Independent Low-Rank Tensor Analysis for Audio Source Separation

Kazuyoshi Yoshii*† Koichi Kitamura* Yoshiaki Bando* Eita Nakamura* Tatsuya Kawahara*
*Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
Email: {yoshii, kitamura, bando, enakamura, kawahara}@sap.ist.i.kyoto-u.ac.jp
†Center for Advanced Intelligence Project (AIP), RIKEN, Chuo-ku, Tokyo 103-0027, Japan

***Abstract*—This paper describes a versatile tensor factorization technique called independent low-rank tensor analysis (ILRTA) and its application to single-channel audio source separation. In general, audio source separation has been conducted in the short-time Fourier transform (STFT) domain under an unrealistic but conventional assumption of the independence of time-frequency (TF) bins. Nonnegative matrix factorization (NMF) is a typical technique of single-channel source separation based on the low-rankness of source spectrograms. In a multichannel setting, independent component analysis (ICA) and its multivariate extension called independent vector analysis (IVA) have often been used for blind source separation based on the independence of source spectrograms. Integrating NMF and IVA, independent low-rank matrix analysis (ILRMA) was recently proposed. To deal with the covariance of TF bins, in this paper we propose ILRTA as a new extension of NMF. Both ILRMA and ILRTA aim to find independent and low-rank sources. A key difference is that while ILRMA estimates demixing filters that decorrelate the channels for multichannel source separation, ILRTA finds optimal transforms that decorrelate the time frames and frequency bins of a STFT representation for single-channel source separation in a way that the bin-wise independence assumed by NMF holds true as much as possible. We report evaluation results of ILRTA and discuss extension of ILRTA to multichannel source separation.**

I. Introduction

Audio source separation is a fundamental technique for audio event detection and identification [1], automatic speech recognition [2], and automatic music transcription [3]. Single-channel or multichannel source separation has commonly been performed in the short-time Fourier transform (STFT) domain. Since single-channel source separation is an underdetermined ill-posed problem, one needs to assume some *time-frequency* (TF) statistics of source spectrograms for evaluating the optimality of a solution. Multichannel source separation without the aid of prior information about sources, a.k.a. blind source separation (BSS), can yield reasonable results under a determined or overdetermined condition (the number of sources is no more than that of microphones) by leveraging the *spatial* statistics of source spectrograms.

Nonnegative matrix factorization (NMF) [4] is a well-known technique of single-channel source separation based on the low-rankness of source spectrograms. A given nonnegative matrix (mixture spectrogram) is approximated as the product of two nonnegative matrices (a set of basis spectra and a set of temporal activations). If the TF bins of audio spectrograms are *independently* complex Gaussian distributed, a variant of NMF based on the Itakura-Saito (IS) divergence (IS-NMF) [5] is theoretically justified. The mixture spectrogram is decomposed into the sum of source spectrograms via *bin-wise* Wiener filtering, in which the phase information of each source remains the same as that of the mixture. Note that estimation of consistent phase information [6], [7] cannot always improve the perceptual quality of the synthesized source signals.

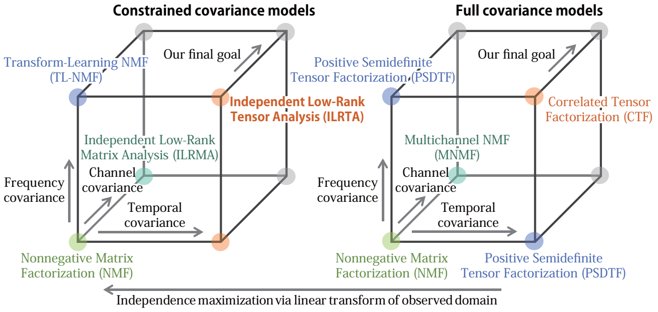
Constrained covariance models
Full covariance models
Our final goal
Transform-Learning NMF (TL-NMF)
Independent Low-Rank Tensor Analysis (ILRTA)
Independent Low-Rank Matrix Analysis (ILRMA)
Frequency covariance
Channel covariance
Temporal covariance
Nonnegative Matrix Factorization (NMF)
Our final goal
Positive Semidefinite Tensor Factorization (PSDTF)
Correlated Tensor Factorization (CTF)
Multichannel NMF (MNMF)
Frequency covariance
Channel covariance
Temporal covariance
Nonnegative Matrix Factorization (NMF)
Positive Semidefinite Tensor Factorization (PSDTF)
Independence maximization via linear transform of observed domain

Fig. 1. A new look of low-rank decomposition methods of single-channel or multichannel audio source separation in terms of covariance modeling.

A fundamental problem causing the phase inconsistency of NMF lies in the unrealistic assumption of the independence of TF bins. In theory, the independence of frequency bins holds true in the Fourier transform of an infinite stationary signal. In reality, the frequency bins are correlated with each other in the short-time Fourier transform (STFT) of a finite non-stationary signal. The time frames at which the same sounds occur are also correlated with each other.

Correlated tensor factorization (CTF) [8] is an ultimate approach to dealing with the full covariance of TF bins. A given positive semidefinite (PSD) matrix (the huge covariance matrix over all the TF bins of a mixture spectrogram) is approximated as the sum of the Kronecker products between multiple sets of PSD matrices (a set of frequency covariance matrices and at set of temporal covariance matrices). All the TF bins of each source spectrogram can be estimated jointly in a correlated manner via Wiener filtering. CTF includes as its special cases NMF, positive semidefinite tensor factorization (PSDTF) [9], [10] dealing with either frequency or temporal covariance matrices, and nonnegative tensor factorization (NTF) [11]. The major limitation of CTF, however, lies in the prohibitively huge computational cost. For a mixture spectrogram of F bins and T frames consisting of K sources, the computational cost of CTF is $\mathcal{O}(KF^3T^3)$ while that of NMF is $\mathcal{O}(KFT)$.

The covariance of channels plays a central role in multichannel source separation based on the independence of source

ISBN 978-90-827970-1-5 © EURASIP 2018

spectrograms. Assuming an instantaneous mixing process in the frequency domain, independent component analysis (ICA) [12] has often been used for estimating source components in a frequency-wise manner. To avoid the permutation problem, independent vector analysis (IVA) [13], [14] assumes source spectra to follow multivariate probability distributions. To use the low-rankness of source spectrograms, independent low-rank matrix analysis (ILRMA) [15] was derived by integrating IVA and NMF. These methods aim to estimate demixing matrices that make the channels independent (*i.e.*, transform the channel domain into the source domain) under a determined condition in a way that an assumption on source spectrograms (*e.g.*, super-Gaussianity and low-rankness) holds true as much as possible. Another approach to demixing matrix estimation is to perform approximate joint diagonalization (AJD) of spatial covariance matrices at different frames [16]–[18].

Using the techniques of multichannel source separation, we propose a constrained version of CTF called *independent low-rank tensor analysis* (ILRTA) with a feasible complexity of $\mathcal{O}(F^2T)+\mathcal{O}(FT^2)+\mathcal{O}(F^4)+\mathcal{O}(T^4)+\mathcal{O}(KFT)$ for single-channel source separation based on the independence and low-rankness of source spectrograms. ILRTA aims to find optimal transforms that make the time frames and frequency bins independent by constraining the temporal and frequency covariance matrices of CTF to be jointly diagonalizable, respectively. CTF in the STFT domain can thus be approximated by NMF in a transformed domain. Such transforms can be estimated by using an iterative projection (IP) algorithm proposed for estimating demixing matrices of IVA [14]. In summary, ILRTA iterates until convergence the optimization of decorrelation transforms and NMF of a decorrelated spectrogram, in the same way that ILRMA iterates the optimization of demixing matrices and NMF of demixed spectrograms.

A major contribution of this study is to situate low-rank decomposition methods in terms of multiway covariance modeling (Fig. 1). ILRTA is a special case of CTF with jointly diagonalizable temporal and frequency covariance matrices while ILRMA is that of multichannel NMF (MNMF) [19] with rank-1 spatial covariance matrices. ILRTA is a multiway extension of transform-learning NMF (TL-NMF) [20] that estimates unirary frequency covariance matrices.

II. Correlated Tensor Factorization

We briefly review correlated tensor factorization (CTF) [8] and its application to single-channel audio source separation.

A. General Formulation

Let $\mathbf{X} \in \mathbb{S}_+^{D_1 D_2 \cdots D_M}$ be a PSD matrix, where $\mathbb{S}_+^D$ denotes a $D \times D$ PSD matrix [21] and the dimension of $\mathbf{X}$ is assumed to be factorized as the product of M integers $\{D_m\}_{m=1}^M$. CTF aims to approximate $\mathbf{X}$ as a PSD matrix $\mathbf{Y}$ as follows:

$$\mathbf{X} \approx \mathbf{Y} \overset{\text{def}}{=} \sum_{k=1}^{K} \bigotimes_{m=1}^{M} \mathbf{V}_{km} \overset{\text{def}}{=} \sum_{k=1}^{K} \mathbf{V}_{k1} \otimes \cdots \otimes \mathbf{V}_{kM}, \qquad (1)$$

where $\{\mathbf{V}_{km} \in \mathbb{S}_+^{D_m}\}_{m=1}^M$ is a set of PSD matrices related to component k and $\otimes$ denotes the Kronecker product.

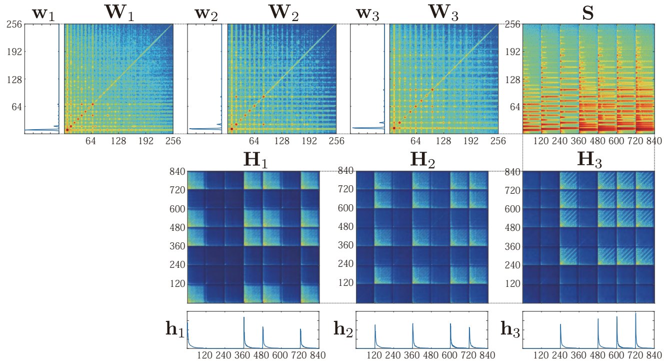

Fig. 2. Comparison between IS-NMF, LD-PSDTF, and LD-CTF.

To evaluate the approximation error between $\mathbf{X}$ and $\mathbf{Y}$, one can use the Bregman matrix divergence [22] defined as

$$\mathcal{D}_\phi(\mathbf{X}|\mathbf{Y}) = \phi(\mathbf{X}) - \phi(\mathbf{Y}) - \mathrm{tr}\big(\nabla\phi(\mathbf{Y})^{\mathrm{T}}(\mathbf{X}-\mathbf{Y})\big), \qquad (2)$$

where ϕ is a strictly convex function on $\mathbb{S}_+^{D_1 D_2 \cdots D_M}$ and $*^{\mathrm{T}}$ denotes the matrix transpose (cf., $*^{\mathrm{H}}$ denotes the Hermitian conjugate). For audio source separation, it is theorerically reasonable to use the log-det (LD) divergence [23] with $\phi(\mathbf{Z}) = -\log|\mathbf{Z}|$, which is given by

$$\mathcal{D}_{\text{LD}}(\mathbf{X}|\mathbf{Y}) = -\log\left|\mathbf{X}\mathbf{Y}^{-1}\right| + \mathrm{tr}\left(\mathbf{X}\mathbf{Y}^{-1}\right) - D_1 \cdots D_M. \; (3)$$

To estimate $\mathbf{V}_{km}$ that minimizes $\mathcal{D}_{\text{LD}}(\mathbf{X}|\mathbf{Y})$, we consider only $M = 2$ because Eq. (1) can be written as $\mathbf{X}^{(m)} \approx \sum_{k=1}^{K} \mathbf{V}_{km} \otimes (\bigotimes_{m' \neq m} \mathbf{V}_{km'})$, where $\mathbf{X}^{(m)}$ is a permutation of $\mathbf{X}$.

B. Relationships to Conventional Methods

Given a PSD matrix $\mathbf{X} \in \mathbb{S}_+^{FT}$, we aim to estimate two sets of PSD matrices $\{\mathbf{W}_k \in \mathbb{S}_+^F\}_{k=1}^K$ and $\{\mathbf{H}_k \in \mathbb{S}_+^{\mathrm{T}}\}_{k=1}^K$ such that

$$\mathbf{X} \approx \mathbf{Y} \overset{\text{def}}{=} \sum_{k=1}^{K} \mathbf{W}_k \otimes \mathbf{H}_k, \qquad (4)$$

where F and T are positive integers (the number of frequency bins and that of frames). Let $\mathbf{Y}_k = \mathbf{W}_k \otimes \mathbf{H}_k$ such that $\mathbf{Y} = \sum_k \mathbf{Y}_k$. Let $[\mathbf{z}]$ be a diagonal matrix whose diagonal elements form a nonnegative vector $\mathbf{z}$. As shown in Fig. 2, if all the PSD matrices are restricted to diagonal matrices such that $\mathbf{X} = [\mathbf{x}]$, $\mathbf{W}_k = [\mathbf{w}_k]$, and $\mathbf{H}_k = [\mathbf{h}_k]$, LD-CTF reduces to IS-NMF [5]:

$$x_{ft} \approx y_{ft} \overset{\text{def}}{=} \sum_{k=1}^{K} w_{kf} h_{kt}, \qquad (5)$$

where x_{ft} is a nonnegative element of $\mathbf{x}$ and y_{ft} is that of $\mathbf{y}$. If either $\{\mathbf{W}_k \in \mathbb{S}_+^F\}_{k=1}^K$ or $\{\mathbf{H}_k \in \mathbb{S}_+^T\}_{k=1}^K$ are restricted to diagonal matrices, LD-CTF reduces to LD-PSDTF [9], [10]:

$$\hat{\mathbf{X}}_f \approx \sum_{k=1}^{K} w_{kf} \mathbf{H}_k \quad \text{or} \quad \check{\mathbf{X}}_t \approx \sum_{k=1}^{K} \mathbf{W}_k h_{kt}, \qquad (6)$$

where the PSD matrix $\hat{\mathbf{X}}_f \in \mathbb{S}_+^T$ is obtained by extracting the rows and columns of $\mathbf{X}$ related to f, and $\check{\mathbf{X}}_t \in \mathbb{S}_+^F$ is obtained similarly. While LD-PSDTF can deal with the covariance of a particular dimension (*e.g.*, frequency or time axis), LD-CTF can deal with both covariances. LD-PSDTF and LD-CTF can thus achieve better phase-aware source separation.

ISBN 978-90-827970-1-5 © EURASIP 2018

C. Parameter Estimation

A convergence-guaranteed iterative algorithm was proposed for estimating $\{\mathbf{W}_k \in \mathbb{S}_+^F\}_{k=1}^K$ and $\{\mathbf{H}_k \in \mathbb{S}_+^T\}_{k=1}^K$ from $\mathbf{X}$ in an unsupervised manner [8]. Let $\#$ denote the geometric mean of two PSD matrices defined as follows [24]–[26]:

$$\mathbf{A}\#\mathbf{B} = \mathbf{A}^{\frac{1}{2}} \left(\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}\right)^{\frac{1}{2}} \mathbf{A}^{\frac{1}{2}} = \mathbf{A}(\mathbf{A}^{-1}\mathbf{B})^{\frac{1}{2}}. \tag{7}$$

The updating formulas of $\mathbf{W}_k$ and $\mathbf{H}_k$ are given by

$$\mathbf{W}_k \leftarrow \mathbf{A}_k^{-1}\#(\mathbf{W}_k\mathbf{B}_k\mathbf{W}_k), \tag{8}$$
$$\mathbf{H}_k \leftarrow \mathbf{C}_k^{-1}\#(\mathbf{H}_k\mathbf{D}_k\mathbf{H}_k), \tag{9}$$

where $\mathbf{A}_k \in \mathbb{S}_+^F$, $\mathbf{B}_k \in \mathbb{S}_+^F$, $\mathbf{C}_k \in \mathbb{S}_+^T$, and $\mathbf{D}_k \in \mathbb{S}_+^T$ are temporary PSD matrices given by

$$\begin{aligned}
\mathbf{A}_k &= (\mathbf{I}_F \otimes \mathbf{1}_T^{\mathrm{T}}) \left((\mathbf{1}_F \otimes \mathbf{H}_k^{\mathrm{T}}) \odot \mathbf{Y}^{-1}\right) (\mathbf{I}_F \otimes \mathbf{1}_T), \\
\mathbf{B}_k &= (\mathbf{I}_F \otimes \mathbf{1}_T^{\mathrm{T}}) \left((\mathbf{1}_F \otimes \mathbf{H}_k^{\mathrm{T}}) \odot \mathbf{Y}^{-1}\mathbf{X}\mathbf{Y}^{-1}\right) (\mathbf{I}_F \otimes \mathbf{1}_T), \\
\mathbf{C}_k &= (\mathbf{1}_F^{\mathrm{T}} \otimes \mathbf{I}_T) \left((\mathbf{W}_k^{\mathrm{T}} \otimes \mathbf{1}_T) \odot \mathbf{Y}^{-1}\right) (\mathbf{1}_F \otimes \mathbf{I}_T), \\
\mathbf{D}_k &= (\mathbf{1}_F^{\mathrm{T}} \otimes \mathbf{I}_T) \left((\mathbf{W}_k^{\mathrm{T}} \otimes \mathbf{1}_T) \odot \mathbf{Y}^{-1}\mathbf{X}\mathbf{Y}^{-1}\right) (\mathbf{1}_F \otimes \mathbf{I}_T),
\end{aligned}$$

where $\mathbf{I}_D$ and $\mathbf{1}_D$ indicate the identity matrix of size D and the all-one vector of length D, respectively, and $\odot$ indicates the element-wise product (Hadamard product). The computational complexity of this algorithm is $\mathcal{O}(KF^3T^3)$, which prohibits the practical application of LD-CTF.

D. Audio Source Separation

We explain a probabilistic generative model underlying LD-CTF in single-channel audio source separation. Let $\mathbf{s} \in \mathbb{C}^{FT}$ be a complex vector obtained by listing in a row-major manner all the TF bins of the complex spectrogram $\mathbf{S} \in \mathbb{C}^{F\times T}$ of a mixture signal over F bins and T frames. Let $\mathbf{X} \overset{\text{def}}{=} \mathbf{s}\mathbf{s}^{\mathrm{H}}$ be the rank-1 covariance matrix over $\mathbf{S}$. Similarly, let $\mathbf{s}_k \in \mathbb{C}^{FT}$ be a complex vector obtained by listing the complex spectrogram $\mathbf{S}_k \in \mathbb{C}^{F\times T}$ of source k. If the linear additivity of complex spectrograms holds true, we can say $\mathbf{s} = \sum_k \mathbf{s}_k$.

In LD-CTF, each $\mathbf{s}_k$ is assumed to follow a centered multivariate complex Gaussian distribution with a covariance matrix $\mathbf{Y}_k \in \mathbb{S}_+^{FT}$ as follows:

$$\mathbf{s}_k \mid \mathbf{Y}_k \sim \mathcal{N}_c(\mathbf{s}_k|\mathbf{0}, \mathbf{Y}_k). \tag{10}$$

The full covariance structure over all the TF bins can be taken into account unlike IS-NMF and LD-PSDTF. The reproductive property of the complex Gaussian distribution gives

$$\mathbf{s} \mid \mathbf{Y} \sim \mathcal{N}_c(\mathbf{s}|\mathbf{0}, \mathbf{Y}). \tag{11}$$

The log-likelihood for the observed data $\mathbf{s}$ is thus given by

$$\log p(\mathbf{s}|\mathbf{Y}) \overset{c}{=} -\log|\mathbf{Y}| - \mathrm{tr}(\mathbf{X}\mathbf{Y}^{-1}) \overset{c}{=} -\mathcal{D}_{\mathrm{LD}}(\mathbf{X}|\mathbf{Y}). \tag{12}$$

This shows that LD-CTF is equivalent to maximum likelihood estimation of a probabilistic model given by Eq. (11).

Once $\mathbf{W}_k$'s and $\mathbf{H}_k$'s are estimated by LD-CTF, the latent variable $\mathbf{s}_k$ can be inferred by Wiener filtering as follows:

$$p(\mathbf{s}_k|\mathbf{s}, \mathbf{W}, \mathbf{H}) = \mathcal{N}_c\left(\mathbf{s}_k \middle| \mathbf{Y}_k\mathbf{Y}^{-1}\mathbf{s}, \mathbf{Y} - \mathbf{Y}_k\mathbf{Y}^{-1}\mathbf{Y}_k\right). \tag{13}$$

The time-domain signal of source k is obtained by applying the inverse STFT to the complex spectrogram $\mathbb{E}[\mathbf{s}_k] = \mathbf{Y}_k\mathbf{Y}^{-1}\mathbf{s}$.

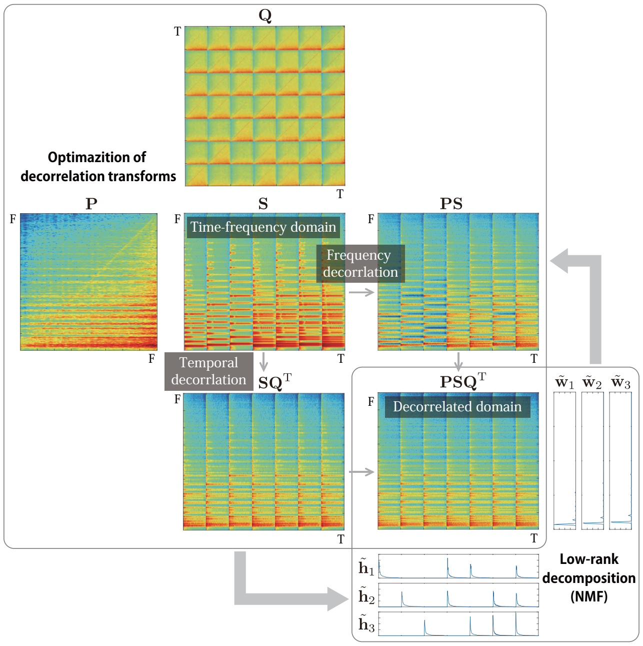
Q
T
T
Optimazition of decorrelation transforms
P
F
F
S
F
Time-frequency domain
T
PS
F
Frequency decorrlation
T
Temporal decorrlation
$\mathbf{SQ}^{\mathrm{T}}$
F
T
$\mathbf{PSQ}^{\mathrm{T}}$
F
Decorrelated domain
T
$\tilde{\mathbf{w}}_1$ $\tilde{\mathbf{w}}_2$ $\tilde{\mathbf{w}}_3$
$\tilde{\mathbf{h}}_1$
$\tilde{\mathbf{h}}_2$
$\tilde{\mathbf{h}}_3$
Low-rank decomposition (NMF)

Fig. 3. ILRTA based on iteration of multiway transform learning and NMF.

III. Independent Low-Rank Tensor Analysis

We explain independent low-rank tensor analysis (ILRTA), a constrained version of LD-CTF with jointly diagonalizable frequency and temporal covariance matrices (Fig. 3).

A. Mathematical Formulation

ILRTA is given by putting on LD-CTF (Eq. (4)) a constraint that $\{\mathbf{W}_k \in \mathbb{S}_+^F\}_{k=1}^K$ and $\{\mathbf{H}_k \in \mathbb{S}_+^T\}_{k=1}^K$ are jointly diagonalizable, respectively, as follows:

$$\forall k \ \mathbf{W}_k = \mathbf{P}^{-1}[\tilde{\mathbf{w}}_k]\mathbf{P}^{-\mathrm{H}}, \tag{14}$$
$$\forall k \ \mathbf{H}_k = \mathbf{Q}^{-1}[\tilde{\mathbf{h}}_k]\mathbf{Q}^{-\mathrm{H}}, \tag{15}$$

where $\tilde{\mathbf{w}}_k \in \mathbb{R}_+^F$ and $\tilde{\mathbf{h}}_k \in \mathbb{R}_+^T$ are nonnegative vectors and $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_F]^{\mathrm{H}} \in \mathbb{C}^{F\times F}$ and $\mathbf{Q} = [\mathbf{q}_1, \cdots, \mathbf{q}_T]^{\mathrm{H}} \in \mathbb{C}^{T\times T}$ are non-singular matrices not limited to unitary matrices unlike TL-NMF [20]. If $\mathbf{P}$ and $\mathbf{Q}$ are identity matrices, ILRTA reduces to IS-NMF. If either of $\mathbf{P}$ and $\mathbf{Q}$ is an identity matrix, ILRTA reduces to a constrained version of LD-PSDTF (Fig. 1). The reconstruction matrix $\mathbf{Y}$ is given by

$$\begin{aligned}
\mathbf{Y} &= \sum_{k=1}^K \mathbf{W}_k \otimes \mathbf{H}_k \\
&= (\mathbf{P}\otimes\mathbf{Q})^{-1} \left(\sum_{k=1}^K [\tilde{\mathbf{w}}_k] \otimes [\tilde{\mathbf{h}}_k]\right) (\mathbf{P}\otimes\mathbf{Q})^{-\mathrm{H}}, \qquad (16)
\end{aligned}$$

For brevity, we define $\tilde{x}_{ft}$ and $\tilde{y}_{ft}$ as follows:

$$\begin{aligned}
\tilde{x}_{ft} &= (\mathbf{p}_f^{\mathrm{H}} \otimes \mathbf{q}_t^{\mathrm{H}})\mathbf{X}(\mathbf{p}_f \otimes \mathbf{q}_t) \\
&= \mathbf{p}_f^{\mathrm{H}}(\mathbf{I}_F \otimes \mathbf{q}_t^{\mathrm{H}})\mathbf{X}(\mathbf{I}_F \otimes \mathbf{q}_t)\mathbf{p}_f \\
&= \mathbf{q}_t^{\mathrm{H}}(\mathbf{p}_f^{\mathrm{H}} \otimes \mathbf{I}_T)\mathbf{X}(\mathbf{p}_f \otimes \mathbf{I}_T)\mathbf{q}_t, \qquad (17)
\end{aligned}$$

$$\tilde{y}_{ft} = \sum_{k=1}^K \tilde{w}_{kf}\tilde{h}_{kt}. \tag{18}$$

ISBN 978-90-827970-1-5 © EURASIP 2018

The cost function based on the LD divergence is given by

$$\begin{aligned}\mathcal{D}_{\text{LD}}(\mathbf{X}|\mathbf{Y}) &= -\log|\mathbf{X}\mathbf{Y}^{-1}| + \operatorname{tr}\left(\mathbf{X}\mathbf{Y}^{-1}\right) - FT \\ &\overset{c}{=} -T\log|\mathbf{P}\mathbf{P}^{\mathrm{H}}| - F\log|\mathbf{Q}\mathbf{Q}^{\mathrm{H}}| \\ &\quad + \sum_{f=1}^{F}\sum_{t=1}^{T}\log\tilde{y}_{ft} + \sum_{f=1}^{F}\sum_{t=1}^{T}\tilde{x}_{ft}\tilde{y}_{ft}^{-1}. \qquad (19)\end{aligned}$$

B. Parameter Estimation

To estimate latent patterns $\{\tilde{\mathbf{w}}_k\}_{k=1}^{K}$ and $\{\tilde{\mathbf{h}}_k\}_{k=1}^{K}$ and transform matrices $\mathbf{P}$ and $\mathbf{Q}$ from a PSD matrix $\mathbf{X}$, we propose an iterative optimization algorithm. The scale ambiguity between these parameters is solved in each iteration as in NMF.

1) Updating Patterns: To minimize Eq. (19) w.r.t. $\tilde{\mathbf{w}}_k$ and $\tilde{\mathbf{h}}_k$, we focus on the sum of the last two terms. This sum is equal to $\sum_{ft}\mathcal{D}_{\text{IS}}(\tilde{x}_{ft}|\tilde{y}_{ft})$ except for constant terms, where $\mathcal{D}_{\text{IS}}(x|y) = -\log x/y + x/y - 1$ denotes the IS divergence between x and y. The updating formulas of $\tilde{\mathbf{w}}_k$ and $\tilde{\mathbf{h}}_k$ are given in a multiplicative form in the same way as IS-NMF [27]:

$$\tilde{w}_{kf} \leftarrow \tilde{a}_{kf}^{-1}\#(\tilde{w}_{kf}\tilde{b}_{kf}\tilde{w}_{kf}) = \tilde{a}_{kf}^{-\frac{1}{2}}\tilde{b}_{kf}^{\frac{1}{2}}\tilde{w}_{kf}, \qquad (20)$$

$$\tilde{h}_{kt} \leftarrow \tilde{c}_{kt}^{-1}\#(\tilde{h}_{kt}\tilde{d}_{kt}\tilde{h}_{kt}) = \tilde{c}_{kt}^{-\frac{1}{2}}\tilde{d}_{kt}^{\frac{1}{2}}\tilde{h}_{kt}, \qquad (21)$$

where $\tilde{a}_{kf}$, $\tilde{b}_{kf}$, $\tilde{c}_{kt}$, and $\tilde{d}_{kt}$ are temporary scalars given by

$$\tilde{a}_{kf} = \sum_{t=1}^{T}\tilde{h}_{kt}\tilde{y}_{ft}^{-1}, \quad \tilde{b}_{kf} = \sum_{t=1}^{T}\tilde{h}_{kt}\tilde{x}_{ft}\tilde{y}_{ft}^{-2}, \qquad (22)$$

$$\tilde{c}_{kt} = \sum_{f=1}^{F}\tilde{w}_{kf}\tilde{y}_{ft}^{-1}, \quad \tilde{d}_{kt} = \sum_{f=1}^{F}\tilde{w}_{kf}\tilde{x}_{ft}\tilde{y}_{ft}^{-2}. \qquad (23)$$

Note that LD-CTF calculates the geometric mean of two PSD matrices (Eq. (8) and Eq. (9)) while IS-NMF calculates that of two nonnegative scalars (Eq. (20) and Eq. (21)).

2) Updating Transforms: To minimize Eq. (19) w.r.t. $\mathbf{P}$, we focus on the sum of the first and last terms. Using Eq. (17), this sum is found to have the same form of the cost function of IVA based on the majorization-minimization (MM) principle [14]. The updating formula of $\mathbf{P}$ is thus given by an iterative projection (IP) algorithm as follows:

$$\text{updating direction: } \mathbf{p}_f \leftarrow (\mathbf{P}\mathbf{U}_f)^{-1}\mathbf{e}_f, \qquad (24)$$

$$\text{updating norm: } \mathbf{p}_f \leftarrow (\mathbf{p}_f^{\mathrm{H}}\mathbf{U}_f\mathbf{p}_f)^{-\frac{1}{2}}\mathbf{p}_f, \qquad (25)$$

where $\mathbf{e}_f \in \mathbb{R}^F$ is a unit vector that takes 1 in dimension f and $\mathbf{U}_f \in \mathbb{S}_+^F$ is a temporary PSD matrix given by

$$\mathbf{U}_f = \sum_{t=1}^{T}(\mathbf{I}_F \otimes \mathbf{q}_t^{\mathrm{H}})\mathbf{X}(\mathbf{I}_F \otimes \mathbf{q}_t)\tilde{y}_{ft}^{-1}. \qquad (26)$$

Similarly, the updating formula of $\mathbf{Q}$ is given by

$$\text{updating direction: } \mathbf{q}_t \leftarrow (\mathbf{Q}\mathbf{V}_t)^{-1}\mathbf{e}_t, \qquad (27)$$

$$\text{updating norm: } \mathbf{q}_t \leftarrow (\mathbf{q}_t^{\mathrm{H}}\mathbf{V}_t\mathbf{q}_t)^{-\frac{1}{2}}\mathbf{q}_t, \qquad (28)$$

where $\mathbf{e}_t \in \mathbb{R}^{\mathrm{T}}$ is a unit vector that takes 1 in dimension t and $\mathbf{V}_t \in \mathbb{S}_+^T$ is a temporary PSD matrix given by

$$\mathbf{V}_t = \sum_{f=1}^{F}(\mathbf{p}_f^{\mathrm{H}} \otimes \mathbf{I}_T)\mathbf{X}(\mathbf{p}_f \otimes \mathbf{I}_T)\tilde{y}_{ft}^{-1}. \qquad (29)$$

C. Audio Source Separation

We investigate how ILRTA works for single-channel audio source separation. Substituting Eq. (16) into Eq. (11), we obtain the probabilistic model of ILRTA as follows:

$$\begin{aligned}&\mathbf{s} \mid \mathbf{Y} \\ &\sim \mathcal{N}_c\left(\mathbf{s}\,\middle|\,\mathbf{0}, (\mathbf{P}\otimes\mathbf{Q})^{-1}\left(\sum_{k=1}^{K}[\tilde{\mathbf{w}}_k]\otimes[\tilde{\mathbf{h}}_k]\right)(\mathbf{P}\otimes\mathbf{Q})^{-\mathrm{H}}\right). \quad (30)\end{aligned}$$

A linear transform of $\mathbf{s} \in \mathbb{C}^{FT}$ using $\mathbf{P} \otimes \mathbf{Q}$ as a transform matrix also follows a Gaussian distribution given by

$$(\mathbf{P}\otimes\mathbf{Q})\mathbf{s} \mid \mathbf{Y} \sim \mathcal{N}_c\left((\mathbf{P}\otimes\mathbf{Q})\mathbf{s}\,\middle|\,\mathbf{0}, \sum_{k=1}^{K}[\tilde{\mathbf{w}}_k]\otimes[\tilde{\mathbf{h}}_k]\right), \quad (31)$$

where $(\mathbf{P}\otimes\mathbf{Q})\mathbf{s} \in \mathbb{C}^{FT}$ is a complex vector obtained by serializing a *transformed* spectrogram $\mathbf{PSQ}^{\mathrm{T}} \in \mathbb{C}^{F\times T}$ in a row-major manner. Eq. (31) means that all the bins of $\mathbf{PSQ}^{\mathrm{T}}$ are independent (uncorrelated) because the covariance matrix is diagonal while those of $\mathbf{S}$ are correlated in the STFT domain. Therefore, $\mathbf{PSQ}^{\mathrm{T}}$ is more suitable to IS-NMF than $\mathbf{S}$. $\mathbf{P}$ and $\mathbf{Q}$ are optimized in a way that the bin-wise independence and low-rankness of $\mathbf{PSQ}^{\mathrm{T}}$ hold true as much as possible. In ILRTA, decorrelation transforms and NMF in a transformed domain are iterated until convergence. This drastically reduces the computational cost of LD-CTF to a manageable level in exchange for the joint diagonalizability constraint.

D. Open Problems

For future research, we discuss some technical difficulties of ILRTA. Since ILRTA is based on an overparametrized model, it is very sensitive to the initialization of iterative optimization. The degree of freedom (DOF) of ILRTA (the number of free parameters) is $K(F+T)+F^2+T^2$, which is only a fraction of the DOF of LD-CTF, $K(F^2+T^2)$, but still larger than the number of observed TF bins, FT. It is effective to initialize ILRTA by using $\mathbf{w}_k$ and $\mathbf{h}_k$ obtained by IS-NMF, *i.e.*, $\tilde{\mathbf{w}}_k \leftarrow \mathbf{w}_k$, $\tilde{\mathbf{h}}_k \leftarrow \mathbf{h}_k$, $\mathbf{P} \leftarrow \mathbf{I}_F$, and $\mathbf{Q} \leftarrow \mathbf{I}_T$.

The IP algorithm (Section III-B2) has non-trivial problems. Since $\mathbf{X} = \mathbf{s}\mathbf{s}^{\mathrm{H}}$ is a rank-1 matrix in audio source separation, Eq. (26) and Eq. (29) can be efficiently calculated as follows:

$$\mathbf{U}_f = \underbrace{(\mathbf{SQ}^{\mathrm{T}})}_{F\times T}\underbrace{\left[[\tilde{y}_{f1},\cdots,\tilde{y}_{fT}]^{\mathrm{T}}\right]}_{T\times T}\underbrace{(\mathbf{SQ}^{\mathrm{T}})^{\mathrm{H}}}_{T\times F}, \qquad (32)$$

$$\mathbf{V}_t = \underbrace{(\mathbf{PS})^{\mathrm{H}}}_{T\times F}\underbrace{\left[[\tilde{y}_{1t},\cdots,\tilde{y}_{Ft}]^{\mathrm{T}}\right]}_{F\times F}\underbrace{(\mathbf{PS})}_{F\times T}. \qquad (33)$$

When $F < T$ as is often the case with audio source separation, $\mathbf{V}_t$ is not invertible due to the rank deficiency. The rank of $\mathbf{V}_t$ is F and the pseudo-inverse operation does not work. Dimensionality reduction techniques such as principal component analysis (PCA) would be a remedy to this problem. Another possibility would be to perform AJD of the temporal covariance matrices $\{\mathbf{H}_k\}_{k=1}^{K}$ obtained by LD-PSDTF for estimating $\mathbf{Q}$. Since the IP algorithm is numerically unstable because of the high dimensionality, we currently update only $\mathbf{P}$ in a few iterations.

TABLE I
SOURCE SEPARATION PERFORMANCE [dB]

Method	SDR	SIR	SAR
IS-NMF	18.9	24.2	20.4
LD-PSDTF	22.8	28.5	24.2
ILRTA (Fast LD-PSDTF)	24.3	31.4	25.2

IV. EVALUATION

We report comparative evaluation of ILRTA with its special cases such as IS-NMF and LD-PSDTF.

A. Experimental Conditions

We synthesized a mixture signal of 8.4 s sampled at 16 [kHz] by concatenating three isolated piano tones (C4, E4, and G4) and four chords (C4+E4, C4+G4, E4+G4, and C4+E4+G4) of 1.2 s ($K = 3$). The STFT with a Gaussian window of 512 pts and a shifting interval of 160 pts was used for calculating the complex spectrogram $\mathbf{S} \in \mathbb{C}^{F \times T}$ with $F = 256$ and $T = 840$. As discussed in Section III-D, we updated $\mathbf{P}$, $\mathbf{W}$, and $\mathbf{H}$ while keeping $\mathbf{Q} = \mathbf{I}_T$. This is equivalent to a fast approximation of LD-PSDTF based on jointly-diagonalizable frequency covariance matrices. Both ILRTA and LD-PSDTF were initialized with the results of IS-NMF. BSS Eval Toolbox [28] was used for measuring the source-to-distortion ratio (SDR), source-to-interferences ratio (SIR), and sources-to-artifacts ratio (SAR) of separated signals.

B. Experimental Results

Table I shows the experimental results. ILRTA outperformed IS-NMF and LD-PSDTF in terms of all the evaluation measures. Interestingly, although ILRTA with fixed $\mathbf{Q} = \mathbf{I}_T$ is an approximation of LD-PSDTF, it was better than LD-PSDTF. The covariance constraint was found to be effective for regularizing an overparametrized model such as LD-PSDTF and LD-CTF. We confirmed that the jointly-diagonalizable frequency covariance matrices found by ILRTA looked similar to unconstrained ones found by LD-PSDTF (Fig. 2).

V. CONCLUSION

This paper described a new low-rank decomposition method called ILRTA and its application to single-channel source separation. It iterates the decorrelation of the TF bins of a mixture spectrogram and the low-rank decomposition of the decorrelated spectrogram in a way that a unified cost function is minimized. We showed that ILRTA outperformed computationally-intensive LD-PSDTF, a special case of LD-CTF. To draw the full potential of ILRTA as fast approximate LD-CTF, we need to develop a stable and rank-deficiency-free optimization algorithm. Since ILRTA is a general framework based on multiway (*e.g.*, temporal, frequency, and channel) covariance modeling, ILRTA could be straightforwardly extended for multichannel source separation by integrating IVA in the same way that NMF was extended to ILRMA (Fig. 1).

ACKNOWLEDGMENT

This study was partially supported by JSPS KAKENHI No. 26700020 and No. 16H01744, JSPS Grant-in-Aid for Fellows No. 16J05486, and JST ACCEL No. JPMJAC1602.

REFERENCES

[1] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. Plumbley, “Detection and classification of acoustic scenes and events,” *IEEE Trans. on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[2] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines,” in *ASRU*, 2015, pp. 504–511.

[3] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, “Automatic music transcription: Challenges and future directions,” *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

[4] D. Lee and H. Seung, “Algorithms for non-negative matrix factorization,” in *NIPS*, 2000, pp. 556–562.

[5] 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.

[6] D. W. Griffin and J. S. Lim, “Signal estimation from modified short-time Fourier transform,” *IEEE TASLP*, vol. 32, no. 2, pp. 236–243, 1984.

[7] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, “Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction,” in *SAPA*, 2008, pp. 23–28.

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

[9] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto, “Infinite positive semidefinite tensor factorization for source separation of mixture signals,” in *ICML*, 2013, pp. 576–584.

[10] ——, “Beyond NMF: Time-domain audio source separation without phase reconstruction,” in *ISMIR*, 2013, pp. 369–374.

[11] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. John Wiley & Sons, 2009.

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

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

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

[15] 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 TASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[16] E. Weinstein, M. Feder, and A. Oppenheim, “Multi-channel signal separation by decorrelation,” *IEEE TSAP*, vol. 1, no. 4, pp. 405–413, 1993.

[17] L. Molgedey and H. G. Schuster, “Separation of a mixture of independent signals using time delayed correlations,” *Physical Review Letters*, vol. 72, no. 23, pp. 3634–3636, 1994.

[18] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, “A blind source separation technique using second-order statistics,” *IEEE TSP*, vol. 45, no. 2, pp. 434–444, 1997.

[19] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, “Multichannel extensions of non-negative matrix factorization with complex-valued data,” *IEEE TASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[20] D. Fagot, H. Wendt, and C. Févotte, “Nonnegative matrix factorization for transform learning,” in *ICASSP*, 2018, pp. 2431–2435.

[21] R. Bhatia, *Positive Definite Matrices*. Princeton University Press, 2007.

[22] L. M. Bregman, “The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming,” *USSR CMMP*, vol. 7, no. 3, pp. 200–217, 1967.

[23] B. Kulis, M. Sustik, and I. Dhillon, “Low-rank kernel learning with Bregman matrix divergences,” *JMLR*, vol. 10, pp. 341–376, 2009.

[24] T. Ando, “Topics on operator inequalities,” Division of Applied Mathematics, Hokkaido University, Japan, Tech. Rep., 1974.

[25] T. Andoa, C.-K. Li, and R. Mathias, “Geometric means,” *Linear Algebra and its Applications*, vol. 385, no. 1, pp. 305–334, 2004.

[26] M. Congedo, B. Afsari, A. Barachant, and M. Moakher, “Approximate joint diagonalization and geometric mean of symmetric positive definite matrices,” *PLoS ONE*, vol. 10, no. 4, pp. 1–25, 2015.

[27] M. Nakano, H. Kameoka, J. L. Roux, Y. Kitano, N. Ono, and S. Sagayama, “Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta divergence,” in *MLSP*, 2010, pp. 283–288.

[28] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.