Efficient Full-Rank Spatial Covariance Estimation Using Independent Low-Rank Matrix Analysis for Blind Source Separation

Yuki Kubo†, Norihiro Takamune†, Daichi Kitamura‡, Hiroshi Saruwatari†
†*The University of Tokyo, Graduate School of Information Science and Technology,*
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
‡*National Institute of Technology, Kagawa College,*
355 Chokushi-cho, Takamatsu, Kagawa 761-8058, Japan

***Abstract*—In this paper, we propose a new algorithm that efficiently separates a directional source and diffuse background noise based on independent low-rank matrix analysis (ILRMA). ILRMA is one of the state-of-the-art techniques of blind source separation (BSS) and is based on a rank-1 spatial model. Although such a model does not hold for diffuse noise, ILRMA can accurately estimate the spatial parameters of the directional source. Motivated by this fact, we utilize these estimates to restore the lost spatial basis of diffuse noise, which can be considered as an efficient full-rank spatial covariance estimation. BSS experiments show the efficacy of the proposed method in terms of the computational cost and separation performance.**

***Index Terms*—Blind source separation, independent low-rank matrix analysis, full-rank spatial covariance model, diffuse noise**

I. Introduction

Blind source separation (BSS) is a technique for separating an observed multichannel signal, which is a mixture of multiple sources, into each source without any prior information about the sources or the mixing system. In a determined or overdetermined situation (number of sensors $\geq$ number of sources), frequency-domain independent component analysis (FDICA) [1], [2], independent vector analysis (IVA) [3], [4], and independent low-rank matrix analysis (ILRMA) [5], [6] have been proposed for audio BSS problems. In particular, ILRMA assumes low-rankness for the power spectrogram of each source using nonnegative matrix factorization (NMF) [7], [8] in addition to statistical independence between sources, and achieves efficient and accurate separation [5]. These methods assume a rank-1 spatial model; the frequency-wise acoustic path of each source can be represented by a single time-invariant spatial basis, which is often called a steering vector. Under this assumption, the determined BSS problem reduces to the estimation of a demixing matrix for each frequency. However, the assumption in the rank-1 spatial model becomes invalid in actual situations. For instance, when a target source (directional source) and diffuse noise that arrives from all directions are mixed, FDICA, IVA, and ILRMA cannot extract only the target source in principle [9], and the estimated target source includes residual diffuse noise.

Multichannel NMF (MNMF) [10], [11] is theoretically equivalent to ILRMA except for the mixing model, namely, MNMF employs a full-rank spatial covariance matrix [12]. This model can represent not only the acoustic path but also the spatial spread of each source or diffuse noise, while its optimization has a huge computational cost and lacks robustness against the initialization [5]. To accelerate the parameter estimation, FastMNMF has been proposed [13], [14]. It assumes a jointly diagonalizable spatial covariance matrix to greatly reduce the computational cost of the update algorithm, although its performance still depends on the initial values of parameters. To increase the stability of its performance, ILRMA-based initialization was utilized for MNMF in [15]. However, the improvement is still limited because of the complexity of optimization with a large number of parameters.

In this paper, we treat the BSS problem with one directional target source and diffuse background noise, where more than or equal to two microphones are available. In this case, the target source can be expressed using the rank-1 spatial covariance (one steering vector), but diffuse noise requires the full-rank spatial covariance because of its spatial spread. To achieve robust and computationally efficient BSS in this situation, we propose a new approach based on ILRMA: (a) rank-1 target covariance and rank-$(M-1)$ diffuse noise covariance matrices are simultaneously estimated by ILRMA, where M is the number of microphones, (b) one lost spatial basis for diffuse noise is restored to obtain the rank-M (full-rank) noise covariance via the expectation-maximization (EM) algorithm, and (c) a multichannel Wiener filter is applied to enhance only the target source. The efficacy of the proposed method is confirmed through BSS experiments using a mixture of speech and diffuse noise.

Regarding its relation to prior works, the proposed method is considered as a spatial model extension of FDICA, IVA, and ILRMA, which are the conventional independence-based BSS algorithms utilizing the rank-1 spatial model. Compared with conventional MNMF and FastMNMF based on the full-rank spatial model, the proposed method is regarded as a computationally efficient algorithm with higher separation accuracy.

978-9-0827-9703-9/19/$31.00

II. Independent Low-Rank Matrix Analysis

A. Formulation

Let us denote a multichannel observed signal as $\boldsymbol{x}_{ij} = (x_{ij,1}, \ldots, x_{ij,m}, \ldots, x_{ij,M})^{\mathsf{T}} \in \mathbb{C}^M$ that is obtained via a short-time Fourier transform (STFT), where $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $m = 1, \ldots, M$ are the indices of the frequency bins, time frames, and microphones, respectively, and $^{\mathsf{T}}$ denotes the transpose. Also, source signals (dry sources) are denoted as $\boldsymbol{s}_{ij} = (s_{ij,1}, \ldots, s_{ij,n}, \ldots, s_{ij,N})^{\mathsf{T}} \in \mathbb{C}^N$, where $n = 1, \ldots, N$ is the index of the sources and N is the number of sources. If each source in $\boldsymbol{x}_{ij}$ can be represented by a time-invariant steering vector $\boldsymbol{a}_{i,n} \in \mathbb{C}^M$, the following mixing system holds:

$$\boldsymbol{x}_{ij} = \mathbf{A}_i \boldsymbol{s}_{ij}, \tag{1}$$

where $\mathbf{A}_i = (\boldsymbol{a}_{i,1} \ \cdots \ \boldsymbol{a}_{i,N})$ is called a mixing matrix. If $M = N$ and $\mathbf{A}_i$ is invertible, the separated signal $\boldsymbol{y}_{ij} = (y_{ij,1}, \ldots, y_{ij,N})^{\mathsf{T}} \in \mathbb{C}^N$ can be obtained by estimating the demixing matrix $\mathbf{W}_i = (\boldsymbol{w}_{i,1} \ \cdots \ \boldsymbol{w}_{i,N})^{\mathsf{H}} = \mathbf{A}_i^{-1}$ as

$$\boldsymbol{y}_{ij} = \mathbf{W}_i \boldsymbol{x}_{ij}, \tag{2}$$

where $^{\mathsf{H}}$ denotes the Hermitian transpose.

B. Generative Model and Update Rules

In ILRMA, as the generative model of source signals, the following complex Gaussian distribution is assumed:

$$s_{ij,n} \sim \mathcal{N}_c\left(0, r_{ij,n}\right), \tag{3}$$

where $r_{ij,n}$ is the time-frequency-varying variance (power spectrogram model of $s_{ij,n}$). Also, $r_{ij,n}$ is modeled by NMF [16] as $r_{ij,n} = \sum_l t_{il,n} v_{lj,n}$, where $t_{il,n} \geq 0$ and $v_{lj,n} \geq 0$ are the NMF variables, $l = 1, \ldots, L$ is the index of the NMF bases, and L is the number of bases. From (1) and (3), the generative model of the observed signal becomes

$$\boldsymbol{x}_{ij} \sim \mathcal{N}_c\left(\mathbf{0}, \sum_n r_{ij,n} \boldsymbol{a}_{i,n} \boldsymbol{a}_{i,n}^{\mathsf{H}}\right). \tag{4}$$

Since the mixing system (1) is assumed in ILRMA, the spatial covariance is represented by a rank-1 matrix as $\boldsymbol{a}_{i,n}\boldsymbol{a}_{i,n}^{\mathsf{H}}$, which is called the rank-1 spatial model.

The cost function in ILRMA is defined as the negative log-likelihood function of (4) as

$$\mathcal{L} = -2J \sum_i \log |\det \mathbf{W}_i| + \sum_{i,j,n} \left(\frac{|y_{ij,n}|^2}{r_{ij,n}} + \log r_{ij,n} \right), \tag{5}$$

where $y_{ij,n} = \boldsymbol{w}_{i,n}^{\mathsf{H}} \boldsymbol{x}_{ij}$. Both the separation filter $\boldsymbol{w}_{i,n}$ and the NMF variables $t_{il,n}$ and $v_{lj,n}$ can be optimized in the maximum likelihood sense (minimization of (5)) by iterating the following iterative update rules [5]:

$$\mathbf{G}_{i,n} = \frac{1}{J} \sum_j \frac{1}{r_{ij,n}} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathsf{H}}, \tag{6}$$

$$\boldsymbol{w}_{i,n} \leftarrow (\mathbf{W}_i \mathbf{G}_{i,n})^{-1} \boldsymbol{e}_n, \tag{7}$$

$$\boldsymbol{w}_{i,n} \leftarrow \boldsymbol{w}_{i,n} (\boldsymbol{w}_{i,n}^{\mathsf{H}} \mathbf{G}_{i,n} \boldsymbol{w}_{i,n})^{-\frac{1}{2}}, \tag{8}$$

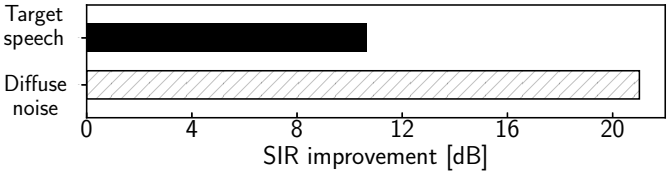

Fig. 1. SIR improvement for directional speech and diffuse noise.

where $\boldsymbol{e}_n$ denotes the unit vector with the nth element equal to unity. The update rules for $\boldsymbol{w}_{i,n}$ are called the iterative projection [17], which promises convergence-guaranteed efficient optimization. Also, we can update $t_{il,n}$ and $v_{lj,n}$ by minimizing the Itakura–Saito divergence between $\sum_l t_{il,n} v_{lj,n}$ and $r_{ij,n}$ (see [5] for details).

III. Proposed Method

A. Motivation and Strategy

In this paper, we deal with a mixture signal that includes one directional target source and diffuse background noise. Since diffuse noise cannot be expressed by the rank-1 spatial model (one steering vector), BSS based on a full-rank covariance model, such as MNMF, should be applied in this situation. However, estimation of the full-rank covariance has a huge computational cost, and its performance is always more unstable than ILRMA [5] because of the large number of spatial parameters, INM^2, which can be reduced to INM using the rank-1 spatial model (ILRMA).

For this reason, to achieve efficient and stable BSS, we propose a new ILRMA-based full-rank covariance estimation using more than or equal to two microphones. Although the sources are categorized into two groups (target and noise), we assume that one target source and $M-1$ noise components are mixed ($N = M$). This assumption allows us to model the diffuse noise using $M-1$ spatial bases (rank-$(M-1)$ spatial covariance). The extraction of the target source in this manner is still difficult because noise components exist even in the same direction as the target source. However, FDICA or ILRMA can separate the diffuse noise with high accuracy even if one spatial basis for diffuse noise is lacking. Figure 1 shows an example of the separation performance (source-to-interference ratio (SIR) [18]) obtained by ILRMA, where directional speech and diffuse noise are mixed and the experimental conditions are described in Sect. IV. It can be seen that diffuse noise is accurately estimated (almost perfectly with more than 20 dB accuracy) rather than the target speech, where diffuse noise is modeled using the rank-$(M-1)$ spatial covariance. This is because the demixing filters for the diffuse noise can precisely cancel the target speech, which is a point source [19], meaning that the steering vector of the directional source $\boldsymbol{a}_{i,n_h}$ can be estimated by ILRMA with high accuracy, where n_h denotes the index of the target source. This implies that we can fix some spatial parameters in the full-rank spatial model for diffuse noise by utilizing the estimates obtained by ILRMA in advance.

On the basis of the above motivation, we propose the following new estimation method for the full-rank spatial

covariance of diffuse noise: (a) the rank-1 spatial covariance for the target source, $\boldsymbol{a}_{i,n_h}\boldsymbol{a}_{i,n_h}^{\mathsf{H}}$, and rank-$(M-1)$ covariance for diffuse noise, $\sum_{n\neq n_h}\boldsymbol{a}_{i,n}\boldsymbol{a}_{i,n}^{\mathsf{H}}$, are estimated by ILRMA, (b) the lost spatial basis for diffuse noise is restored via the EM algorithm to estimate the noise components in the direction of the target source, and (c) a multichannel Wiener filter is applied to suppress the noise components remaining in the separated target source.

B. Model of Target Source and Diffuse Noise

The observed signal $\boldsymbol{x}_{ij}$ is assumed to be the sum of two components, as

$$\boldsymbol{x}_{ij} = \boldsymbol{h}_{ij} + \boldsymbol{u}_{ij}, \tag{9}$$

where $\boldsymbol{h}_{ij} = (h_{ij,1}, \ldots, h_{ij,M})^{\mathsf{T}} \in \mathbb{C}^M$ is the spatial image of the target source and $\boldsymbol{u}_{ij} = (u_{ij,1}, \ldots, u_{ij,M})^{\mathsf{T}} \in \mathbb{C}^M$ is that of the diffuse noise. The target source $\boldsymbol{h}_{ij}$ is modeled as

$$\boldsymbol{h}_{ij} = \boldsymbol{a}_i^{(h)} s_{ij}^{(h)}, \tag{10}$$

$$s_{ij}^{(h)} \sim \mathcal{N}_c(0, r_{ij}^{(h)}), \tag{11}$$

where $\boldsymbol{a}_i^{(h)}$, $s_{ij}^{(h)}$, and $r_{ij}^{(h)}$ are the n_hth steering vector $\boldsymbol{a}_{i,n_h}$, the dry source component, and the power spectrogram of the n_hth source, respectively. As mentioned in Sect. III-A, $\boldsymbol{a}_i^{(h)}$ can be accurately estimated by ILRMA. Thus, we hereafter consider $\boldsymbol{a}_i^{(h)}$ as a given and fixed parameter in the following processes. In addition to (11), to improve the estimation performance, we introduce an a priori distribution for the variance $r_{ij}^{(h)}$ using the inverse gamma distribution,

$$p(r_{ij}^{(h)}; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(r_{ij}^{(h)}\right)^{-\alpha-1} \exp\left(-\frac{\beta}{r_{ij}^{(h)}}\right), \tag{12}$$

where $\alpha > 0$ and $\beta > 0$ are shape and scale parameters, respectively, and a large α with a small β induces the sparseness of $r_{ij}^{(h)}$.

Since diffuse noise should have a full-rank spatial covariance, the generative model of $\boldsymbol{u}_{ij}$ is expressed by a multivariate complex Gaussian distribution as

$$\boldsymbol{u}_{ij} \sim \mathcal{N}_c(\boldsymbol{0}, r_{ij}^{(u)}\mathbf{R}_i^{(u)}), \tag{13}$$

where $r_{ij}^{(u)}$ and $\mathbf{R}_i^{(u)}$ are the variance and spatial covariance for the diffuse noise, respectively. From the estimated demixing filter $\boldsymbol{w}_{i,n}$ obtained by ILRMA, we can model the full-rank spatial covariance of the diffuse noise as follows:

$$\mathbf{R}_i^{(u)} = \mathbf{R}_i^{\prime(u)} + \lambda_i \boldsymbol{b}_i \boldsymbol{b}_i^{\mathsf{H}}, \tag{14}$$

$$\mathbf{R}_i^{\prime(u)} = \frac{1}{J}\sum_j \mathbf{W}_i^{-1} \operatorname{diag}\Big(|\boldsymbol{w}_{i,1}^{\mathsf{H}}\boldsymbol{x}_{ij}|^2, \ldots, |\boldsymbol{w}_{i,n_h-1}^{\mathsf{H}}\boldsymbol{x}_{ij}|^2, 0, |\boldsymbol{w}_{i,n_h+1}^{\mathsf{H}}\boldsymbol{x}_{ij}|^2, \ldots, |\boldsymbol{w}_{i,N}^{\mathsf{H}}\boldsymbol{x}_{ij}|^2\Big) \left(\mathbf{W}_i^{-1}\right)^{\mathsf{H}}, \tag{15}$$

where $\boldsymbol{b}_i$ is the unit eigenvector of $\mathbf{R}_i^{\prime(u)}$ that corresponds to the zero eigenvalue and λ_i is a scalar weight used to complement the lost spatial basis, namely, the direction of the target source. Note that (15) includes a back-projection operation to compensate the scales of the signals [20]. Since $\mathbf{R}_i^{\prime(u)}$ consists of $M-1$ noise estimates, its rank is $M-1$. Therefore, to restore the lost spatial basis in $\mathbf{R}_i^{\prime(u)}$, we must simultaneously estimate the eigenvalue λ_i, the variance of the target source $r_{ij}^{(h)}$, and the variance of the diffuse noise $r_{ij}^{(u)}$ with $\boldsymbol{a}_i^{(h)}$ and the rank-$(M-1)$ spatial covariance $\mathbf{R}_i^{\prime(u)}$ fixed. In summary, the number of spatial parameters to be estimated in the proposed method is INM (for ILRMA) $+ I$ (for λ_i), i.e., $I(NM+1)$, which is much less than that of MNMF (INM^2) and FastMNMF ($IM^2 + INM$).

C. Update Rules Based on EM Algorithm

The parameters λ_i, $r_{ij}^{(h)}$, and $r_{ij}^{(u)}$ are optimized by a maximum a posteriori estimation based on the EM algorithm. A Q function is defined by the expected value of the complete-data log-likelihood w.r.t. $p(s_{ij}^{(h)}, \boldsymbol{u}_{ij}|\boldsymbol{x}_{ij}; \tilde{\Theta})$ as

$$Q(\Theta; \tilde{\Theta}) = \sum_{i,j}\left[-(\alpha+2)\log r_{ij}^{(h)} - \frac{\hat{r}_{ij}^{(h)} + \beta}{r_{ij}^{(h)}} - M\log r_{ij}^{(u)} - \log\det\mathbf{R}_i^{(u)} - \frac{\operatorname{tr}\left((\mathbf{R}_i^{(u)})^{-1}\hat{\mathbf{R}}_{ij}^{(u)}\right)}{r_{ij}^{(u)}}\right] + \text{const.}, \tag{16}$$

where const. includes the constant terms that do not depend on the parameters, $\Theta = \left\{r_{ij}^{(h)}, r_{ij}^{(u)}, \lambda_i\right\}$ is the set of parameters to be updated, $\tilde{\Theta} = \left\{\tilde{r}_{ij}^{(h)}, \tilde{r}_{ij}^{(u)}, \tilde{\lambda}_i\right\}$ is the set of up-to-date parameters, and $\hat{r}_{ij}^{(h)}$ and $\hat{\mathbf{R}}_{ij}^{(u)}$ are the sufficient statistics obtained by the E-step. The update rules in the E-step are as follows:

$$\tilde{\mathbf{R}}_i^{(u)} = \mathbf{R}_i^{\prime(u)} + \tilde{\lambda}_i \boldsymbol{b}_i \boldsymbol{b}_i^{\mathsf{H}}, \tag{17}$$

$$\mathbf{R}_{ij}^{(x)} = \tilde{r}_{ij}^{(h)} \boldsymbol{a}_i^{(h)} \left(\boldsymbol{a}_i^{(h)}\right)^{\mathsf{H}} + \tilde{r}_{ij}^{(u)}\tilde{\mathbf{R}}_i^{(u)}, \tag{18}$$

$$\hat{r}_{ij}^{(h)} = \tilde{r}_{ij}^{(h)} - \left(\tilde{r}_{ij}^{(h)}\right)^2 \left(\boldsymbol{a}_i^{(h)}\right)^{\mathsf{H}} \left(\mathbf{R}_{ij}^{(x)}\right)^{-1} \boldsymbol{a}_i^{(h)} + \left|\tilde{r}_{ij}^{(h)} \boldsymbol{x}_{ij}^{\mathsf{H}} (\mathbf{R}_{ij}^{(x)})^{-1} \boldsymbol{a}_i^{(h)}\right|^2, \tag{19}$$

$$\hat{\mathbf{R}}_{ij}^{(u)} = \tilde{r}_{ij}^{(u)}\tilde{\mathbf{R}}_i^{(u)} - \left(\tilde{r}_{ij}^{(u)}\right)^2 \tilde{\mathbf{R}}_i^{(u)} \left(\mathbf{R}_{ij}^{(x)}\right)^{-1} \tilde{\mathbf{R}}_i^{(u)} + \left(\tilde{r}_{ij}^{(u)}\right)^2 \tilde{\mathbf{R}}_i^{(u)} \left(\mathbf{R}_{ij}^{(x)}\right)^{-1} \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathsf{H}} \left(\mathbf{R}_{ij}^{(x)}\right)^{-1} \tilde{\mathbf{R}}_i^{(u)}. \tag{20}$$

In the M-step, we employ a coordinate ascent algorithm to the Q function. The update rules are as follows:

$$r_{ij}^{(h)} \leftarrow \frac{\hat{r}_{ij}^{(h)} + \beta}{\alpha + 2}, \tag{21}$$

$$\mathbf{K}_i = \frac{1}{J}\sum_j \frac{1}{\tilde{r}_{ij}^{(u)}} \hat{\mathbf{R}}_{ij}^{(u)}, \tag{22}$$

$$\lambda_i \leftarrow \boldsymbol{b}_i^{\mathsf{H}} \mathbf{K}_i \boldsymbol{b}_i, \tag{23}$$

$$\mathbf{R}_i^{(u)} \leftarrow \mathbf{R}_i^{\prime(u)} + \lambda_i \boldsymbol{b}_i \boldsymbol{b}_i^{\mathsf{H}}, \tag{24}$$

$$r_{ij}^{(u)} \leftarrow \frac{1}{M}\operatorname{tr}\left(\left(\mathbf{R}_i^{(u)}\right)^{-1} \hat{\mathbf{R}}_{ij}^{(u)}\right). \tag{25}$$

TABLE I
EXPERIMENTAL CONDITIONS

Sampling frequency	16 kHz
STFT	256-ms-long Hamming window with 128 ms shift
Number of NMF bases L	10 for source model
Number of iterations in ILRMA	50
Number of iterations in methods except ILRMA	200

D. Multichannel Wiener Filter

After the estimation of all the parameters, the following multichannel Wiener filter is employed:

$$\hat{\boldsymbol{h}}_{ij} = r_{ij}^{(h)} \boldsymbol{a}_i^{(h)} \left(\boldsymbol{a}_i^{(h)}\right)^{\mathsf{H}} \left(\mathbf{R}_{ij}^{(x)}\right)^{-1} \boldsymbol{x}_{ij}, \tag{26}$$

$$\hat{\boldsymbol{u}}_{ij} = r_{ij}^{(u)} \mathbf{R}_i^{(u)} \left(\mathbf{R}_{ij}^{(x)}\right)^{-1} \boldsymbol{x}_{ij}. \tag{27}$$

E. Initialization of Source Variances

Since the EM algorithm strongly depends on the initial values of the parameters, we employ the ILRMA estimates to initialize the source variances $r_{ij}^{(h)}$ and $r_{ij}^{(u)}$ to avoid trapping at a poor local solution as follows:

$$r_{ij}^{(h)} = \sum_l t_{il,n_h} v_{lj,n_h}, \tag{28}$$

$$r_{ij}^{(u)} = \frac{1}{M} \left(\hat{\boldsymbol{y}}_{ij}^{(u)}\right)^{\mathsf{H}} \left(\mathbf{R}_i^{\prime(u)}\right)^{+} \hat{\boldsymbol{y}}_{ij}^{(u)}, \tag{29}$$

where t_{il,n_h} and v_{lj,n_h} are the low-rank source model of the target source obtained by ILRMA, $^+$ denotes the pseudoinverse, and $\hat{\boldsymbol{y}}_{ij}^{(u)}$ is the scale-fixed source image of diffuse noise obtained as $\sum_{n \neq n_h} \mathbf{W}_i^{-1}(0, \ldots, 0, \boldsymbol{w}_{i,n}^{\mathsf{H}} \boldsymbol{x}_{ij}, 0, \ldots, 0)^{\mathsf{T}}$. Also, λ_i is initialized by the minimum nonzero eigenvalue of $\mathbf{R}_i^{\prime(u)}$.

IV. EXPERIMENTS

A. Experimental Conditions

To confirm the efficacy of the proposed method, we conducted a BSS experiment using a simulated mixture of a target speech source and diffuse noise. We compared seven methods, namely, ILRMA [5], BSSA [19], the original MNMF [11], MNMF initialized by ILRMA (ILRMA+MNMF) [5], [15], the original FastMNMF [14], FastMNMF initialized by ILRMA (ILRMA+FastMNMF), and the proposed method ($\alpha = 0.7$ and $\beta = 10^{-16}$ were selected experimentally). In ILRMA, the observation $\boldsymbol{x}_{ij}$ was preprocessed via a sphering transformation using PCA. For BSSA, we replaced FDICA in [19] with ILRMA and set the oversubtraction and flooring parameters to 1.4 and 0, respectively. For ILRMA, the original MNMF, and the original FastMNMF, all the NMF variables were initialized by nonnegative random values. The demixing matrix $\mathbf{W}_i$ in ILRMA and the spatial covariance matrix in the original MNMF and the original FastMNMF were initialized by the identity matrix $\mathbf{I}$. For ILRMA+MNMF and ILRMA+FastMNMF, the NMF variables were taken from ILRMA. Also, the spatial covariance matrix was initialized using $\boldsymbol{a}_{i,n}\boldsymbol{a}_{i,n}^{\mathsf{H}} + \varepsilon\mathbf{I}$ for ILRMA+MNMF and $\boldsymbol{a}_{i,n}\boldsymbol{a}_{i,n}^{\mathsf{H}} + \varepsilon \sum_{n' \neq n} \boldsymbol{a}_{i,n'}\boldsymbol{a}_{i,n'}^{\mathsf{H}}$ for ILRMA+FastMNMF, where $\boldsymbol{a}_{i,n}$ was estimated by ILRMA and ε was set to 10^{-5}.

We used speech signals obtained from the JNAS speech corpus [21] to produce the target speech source and diffuse babble noise. The station and traffic noise signals were obtained from DEMAND [22]. These dry sources were convoluted with the impulse responses shown in Fig. 2 to simulate the mixture, where the target source was located at 30°, 20°, 10°, or 0° clockwise from the normal to a microphone array, the 18 loudspeakers used to simulate diffuse noise were arranged at intervals of 10° except in the target source direction, the size of the recording room for these impulse responses was 3.9 m $\times$ 3.9 m, and its reverberation time was about 200 ms. Note that the diffuse babble noise was produced by convoluting 18 independent speakers with each impulse response, and the diffuse station and traffic noises were produced by splitting the dry source into 18 short-time periods and convoluting them with each impulse response. The speech-to-noise ratio was set to 0 dB. The other conditions are shown in Table I.

B. Results

Source-to-distortion ratio (SDR) [18] is used as a total evaluation score in terms of separation performance and sound distortion. The SDR behaviors for each of the methods, which are the averaged results over 10 parameter-initialization random seeds and four target directions, are shown in Fig. 3, where those of ILRMA-initialized methods are depicted except for their initializing iterations of ILRMA. The proposed method outperformed the other methods. In particular, the full-rank spatial model in the proposed method showed an improvement of more than 3 dB compared with the rank-1 spatial model in ILRMA, and the efficacy of the proposed spatial model extension was confirmed. Also, we reveal that, even with the assistance of ILRMA-based initialization, the SDRs of the conventional MNMFs and FastMNMFs with the full-rank spatial model cannot reach that of the proposed method.

As regards the optimization cost, the EM algorithm in the proposed method converged within five iterations, which was greatly reduced from the number of iterations required for MNMFs and FastMNMFs. In addition, the actual computational times of MNMF, FastMNMF, and the proposed EM algorithm for each iteration were 10.18 s, 0.87 s, and 0.005 s, respectively, further illustrating the advantageousness of the proposed method.

On the other hand, the unbiased sample standard deviations of SDR improvements just after 200 iterations of ILRMA, original MNMF, ILRMA+MNMF, original FastMNMF, ILRMA+FastMNMF, and the proposed method are 0.19, 2.37, 0.38, 5.75, 0.26, and 0.22, respectively. This means that the proposed method is a more stable algorithm than MNMF and FastMNMF in terms of initialization dependency.

V. CONCLUSION

We proposed a new algorithm that accurately and efficiently extracts a directional target source in diffuse background noise.

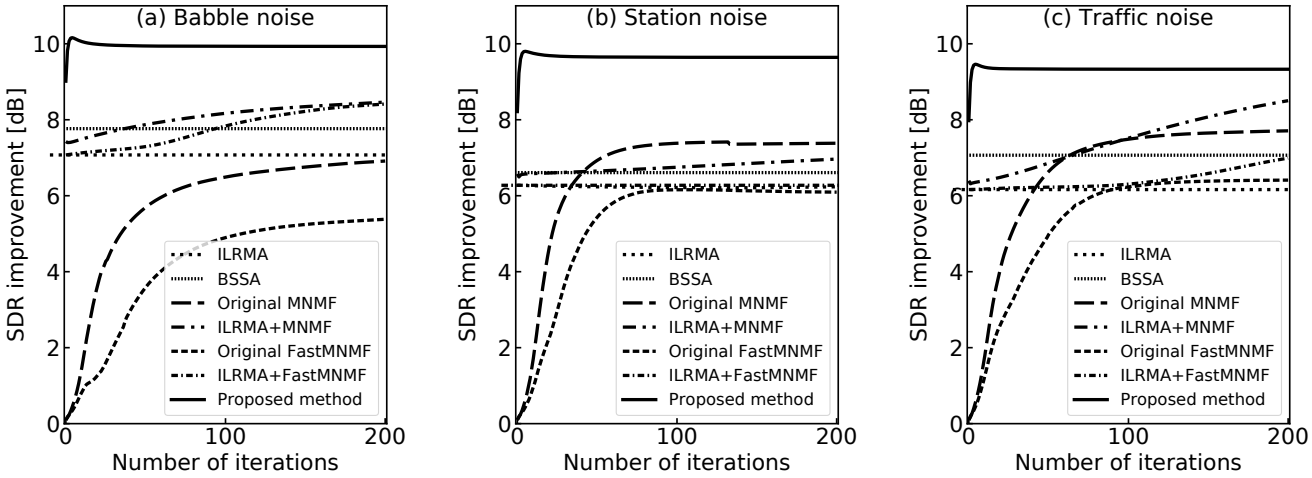

Fig. 3. SDR behaviors averaged over 10 parameter-initialization random seeds and four target directions in separation of target speech and diffuse (a) babble, (b) station, (c) traffic noises, where speech-to-noise ratio is 0 dB.

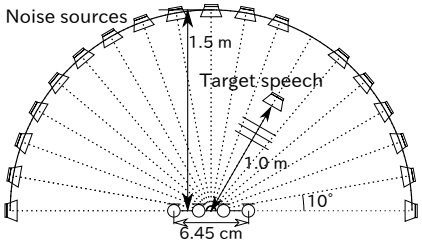

Fig. 2. Recording conditions of impulse responses (when target source is located at 30°), where reverberation time T_{60} is 200 ms.

The proposed method is based on ILRMA and restores the lost spatial basis by using the EM algorithm to extend the spatial covariance of the noise from a rank-$(M-1)$ matrix to the full-rank matrix. In an experiment, we confirmed that the proposed method outperforms the conventional methods in terms of accuracy and computational efficiency.

Acknowledgment

This work was partly supported by SECOM Science and Technology Foundation and JSPS KAKENHI Grant Numbers 17H06101, 19H01116, and 19K20306.

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 Trans. ASLP*, vol. 14, no. 2, pp. 666–678, 2006.

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

[4] T. Kim et al., "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2007.

[5] D. Kitamura et al., "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[6] D. Kitamura et al., "Determined blind source separation with independent low-rank matrix analysis," in *Audio Source Separation*, S. Makino, Ed., pp. 125–155. Springer, Cham, 2018.

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

[8] D. D. Lee, H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, 2000, pp. 556–562.

[9] S. Araki et al., "Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures," *EURASIP JASP*, vol. 2003, no. 11, pp. 1–10, 2003.

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

[11] H. Sawada et al., "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. ASLP*, vol. 21, no. 5, pp. 971–982, 2013.

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

[13] N. Ito, T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *Proc. ICASSP*, 2019, pp. 371–375.

[14] K. Sekiguchi et al., "Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices," *CoRR*, vol. abs/1903.03237, 2019.

[15] K. Shimada et al., "Unsupervised beamforming based on multichannel nonnegative matrix factorization for noisy speech recognition," in *Proc. ICASSP*, 2018, pp. 5734–5738.

[16] C. Févotte et al., "Nonnegative matrix factorization with the Itakura–Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, 2009.

[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] E. Vincent et al., "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] Y. Takahashi et al., "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Trans. ASLP*, vol. 17, no. 4, pp. 650–664, 2009.

[20] N. Murata et al., "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, 2001.

[21] K. Itou et al., "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *J. Acoust. Soc. Jpn. (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[22] J. Thiemann et al., "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," June 2013, Supported by Inria under the Associate Team Program VERSAMUS.