Independent Positive Semidefinite Tensor Analysis in Blind Source Separation

Rintaro Ikeshita
Hitachi, Ltd. Research & Development Group, Japan

***Abstract*—The issue of convolutive blind source separation (BSS) is addressed in this paper. Independent low-rank matrix analysis (ILRMA), unifying frequency-domain independent component analysis (FDICA) and nonnegative matrix factorization (NMF), is a method that has recently proposed to model low-rank structure of source spectra by using NMF in addition to independence between sources used in FDICA and independent vector analysis (IVA). Although ILRMA has been shown to provide better separation performance than FDICA and IVA, the frequency components of each source are assumed to be independent in ILRMA due to NMF modeling of source spectra, which may degrade its performance when the short-term Fourier transform (STFT) is unable to decorrelate the frequency components for each source. This paper therefore presents a new BSS method that unifies IVA and positive semidefinite tensor factorization (PSDTF). PSDTF models not only power spectra in the same way NMF does but also models the correlations between frequency bins in each source. The proposed method can be viewed as a multichannel extension of PSDTF and exploits both the independence between sources and the inter-frequency correlations as a clue for separating mixtures. Experimental results indicate the improved performance of our approach.**

***Index Terms*—Blind source separation, nonnegative matrix factorization, positive semidefinite tensor factorization, independent component analysis, independent vector analysis**

I. Introduction

Multichannel blind source separation (BSS) is a task designed to recover the original source signals from an observed mixture without having any knowledge of mixing systems or microphone positions. Frequency-domain independent component analysis (FDICA [1]–[3]) and independent vector analysis (IVA [4]–[6]) are common BSS approaches for convolutive mixtures, where the number of sources is no greater than that of microphones. Both approaches address BSS in the time-frequency (TF) domain using the short-term Fourier transform (STFT) and rely mainly on statistical independence between source signals as a clue for separating mixtures.

FDICA and independent vector analysis (IVA) have been extended in recent years to take advantage of specific TF structures of source spectra to be separated, in addition to the independence between source signals. The technique in which low-rank approximation of source spectra by nonnegative matrix factorization (NMF [7]–[9]) is applied is one such extension that has been very successful. This approach is called independent low-rank matrix analysis (ILRMA [10]–[13]) and was reported to outperform the separation performance of FDICA and IVA. Besides that, Gaussian ILRMA [11], unifying FDICA and Itakura-Saito NMF (IS-NMF [8]), is interesting from another perspective. This is, it can be regarded as the method that transforms the multichannel extension of IS-NMF (MNMF [14]–[16]) into an optimization problem of the separation filters in a demixing system [11]. From an optimization point of view, ILRMA converges much faster and is more stable than MNMF since it has fewer model parameters than MNMF. It can also utilize the fast and stable algorithm for the auxiliary-function-based IVA [17], called an iterative projection (IP [10]–[12], [17]) method, to optimize the separation filters.

While ILRMA has the above-mentioned advantages, note that frequency components are assumed to be independent in ILRMA as well as MNMF due to its NMF modeling of source spectra. If STFT perfectly decorrelates the frequency components for each source (and also, if each TF component is Gaussian distributed), then the model of Gaussian ILRMA is theoretically justified. Many signals such as speech, however, have high non-stationarity and cannot be decorrelated by STFT, resulting in degraded performance of ILRMA.

The same problem has been discussed in the single-channel BSS scenario, and consequently, positive semidefinite tensor factorization (PSDTF [18]–[20]) has been developed as an extension of NMF. In PSDTF based on log-determinant divergence (LD-PSDTF [18]), each frame of the source spectra is assumed to have a multivariate complex Gaussian distribution, and its covariance matrix is represented by a conic sum of Hermitian positive semidefinite basis matrices. (Note that IS-NMF can be achieved as the LD-PSDTF without non-diagonal bases.) This modeling enables PSDTF to take inter-frequency correlations in each source spectrum into account and to achieve better separation performance than NMF [19].

We thus propose in this paper a new BSS method that unifies IVA and PSDTF to improve the separation performance of Gaussian ILRMA. We call this method *independent positive semidefinite tensor analysis (IPSDTA)* (see Section II). By modeling source spectra with PSDTF, the proposed method can rely not only on the independence between source signals but also on the inter-frequency correlations in each source as a clue for separating mixtures, unlike previous methods including ILRMA and MNMF. We also develop a new optimization algorithm of the separation filters for IPSDTA by extending the original IP method (see Subsection III-A), and we reveal the relationship between them (see Section IV). Furthermore, we propose an approximation approach for IPSDTA to speed up the optimization of the model because IPSDTA is a multichannel extension of PSDTF and inevitably suffers from heavy computational cost (see Section V). The effectiveness

ISBN 978-90-827970-1-5 © EURASIP 2018

of the proposed approach is confirmed experimentally.

II. Proposed Generative Model

A. Blind source separation in the time-frequency domain

Assume that N sources are observed by N microphones. The source signals and the observations in each time-frequency slot $(f,t) \in [F] \times [T]$ are denoted as

$$\boldsymbol{s}_{f,t} = [s_{1,f,t}, \dots, s_{N,f,t}]^\top \in \mathbb{C}^N \tag{1}$$
$$\boldsymbol{x}_{f,t} = [x_{1,f,t}, \dots, x_{N,f,t}]^\top \in \mathbb{C}^N \tag{2}$$

respectively, where $f \in [F] := \{1, \dots, F\}$ and $t \in [T] := \{1, \dots, T\}$ indicate the frequency bin and time frame indices. Also, $(\cdot)^\top$ denotes the matrix transpose. In this paper, the linear mixing system

$$\boldsymbol{x}_{f,t} = A_f \boldsymbol{s}_{f,t}, \quad \boldsymbol{s}_{f,t} = W_f^h \boldsymbol{x}_{f,t}, \quad W_f^h = A_f^{-1} \tag{3}$$

is considered. Here, $(\cdot)^h$ means the Hermitian transpose, and

$$A_f = [\boldsymbol{a}_{1,f}, \dots, \boldsymbol{a}_{N,f}] \in \mathbb{C}^{N \times N} \tag{4}$$
$$W_f = [\boldsymbol{w}_{1,f}, \dots, \boldsymbol{w}_{N,f}] \in \mathbb{C}^{N \times N} \tag{5}$$

denote the mixing and demixing matrices whose columns are composed of the steering vectors $\boldsymbol{a}_{n,f} \in \mathbb{C}^N$ and the separation filters $\boldsymbol{w}_{n,f} \in \mathbb{C}^N$ for each source $n \in [N] := \{1, \dots, N\}$ and frequency $f \in [F]$, respectively.

In what follows, for the sake of simplicity, the following notations are used:

$$\boldsymbol{x}_t := [\boldsymbol{x}_{1,t}^\top, \dots, \boldsymbol{x}_{F,t}^\top]^\top \in \mathbb{C}^{NF} \tag{6}$$
$$\boldsymbol{s}_{n,t} := [s_{n,1,t}, \dots, s_{n,F,t}]^\top \in \mathbb{C}^F \tag{7}$$
$$\boldsymbol{a}_n := [\boldsymbol{a}_{n,1}^\top, \dots, \boldsymbol{a}_{n,F}^\top]^\top \in \mathbb{C}^{NF} \tag{8}$$
$$\boldsymbol{w}_n := [\boldsymbol{w}_{n,1}^\top, \dots, \boldsymbol{w}_{n,F}^\top]^\top \in \mathbb{C}^{NF} \tag{9}$$
$$\boldsymbol{W} := \begin{bmatrix} W_1 & O & \cdots & O \\ O & W_2 & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & W_F \end{bmatrix} \in \mathbb{C}^{NF \times NF}, \tag{10}$$

where $\boldsymbol{W}$ denotes the block diagonal matrix whose diagonal blocks are demixing matrices $\{W_f\}_{f=1}^F \subseteq \mathbb{C}^{N \times N}$.

B. Independence assumption of source signals

By using the relation (3), the likelihood of the observed signals $\{\boldsymbol{x}_t\}_t$ is represented as

$$p(\{\boldsymbol{x}_t\}_t) = p(\{\boldsymbol{s}_{n,t}\}_{n,t}) \cdot \prod_{f,t} |\det W_f|^2. \tag{11}$$

In this paper, in the same way as in the conventional IVA [4]–[6], the decomposition of

$$p(\{\boldsymbol{s}_{n,t}\}_{n,t}) = \prod_{n,t} p(\boldsymbol{s}_{n,t}) \tag{12}$$

is assumed. This implies that the source signals are independent of each other, and each signal is also independent in the time direction. Note that FDICA and ILRMA further assume the independence along the frequency direction for each source, i.e.,

$$p(\{\boldsymbol{s}_{n,t}\}_{n,t}) = \prod_{n,f,t} p(s_{n,f,t}), \tag{13}$$

which is not necessarily suited for real-world signals such as speech signals having strong correlations between neighboring frequency bins [21], [22].

C. Source spectrum model based on positive semidefinite tensor factorization

The source generative model of the Gaussian ILRMA [11] is characterized by the spatial model (3) and (11), the independence assumption (13), and the following source spectrum model (i) and (ii):

(i) Each time-frequency component $s_{n,f,t} \in \mathbb{C}$ of source $n \in [N]$ obeys the complex Gaussian distribution having zero mean and the variance $v_{n,f,t} \in \mathbb{R}_{>0}$.

(ii) For each source $n \in [N]$, the variances $\{v_{n,f,t}\}_{f,t}$, which encode the information of source spectra, are low-rank approximated by NMF.

The source spectrum model of ILRMA is based on NMF, and it thus inevitably ignores the correlations between frequency bins. We therefore employ PSDTF, which was first proposed as a monaural source separation method to overcome the independence assumption in NMF, to model source spectra in multi-channel source separation tasks. PSDTF is a factorization method that includes NMF as a special case, and was reported to outperform the separation performance of NMF in monaural source separation tasks [19].

The generative model in the proposed independent positive semidefinite tensor analysis (IPSDTA) is characterized by the spatial model (3) and (11), the independence assumption (12), and the following source spectrum model (i') and (ii'):

(i') Each time frame component $\boldsymbol{s}_{n,t} \in \mathbb{C}^F$ of source $n \in [N]$ obeys the multivariate complex Gaussian distribution having zero mean and the covariance matrix $R_{n,t} \in \mathbb{C}^{F \times F}$.

(ii") For each source $n \in [N]$, the covariance matrices $\{R_{n,t}\}_t$, which encode the information of the source spectrum as well as the correlation between frequency components, are modeled by PSDTF:

$$R_{n,t} = \sum_{k=1}^{K_n} U_{n,k} v_{n,k,t}. \tag{14}$$

Here, for each source $n \in [N]$, K_n is the number of bases in PSDTF, and $U_{n,k} \in \mathbb{C}^{F \times F}$ is a time-invariant Hermitian positive semidefinite matrix of the k-th basis, and $v_{n,k,t} \in \mathbb{R}_{>0}$ is a time-variant activation for the k-th basis and t-th frame.

Note that in the proposed generative model defined by (3), (11), (12), (i'), and (ii'), the set of model parameters Θ is given by

$$\Theta = \{W_f,\, U_{n,k},\, v_{n,k,t},\}_{n,f,t,k}. \tag{15}$$

The optimization of the parameters is based on the maximization of the log-likelihood, which is equivalent to the minimization of the cost function $J(\Theta)$ defined by

$$
\begin{aligned}
J(\Theta) &:= -\frac{1}{T}\sum_{n,t}\log p(\boldsymbol{s}_{n,t}) - 2\sum_{f}\log|\det W_f| \\
&= \frac{1}{T}\sum_{n,t}\left\{\boldsymbol{s}_{n,t}^h R_{n,t}^{-1}\boldsymbol{s}_{n,t} + \log\det R_{n,t}\right\} \\
&\quad - 2\sum_{f}\log|\det W_f| + C, \qquad (16)
\end{aligned}
$$

where C is a constant independent of Θ.

III. Optimization Algorithm of the model

In this section, an algorithm derived to minimize the cost function (16) is presented. It is based on a block coordinate descent method that alternately updates the separation filters $\{W_f\}_f$ and the PSDTF parameters $\{U_{n,k}, v_{n,k,t}\}_{n,k,t}$. After the convergence of the algorithm, the amplitude ambiguities of separated signals are restored by using the projection back technique [23], [24] as follows:

$$
s_{n,f,t}\boldsymbol{a}_{n,f} = (\boldsymbol{w}_{n,f}^h \boldsymbol{x}_{f,t})(W_f^h)^{-1}e_n \in \mathbb{C}^N, \qquad (17)
$$

where $e_n \in \mathbb{R}^N$ denotes the unit vector with the n-th element equal to one and the others equal to zero.

A. Optimization of separation filters

The natural gradient method [25] has conventionally been used to optimize the separation filters in FDICA and IVA. In recent years, however, the iterative projection (IP) method, proposed for auxiliary-function based IVA [17] and used in ILRMA [10]–[13] has been attracting a lot of attention because it is more stable and rather faster to converge. We therefore extend the IP method to optimize the separation filters in IPSDTA. The proposed IP method is identical to the block coordinate descent that iteratively optimizes the separation filter $\boldsymbol{w}_n$ for each source $n \in [N]$.

The stationary points of the cost function (16) with respect to $\boldsymbol{w}_n$ satisfy

$$
W^h G_n \boldsymbol{w}_n = g_n := [e_n^\top \mid \cdots \mid e_n^\top]^\top \in \mathbb{C}^{NF} \qquad (18)
$$

$$
G_n = \frac{1}{T}\sum_t (\boldsymbol{x}_t \boldsymbol{x}_t^h) \odot (J_N \otimes (\overline{R}_{n,t})^{-1}) \in \mathbb{C}^{NF\times NF}, \quad (19)
$$

where $g_n \in \mathbb{R}^{NF}$ is a vector obtained by arranging F e_n's in the vertical direction, $\odot$ denotes a Hadamard product, $J_N \in \mathbb{R}^{N\times N}$ is the matrix whose elements are all equal to 1, $\overline{R}$ is defined as $(R^h)^\top$ for matrix R, and $\otimes$ is a Kronecker product and calculated by $A \otimes [b_{ij}]_{ij} = [Ab_{ij}]_{ij}$. Note that G_n is positive semidefinite in equation (19), since both $\boldsymbol{x}_t\boldsymbol{x}_t^h$ and $J_N \otimes (\overline{R}_{n,t})^{-1}$ are positive semidefinite and the Hadamard product of two positive semidefinite matrix is also positive semidefinite.

The solution $\tilde{\boldsymbol{w}}_n$ of (18) with respect to $\boldsymbol{w}_n$ is written by

$$
\tilde{\boldsymbol{w}}_n = (W^h G_n)^{-1}\Lambda_n g_n = G_n^{-1}\Lambda_n \boldsymbol{a}_n \in \mathbb{C}^{NF} \qquad (20)
$$

$$
\begin{aligned}
\Lambda_n &= \begin{bmatrix} \lambda_{n,1}I_N & O & \cdots & O \\ O & \lambda_{n,2}I_N & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & \lambda_{n,F}I_N \end{bmatrix} \\
&= I_N \otimes \mathrm{diag}\{\lambda_{n,1},\ldots,\lambda_{n,F}\} \in \mathbb{C}^{NF\times NF}, \qquad (21)
\end{aligned}
$$

where I_N is an identity matrix of size $N\times N$ and $\{\lambda_{n,f}\}_{f=1}^F \subseteq \mathbb{C}$ are complex values satisfying (22) below:

$$
1 = \lambda_{n,f}\sum_{f'=1}^{F}\overline{\lambda}_{n,f'} b_{n,f',f} \qquad (22)
$$

$$
b_{n,f',f} = \boldsymbol{a}_{n,f'}^h (G_n^{-h})_{f',f}\boldsymbol{a}_{n,f}. \qquad (23)
$$

In (23), $(G_n^{-h})_{f,f'} \in \mathbb{C}^{N\times N}$ means the (f,f')-th block in the block matrix $G_n^{-h} \in \mathbb{C}^{NF\times NF}$ having F^2 matrices of size $N \times N$. As an alternative way to solve (22) rigorously, we propose a fixed point iteration to optimize $\{\lambda_{n,f}\}_{f=1}^F$, whose update rules are represented by

$$
\boldsymbol{\lambda}_n \leftarrow \boldsymbol{1} \oslash (B_n^\top \overline{\boldsymbol{\lambda}}_n), \qquad (24)
$$

where $\boldsymbol{\lambda}_n := [\lambda_{n,1},\ldots,\lambda_{n,F}]^\top \in \mathbb{C}^F$, $\boldsymbol{1} := [1,\ldots,1]^\top \in \mathbb{R}^F$, $\oslash$ is a coordinate-wise quotient of two vectors and $B_n \in \mathbb{C}^{F\times F}$ is a matrix whose (i,j)-th element is equal to $b_{n,i,j}$.

B. Optimization of PSDTF parameters

As for the optimization of PSDTF parameters, observe that the cost function (16) is separable for each source $n \in [N]$. (We consider $\{W_f\}_f$ as constants in this stage.) Then, we can immediately apply the EM-algorithm for PSDTF proposed in [20]. The update rules are expressed as follows:

$$
v_{n,k,t} = \frac{1}{F}\cdot \mathrm{tr}\left(U_{n,k}^{-1}\Phi_{n,k,t}\right) \qquad (25)
$$

$$
U_{n,k} = \frac{1}{T}\sum_{t=1}^{T}\frac{\Phi_{n,k,t}}{v_{n,k,t}}, \qquad (26)
$$

where

$$
\Phi_{n,k,t} = \hat{\boldsymbol{s}}_{n,k,t}\hat{\boldsymbol{s}}_{n,k,t}^h + \hat{R}_{n,k,t} \qquad (27)
$$

$$
\hat{\boldsymbol{s}}_{n,k,t} = R_{n,k,t}(R_{n,t})^{-1}\boldsymbol{s}_{n,t} \qquad (28)
$$

$$
\hat{R}_{n,k,t} = R_{n,k,t} - R_{n,k,t}\left(R_{n,t}\right)^{-1}R_{n,k,t} \qquad (29)
$$

$$
R_{n,t} = \sum_{k=1}^{K_n} R_{n,k,t} = \sum_{k=1}^{K_n} v_{n,k,t}U_{n,k}. \qquad (30)
$$

C. Summary of the proposed algorithm

The procedure of the proposed algorithm is as follows:

1) Initialize the parameters Θ and $\{\lambda_{n,f}\}_{n,f}$.
2) Iterate the following steps until convergence.
   a) Calculate $\{\boldsymbol{s}_{n,t}\}_{n,t}$ and $\{\boldsymbol{a}_n\}_n$ by (3).
   b) Iteratively update $\{v_{n,t,k}, U_{n,k}\}_{n,t,k}$ by (25)–(30).
   c) Calculate $\{G_n\}_n$ by (14) and (19).
   d) Iteratively update $\{\lambda_{n,f}\}_{n,f}$ by (23)–(24).
   e) Iteratively update $\{\boldsymbol{w}_{n,f}\}_{n,f}$ by (20)–(21).
3) Calculate the separated signals by (3) and (17).

IV. Relation to prior work

Let us confirm that the proposed IPSDTA is identical to the conventional Gaussian ILRMA [11] when the basis matrices $\{U_{n,k}\}_{n,k}$ of PSDTF are limited to diagonal matrices.

If all the basis matrices are diagonals, then $\{R_{n,t}\}_{n,t}$ are also diagonals, and hence the factorization of (14) is simply the NMF decomposition: $(R_{n,t})_{f,f} = \sum_{k=1}^{K_n}(U_{n,k})_{f,f} \cdot v_{n,k,t}$, where $(R_{n,t})_{f,f}$ and $(U_{n,k})_{f,f}$ are the (f, f)-th elements of $R_{n,t}$ and $U_{n,k}$, respectively. Then, the cost function (16) turns out to be identical to that of the Gaussian ILRMA [11]. In this sense, the proposed IPSDTA can be considered as an extension of the Gaussian ILRMA.

Next, we check that the proposed optimization algorithm of the separation filters derived in Subsection III-A is the same as the original IP method for IVA [17] and ILRMA [11] if all the $\{R_{n,t}\}_{n,t}$ are diagonals. If $\{R_{n,t}\}_{n,t}$ are diagonals, the G_n in (19) can be simplified to the block diagonal matrix whose diagonal blocks are expressed as

$$G_{n,f} \coloneqq \frac{1}{T}\sum_t \frac{\boldsymbol{x}_{f,t}\boldsymbol{x}_{f,t}^h}{(R_{n,t})_{f,f}} \in \mathbb{C}^{N\times N}, \quad f \in [F]. \tag{31}$$

Then, the equations (20) and (22) are also decomposed into

$$\tilde{\boldsymbol{w}}_{n,f} = \lambda_{n,f}(W_f^h G_{n,f})^{-1} e_n \tag{32}$$

$$|\lambda_{n,f}|^2 = \frac{1}{\boldsymbol{a}_{n,f}^h G_{n,f}^{-1} \boldsymbol{a}_{n,f}} \tag{33}$$

for each $f \in [F]$, which are identical to the correspondences in the original IP method [11], [17].

V. Approximation for computational cost reduction

As an adverse effect of increasing the model flexibility by using PSDTF, IPSDTA suffers from heavy computational cost in each iteration (Step 2 in Subsection III-C). In particular, the following computations are cumbersome:

- Computation of $\{G_n\}_n$ in (19): $O(N^3F^2T)$
- Computation of $\{\boldsymbol{w}_n\}_n$ in (20): $O(N^4F^3)$
- Computation of $\{R_{n,k,t}\}_{n,k,t}$ in (29): $O(KF^3T)$,

where we define $K \coloneqq \sum_{n=1}^N K_n$. We therefore propose an approximation of the PSDTF modeling in the following two steps: 1) First, we divide the set of all frequency bins $[F]$ into a family of sets of frequency bins $\mathcal{F}$ as follows [26]:

$$\mathcal{F} \subseteq 2^{[F]} \quad \text{s.t.} \quad \sqcup_{E\in\mathcal{F}} E = [F], \tag{34}$$

where $\sqcup$ denotes the disjoint union of sets; and 2) we impose the independence assumption, instead of (12), as follows:

$$p(\{\boldsymbol{s}_{n,t}\}_{n,t}) = \prod_{n,t}\prod_{E\in\mathcal{F}} p(\{\boldsymbol{s}_{n,f,t}\}_{f\in E}), \tag{35}$$

which is equivalent to the block decomposition of the bases $\{U_{n,k}\}_{n,k}$ in PSDTF by using $\mathcal{F}$. This approximation dramatically reduces the computational load, resulting in

- Computation of $\{G_n\}_n$ in (19): $O(\sum_{E\in\mathcal{F}} N^3E^2T)$
- Computation of $\{\boldsymbol{w}_n\}_n$ in (20): $O(\sum_{E\in\mathcal{F}} N^4E^3)$
- Computation of $\{R_{n,k,t}\}_{n,k,t}$ in (29): $O(\sum_{E\in\mathcal{F}} KE^3T)$.

TABLE I
Experimental conditions

Sampling rate	16 kHz
Frame length	4096 points (256 ms)
Frame shift	1024 points (64 ms)
Window function	Hanning
Signal length	10 s
Mixture signal ($N = 2$)	(female, female) or (male, male)
Reverberation time (RT_{60})	130 ms/250 ms
Microphone spacing	5 cm/1 m

VI. Experiment

A. Conditions

The performance of the proposed method was evaluated in an experiment we carried out using live recorded speech data in the *dev1* dataset provided by SiSEC2008 [27]. We obtained 16 determined stereo ($N = 2$) mixtures in total by adding clean spatial images in the dataset. We tested three methods in the experiment: Gaussian ILRMA [11] and the two proposed IPSDTA with $\mathcal{F}_k = \{E_i \subseteq [F] \mid i = 1,\ldots,k\}$ for $k \in \{512, 1024\}$ in (34), where we define

$$E_i = \{\lfloor \frac{F}{k}\cdot(i-1)\rfloor + 1, \ldots, \lfloor \frac{F}{k}\cdot i\rfloor\} \tag{36}$$

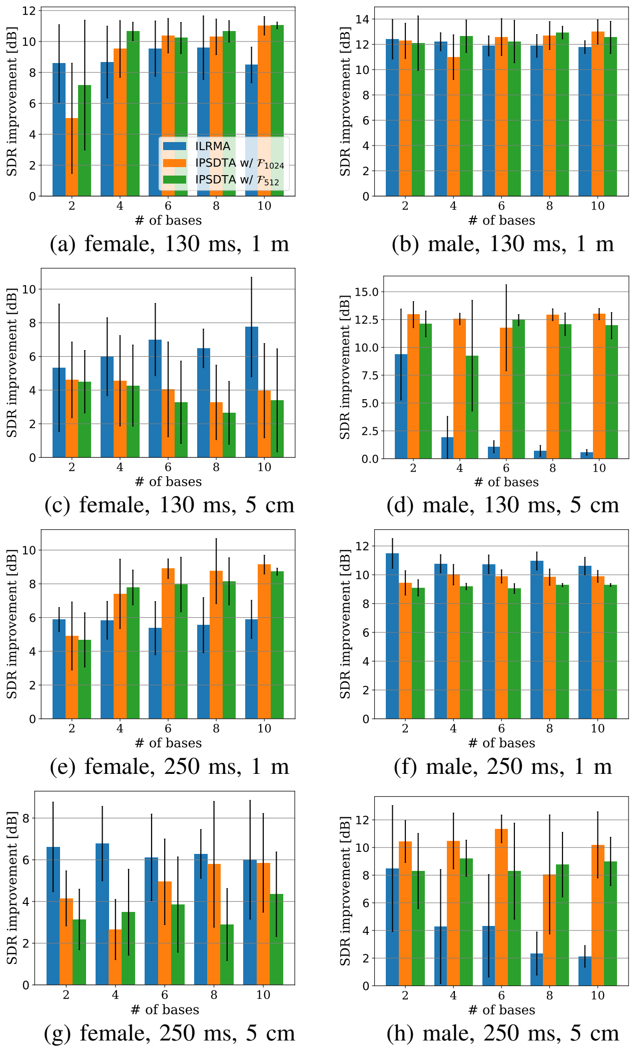
for $i = 1,\ldots,k$.

The number of iterations in the optimization, corresponding to Step 2 in Subsection III-C, was set to 400 for all methods, and each iterative update (Step 2 (b)(d)(e) in Subsection III-C) was performed once in each iteration. The number of bases in NMF and PSDTF was set to $K_n \in \{2, 4, 6, 8, 10\}$ for each source $n \in [N]$. The separation filters $\{W_f\}_f$ were initialized by the identity matrix, while each parameter of NMF and PSDTF was randomly initialized from the uniform distribution over $(0, 1)$, except that the bases of PSDTF were chosen to be diagonal matrices in the initialization. We also initialized $\boldsymbol{\lambda}_n = \boldsymbol{1}$ for all $n \in [N]$ in (24). The performance was evaluated by averaging the signal-to-distortion ratio (SDR) improvements [28] for all mixtures and trials. The other experimental conditions are listed in Table I.

B. Results

Figure 1 shows the average SDR improvements for the recording conditions in Table I. The average was taken over 10 samples, namely, 2 mixtures and 5 trials with the random initialization. Although the proposed IPSDTA gives somewhat low scores in (c) and (g), it generally outperforms or gives comparable results to ILRMA, implying the validity of the proposed approach. The reason IPSDTA performance degrades in (c) and (g) may be because the optimizations were trapped in bad local optima since IPSDTA has a larger number of parameters than ILRMA. When ILRMA is used to separate speech mixtures, the number of NMF bases should be set at 2 because NMF has difficulty capturing complex speech spectrograms [11]. For IPSDTA, however, it is considered from (a), (e) and (g) that the number of bases should not be too few. This suggests that PSDTF in IPSDTA can capture inter-frequency correlations of source spectra and can potentially further improve the separation performance of ILRMA.

ISBN 978-90-827970-1-5 © EURASIP 2018



(a) female, 130 ms, 1 m

(b) male, 130 ms, 1 m

(c) female, 130 ms, 5 cm

(d) male, 130 ms, 5 cm

(e) female, 250 ms, 1 m

(f) male, 250 ms, 1 m

(g) female, 250 ms, 5 cm

(h) male, 250 ms, 5 cm

Fig. 1. Average SDR improvements and their standard deviations for Gaussian ILRMA and the proposed IPSDTA with $\mathcal{F}_{512}$ or $\mathcal{F}_{1024}$, for the recording conditions in Table I: (mixture signal, reverberation time, microphone spacing). The horizontal axis denotes the number of bases K_n for each source in NMF and PSDTF. The legend is the same for all graphs.

VII. Conclusion

This paper presented a new BSS approach that estimates a source spatial model by IVA and a source spectrum model by PSDTF, which enables us to exploit both the independence between sources and the inter-frequency correlations in each source spectrum as a clue for separating mixtures. The experimental results show the validity of the proposed approach.

References

[1] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.

[2] H. Saruwatari *et al.*, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Transactions on Audio, speech, and language processing*, vol. 14, no. 2, pp. 666–678, 2006.

[3] H. Sawada *et al.*, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE transactions on speech and audio processing*, vol. 12, no. 5, pp. 530–538, 2004.

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

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

[6] A. Hiroe, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," in *Proc. ICA*, 2006, pp. 601–608.

[7] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788, 1999.

[8] 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.

[9] A. Liutkus, D. Fitzgerald, and R. Badeau, "Cauchy nonnegative matrix factorization," in *Proc. WASPAA*, 2015, pp. 1–5.

[10] D. Kitamura *et al.*, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in *Proc. ICASSP*, 2015, pp. 276–280.

[11] D. Kitamura *et al.*, "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.

[12] S. Mogami *et al.*, "Independent low-rank matrix analysis based on complex Student's t-distribution for blind audio source separation," in *Proc. MLSP*, 2017.

[13] D. Kitamura, N. Ono, and H. Saruwatari, "Experimental analysis of optimal window length for independent low-rank matrix analysis," in *Proc. EUSIPCO*, 2017, pp. 1170–1174.

[14] 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.

[15] H. Sawada *et al.*, "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.

[16] S. Arberet *et al.*, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. ISSPA*. IEEE, 2010, pp. 1–4.

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

[18] K. Yoshii *et al.*, "Infinite positive semidefinite tensor factorization for source separation of mixture signals," in *Proc. ICML*, 2013, pp. 576–584.

[19] K. Yoshii, K. Itoyama, and M. Goto, "Student's t nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation," in *Proc. ICASSP*, 2016, pp. 51–55.

[20] A. Liutkus and K. Yoshii, "A diagonal plus low-rank covariance model for computationally efficient source separation," in *Proc. MLSP*, 2017, pp. 1–6.

[21] G.-J. Jang, I. Lee, and T.-W. Lee, "Independent vector analysis using non-spherical joint densities for the separation of speech signals," in *Proc. ICASSP*, 2007, vol. 2, pp. II–629.

[22] I. Lee and G.-J. Jang, "Independent vector analysis based on overlapped cliques of variable width for frequency-domain blind signal separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 113, 2012.

[23] 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, pp. 1–24, 2001.

[24] K. Matsuoka, "Minimal distortion principle for blind source separation," in *Proc. SICE*, 2002, vol. 4, pp. 2138–2143.

[25] S.-I. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Proc. NIPS*, 1996, pp. 757–763.

[26] R. Ikeshita *et al.*, "Independent vector analysis with frequency range division and prior switching," in *Proc. EUSIPCO*, 2017, pp. 2329–2333.

[27] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. ICA*, 2009, pp. 734–741.

[28] 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.