BLIND SOURCE SEPARATION BASED ON INDEPENDENT LOW-RANK MATRIX ANALYSIS WITH SPARSE REGULARIZATION FOR TIME-SERIES ACTIVITY

Yoshiki Mitsui[1] *Daichi Kitamura*[2] *Shinnosuke Takamichi*[1] *Nobutaka Ono*[3,2] *Hiroshi Saruwatari*[1]

[1] The University of Tokyo, Tokyo, Japan
[2] SOKENDAI (The Graduate University for Advanced Studies), Kanagawa, Japan
[3] National Institute of Information, Tokyo, Japan

ABSTRACT

In this paper, we propose a new blind source separation (BSS) method based on independent low-rank matrix analysis (ILRMA) with novel sparse regularization. ILRMA is a recently proposed BSS algorithm that simultaneously estimates a demixing matrix and source spectrogram models based on nonnegative matrix factorization (NMF). To improve the separation accuracy and stability, an additional constraint such as sparseness is needed but there have been no studies on this so far. In this study, we introduce an a priori statistical model for time-series amplitudes of source spectrograms, employing a new frequency-wise sparse regularization using estimates from the Bayesian postfilter to enhance the modeling accuracy. This regularization results in a bilevel optimization problem that consists of the estimation of a sparsity-emphasized source model using NMF and the separation of sources by ILRMA. In this paper, we present two approximated optimization schemes and their combination for performing regularized ILRMA. The efficacy of the proposed method is confirmed in a BSS experiment.

Index Terms— Blind source separation, nonnegative matrix factorization, independent low-rank matrix analysis, reguralization

1. INTRODUCTION

Blind source separation (BSS) is a technique for estimating hidden sources from a mixture without knowing any information about the mixing conditions and the characteristics of the sources. For an overdetermined or determined situation, frequency-domain independent component analysis (FDICA) [1, 2, 3, 4] and independent vector analysis (IVA) [5, 6, 7] are widely used. Nonnegative matrix factorization (NMF) [8] is also a popular method for BSS, and multichannel NMF (MNMF) [9, 10, 11, 12, 13] has been proposed as a more flexible BSS technique. Some of the authors have also proposed a unified algorithm involving IVA and NMF, referred to as *independent low-rank matrix analysis* (ILRMA) [14]. ILRMA achieves more accurate and stable separation than FDICA, IVA, and MNMF. However, the performance of ILRMA is still not sufficient for a mixture including speech signals. This is due to the difficulty of capturing complicated spectrograms using NMF.

To cope with this problem, in this study we introduce an a priori statistical model for time-series amplitudes of source spectrograms. Since many acoustic sources have sparse time-series activity, we propose a new frequency-wise sparse regularization using estimates from the Bayesian postfilter MOSIE (Minimum mean-square error estimation with Optimizable Speech model and Inhomogeneous Error criterion) [15] to enhance the modeling accuracy of the source spectrograms. This regularization results in a *bilevel optimization* [16, 17, 18] problem that consists of the estimation of a sparsity-emphasized source model using NMF and the separation of sources by ILRMA. In this paper, we present two approximated optimization schemes and their combination for performing regularized ILRMA.

In this study, the main contributions to the prior works are as follows. First, ILRMA is a newly proposed BSS method and an additional constraint such as sparseness is needed but there have been no studies on this so far. The method proposed in this paper is the first attempt to introduce sparsity-emphasized regularization in ILRMA. Next, in contrast to naive regularization such as L1-norm minimization [19, 20], the proposed method directly employs a sparse statistical model in each frequency subband and the improved performance is revealed via experimental evaluation. Thirdly, we formulate the regularization as a bilevel optimization, which is generally difficult to solve; however, we give effective approximated solutions.

2. CONVENTIONAL BSS ALGORITHM

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 signals, observed signals, and separated signals are defined as

$$\boldsymbol{s}_{ij} = (s_{ij,1}, s_{ij,2}, \ldots, s_{ij,N})^{\mathsf{T}}, \tag{1}$$

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

$$\boldsymbol{y}_{ij} = (y_{ij,1}, y_{ij,2}, \ldots, y_{ij,N})^{\mathsf{T}}, \tag{3}$$

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 window length in the STFT is sufficiently longer than the impulse responses between sources and microphones, the following instantaneous mixture model in a frequency domain holds:

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

where $\boldsymbol{A}_i = (\boldsymbol{a}_{i,1}, \ldots, \boldsymbol{a}_{i,N})$ is the mixing matrix and $\boldsymbol{a}_{i,n}$ is the steering vector for the nth source. If the number of sources equals the number of channels ($M = N$), the demixing matrix $\boldsymbol{W}_i$ ($= \boldsymbol{A}_i^{-1}$) can be defined, and the separated signals are represented as

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

The BSS problem is to estimate $\boldsymbol{W}_i$ using only $\boldsymbol{x}_{ij}$.

2.2. ILRMA

In this section, we give a brief introduction to ILRMA, which was recently proposed as a method for efficiently solving the BSS problem, and clarify the relationship among FDICA, IVA, MNMF, and

ILRMA. In IVA, the independent scalars in FDICA are extended to independent frequency vectors, where higher-order correlations between the frequency components in each vector are assumed using a spherical multivariate prior [6]. Similarly, in ILRMA, the independent frequency vectors in IVA are extended to low-rank matrices, which correspond to the power spectrograms of estimated sources, using NMF decomposition [14]. Therefore, ILRMA can simultaneously estimate both the demixing matrix and the power spectrogram of each separated source in a fully blind fashion. This signal model (independence between sources and low-rank decomposition of source spectrograms) is theoretically equivalent to conventional MNMF only when the spatial covariance matrix of each source in MNMF is constrained to a rank-1 matrix, which yields a computationally efficient algorithm for ILRMA.

The cost function in ILRMA is defined as follows:

$$Q_{\mathrm{ILRMA}} = \sum_{i,j}\left[\sum_{n}\frac{|y_{ij,n}|^2}{\sum_l t_{il,n}v_{lj,n}} - 2\log|\det \boldsymbol{W}_i| + \sum_{n}\log\sum_{l} t_{il,n}v_{lj,n}\right], \tag{6}$$

where $t_{il,n}$ and $v_{lj,n}$ are nonnegative elements of basis matrix $\boldsymbol{T}_n \in \mathbb{R}_{\geq 0}^{I\times L}$ and activation matrix $\boldsymbol{V}_n \in \mathbb{R}_{\geq 0}^{L\times J}$, respectively, and $l = 1,\ldots,L$ is an integral index for the NMF bases. Therefore, $\boldsymbol{T}_n\boldsymbol{V}_n$ corresponds to the NMF decomposition and represents an estimated power spectrogram of the nth source. The first and second terms in (6) are equivalent to the cost function in IVA, which evaluates the independence of the sources, and the first and third terms in (6) are equivalent to the cost function in simple NMF based on Itakura–Saito (IS) divergence. Efficient update rules for optimizing (6) were proposed in [14].

3. PROPOSED METHOD

3.1. Motivation and strategy

ILRMA often fails to separate sources that have complicated time-frequency structures. Indeed, it has been reported that the separation performance for mixtures including speech becomes unstable, and the improvement of the performance from that in the IVA or MNMF is less than outstanding [14]. This is because NMF decomposition is not suitable for capturing complicated time-frequency structures. When we set the number of bases, L, to a large value, the source model $\boldsymbol{T}_n\boldsymbol{V}_n$ has the potential to represent complicated structures; however, it also makes the separation unstable because the number of variables markedly increases.

To cope with this problem, in this study we introduce an a priori statistical model for time-series amplitudes of source spectrograms, in which the source model $\boldsymbol{T}_n\boldsymbol{V}_n$ in ILRMA is retrained to enhance the modeling accuracy. Fig. 1 shows the concept of the proposed method. Since many acoustic sources have sparse time-series activity in each frequency component, we propose to apply a sparse regularization to the spectral amplitudes in a frequency-wise manner. We here assume a chi distribution with various shape parameters $\rho_{i,n}$ as a prior. The shape parameters $\rho_{i,n}$ correspond to the sparsity of the prior distribution at the ith frequency of the nth source, and we can calculate them using the estimates from ILRMA and the moment-matching method [21]. On the basis of the chi distribution prior, the Bayesian postfilter MOSIE outputs the sparsity-emphasized power spectrogram of the nth source, $\boldsymbol{D}_n$, and such outputs should be more accurate estimates than the outputs of simple ILRMA, $y_{ij,n}$, if the prior is appropriate. Finally, we retrain the activation matrix $\boldsymbol{V}_n$ along with $\boldsymbol{D}_n$ and update ILRMA with the new regularized

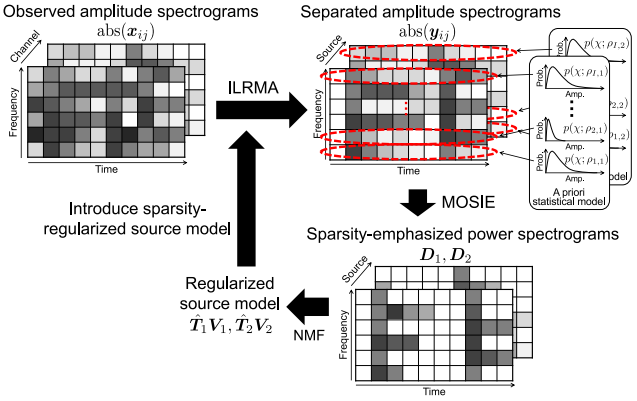

Fig. 1. Concept of proposed method, where $M = N = 2$.

source model. Note that we do not utilize a naive sparse regularization for the activation matrix $\boldsymbol{V}_n$ (e.g., L1-norm minimization of $\boldsymbol{V}_n$) because there should be a large variation of the sparsity depending on the frequency, and the simple sparse regularization for $\boldsymbol{V}_n$ cannot control such differences in sparsity among different frequency subbands.

3.2. Cost function

Since the proposed regularization includes another NMF outside ILRMA as shown in Fig. 1, the cost function is formulated as follows:

$$\min_{\boldsymbol{W}_i,\boldsymbol{T}_n,\boldsymbol{V}_n} Q_{\mathrm{ILRMA}} \tag{7}$$

$$\text{s.t.}\quad \boldsymbol{V}_n = \arg\min_{\boldsymbol{V}_n'} \sum_n \mathcal{D}_{\mathrm{IS}}\left(\boldsymbol{D}_n \mid \hat{\boldsymbol{T}}_n\boldsymbol{V}_n'\right) \quad \text{for given } \hat{\boldsymbol{T}}_n, \tag{8}$$

where the $\mathcal{D}_{\mathrm{IS}}(\cdot|\cdot)$ denotes IS divergence, $\boldsymbol{D}_n \in \mathbb{R}_{\geq 0}^{I\times J}$ is the sparsity-emphasized power spectrogram of the nth source, mentioned later in Sect. 3.3, and $[\hat{\boldsymbol{T}}_n]_{il} = |a_{i,1n}|^2 t_{il,n}$ is the scale-fitted basis matrix of the nth source (hereafter $[\cdot]_{ij}$ denotes the ijth element of matrix $\cdot$), where $a_{i,1n}$ is given by the inverse of the estimated $\boldsymbol{W}_i$. This type of cost function is referred to as bilevel optimization [16, 17, 18] because the constraint also includes another *parameter-nested* optimization. It is difficult to directly solve (7) and (8); instead, we propose two approximated optimization schemes in the following section.

3.3. Sparse regularization by MOSIE

MOSIE is a Bayesian postfilter based on an a priori statistical model [15]. It assumes a chi distribution as the a priori statistical model of the target amplitude spectra as follows:

$$p(\chi) = \frac{2}{\Gamma(\rho_{i,n})}\left(\frac{\rho_{i,n}}{\mathrm{E}[\chi^2]}\right)^{\rho_{i,n}} \chi^{2\rho_{i,n}-1}\exp\left(-\frac{\rho_{i,n}}{\mathrm{E}[\chi^2]}\chi^2\right), \tag{9}$$

where $\chi \in \mathbb{R}_{\geq 0}$ is a random variable that corresponds to the amplitude spectrum of the target source, and $\mathrm{E}[\cdot]$ denotes an expectation operator. The shape parameter $\rho_{i,n}$ can be calculated from the observed and estimated signals, $\boldsymbol{x}_{ij}$ and $\boldsymbol{y}_{ij}$, respectively, by the moment-matching method [21] as

$$\rho_{i,n} = \eta\cdot\bar{\rho}_{i,n} \tag{10}$$

$$= \eta\cdot\{\mu_2^2(|x_{ij,1}|) + \mu_2^2(|q_{ij,n}|) - 2\mu_2(|x_{ij,1}|)\mu_2(|q_{ij,n}|)\} \cdot\{\mu_4(|x_{ij,1}|) - \mu_4(|q_{ij,n}|) - \mu_2^2(|x_{ij,1}|) - 2\mu_2(|x_{ij,1}|)\mu_2(|q_{ij,n}|) + 3\mu_2^2(|q_{ij,n}|)\}^{-1}, \tag{11}$$

where $\mu_k(\cdot)$ denotes the kth-order moment w.r.t. time frame j, namely, $\mu_k(\alpha_{ij}) = (\sum_j \alpha_{ij}^k)/J$, $q_{ij,n} = \sum_{n' \neq n} \hat{y}_{ij,n'}$ is the sum of the interfering sources, and $\hat{y}_{ij,n}$ is a scale-fitted estimated signal using a projection-back technique [22]. Also, $0 < \eta \leq 1$ is a hyperparameter for discounting the empirical shape parameter $\bar{\rho}_{i,n}$ given by the moment-matching method. Since the separated signal $\boldsymbol{y}_{ij}$ includes residual interference at the beginning of optimization, the empirical shape parameter $\bar{\rho}_{i,n}$ is always overestimated owing to the central limit theorem; this is equivalent to an underestimation of sparsity. The discounting η will enhance the sparse regularization and separation. Note that η should not be determined by a complicated function but a simple monotonically increasing progression converging to 1 because we expect that the separation performance will be gradually improved with increasing iterations. In this study, we set η to c/C, where c and C are the current and total iteration numbers, respectively.

The sparsity-emphasized target power spectrogram $d_{ij,n} = [\boldsymbol{D}_n]_{ij}$ that minimizes the mean-squared error criterion can be calculated as [15]

$$d_{ij,n} = \left[\frac{\Gamma(\frac{\beta}{2} + \rho_{i,n}) \Phi\left(1 - \frac{\beta}{2} - \rho_{i,n}, 1; -\frac{|x_{ij,1}|^2}{\mu_2(|q_{ij,n}|)} \frac{\xi_{ij}}{\rho_{i,n} + \xi_{ij}}\right)}{\Gamma(\rho_{i,n}) \Phi\left(1 - \rho_{i,n}, 1; -\frac{|x_{ij,1}|^2}{\mu_2(|q_{ij,n}|)} \frac{\xi_{ij}}{\rho_{i,n} + \xi_{ij}}\right)} \right]^{\frac{2}{\beta}} \cdot \frac{\xi_{ij}}{\rho_{i,n} + \xi_{ij}} \cdot \mu_2(|q_{ij,n}|), \tag{12}$$

where β is a domain parameter, $\Gamma(\cdot)$ is the complete gamma function, $\Phi(e, f; g) = {}_1F_1(e; f; g)$ is the confluent hypergeometric function of the first kind, and ξ_{ij} is the a priori SNR given by

$$\xi_{ij} = \kappa \frac{d_{i(j-1),n}}{\mu_2(|q_{ij,n}|)} + (1 - \kappa) \max\left[\frac{|x_{ij,1}|^2}{\mu_2(|q_{ij,n}|)} - 1, 0 \right], \tag{13}$$

where κ is the forgetting factor.

3.4. Approximated optimizations in proposed method

3.4.1. Alternating optimization

As the simplest way to optimize (7) and (8), we can employ an alternating minimization. In this method, the update rules in ILRMA (for $\boldsymbol{T}_n$, $\boldsymbol{V}_n$, and $\boldsymbol{W}_i$) and IS-divergence NMF (ISNMF) [23] (for $\boldsymbol{V}_n$) can be used alternately for the optimization. First, to solve the ILRMA problem in (7), we update $\boldsymbol{W}_i$, $\boldsymbol{T}_n$, and $\boldsymbol{V}_n$ R times. Since the update rule of $\boldsymbol{W}_i$ is complicated and there is insufficient space to state it here, please refer to [14]. The update rules of $\boldsymbol{T}_n$ and $\boldsymbol{V}_n$ are given as follows:

$$t_{il,n} \leftarrow t_{il,n} \sqrt{\frac{\sum_j |y_{ij,n}|^2 v_{lj,n} \left(\sum_{l'} t_{il',n} v_{l'j,n}\right)^{-2}}{\sum_j v_{lj,n} \left(\sum_{l'} t_{il',n} v_{l'j,n}\right)^{-1}}}, \tag{14}$$

$$v_{lj,n} \leftarrow v_{lj,n} \sqrt{\frac{\sum_i |y_{ij,n}|^2 t_{il,n} \left(\sum_{l'} t_{il',n} v_{l'j,n}\right)^{-2}}{\sum_i t_{il,n} \left(\sum_{l'} t_{il',n} v_{l'j,n}\right)^{-1}}}. \tag{15}$$

Second, to solve the ISNMF problem in (8), we update $\boldsymbol{V}_n$ 100 times as follows:

$$v_{lj,n} \leftarrow v_{lj,n} \sqrt{\frac{\sum_i \hat{t}_{il,n} d_{ij,n} \left(\sum_{l'} \hat{t}_{il',n} v_{l'j,n}\right)^{-2}}{\sum_i \hat{t}_{il,n} \left(\sum_{l'} \hat{t}_{il',n} v_{l'j,n}\right)^{-1}}}. \tag{16}$$

The retrained $\boldsymbol{V}_n$ along with $\boldsymbol{D}_n$ is applied to the source model in the next ILRMA update. In this paper, MOSIE and the retraining of $\boldsymbol{V}_n$ are performed at intervals of R iterative updates in ILRMA. Note that this alternating scheme does not always guarantee the convergence, having the risk of oscillation.

3.4.2. Simultaneous optimization

Another optimization scheme is a concatenation of (7) and (8) as follows:

$$\min_{\boldsymbol{W}_i, \boldsymbol{T}_n, \boldsymbol{V}_n} \left[Q_{\mathrm{ILRMA}} + \lambda \sum_n \mathcal{D}_{\mathrm{IS}}\left(\boldsymbol{D}_n \mid \hat{\boldsymbol{T}}_n \boldsymbol{V}_n\right) \right] \text{ for given } \hat{\boldsymbol{T}}_n, \tag{17}$$

where λ is the weight parameter for the regularization term. We can consider that this minimization is more accurate than that in Sect. 3.4.1 because the gradient of $\boldsymbol{V}_n$ depends on both Q_{ILRMA} and the regularization term. Strictly speaking, the estimate of MOSIE, $\boldsymbol{D}_n$, includes $\boldsymbol{W}_i$ and can be used as a variable for the optimization. However, in this method, we assume that $\boldsymbol{D}_n$ is constant and defined by $\boldsymbol{W}_i$ in the previous iteration for easy derivation (this approximately holds when the change of $\boldsymbol{D}_n$ is small, e.g., in the point close to convergence). Also, we consider that $\hat{\boldsymbol{T}}_n$ is constant, whereas it includes $\boldsymbol{T}_n$. Here we derive the update rules for $\boldsymbol{T}_n$ and $\boldsymbol{V}_n$ based on the *auxiliary function method* [24], which guarantees monotonic decrease of the cost in (17). Similarly to in Sect. 3.4.1, the update rule for $\boldsymbol{W}_i$ is the same as that in [14]. First, we define the cost function in (17) as Q_{sim}. This cost function can be rewritten as

$$Q_{\mathrm{sim}} = \sum_{i,j,n} \left[\frac{|y_{ij,n}|^2}{\sum_l t_{il,n} v_{lj,n}} + \log \sum_l t_{il,n} v_{lj,n} + \frac{\lambda d_{ij,n}}{\sum_l \hat{t}_{il,n} v_{lj,n}} + \lambda \log \sum_l \hat{t}_{il,n} v_{lj,n} \right] - 2 \sum_{i,j} \log |\det \boldsymbol{W}_i|. \tag{18}$$

The first and third terms in (18) are convex functions and the second and fourth terms are concave functions. Applying Jensen's inequality to the first and third terms with auxiliary variables $\alpha_{ijl,n}, \gamma_{ijl,n} \geq 0$ that satisfy $\sum_l \alpha_{ijl,n} = \sum_l \gamma_{ijl,n} = 1$, we have

$$\frac{1}{\sum_l t_{il,n} v_{lj,n}} \leq \sum_l \frac{\alpha_{ijl,n}^2}{t_{il,n} v_{lj,n}}, \tag{19}$$

$$\frac{1}{\sum_l \hat{t}_{il,n} v_{lj,n}} \leq \sum_l \frac{\gamma_{ijl,n}^2}{\hat{t}_{il,n} v_{lj,n}}. \tag{20}$$

Also, applying the tangent-line inequality to the second and fourth terms with auxiliary variables $\beta_{ij,n}, \delta_{ij,n} \geq 0$, we have

$$\log \sum_l t_{il,n} v_{lj,n} \leq \frac{\sum_l t_{il,n} v_{lj,n} - \beta_{ij,n}}{\beta_{ij,n}} + \log \beta_{ij,n}, \tag{21}$$

$$\log \sum_l \hat{t}_{il,n} v_{lj,n} \leq \frac{\sum_l \hat{t}_{il,n} v_{lj,n} - \delta_{ij,n}}{\delta_{ij,n}} + \log \delta_{ij,n}. \tag{22}$$

The inequalities of (19)–(22) hold if and only if

$$\alpha_{ijl,n} = \frac{t_{il,n} v_{lj,n}}{\sum_{l'} t_{il',n} v_{l'j,n}}, \tag{23}$$

$$\beta_{ij,n} = \sum_l t_{il,n} v_{lj,n}, \tag{24}$$

$$\gamma_{ijl,n} = \frac{\hat{t}_{il,n} v_{lj,n}}{\sum_{l'} \hat{t}_{il',n} v_{l'j,n}}, \tag{25}$$

$$\delta_{ij,n} = \sum_l \hat{t}_{il,n} v_{lj,n}. \tag{26}$$

Therefore, we can define an upper-bound function Q_{sim}^+ for Q_{sim} as

$$Q_{\mathrm{sim}} \leq Q_{\mathrm{sim}}^+ = \sum_{i,j,n,l} \left[\frac{\alpha_{ijl,n}^2 |y_{ij,n}|^2}{t_{il,n} v_{lj,n}} + \frac{t_{il,n} v_{lj,n}}{\beta_{ij,n}} \right.$$

Table 1. Experimental conditions

Sampling frequency	16000 Hz
FFT length	256 ms (4096 points)
Window shift	64 ms (1024 points)
Number of bases L	10
Total number of ILRMA iterations C	200
Regularization intervals R	20
Number of ISNMF iterations	100
Domain parameter β	1.0
Forgetting factor κ	0.98

$$+\lambda\frac{\gamma_{ijl,n}^2 d_{ij,n}}{\hat{t}_{il,n}v_{lj,n}}+\lambda\frac{\hat{t}_{il,n}v_{lj,n}}{\delta_{ij,n}}\Bigg]+\mathcal{C}, \tag{27}$$

where $\mathcal{C}$ denotes terms independent of $t_{il,n}$ and $v_{lj,n}$. The update rule for Q_{sim}^{+} is determined by setting the gradients to zero. From $\partial Q_{\mathrm{sim}}^{+}/\partial t_{il,n}=0$ and $\partial Q_{\mathrm{sim}}^{+}/\partial v_{lj,n}=0$, we obtain

$$\sum_j\left[-\frac{\alpha_{ijl,n}^2|y_{ij,n}|^2}{t_{il,n}^2 v_{lj,n}}+\frac{v_{lj,n}}{\beta_{ij,n}}\right]=0, \tag{28}$$

$$\sum_i\left[-\frac{\alpha_{ijl,n}^2|y_{ij,n}|^2}{t_{il,n}v_{lj,n}^2}+\frac{t_{il,n}}{\beta_{ij,n}}-\lambda\frac{\gamma_{ijl,n}^2 d_{ij,n}}{\hat{t}_{il,n}v_{lj,n}^2}+\lambda\frac{\hat{t}_{il,n}}{\delta_{ij,n}}\right]=0. \tag{29}$$

By substituting (23)–(26) into (28) and (29) and simplifying, the multiplicative update rules for $\boldsymbol{T}_n$ and $\boldsymbol{V}_n$ can be obtained as

$$t_{il,n}\leftarrow t_{il,n}\sqrt{\frac{\sum_j|y_{ij,n}|^2 v_{lj,n}\left(\sum_{l'}t_{il',n}v_{l'j,n}\right)^{-2}}{\sum_j v_{lj,n}\left(\sum_{l'}t_{il',n}v_{l'j,n}\right)^{-1}}}, \tag{30}$$

$$v_{lj,n}\leftarrow v_{lj,n}\sqrt{\frac{\sum_i\left[\frac{|y_{ij,n}|^2 t_{il,n}}{\left(\sum_{l'}t_{il',n}v_{l'j,n}\right)^2}+\frac{\lambda d_{ij,n}\hat{t}_{il,n}}{\left(\sum_{l'}\hat{t}_{il',n}v_{l'j,n}\right)^2}\right]}{\sum_i\left[\frac{t_{il,n}}{\sum_{l'}t_{il',n}v_{l'j,n}}+\frac{\lambda\hat{t}_{il,n}}{\sum_{l'}\hat{t}_{il',n}v_{l'j,n}}\right]}}. \tag{31}$$

In this method, the activation matrix $\boldsymbol{V}_n$ is simultaneously optimized using $\boldsymbol{D}_n$ in each iteration.

4. EXPERIMENT

4.1. Experimental conditions

To confirm the efficacy of the proposed method, we conducted BSS experiments with two sources and two microphones. We compared seven methods: *IVA*, *Ozerov's MNMF* [9], *ILRMA*, *ILRMA (L1)*, *proposed method (alt)*, *proposed method (sim)*, and *proposed method (alt+sim)*. ILRMA (L1) is based on a naive sparse regularization, namely, the L1 norm of $\boldsymbol{V}_n$ to solve $\min_{\boldsymbol{W}_i,\boldsymbol{T}_n,\boldsymbol{V}_n}Q_{\mathrm{ILRMA}}+\lambda\sum_n|\boldsymbol{V}_n|_1$. This regularization only controls the sparsity of $\boldsymbol{V}_n$, and cannot treat the difference in sparsity among frequencies. Proposed method (alt) and Proposed method (sim) are the methods described in Sects. 3.4.1 and 3.4.2, respectively. Proposed method (alt+sim) is a hybrid method, where Proposed method (alt) is performed in the former half of the updates, and Proposed method (sim) is performed in the latter half (the motivation for using this concatenation will be described in Sect. 4.2). The observed signals were convoluted with the E2A impulse response [25] and one speech and two music sources obtained from SiSEC2011 [26], where the reverberation time was 300 ms. The spatial positions of the sources were set to the following three pairs: 50° & 70°, 50° & 130°, and 70° & 110°. As the evaluation criterion, we used the signal-to-distortion ratio (SDR) [27]. The weight parameter λ was experimentally set to the optimal value. Initial values of all the estimates were set at random and we performed 10 different BSS trials for each experimental condition. The other conditions used are shown in Table 1.

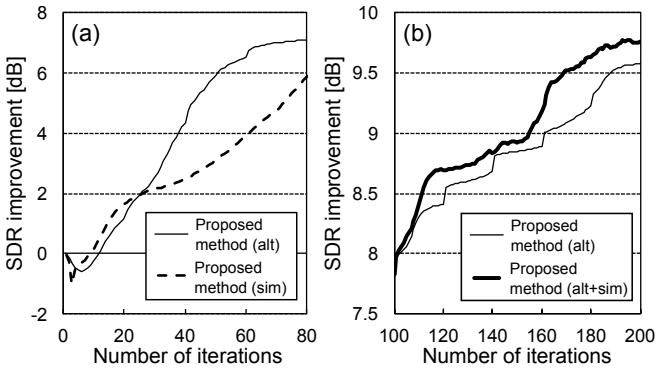

Fig. 2. Example of separation results of speech source with spatial positions 50° & 130° in (a) former and (b) latter half of iterations.

Table 2. Average SDR improvements [dB]

	Speech	Music
IVA	3.45	5.47
Ozerov's MNMF	3.62	4.81
ILRMA	7.06	9.52
ILRMA (L1)	7.12	9.62
Proposed method (alt)	7.59	10.45
Proposed method (sim)	7.66	10.15
Proposed method (alt+sim)	**7.86**	**10.49**

4.2. Results

Fig. 2 shows a typical example of SDR convergences for the speech source with the 50° & 130° setting. In the case that the number of iterations is small (Fig. 2(a)), the SDR convergence of Proposed method (alt) is much faster than that of Proposed method (sim). However, in the latter half of the iterations (Fig. 2(b)), the convergence curve of Proposed method (alt) appears to be unstable and have a sawtooth shape owing to the alternation of optimization, whereas that of Proposed method (alt+sim) continues to increase relatively smoothly, thanks to the suitability of auxiliary-function-based optimization.

Table 2 shows the average SDR improvements for all the spatial conditions in each source. From this result, we can confirm that the proposed regularizations outperform the conventional methods and simple ILRMA. This indicates that the sparse regularization can induce better NMF capture of the source spectrograms. In particular, the hybrid regularization achieves the best separation performance.

5. CONCLUSION

In this paper, to improve the separation performance in ILRMA, we introduced a new frequency-wise sparse regularization based on an a priori statistical model. Here, we proposed new approximated algorithms for the bilevel optimization problem. The efficacy of the proposed methods was confirmed by performing a BSS task with speech and music mixtures.

Acknowledgments This work was supported by ImPACT Program of Council for Science, Technology and Innovation and SECOM Science and Technology Foundation.

6. REFERENCES

[1] S. Araki, R. Mukai, S. Makino, T. Nishikawa and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 2, pp. 109–116, 2003.

[2] H. Sawada, R. Mukai, S. Araki and S. Makino, "Convolutive blind source separation for more than two sources in the frequency domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, vol. 3, pp. 885–888.

[3] H. Buchner, R. Aichner and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. on Speech and Audio Process.*, vol. 13, no. 1, pp. 120–134, 2005.

[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. Audio, Speech, Lang. Process.*, 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. Int. Conf. Independent Compon. Anal. Blind Source Separation*, 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. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70–79, 2007.

[7] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," *in Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2011, pp. 189–192.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inform. Process. Syst.*, 2001, pp. 556–562.

[9] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, 2010.

[10] 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. Int. Conf. Latent Variable Anal. Signal Separation*, 2010, pp. 245–253.

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

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

[13] Y. Mitsufuji, S. Koyama and H. Saruwatari, "Multichannel blind source separation based on non-negative tensor factorization in wavenumber domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 56–60.

[14] 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. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[15] C. Breithaupt and R. Martin, "Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 277–289, 2011.

[16] B. Colson, P. Marcotte and G. Savard, "An overview of bilevel optimization," *Ann. Operat. Res.*, vol. 153, no. 1, pp. 235-256, 2007.

[17] H. Nakajima, D. Kitamura, N. Takamune, S. Koyama, H. Saruwatari, N. Ono, Y. Takahashi and K. Kondo, "Music signal separation using supervised NMF with all-pole-model-based discriminative basis deformation," in *Proc. Eur. Signal Process. Conf.*, 2016, pp. 1143–1147.

[18] D. Kitamura, N. Ono, H. Saruwatari, Y. Takahashi and K. Kondo, "Discriminative and reconstructive basis training for audio source separation with semi-supervised nonnegative matrix factorization," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2016.

[19] W. Liu, N. Zheng and X. Lu, "Non-negative matrix factorization for visual coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, pp. 293–296, 2003.

[20] J. Le Roux, F. Weninger and J. R. Hershey, "Sparse NMF - half-baked or well done?," *MERL Technical Report*, TR 2015-023, 2015.

[21] Y. Murota, D. Kitamura, H. Saruwatari, S. Nakamura, Y. Takahashi and K. Kondo, "Music signal separation based on Bayesian spectral amplitude estimator with automatic target prior adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 7540–7544.

[22] N. Murata, S. Ikeda and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, 2001.

[23] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2010, pp. 283–288.

[24] D. R. Hunter, K. Lange, "A Tutorial on MM Algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[25] 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. Int. Conf. Lang. Resources Evaluation*, 2000, pp. 965–968.

[26] S. Araki, F. Nesta, E. Vincent, Z. Koldovský, G. Nolte, A. Ziehe and A. Benichoux, "The 2011 Signal Separation Evaluation Campaign (SiSEC2011): -Audio Source Separation-," in *Proc. Latent Variable Anal. Signal Separation*, 2012, pp. 414–422.

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