VECTORWISE COORDINATE DESCENT ALGORITHM FOR SPATIALLY REGULARIZED INDEPENDENT LOW-RANK MATRIX ANALYSIS

Yoshiki Mitsui[1], *Norihiro Takamune*[1], *Daichi Kitamura*[1],
Hiroshi Saruwatari[1], *Yu Takahashi*[2], *Kazunobu Kondo*[2]

[1] The University of Tokyo, Tokyo, Japan [2] Yamaha Corporation, Shizuoka, Japan

ABSTRACT

Audio source separation is an important problem for many audio applications. Independent low-rank matrix analysis (ILRMA) is a recently proposed algorithm that employs the statistical independence between sources and the low-rankness of the time-frequency structure in each source. As reported in this paper, we have developed a new framework that enables us to introduce a spatial regularization of the demixing matrix in ILRMA. Since the conventional optimization cannot be applied to this regularized ILRMA, we derive a novel approach based on vectorwise coordinate descent, which does not require a step-size parameter and guarantees convergence. In experiments, ILRMA with beamforming-based regularization is evaluated as an application of the proposed framework.

***Index Terms*—** Audio source separation, independent low-rank matrix analysis, spatial regularization, vectorwise coordinate descent.

1. INTRODUCTION

Audio source separation is a technique for estimating individual audio sources from an observed mixture signal. In particular, blind source separation (BSS) aims to separate the sources without knowing their spatial arrangements, and many methods based on frequency-domain independent component analysis (FDICA) have been proposed so far [1, 2, 3, 4]. FDICA estimates the frequency-wise demixing matrix by assuming statistical independence between sources, where the permutation problem (alignment of estimated components over all frequency bins) must be solved. Independent vector analysis (IVA) [5, 6] and independent low-rank matrix analysis (ILRMA) [7, 8] are more sophisticated approaches that simultaneously estimate the demixing matrix and solve the permutation problem. IVA employs a generative model of source frequency vectors, which ensures higher-order correlation among frequency components. ILRMA extends the vector model to a low-rank time-frequency matrix model incorporating nonnegative matrix factorization (NMF) [9, 10]. Since co-occurrence among frequency or time-frequency slots in each source is ensured, the permutation problem can be avoided. Also, for IVA and ILRMA, a fast and stable optimization algorithm called iterative projection (IP) [11] has been derived that does not require a step-size parameter and guarantees theoretical convergence. However, IVA sometimes causes the block permutation problem [12], which is a misalignment in the low- and high-frequency bands, and ILRMA often fails to separate speech mixtures because the spectrogram of speech signals is not low-rank and the NMF optimization is trapped at a poor local minimum [7].

This work was partly supported by SECOM Science and Technology Foundation and JSPS KAKENHI Grant Numbers JP17H06101 and JP17H06572.

In FDICA, a regularization term of the demixing matrix is often introduced into the cost function to improve the separation accuracy or increase the speed of optimization. For example, in [4, 13, 14, 15], a spatial regularizer based on a fixed beamforming technique [16] was introduced, where it was assumed that either or both the microphone spacing and the locations of sources are known. In [17], a penalty term that solves the scale ambiguity of the demixing matrix was imposed. In these methods, the optimization algorithm is derived by a naive gradient-based method. Since the regularizer is linearly combined with the main cost function in FDICA, the gradient-based update rules can be easily implemented by concatenating the gradients for both the cost and the regularizer. However, its convergence is not guaranteed and the step-size parameter must be carefully tuned. Regarding IP-based algorithms, to the best of our knowledge, the update rules of the ICA or IVA cost function with an arbitrary regularizer have not yet been derived.

In this paper, we develop a new IP-like optimization algorithm called *vectorwise coordinate descent (VCD)* for ILRMA, which can theoretically guarantee convergence even if the regularizer of the demixing matrix is imposed on the original cost function. VCD can optimize matrix variables in a vectorwise manner, which is performed by a computationally efficient closed-form solution. On the basis of this new framework, in this paper we introduce a spatial regularizer obtained by the beamforming technique into ILRMA and show that the spatial regularizer can markedly improve both the separation accuracy and the stability of the optimization in the case of speech mixture separation. Also, the proposed algorithm can be used not only for a beamforming-based regularizer but also for many types of regularization of the demixing matrix. This paper provides an example of the application of this algorithm when the microphone spacing is known but the source locations are unknown in advance.

2. CONVENTIONAL ILRMA

2.1. Formulation

Let N and M be the numbers of sources and microphones, respectively. The complex-valued short-time Fourier transform (STFT) coefficients of source, observed, and separated signals are defined as

$$\boldsymbol{s}_{ij} = (s_{ij,1}, \ldots, s_{ij,n}, \ldots, s_{ij,N})^{\mathsf{T}}, \tag{1}$$

$$\boldsymbol{x}_{ij} = (x_{ij,1}, \ldots, x_{ij,m}, \ldots, x_{ij,M})^{\mathsf{T}}, \tag{2}$$

$$\boldsymbol{y}_{ij} = (y_{ij,1}, \ldots, y_{ij,n}, \ldots, y_{ij,N})^{\mathsf{T}}, \tag{3}$$

where $i = 1, \ldots, I$; $j = 1, \ldots, J$; $n = 1, \ldots, N$; and $m = 1, \ldots, M$ are the integral indexes of the frequency bins, time frames, sources, and channels, respectively, and $^{\mathsf{T}}$ denotes a transpose. When the mixing system is time-invariant and the window length in the STFT is sufficiently longer than the impulse responses between sources and microphones, the following instantaneous

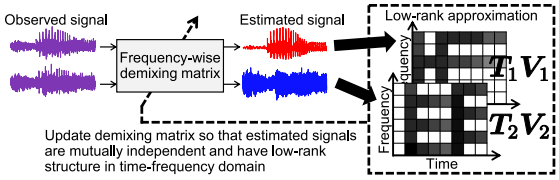

Fig. 1. Principle of source separation based on ILRMA.

mixture model holds in a frequency domain:

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij}, \tag{4}$$

where $\boldsymbol{A}_i$ is the mixing matrix. This assumption of a mixing system is often called the rank-1 spatial model [18]. If the number of sources equals the number of channels ($M = N$), the demixing matrix $\boldsymbol{W}_i = (\boldsymbol{w}_{i,1}, \ldots, \boldsymbol{w}_{i,N})^{\mathsf{H}} = \boldsymbol{A}_i^{-1}$ can be defined, where $^{\mathsf{H}}$ denotes a Hermitian transpose, and the separated signals are represented as

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}. \tag{5}$$

The goal of FDICA, IVA, or ILRMA is to estimate $\boldsymbol{W}_i$ with the correct source permutations over all frequency bins.

2.2. Cost function

ILRMA [7] is a method unifying IVA and NMF, namely, the demixing matrix $\boldsymbol{W}_i$ is estimated by maximizing the independence between sources while the power spectrogram of the estimated source $|y_{ij,n}|^2 = |\boldsymbol{w}_{i,n}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2$ is modeled by the low-rank decomposition in NMF. Fig. 1 shows the separation mechanism of ILRMA, where $\boldsymbol{T}_n \in \mathbb{R}_{\geq 0}^{I \times L}$ and $\boldsymbol{V}_n \in \mathbb{R}_{\geq 0}^{L \times J}$ are the basis and activation matrices for the nth estimated source, respectively, and L is the number of bases for the nth source. The demixing matrix $\boldsymbol{W}_i$ and the separated signal $\boldsymbol{y}_{ij}$ are optimized so that the spectrogram of each source tends to be a low-rank matrix. ILRMA can avoid the permutation problem by assuming co-occurrence among the time-frequency slots of the same source and restricting them to be a low-rank matrix represented as $\boldsymbol{T}_n \boldsymbol{V}_n$. It has been revealed that ILRMA is equivalent to multichannel NMF (MNMF) [19, 20, 21] only when the rank-1 spatial model (4) is assumed, which yields a more stable and efficient algorithm than MNMF.

The cost function in ILRMA is defined as follows:

$$\begin{aligned} \mathcal{J} = \sum_i \left(\sum_n \boldsymbol{w}_{i,n}^{\mathsf{H}} \boldsymbol{D}_{i,n} \boldsymbol{w}_{i,n} - \log |\det \boldsymbol{W}_i|^2 \right) \\ + \frac{1}{J} \sum_{i,j,n} \log \sum_l t_{il,n} v_{lj,n}, \end{aligned} \tag{6}$$

$$\boldsymbol{D}_{i,n} = \frac{1}{J} \sum_j \frac{\boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathsf{H}}}{\sum_l t_{il,n} v_{lj,n}}, \tag{7}$$

where $l = 1, \ldots, L$ is the integral index of the bases and $t_{il,n}$ and $v_{lj,n}$ are the nonnegative elements of $\boldsymbol{T}_n$ and $\boldsymbol{V}_n$, respectively. The rank-L matrix $\boldsymbol{T}_n \boldsymbol{V}_n$ corresponds to the NMF decomposition and represents a power spectrogram model of the nth source. The first and second terms in (6) are equivalent to the cost function in IVA, which evaluates the independence between sources, and the first and third terms in (6) are equivalent to the cost function in NMF based on Itakura–Saito divergence [22]. Regarding the demixing matrix, the cost function (6) is the sum of the quadratic form of $\boldsymbol{w}_{i,n}$ and the negative log-determinant of $\boldsymbol{W}_i$. This type of cost function can be efficiently optimized by IP [11, 7], which is given by

$$\boldsymbol{u}_{i,n} = \boldsymbol{D}_{i,n}^{-1} \boldsymbol{W}_i^{-1} \boldsymbol{e}_n, \tag{8}$$

$$\boldsymbol{w}_{i,n} \leftarrow \frac{\boldsymbol{u}_{i,n}}{\sqrt{\boldsymbol{u}_{i,n}^{\mathsf{H}} \boldsymbol{D}_{i,n} \boldsymbol{u}_{i,n}}}, \tag{9}$$

where $\boldsymbol{e}_n$ denotes the unit vector with the nth element equal to unity.

3. PROPOSED METHOD

3.1. Motivation

In independence-based BSS, the regularization of $\boldsymbol{w}_{i,n}$ has the potential to improve its optimization stability. In [4, 13, 14], a beamforming-based regularizer was imposed to precisely solve the permutation problem. In [15], a similar regularizer was utilized for annealing so that $\boldsymbol{w}_{i,n}$ rapidly converges to a better solution. In [17], a penalty term of $\boldsymbol{w}_{i,n}$ was added to avoid the scale ambiguity of the estimated signal $y_{ij,n}$. However, these methods are based on the gradient-based algorithm without guaranteeing its convergence, and, to the best of our knowledge, convergence-guaranteed updates for the regularized optimization problem of $\boldsymbol{w}_{i,n}$ have never been considered. Motivated by this issue, as reported in this section, we propose the addition of a regularizer to ILRMA and develop a new algorithm called VCD that guarantees the convergence without a step-size parameter.

3.2. Cost function of spatially regularized ILRMA

Let $\widehat{\boldsymbol{W}}_i = (\widehat{\boldsymbol{w}}_{i,1}, \ldots, \widehat{\boldsymbol{w}}_{i,N})^{\mathsf{H}}$ be the supervisor of the demixing matrix, and we consider the problem of finding the optimal $\boldsymbol{W}_i$ around $\widehat{\boldsymbol{W}}_i$ in ILRMA-based BSS. This problem can be solved by imposing the regularizer of $\boldsymbol{w}_{i,n}$ on the original cost function (6) as follows:

$$\begin{aligned} \mathcal{J}_{\mathrm{R}} &= \mathcal{J} + \sum_{i,n} \lambda_n \|\boldsymbol{w}_{i,n} - \widehat{\boldsymbol{w}}_{i,n}\|^2 \\ &= \sum_i \Bigg[\sum_n \Big(\boldsymbol{w}_{i,n}^{\mathsf{H}} \widehat{\boldsymbol{D}}_{i,n} \boldsymbol{w}_{i,n} - \lambda_n \widehat{\boldsymbol{w}}_{i,n}^{\mathsf{H}} \boldsymbol{w}_{i,n} \\ &\quad - \lambda_n \boldsymbol{w}_{i,n}^{\mathsf{H}} \widehat{\boldsymbol{w}}_{i,n} \Big) - \log |\det \boldsymbol{W}_i|^2 \Bigg] + \mathcal{C}, \end{aligned} \tag{10}$$

where λ_n is the weight parameter of the regularizer, $\widehat{\boldsymbol{D}}_{i,n} = \boldsymbol{D}_{i,n} + \lambda_n \boldsymbol{I}_N$, $\boldsymbol{I}_N$ is the $N \times N$ identity matrix, and $\mathcal{C}$ denotes terms independent of $\boldsymbol{w}_{i,n}$. Since (10) includes the linear terms $\widehat{\boldsymbol{w}}_{i,n}^{\mathsf{H}} \boldsymbol{w}_{i,n}$ and $\boldsymbol{w}_{i,n}^{\mathsf{H}} \widehat{\boldsymbol{w}}_{i,n}$, IP cannot be applied unlike in the case of (6).

3.3. Derivation of vectorwise coordinate descent

To derive the novel IP-like optimization algorithm for (10), we calculate the partial derivative of (10) w.r.t. $\boldsymbol{w}_{i,n}^*$, where * denotes the complex conjugate. Therefore, $\boldsymbol{w}_{i,n}$ in the matrix variable $\boldsymbol{W}_i$ is cyclically updated, resulting in VCD to find the optimal $\boldsymbol{w}_{i,n}$ with a computationally efficient closed-form solution.

First, we arrange the term $\log |\det \boldsymbol{W}_i|^2$ in (10) by using $\boldsymbol{B}_i = (\boldsymbol{b}_{i,1}, \ldots, \boldsymbol{b}_{i,N})$, which is the adjugate matrix of $\boldsymbol{W}_i$ and defined as

$$[\boldsymbol{B}_i]_{pq} = (-1)^{p+q} \breve{\boldsymbol{W}}_{i,qp}, \tag{11}$$

where $[\boldsymbol{B}_i]_{pq}$ is the (p, q) entry of $\boldsymbol{B}_i$ and $\breve{\boldsymbol{W}}_{i,qp}$ is the (q, p) minor determinant of $\boldsymbol{W}_i$. Note that the column vector of $\boldsymbol{B}_i$, $\boldsymbol{b}_{i,n}$, only depends on $\boldsymbol{w}_{i,n'}$ $(n' \neq n)$ and becomes independent of $\boldsymbol{w}_{i,n}$ from its definition. By using the property of cofactor expansion, we obtain $\det \boldsymbol{W}_i = \boldsymbol{w}_{i,n}^{\mathsf{H}} \boldsymbol{b}_{i,n}$. Then, we can calculate the partial derivative of $\log |\det \boldsymbol{W}_i|^2$ w.r.t. $\boldsymbol{w}_{i,n}^*$ as

$$\frac{\partial \log |\det \boldsymbol{W}_i|^2}{\partial \boldsymbol{w}_{i,n}^*} = \frac{\partial \log |\boldsymbol{w}_{i,n}^{\mathsf{H}} \boldsymbol{b}_{i,n}|^2}{\partial \boldsymbol{w}_{i,n}^*} = \frac{\boldsymbol{b}_{i,n}}{\boldsymbol{w}_{i,n}^{\mathsf{H}} \boldsymbol{b}_{i,n}}. \tag{12}$$

Next, we derive a stationary point of $\boldsymbol{w}_{i,n}$. By using (12), we can obtain the partial derivative of $\mathcal{J}_{\mathrm{R}}$ w.r.t. $\boldsymbol{w}_{i,n}^*$ as follows:

$$\frac{\partial \mathcal{J}_{\mathrm{R}}}{\partial \boldsymbol{w}_{i,n}^*} = \widehat{\boldsymbol{D}}_{i,n}\boldsymbol{w}_{i,n} - \lambda_n \widehat{\boldsymbol{w}}_{i,n} - \frac{\boldsymbol{b}_{i,n}}{\boldsymbol{w}_{i,n}^{\mathsf{H}}\boldsymbol{b}_{i,n}}. \tag{13}$$

From $\partial \mathcal{J}_{\mathrm{R}}/\partial \boldsymbol{w}_{i,n}^* = \boldsymbol{0}$, we have

$$\boldsymbol{w}_{i,n} = \widehat{\boldsymbol{D}}_{i,n}^{-1}(\beta_{i,n}\boldsymbol{b}_{i,n} + \lambda_n \widehat{\boldsymbol{w}}_{i,n}), \tag{14}$$

where $\beta_{i,n} = 1/(\boldsymbol{w}_{i,n}^{\mathsf{H}}\boldsymbol{b}_{i,n})$. From the definition of $\beta_{i,n}$, we have

$$\beta_{i,n}\boldsymbol{w}_{i,n}^{\mathsf{H}}\boldsymbol{b}_{i,n} - 1 = 0. \tag{15}$$

Therefore, we obtain the following equation in $\beta_{i,n}$ by substituting (14) into (15):

$$\boldsymbol{b}_{i,n}^{\mathsf{H}}\widehat{\boldsymbol{D}}_{i,n}^{-1}\boldsymbol{b}_{i,n}|\beta_{i,n}|^2 + \lambda_n \widehat{\boldsymbol{w}}_{i,n}^{\mathsf{H}}\widehat{\boldsymbol{D}}_{i,n}^{-1}\boldsymbol{b}_{i,n}\beta_{i,n} - 1 = 0. \tag{16}$$

Because the first and third terms in (16) are real numbers, the second term in (16) must satisfy

$$\mathrm{Im}\left[\lambda_n \widehat{\boldsymbol{w}}_{i,n}^{\mathsf{H}}\widehat{\boldsymbol{D}}_{i,n}^{-1}\boldsymbol{b}_{i,n}\beta_{i,n}\right] = 0. \tag{17}$$

From $\beta_{i,n} \neq 0$ and (17), we have

$$\beta_{i,n} = \gamma_{i,n}\left(\lambda_n \widehat{\boldsymbol{w}}_{i,n}^{\mathsf{H}}\widehat{\boldsymbol{D}}_{i,n}^{-1}\boldsymbol{b}_{i,n}\right)^* = \gamma_{i,n}\lambda_n \boldsymbol{b}_{i,n}^{\mathsf{H}}\widehat{\boldsymbol{D}}_{i,n}^{-1}\widehat{\boldsymbol{w}}_{i,n} \tag{18}$$

or

$$\lambda_n \widehat{\boldsymbol{w}}_{i,n}^{\mathsf{H}}\widehat{\boldsymbol{D}}_{i,n}^{-1}\boldsymbol{b}_{i,n} = 0, \tag{19}$$

where $\gamma_{i,n} \in \mathbb{R}\backslash\{0\}$. When (18) holds, we can derive a quadratic equation in $\gamma_{i,n}$ from (16) as follows:

$$\begin{aligned}\lambda_n^2 \boldsymbol{b}_{i,n}^{\mathsf{H}}\widehat{\boldsymbol{D}}_{i,n}^{-1}\boldsymbol{b}_{i,n}|\boldsymbol{b}_{i,n}^{\mathsf{H}}\widehat{\boldsymbol{D}}_{i,n}^{-1}\widehat{\boldsymbol{w}}_{i,n}|^2\gamma_{i,n}^2 \\ + \lambda_n^2|\boldsymbol{b}_{i,n}^{\mathsf{H}}\widehat{\boldsymbol{D}}_{i,n}^{-1}\widehat{\boldsymbol{w}}_{i,n}|^2\gamma_{i,n} - 1 = 0.\end{aligned} \tag{20}$$

By substituting the solution $\gamma_{i,n}$ of (20) into (18), we have

$$\beta_{i,n} = \frac{\lambda_n \boldsymbol{b}_{i,n}^{\mathsf{H}}\widehat{\boldsymbol{D}}_{i,n}^{-1}\widehat{\boldsymbol{w}}_{i,n}}{2\boldsymbol{b}_{i,n}^{\mathsf{H}}\widehat{\boldsymbol{D}}_{i,n}^{-1}\boldsymbol{b}_{i,n}}\left(-1 \pm \sqrt{1 + \frac{4\boldsymbol{b}_{i,n}^{\mathsf{H}}\widehat{\boldsymbol{D}}_{i,n}^{-1}\boldsymbol{b}_{i,n}}{\lambda_n^2|\boldsymbol{b}_{i,n}^{\mathsf{H}}\widehat{\boldsymbol{D}}_{i,n}^{-1}\widehat{\boldsymbol{w}}_{i,n}|^2}}\right), \tag{21}$$

where the $\pm$ sign in (21) should be positive (see Appendix A). On the other hand, when (19) holds, the solution of (16) becomes

$$\beta_{i,n} = \frac{e^{\mathrm{j}\phi_{i,n}}}{\sqrt{\boldsymbol{b}_{i,n}^{\mathsf{H}}\widehat{\boldsymbol{D}}_{i,n}^{-1}\boldsymbol{b}_{i,n}}}, \tag{22}$$

where $\phi_{i,n} \in (-\pi, \pi]$ denotes an arbitrary phase and j is the imaginary unit. Since $\phi_{i,n}$ does not change the value of $\mathcal{J}_{\mathrm{R}}$, we set $\phi_{i,n}$ to satisfy $e^{\mathrm{j}\phi_{i,n}} = (\det \boldsymbol{W}_i)^*/|\det \boldsymbol{W}_i|$. These solutions of $\beta_{i,n}$ give us the minimum of $\mathcal{J}_{\mathrm{R}}$ w.r.t. $\boldsymbol{w}_{i,n}$, which guarantee the monotonic nonincrease of $\mathcal{J}_{\mathrm{R}}$. From (14), (21), (22), and the relation $\boldsymbol{b}_{i,n} = (\det \boldsymbol{W}_i)\boldsymbol{W}_i^{-1}\boldsymbol{e}_n$, the update rules of $\boldsymbol{w}_{i,n}$ are obtained as

$$\boldsymbol{u}_{i,n} = \widehat{\boldsymbol{D}}_{i,n}^{-1}\boldsymbol{W}_i^{-1}\boldsymbol{e}_n, \tag{23}$$

$$\widehat{\boldsymbol{u}}_{i,n} = \lambda_n \widehat{\boldsymbol{D}}_{i,n}^{-1}\widehat{\boldsymbol{w}}_{i,n}, \tag{24}$$

$$r_{i,n} = \boldsymbol{u}_{i,n}^{\mathsf{H}}\widehat{\boldsymbol{D}}_{i,n}\boldsymbol{u}_{i,n}, \tag{25}$$

$$\widehat{r}_{i,n} = \boldsymbol{u}_{i,n}^{\mathsf{H}}\widehat{\boldsymbol{D}}_{i,n}\widehat{\boldsymbol{u}}_{i,n}, \tag{26}$$

$$\boldsymbol{w}_{i,n} \leftarrow \begin{cases} \dfrac{\boldsymbol{u}_{i,n}}{\sqrt{r_{i,n}}} + \widehat{\boldsymbol{u}}_{i,n} & (\text{if } \widehat{r}_{i,n} = 0) \\ \dfrac{\widehat{r}_{i,n}}{2r_{i,n}}\left(-1 + \sqrt{1 + \dfrac{4r_{i,n}}{|\widehat{r}_{i,n}|^2}}\right)\boldsymbol{u}_{i,n} + \widehat{\boldsymbol{u}}_{i,n} & (\text{otherwise}) \end{cases}. \tag{27}$$

For the NMF source model with $\boldsymbol{T}_n$ and $\boldsymbol{V}_n$, the update rules proposed in [7] or their generalized form proposed in [23] can be used.

3.4. Regularization based on null beamforming

As an application of the spatially regularized ILRMA proposed in the previous section, similar to [15], the utilization of null beamforming (NBF) can be considered. The steering vector of a given source direction θ_n can be represented as

$$\boldsymbol{h}_{i,n}(\theta_n) = e^{-\mathrm{j}\frac{(M-1)\psi_{i,n}}{2}}[1, e^{\mathrm{j}\psi_{i,n}}, \ldots, e^{\mathrm{j}(M-1)\psi_{i,n}}]^{\mathsf{T}}, \tag{28}$$

$$\psi_{i,n} = \frac{2\pi(i-1)f_{\mathrm{s}}d}{cN_{\mathrm{F}}}\sin\theta_n, \tag{29}$$

where f_{s} is the sampling frequency, d is the microphone spacing, c is the sound speed, and N_{F} is the length of Fourier transform. The NBF coefficients $\boldsymbol{g}_{i,n}(\boldsymbol{\theta})$ that suppress the sound arriving from direction θ_n are defined as

$$\boldsymbol{g}_{i,n}(\boldsymbol{\theta}) = (\det \boldsymbol{H}_i(\boldsymbol{\theta}))\boldsymbol{H}_i(\boldsymbol{\theta})^{-\mathsf{T}}\boldsymbol{e}_n, \tag{30}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)^{\mathsf{T}}$ and $\boldsymbol{H}_i(\boldsymbol{\theta}) = (\boldsymbol{h}_{i,1}(\theta_1), \ldots, \boldsymbol{h}_{i,N}(\theta_N))$. In this paper, we consider the situation that only the microphone spacing d is known and the true directions of the sources $\boldsymbol{\theta}$ are unknown in advance. To obtain the pre-estimated directions $\bar{\boldsymbol{\theta}} = (\bar{\theta}_1, \ldots, \bar{\theta}_N)^{\mathsf{T}}$, we first apply AuxICA [24] to the observed signal $\boldsymbol{x}_{ij}$ at each frequency, then $\bar{\boldsymbol{\theta}}$ is calculated by applying k-means clustering to the spatial null directions of $\boldsymbol{w}_{i,n}$, which is estimated by AuxICA, for all i and n [4, 25]. The reasons why we use AuxICA here are as follows: (1) the permutation problem does not matter in this case because only the source direction need be estimated, which can be achieved by k-means clustering, and (2) this preprocessing should have a low computational cost without any tuning parameters. The centroid direction of each cluster corresponds to $\bar{\theta}_n$. Therefore, $\widehat{\boldsymbol{w}}_{i,n} = \boldsymbol{g}_{i,n}(\bar{\boldsymbol{\theta}})^*$ can be used as the NBF-based supervisor.

The weight parameter of the regularizer in the kth iteration is set to $\lambda_n(k) = \alpha_n \max[0.5 - k/K, 0]$, where K is the total number of iterations in ILRMA. This annealing approach improves the separation speed and accuracy. Note that $\lambda_n(k)$ becomes zero in the last half of the whole iteration because the fixed NBF coefficients do not provide the best separation result owing to room reverberation.

4. EXPERIMENT

4.1. Experimental conditions

We conducted experiments on speech separation with two microphones and two sources. We compared the following three methods.

- Initialize $\boldsymbol{W}_i$ with identity matrix and perform conventional ILRMA (**Method 1**)
- Obtain $\bar{\boldsymbol{\theta}}$ via AuxICA, initialize $\boldsymbol{W}_i$ with $\boldsymbol{g}_{i,n}(\bar{\boldsymbol{\theta}})$, and perform conventional ILRMA (**Method 2**)
- Obtain $\bar{\boldsymbol{\theta}}$ via AuxICA, initialize $\boldsymbol{W}_i$ with $\boldsymbol{g}_{i,n}(\bar{\boldsymbol{\theta}})$, and perform ILRMA with spatial regularizer (**Method 3**)

Method 2 corresponds to Method 3 when $\alpha_n = 0$. In Method 3, we set $\alpha_n = \{0.1, 0.3, 1, 3, 10\}$. The observed signals were produced by convoluting the speech sources shown in Table 1 obtained from the SiSEC2010 dataset [26] and the E2A impulse response ($\mathrm{RT}_{60} = 300$ ms) obtained from the RWCP database [27]. The locations of the two tested sources were $(-40^\circ, +40^\circ)$ and $(-40^\circ, +20^\circ)$, where 0° corresponds to the normal direction to the microphone array. The sampling frequency was 16 kHz and the STFT was performed with a 256-ms-long window and 128 ms shift. In Method 1, the total number of iterations in ILRMA, K, was set to

Table 1. Speech sources obtained from SiSEC database

ID	Speech name	Track name
1	dev1_female4	src_1/src_2
2	dev1_female4	src_3/src_4
3	dev1_male4	src_1/src_2
4	dev1_male4	src_3/src_4

Table 2. Average SDRi [dB], where $(\theta_1, \theta_2) = (-40^\circ, +40^\circ)$

Method	α_n	Number of bases L								
		1	2	3	5	10	15	20	25	30
1	—	6.07	7.35	6.23	6.26	5.60	5.45	5.14	4.97	5.06
2	0.0	8.04	9.38	10.28	10.95	11.34	11.60	11.67	11.66	11.51
3	0.1	8.06	9.49	10.62	11.48	11.95	12.34	12.27	12.18	12.18
	0.3	8.06	9.76	10.95	12.04	12.36	12.55	**12.40**	12.34	12.25
	1.0	8.06	10.13	11.23	**12.16**	**12.40**	**12.62**	12.35	**12.43**	**12.27**
	3.0	8.06	10.43	**11.40**	12.07	12.22	12.40	12.23	12.20	12.04
	10.0	**8.06**	**10.43**	11.26	11.86	11.86	12.02	11.71	11.79	11.59

100. In Methods 2 and 3, AuxICA with 20 iterations was performed to obtain $\bar{\theta}_n$, then we performed regularized ILRMA with $K = 80$. The number of bases, L, was set to $\{1, 2, 3, 5, 10, 15, 20, 25, 30\}$. As the evaluation score, we used the *improvement of the signal-to-distortion ratio* (SDRi) [28]. The initial values of $\boldsymbol{T}_n$ and $\boldsymbol{V}_n$ were generated from uniform random values. In addition, we conducted a comparison with the state-of-the-art BSS methods **AuxIVA** [11], **MNMF** [21], and ***t*-MNMF** [29]. AuxIVA was performed with 100 iterations, whereas MNMF and t-MNMF were performed with 200 iterations. The total number of bases in MNMF and t-MNMF was set to 40. The parameter ν in t-MNMF was set to one (i.e., Cauchy-distribution-based MNMF).

4.2. Results

Tables 2 and 3 show the average results when the source locations are $(-40^\circ, +40^\circ)$ and $(-40^\circ, +20^\circ)$, respectively. In these tables, we show the average results for all the sources. In Method 1, we cannot obtain satisfactory performance when the number of bases, L, increases. This is due to the difficulty of speech spectrogram modeling using the NMF model, as mentioned in [7]. The utilization of NBF (Methods 2 and 3) can markedly improve the separation accuracy, and the NMF source model is correctly estimated even if we increase L. In addition, the spatial regularizer (Method 3) provides further improvement of more than 1 dB compared with Method 2. Fig. 2 shows the typical separation results in terms of the average SDRi and their standard deviations for 20 trials with various random values for $\boldsymbol{T}_n$ and $\boldsymbol{V}_n$. We can confirm that the separation becomes stable owing to the proposed spatial regularizer. Table 4 shows a comparison of the SDRi between Methods 1 and 3 and the state-of-the-art methods, where L was set to two in Method 1 and 15 in Method 3. Also, α_n was 1.0 in Method 3. Method 3 achieves the best SDRi performance for both source locations. The relative computational times per iteration normalized by Method 1 (IP) were 1.61 for Method 3 (VCD), 61.99 for MNMF, and 71.95 for t-MNMF, where the calculations were performed with Intel Core i9-7900X processor and MATLAB 9.1. This indicates that VCD does not require much additional computations compared with IP and sufficiently faster than MNMFs. In summary, all the results show the efficacy of the proposed approach.

5. CONCLUSION

In this paper, we developed a new framework for ILRMA that enables us to introduce a spatial regularizer for the demixing matrix.

Table 3. Average SDRi [dB], where $(\theta_1, \theta_2) = (-40^\circ, +20^\circ)$

Method	α_n	Number of bases L								
		1	2	3	5	10	15	20	25	30
1	—	3.14	3.51	3.13	3.24	2.86	2.57	2.47	2.15	2.19
2	0.0	**7.37**	8.41	8.80	9.41	9.63	9.77	9.73	9.77	9.82
3	0.1	7.32	8.67	9.22	9.89	10.41	10.63	10.68	10.67	10.56
	0.3	7.32	8.79	9.57	10.27	10.82	11.04	10.95	11.02	10.91
	1.0	7.32	9.13	9.89	**10.62**	**10.96**	**11.10**	**11.01**	**11.06**	**10.98**
	3.0	7.33	9.24	10.05	10.60	10.86	10.98	10.89	10.93	10.79
	10.0	7.33	**9.29**	**10.05**	10.57	10.74	10.82	10.76	10.72	10.67

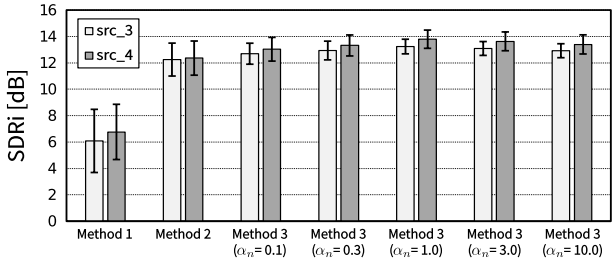

Fig. 2. Average SDRi of src_3 and src_4 in ID2, where $(\theta_1, \theta_2) = (-40^\circ, +40^\circ)$ and $L = 15$.

Table 4. Average SDRi of various BSS methods [dB]

(θ_1, θ_2)	AuxIVA	MNMF	t-MNMF	ILRMA (Method 1)	Regularized ILRMA (Method 3)
$(-40^\circ, +40^\circ)$	3.97	3.84	4.80	7.35	**12.62**
$(-40^\circ, +20^\circ)$	4.15	3.80	4.46	3.51	**11.10**

An efficient optimization algorithm, VCD, was derived, which does not require a step-size parameter and ensures theoretical convergence. NBF-based regularization was newly employed in ILRMA, and we showed that the proposed approach can improve the separation accuracy and stability in speech source separation.

Appendix A. SOLUTION OF SIGN AMBIGUITY IN (21)

The terms containing $\boldsymbol{w}_{i,n}$ in $\mathcal{J}_{\mathrm{R}}$ are

$$\boldsymbol{w}_{i,n}^{\mathsf{H}} \widehat{\boldsymbol{D}}_{i,n} \boldsymbol{w}_{i,n} - \lambda_n \widehat{\boldsymbol{w}}_{i,n}^{\mathsf{H}} \boldsymbol{w}_{i,n} - \lambda_n \boldsymbol{w}_{i,n}^{\mathsf{H}} \widehat{\boldsymbol{w}}_{i,n} - \log |\boldsymbol{w}_{i,n}^{\mathsf{H}} \boldsymbol{b}_{i,n}|^2.$$

After the update of $\boldsymbol{w}_{i,n}$, (13) must be zero. Then we can reformulate the above terms as

$$\begin{aligned} &\boldsymbol{w}_{i,n}^{\mathsf{H}} (\widehat{\boldsymbol{D}}_{i,n} \boldsymbol{w}_{i,n} - \lambda_n \widehat{\boldsymbol{w}}_{i,n}) - \lambda_n \widehat{\boldsymbol{w}}_{i,n}^{\mathsf{H}} \boldsymbol{w}_{i,n} - \log |\boldsymbol{w}_{i,n}^{\mathsf{H}} \boldsymbol{b}_{i,n}|^2 \\ &= \frac{\boldsymbol{w}_{i,n}^{\mathsf{H}} \boldsymbol{b}_{i,n}}{\boldsymbol{w}_{i,n}^{\mathsf{H}} \boldsymbol{b}_{i,n}} - \lambda_n \widehat{\boldsymbol{w}}_{i,n}^{\mathsf{H}} \boldsymbol{w}_{i,n} - \log \left| \frac{1}{\beta_{i,n}} \right|^2 \\ &= 2 \log |\beta_{i,n}| - \lambda_n \widehat{\boldsymbol{w}}_{i,n}^{\mathsf{H}} \boldsymbol{w}_{i,n} + 1. \end{aligned} \tag{31}$$

By focusing on the first term in (31), we have

$$\left| \beta_{i,n}^{(-)} \right| > \left| \beta_{i,n}^{(+)} \right| \Longleftrightarrow \log \left| \beta_{i,n}^{(-)} \right| > \log \left| \beta_{i,n}^{(+)} \right|, \tag{32}$$

where $\beta_{i,n}^{(+)}$ or $\beta_{i,n}^{(-)}$ denotes the solution of (21) with the sign of $+$ or $-$, respectively. Also for the second term in (31), we have

$$-\lambda_n \widehat{\boldsymbol{w}}_{i,n}^{\mathsf{H}} \boldsymbol{w}_{i,n} = -\frac{|\widehat{r}_{i,n}|^2}{2 r_{i,n}} \left(-1 \pm \sqrt{1 + \frac{4 r_{i,n}}{|\widehat{r}_{i,n}|^2}} \right). \tag{33}$$

The right-hand term of (33) becomes smaller when we take $+$ for the $\pm$ ambiguity. From (31)–(33), the $+$ sign must be taken to decrease $\mathcal{J}_{\mathrm{R}}$, and the solution $\beta_{i,n}^{(+)}$ corresponds to the global minimum of (31), which guarantees the monotonic decrease in the VCD.

6. REFERENCES

[1] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

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

[3] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. SAP*, vol. 12, no. 5, pp. 530–538, 2004.

[4] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. ASLP*, vol. 14, no. 2, pp. 666–678, 2006.

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

[6] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2007.

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

[8] Y. Mitsui, D. Kitamura, S. Takamichi, N. Ono, and H. Saruwatari, "Blind source separation based on independent low-rank matrix analysis with sparse regularization for time-series activity," in *Proc. ICASSP*, 2017, pp. 21–25.

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[10] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, 2001, pp. 556–562.

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

[12] Y. Liang, S. Naqvi, and J. Chambers, "Overcoming block permutation problem in frequency domain blind source separation when using AuxIVA algorithm," *Electron. Lett.*, vol. 48, no. 8, pp. 460–462, 2012.

[13] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. ASLP*, vol. 10, no. 6, pp. 352–362, 2002.

[14] Y. Zheng, K. Reindl, and W. Kellermann, "BSS for improved interference estimation for blind speech signal extraction with two microphones," in *Proc. CAMSAP*, 2009, pp. 253–256.

[15] K. Osako, Y. Mori, Y. Takahashi, H. Saruwatari, and K. Shikano, "Fast convergence blind source separation using frequency subband interpolation by null beamforming," *IEICE Trans. Fundam.*, vol. E91-A, no. 6, pp. 1357–1361, 2008.

[16] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer-Verlag Berlin Heidelberg New York, 2001.

[17] T. Nishikawa, H. Abe, H. Saruwatari, and K. Shikano, "Overdetermined blind separation of acoustic signals based on MISO-constrained frequency-domain ICA," in *Proc. Int. Congress Acoust.*, 2004, pp. 3143–3146.

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

[19] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[20] H. Kameoka, T. Yoshioka, M. Hamamura, J. Le Roux, and K. Kashino, "Statistical model of speech signals based on composite autoregressive system with application to blind source separation," in *Proc. LVA/ICA*, 2010, pp. 245–253.

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

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

[23] Y. Mitsui, D. Kitamura, N. Takamune, H. Saruwatari, Y. Takahashi, and K. Kondo, "Independent low-rank matrix analysis based on parametric majorization-equalization algorithm," in *Proc. CAMSAP*, 2017, pp. 98–102.

[24] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," in *Proc. LVA/ICA*, 2010, pp. 165–172.

[25] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures," *Eurasip JASP*, vol. 2003, no. 11, pp. 1157–1166, 2003.

[26] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, and N. Q. K. Duong, "The 2010 signal separation evaluation campaign (SiSEC2010): Audio source separation," in *Proc. LVA/ICA*, 2010, pp. 114–122.

[27] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. LREC*, 2000, pp. 965–968.

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

[29] K. Kitamura, Y. Bando, K. Itoyama, and K. Yoshii, "Student's t multichannel nonnegative matrix factorization for blind source separation," in *Proc. IWAENC*, 2016.