R}_+^T$ and $\tilde{\mathbf{w}}_k \in \mathbb{R}_+^F$ are nonnegative vectors. Eq. (6) is rewritten as

$$\mathbf{X} \approx \mathbf{Y} = \boldsymbol{\Upsilon}^{-1}\left(\sum_{k=1}^K [\tilde{\mathbf{h}}_k] \otimes [\tilde{\mathbf{w}}_k]\right)\boldsymbol{\Upsilon}^{-\mathrm{H}}, \tag{8}$$

where $\boldsymbol{\Upsilon}$ is given by

$$\boldsymbol{\Upsilon} = \mathbf{R} \otimes \mathbf{P} \in \mathbb{C}^{TF\times TF}. \tag{9}$$

Using Eq. (8) and Eq. (9), Eq. (2) gives

$$\boldsymbol{\Upsilon}\mathbf{s} \sim \mathcal{N}_{\mathbb{C}}\big(\boldsymbol{\Upsilon}\mathbf{s} \mid \mathbf{0}, [\tilde{\mathbf{h}}_k] \otimes [\tilde{\mathbf{w}}_k]\big). \tag{10}$$

We can thus say that while all elements of $\mathbf{s}$ or $\mathbf{S}$ are correlated, the elements of $\boldsymbol{\Upsilon}\mathbf{s}$ or $\mathbf{RSP}^{\mathrm{T}}$ are independent. Eq. (4) can be minimized w.r.t. $\mathbf{H}$, $\mathbf{W}$, $\mathbf{P}$, and $\mathbf{R}$ by using a convergence-guaranteed algorithm with a complexity of $O(T^2F+TF^2)$ [13].

B. Multichannel BSS

Our goal is to decompose a multichannel mixture spectrogram $\mathbf{S} \in \mathbb{C}^{T\times F\times M}$ to the sum of N source images $\{\mathbf{S}_n \in \mathbb{C}^{T\times F\times M}\}_{n=1}^N$. Let $\mathbf{s} \in \mathbb{C}^{TFM}$ and $\mathbf{s}_n \in \mathbb{C}^{TFM}$ be vectors obtained by serializing $\mathbf{S}$ and $\mathbf{S}_n$ in the channel- and frequency-major manner, respectively. As in Section III-A, $\mathbf{s}_n$ is assumed to follow a multivariate complex Gaussian distribution with a CM $\mathbf{Y}_n \in \mathbb{S}_+^{TFM}$ as follows:

$$\mathbf{s}_n \mid \mathbf{Y}_n \sim \mathcal{N}_{\mathbb{C}}(\mathbf{s}_n \mid \mathbf{0}, \mathbf{Y}_n). \tag{11}$$

Using the additivity $\mathbf{s} = \sum_n \mathbf{s}_n$, we have

$$\mathbf{s} \mid \mathbf{Y} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{s} \mid \mathbf{0}, \mathbf{Y}), \tag{12}$$

where $\mathbf{Y} = \sum_n \mathbf{Y}_n$. The log-likelihood and cost function of $\mathbf{Y}$ are the same as Eqs. (3) and (4), respectively. Once $\mathbf{Y}$ is estimated, $\mathbf{s}_n$ can be inferred with a Wiener filter as follows:

$$p(\mathbf{s}_n \mid \mathbf{s}, \mathbf{Y}) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{s}_n \mid \mathbf{Y}_n\mathbf{Y}^{-1}\mathbf{s}, \mathbf{Y} - \mathbf{Y}_n\mathbf{Y}^{-1}\mathbf{Y}_n\right). \tag{13}$$

Finally, inverse STFT is applied to $\mathbb{E}[\mathbf{s}_n] = \mathbf{Y}_n\mathbf{Y}^{-1}\mathbf{s}$.

1) Multichannel NMF (MNMF): Assuming the TF independence, channel-axis-aligned slices of $\mathbf{S}_n$, $\{\mathbf{s}_{ntf} \in \mathbb{C}^M\}_{t,f=1}^{T,F}$, are assumed to follow independent multivariate Gaussian distributions. In MNMF, $\mathbf{Y}_n \in \mathbb{S}_+^{TFM}$ is a block-diagonal matrix whose (t,f)-th block, $\mathbf{Y}_{ntf} \in \mathbb{S}_+^M$, is given by

$$\mathbf{Y}_{ntf} = \sum_{k=1}^{K} \tilde{h}_{nkf}\tilde{w}_{nkf}\mathbf{G}_{nf}, \tag{14}$$

where $\sum_{k=1}^K \tilde{h}_{nkf}\tilde{w}_{nkf}$ represents the PSD of source n at time t and frequency f (low-rank source model) and $\mathbf{G}_{nf}$ represents the spatial CM of source n at frequency f (full-rank spatial model). $[\tilde{\mathbf{h}}_{nk}] \in \mathbb{R}_+^T$ and $[\tilde{\mathbf{w}}_{nk}] \in \mathbb{R}_+^F$ are basis vectors and the corresponding activations of source n, respectively. $\mathbf{X}$ is thus approximated by $\mathbf{Y}$ as follows:

$$\mathbf{X} \approx \mathbf{Y} = \sum_{n=1}^{N}\sum_{k=1}^{K}([\tilde{\mathbf{h}}_{nk}] \otimes [\tilde{\mathbf{w}}_{nk}] \otimes \mathbf{I}_M) \odot [\mathbf{G}_{n\cdot}], \tag{15}$$

where $[\mathbf{G}_{n\cdot}]$ denotes a block-diagonal matrix whose diagonal blocks are given by $\{\mathbf{G}_{nf}\}_{f=1}^F$. Eq. (4) can be minimized w.r.t. $\mathbf{H}$, $\mathbf{W}$, and $\mathbf{G}$ by using a convergence-guaranteed algorithm [6], [12] with a complexity of $O(TFM^3)$.

2) FastMNMF: To reduce the model complexity of MNMF, the spatial CMs $\mathbf{G}_{\cdot f} = \{\mathbf{G}_{nf}\}_{n=1}^N$ at each frequency f are assumed to be jointly diagonalizable as follows:

$$\mathbf{G}_{nf} = \mathbf{Q}_f^{-1}[\tilde{\mathbf{g}}_n]\mathbf{Q}_f^{-\mathrm{H}}, \tag{16}$$

where $\mathbf{Q}_f = [\mathbf{q}_{f1}, \cdots, \mathbf{q}_{fM}]^{\mathrm{H}} \in \mathbb{C}^{M\times M}$ is a diagonalizer and $\tilde{\mathbf{g}}_n \in \mathbb{R}_+^M$ is a nonnegative vector. Eq. (15) is rewritten as

$$\mathbf{X} \approx \mathbf{Y} = \boldsymbol{\Upsilon}^{-1}\left(\sum_{n=1}^{N}\sum_{k=1}^{K}([\tilde{\mathbf{h}}_{nk}] \otimes [\tilde{\mathbf{w}}_{nk}] \otimes [\tilde{\mathbf{g}}_n])\right)\boldsymbol{\Upsilon}^{-\mathrm{H}}, \tag{17}$$

where $\boldsymbol{\Upsilon}$ is given by

$$\boldsymbol{\Upsilon} = \mathbf{I}_T \otimes [\mathbf{Q}_\cdot] \in \mathbb{C}^{TFM\times TFM}. \tag{18}$$

Using Eq. (17) and Eq. (18), Eq. (12) gives

$$\boldsymbol{\Upsilon}\mathbf{s} \sim \mathcal{N}_{\mathbb{C}}\big(\boldsymbol{\Upsilon}\mathbf{s} \mid \mathbf{0}, [\tilde{\mathbf{h}}_{nk}] \otimes [\tilde{\mathbf{w}}_{nk}] \otimes [\tilde{\mathbf{g}}_n]\big). \tag{19}$$

This means that all elements of $\boldsymbol{\Upsilon}\mathbf{s}$ (transformed spectrogram $\{\mathbf{Q}_f\mathbf{s}_{tf} \in \mathbb{C}^M\}_{t,f=1}^{T,F}$) are independent. $\mathbf{Q}_f$ is thus considered to serve as a demixing filter. Eq. (4) can be minimized w.r.t. $\mathbf{H}$, $\mathbf{W}$, $\mathbf{G}$, and $\mathbf{Q}$ by using a convergence-guaranteed algorithm with a complexity of $O(TFM^2)$ [10] as in FastCTF.

IV. Proposed Method

This section explains the proposed method called fast multichannel correlated tensor factorization (FastMCTF).

A. Model Formulation

We unify FastCTF [13] based on the jointly-diagonalizable temporal and frequency covariance models (Eq. (7)) and FastMNMF [10] based on the jointly-diagonalizable spatial covariance model (Eq. (16)), *i.e.*, Eq. (17) is extended to

$$\boldsymbol{\Upsilon} = \mathbf{R} \otimes ([\mathbf{Q}_\cdot](\mathbf{P} \otimes \mathbf{I}_M)) \in \mathbb{C}^{TFM\times TFM}. \tag{20}$$

If $N = M = 1$, Eq. (20) reduces to Eq. (9) of FastCTF. If $\mathbf{R} = \mathbf{I}_T$ and $\mathbf{P} = \mathbf{I}_F$, Eq. (20) reduces to Eq. (18) of FastMNMF. $\mathbf{S}$ can thus be transformed to $\tilde{\mathbf{S}}$ consisting of independent and low-rank elements by using Eq. (19), *i.e.*, multiplying $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{R}$ to the axes of $\mathbf{S}$ in this order. Using IS-NMF, the PSDs of $\tilde{\mathbf{S}}$, $\tilde{x}_{tfm} = |\tilde{s}_{tfm}|^2$ are approximated by $\tilde{y}_{tfm}$ as follows:

$$\begin{aligned}\tilde{x}_{tfm} &= (\mathbf{r}_t^{\mathrm{H}} \otimes \mathbf{p}_f^{\mathrm{H}} \otimes \mathbf{q}_{fm}^{\mathrm{H}})\mathbf{X}(\mathbf{r}_t \otimes \mathbf{p}_f \otimes \mathbf{q}_{fm})\\ &= \mathbf{r}_t^{\mathrm{H}}(\mathbf{I}_T \otimes \mathbf{p}_f^{\mathrm{H}} \otimes \mathbf{q}_{fm}^{\mathrm{H}})\mathbf{X}(\mathbf{I}_T \otimes \mathbf{p}_f \otimes \mathbf{q}_{fm})\mathbf{r}_t\\ &= \mathbf{p}_f^{\mathrm{H}}(\mathbf{r}_t^{\mathrm{H}} \otimes \mathbf{I}_F \otimes \mathbf{q}_{fm}^{\mathrm{H}})\mathbf{X}(\mathbf{r}_t \otimes \mathbf{I}_F \otimes \mathbf{q}_{fm})\mathbf{p}_f\\ &= \mathbf{q}_{fm}^{\mathrm{H}}(\mathbf{r}_t^{\mathrm{H}} \otimes \mathbf{p}_f^{\mathrm{H}} \otimes \mathbf{I}_M)\mathbf{X}(\mathbf{r}_t \otimes \mathbf{p}_f \otimes \mathbf{I}_M)\mathbf{q}_{fm}, \end{aligned}\tag{21}$$

$$\tilde{y}_{tfm} = \sum_{n=1}^{N}\sum_{k=1}^{K}\tilde{h}_{nkt}\tilde{w}_{nkf}\tilde{g}_{nm}. \tag{22}$$

B. Parameter Estimation

Given $\mathbf{X} = \mathbf{s}\mathbf{s}^{\mathrm{H}}$, we aim to estimate $\mathbf{H}$, $\mathbf{W}$, $\mathbf{G}$, $\mathbf{R}$, $\mathbf{P}$, and $\mathbf{Q}$, such that $\mathcal{D}_{\mathrm{LD}}(\mathbf{X}|\mathbf{Y})$ (Eqs. (17) and (20)) is minimized.

$$\begin{aligned}\mathcal{D}_{\mathrm{LD}}(\mathbf{X}|\mathbf{Y}) &= -\log|\mathbf{X}\mathbf{Y}^{-1}| + \mathrm{tr}\big(\mathbf{X}\mathbf{Y}^{-1}\big) - TFM\\ &\overset{c}{=} -FM\log|\mathbf{R}\mathbf{R}^{\mathrm{H}}| - TM\log|\mathbf{P}\mathbf{P}^{\mathrm{H}}| - T\sum_{f=1}^{F}\log|\mathbf{Q}_f\mathbf{Q}_f^{\mathrm{H}}|\\ &\quad + \sum_{t=1}^{T}\sum_{f=1}^{F}\sum_{m=1}^{M}\log\tilde{y}_{tfm} + \sum_{t=1}^{T}\sum_{f=1}^{F}\sum_{m=1}^{M}\tilde{x}_{tfm}\tilde{y}_{tfm}^{-1}. \end{aligned}\tag{23}$$

Because Eq. (23) has the same form as FastCTF and FastMNMF, we can derive a convergence-guaranteed optimization algorithm. Specifically, $\mathbf{H}$, $\mathbf{W}$, $\mathbf{G}$ can be updated in a multiplicative manner as in IS-NMF as follows:

$$\tilde{h}_{nkt} \leftarrow \tilde{h}_{nkt}\sqrt{\frac{\sum_{f=1}^F\sum_{m=1}^M \tilde{w}_{nkf}\tilde{g}_{nm}\tilde{y}_{tfm}^{-2}\tilde{x}_{tfm}}{\sum_{f=1}^F\sum_{m=1}^M \tilde{w}_{nkf}\tilde{g}_{nm}\tilde{y}_{tfm}^{-1}}}, \tag{24}$$

$$\tilde{w}_{nkf} \leftarrow \tilde{w}_{nkf}\sqrt{\frac{\sum_{t=1}^T\sum_{m=1}^M \tilde{h}_{nkt}\tilde{g}_{nm}\tilde{y}_{tfm}^{-2}\tilde{x}_{tfm}}{\sum_{t=1}^T\sum_{m=1}^M \tilde{h}_{nkt}\tilde{g}_{nm}\tilde{y}_{tfm}^{-1}}}, \tag{25}$$

$$\tilde{g}_{nm} \leftarrow \tilde{g}_{nm}\sqrt{\frac{\sum_{t=1}^T\sum_{f=1}^F\sum_{k=1}^K \tilde{h}_{nkt}\tilde{w}_{nkf}\tilde{y}_{tfm}^{-2}\tilde{x}_{tfm}}{\sum_{t=1}^T\sum_{f=1}^F\sum_{k=1}^K \tilde{h}_{nkt}\tilde{w}_{nkf}\tilde{y}_{tfm}^{-1}}}. \tag{26}$$

The diagonalizers $\mathbf{R}$, $\mathbf{P}$, and $\mathbf{Q}$ can also be updated with iterative projection as in IVA [3] as follows:

$$\begin{cases}\mathbf{C}_t = \frac{1}{FM}\sum_{f=1}^F\sum_{m=1}^M \frac{(\mathbf{I}_T\otimes\mathbf{p}_f^{\mathrm{H}}\otimes\mathbf{q}_{fm}^{\mathrm{H}})\mathbf{X}(\mathbf{I}_T\otimes\mathbf{p}_f\otimes\mathbf{q}_{fm})}{\tilde{y}_{tfm}},\\ \mathbf{r}_t \leftarrow (\mathbf{R}\mathbf{C}_t)^{-1}\mathbf{e}_t, \quad \mathbf{r}_t \leftarrow (\mathbf{r}_t^{\mathrm{H}}\mathbf{C}_t\mathbf{r}_t)^{-\frac{1}{2}}\mathbf{r}_t,\end{cases}$$

$$\begin{cases}\mathbf{A}_f = \frac{1}{TM}\sum_{t=1}^T\sum_{m=1}^M \frac{(\mathbf{r}_t^{\mathrm{H}}\otimes\mathbf{I}_F\otimes\mathbf{q}_{fm}^{\mathrm{H}})\mathbf{X}(\mathbf{r}_t\otimes\mathbf{I}_F\otimes\mathbf{q}_{fm})}{\tilde{y}_{tfm}},\\ \mathbf{p}_f \leftarrow (\mathbf{P}\mathbf{A}_f)^{-1}\mathbf{e}_f, \quad \mathbf{p}_f \leftarrow (\mathbf{p}_f^{\mathrm{H}}\mathbf{A}_f\mathbf{p}_f)^{-\frac{1}{2}}\mathbf{p}_f,\end{cases}$$

$$\begin{cases}\mathbf{B}_{fm} = \frac{1}{T}\sum_{t=1}^T \frac{(\mathbf{r}_t^{\mathrm{H}}\otimes\mathbf{p}_f^{\mathrm{H}}\otimes\mathbf{I}_M)\mathbf{X}(\mathbf{r}_t\otimes\mathbf{p}_f\otimes\mathbf{I}_M)}{\tilde{y}_{tfm}},\\ \mathbf{q}_{fm} \leftarrow (\mathbf{Q}_f\mathbf{B}_{fm})^{-1}\mathbf{e}_m,\ \mathbf{q}_{fm} \leftarrow (\mathbf{q}_{fm}^{\mathrm{H}}\mathbf{B}_{fm}\mathbf{q}_{fm})^{-\frac{1}{2}}\mathbf{q}_{fm}.\end{cases}$$

Finally, source images are obtained with Eq. (13). While the naive decomposition of $\mathbf{S}$ using Eq. (13) costs $O(T^3F^3M^3)$, it is equivalent to the efficient element-wise decomposition of $\tilde{\mathbf{S}}$ with a complexity of $O(TFM(T+F+M))$.

V. Evaluation

This section reports our preliminary experiment conducted for investigating the effectiveness of considering the temporal and frequency CMs in addition to the frequency CMs in multichannel BSS.

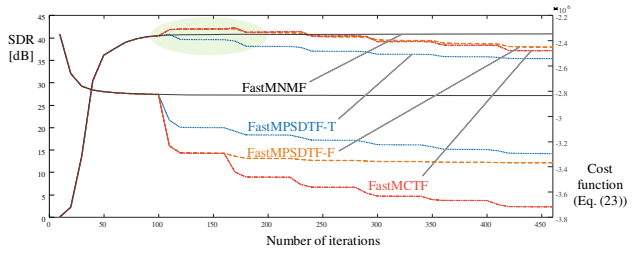

Fig. 3. The change of the SDR and cost function in the iterative optimization.

A. Experimental Conditions

We compared the behaviors of TF-independence-based FastMNMF [10] ($\mathbf{R} = \mathbf{I}_T$ and $\mathbf{P} = \mathbf{I}_F$), frequency- and temporal-covariance-based FastPSDTF-F ($\mathbf{R} = \mathbf{I}_T$) and FastPSDTF-T ($\mathbf{P} = \mathbf{I}_F$), and TF-covariance-based FastMCTF. After randomly initializing $\mathbf{\Theta} = \{\mathbf{H}, \mathbf{W}, \mathbf{G}, \mathbf{Q}\}$ and setting $\mathbf{R} = \mathbf{I}_T$, $\mathbf{P} = \mathbf{I}_F$, and $\mathbf{Q} = \mathbf{I}_M$, we updated only $\mathbf{\Theta}$ (FastMNMF) for the first 100 iterations and then iterates a process of updating $\mathbf{P}$ or $\mathbf{R}$ 10 times and $\mathbf{\Theta}$ 50 times.

For evaluation, an anechoic signal sampled at 16 kHz was synthesized by mixing male and female short utterances taken from the spatialized WSJ0-mix dataset [22] ($N = 2, M = 4$). The STFT with a window size of 512 and a shifting interval of 256 was used for computing $\mathbf{S}$ ($T = 295, F = 257$). The performances of the compared methods were evaluated in term of the cost function and the SDR [23].

B. Experimental Results and Discussions

As shown in Fig. 3, we confirmed that the cost function was monotonically non-increasing in each method, where updating $\mathbf{P}$ or $\mathbf{R}$ led to a large drop of the cost function. Interestingly, we found that although FastMCTF always gave the lowest value of the cost function, a lower cost did not always mean a better SDR. After the first 100 iterations based on FastMNMF, the SDR was slightly improved by updating $\mathbf{P}$ and $\mathbf{\Theta}$, but then gradually degraded by updating $\mathbf{R}$. This indicates that the frequency covariance modeling helps to find an optimal domain other than the frequency domain, but the temporal covariance modeling may conflict with the starring spatial covariance modeling in multichannel BSS, while both techniques were shown to be useful for single-channel BSS (PSDTF-F and PSDTF-T) [13]. The main contribution of this paper lies in the mathematically-solid derivation of FastMCTF and detailed empirical evaluation should be included in future work.

To draw the potential of FastMCTF, we are revisiting the order of frequency- and channel-axis-aligned transforms $\mathbf{P}$ and $\mathbf{Q}$ used for making the elements of $\mathbf{S}$ independent. While FastCTF [13] is invariant w.r.t. the order of axis-aligned transforms, FastMCTF is not because of the frequency-wise spatial modeling. Considering that in theory, the time-domain convolution is equivalent to the frequency-domain product, it would be better to use $\mathbf{Q}$ in the frequency domain before using $\mathbf{P}$.

VI. Conclusion

This paper described a BSS method named fast multichannel correlated tensor factorization (FastMCTF). It includes as its special cases many conventional efficient BSS methods based on the independence, low-rankness, and nonnegativity (positive semidefiniteness) of sources such at NMF [4], FastCTF [13], and FastPSDTF [15] proposed for single-channel BSS and ILRMA [8] and FastMNMF [9], [10] proposed for multichannel BSS. A given multichannel mixture spectrogram can be efficiently decomposed into the sum of source images at once with a Wiener filter considering the temporal, frequency, and spatial CMs estimated by FastMCTF.

Acknowledgment: This study was partially supported by JSPS KAKENHI No. 19H04137 and NII CRIS Collaborative Research Program operated by NII CRIS and LINE Corporation.

References

[1] A. Hyvärinen *et al.*, *Independent Component Analysis*, Wiley, 2004.
[2] T. Kim *et al.*, "Independent vector analysis: An extension of ICA to multivariate components," in *ICA*, 2006, pp. 165–172.
[3] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *WASPAA*, 2011, pp. 189–192.
[4] C. Févotte *et al.*, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.
[5] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE TASLP*, vol. 18, no. 3, pp. 550–563, 2010.
[6] H. Sawada *et al.*, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE TASLP*, vol. 21, no. 5, pp. 971–982, 2013.
[7] N. Q. K. Duong *et al.*, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE TASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.
[8] D. Kitamura *et al.*, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM TASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.
[9] N. Ito and T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *ICASSP*, 2019, pp. 371–375.
[10] K. Sekiguchi *et al.*, "Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices," in *EUSIPCO*, 2019, pp. 1–5.
[11] K. Yoshii *et al.*, "Infinite positive semidefinite tensor factorization for source separation of mixture signals," in *ICML*, 2013, pp. 576–584.
[12] K. Yoshii, "Correlated tensor factorization for audio source separation," in *ICASSP*, 2018, pp. 731–735.
[13] K. Yoshii *et al.*, "Independent low-rank tensor analysis for audio source separation," in *EUSIPCO*, 2018, pp. 1671–1675.
[14] D. Fagot *et al.*, "Nonnegative matrix factorization for transform learning," in *ICASSP*, 2018, pp. 2431–2435.
[15] N. Ito and T. Nakatani, "Multiplicative updates and joint diagonalization based acceleration for under-determined BSS using a full-rank spatial covariance model," in *GlobalSIP*, 2018, pp. 231–235.
[16] ——, "FastFCA-AS: Joint diagonalization based acceleration of full-rank spatial covariance analysis for separating any number of sources," in *IWAENC*, 2018, pp. 151–155.
[17] N. Ito *et al.*, "FastFCA: A joint diagonalization based fast algorithm for audio source separation using a full-rank spatial covariance model," in *EUSIPCO*, 2018, pp. 1681–1685.
[18] R. Ikeshita, "Independent positive semidefinite tensor analysis in blind source separation," in *EUSIPCO*, 2018, pp. 1666–1670.
[19] R. Ikeshita *et al.*, "A unifying framework for blind source separation based on a joint diagonalizability constraint," in *EUSIPCO*, 2019.
[20] T. Nakatani *et al.*, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE TASLP*, vol. 18, no. 7, pp. 1717–1731, 2019.
[21] R. Ikeshita *et al.*, "Independent low-rank matrix analysis with decorrelation learning," in *WASPAA*, 2019, pp. 288–292.
[22] Z.-Q. Wang *et al.*, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *ICASSP*, 2018, pp. 1–5.
[23] E. Vincent *et al.*, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.