NoisyILRMA: Diffuse-Noise-Aware Independent Low-Rank Matrix Analysis for Fast Blind Source Extraction

Koki Nishida[†], Norihiro Takamune[†], Rintaro Ikeshita[‡], Daichi Kitamura[§],
Hiroshi Saruwatari[†], and Tomohiro Nakatani[‡]
[†]The University of Tokyo, Graduate School of Information Science and Technology, Tokyo, Japan
[‡]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
[§]National Institute of Technology, Kagawa College, Kagawa, Japan

***Abstract*—In this paper, we address the multichannel blind source extraction (BSE) of a single source in diffuse noise environments. To solve this problem even faster than by fast multichannel nonnegative matrix factorization (FastMNMF) and its variant, we propose a BSE method called NoisyILRMA, which is a modification of independent low-rank matrix analysis (ILRMA) to account for diffuse noise. NoisyILRMA can achieve considerably fast BSE by incorporating an algorithm developed for independent vector extraction. In addition, to improve the BSE performance of NoisyILRMA, we propose a mechanism to switch the source model with ILRMA-like nonnegative matrix factorization to a more expressive source model during optimization. In the experiment, we show that NoisyILRMA runs faster than a FastMNMF algorithm while maintaining the BSE performance. We also confirm that the switching mechanism improves the BSE performance of NoisyILRMA.**

***Index Terms*—Multichannel blind source extraction, diffuse noise environments, independent low-rank matrix factorization, independent vector extraction, generalized eigenvalue problem**

I. Introduction

Multichannel blind source separation (BSS) is a technique used to separate multiple sources from multichannel observed signals recorded by a microphone array without any prior knowledge of, for example, the characteristics of sources or spatial mixing systems. Among BSS, the technique used to extract the source signal(s) from the background noise is particularly called multichannel blind source extraction (BSE). BSE can be used as a front-end of sound signal processing devices such as hearing aids and smart speakers. This study is particularly focused on extracting one target source in diffuse noise environments.

Independent low-rank matrix analysis (ILRMA) [1] is one of the BSS methods that can separate point sources when the number of sources is less than or equal to that of microphones. ILRMA assumes independence between the sources and the low rankness of the sources in the time-frequency domain using nonnegative matrix factorization (NMF) [2]. In [1], ILRMA was reported to experimentally achieve high and stable performance. Although this method can separate point sources accurately, it is impossible to remove diffuse noise in the same direction as the target source [3].

This research was partly supported by JST Moonshot R&D Grant Number JPMJMS2011 and JSPS KAKENHI Grant Number 19H01116.

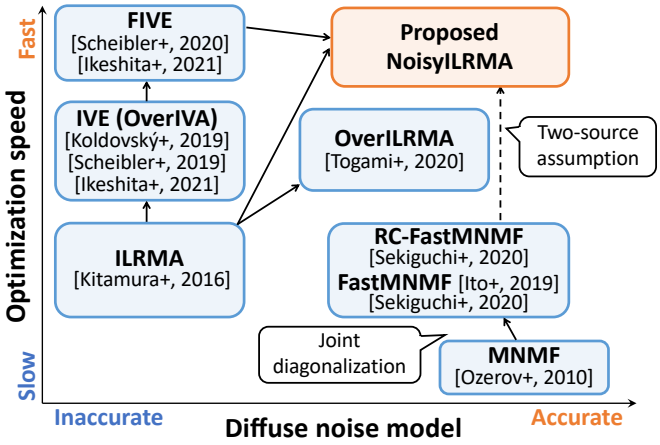

Fig. 1: Relationships between proposed and conventional methods.

Multichannel NMF (MNMF) [4] is a multichannel extension of NMF employing a full-rank spatial covariance matrix (SCM) [5] for each source, which can model diffuse noise. However, its computational complexity is high owing to the large number of parameters for full-rank SCMs. To overcome this problem, FastMNMF [6], [7] has been proposed. FastMNMF assumes that the SCMs for all sources are jointly diagonalizable [8], which allows a faster algorithm called iterative projection (IP) [9] to be used for the estimation of the joint-diagonalization matrix. By restricting the situation where some sources are point sources, rank-constrained FastMNMF (RC-FastMNMF) [7] has been proposed for efficient estimation. This method focuses on the fact that the SCM of a point source can be approximated by a rank-1 matrix and constrains the SCMs corresponding to the point sources as rank-1 matrices, enabling the efficient exploration of the solution space.

In this paper, we propose a method called NoisyILRMA, which is ILRMA modified to account for diffuse noise. When we use ILRMA under diffuse noise environments, the separated signal corresponding to the target source contains the target source and noise, while the other separated signals contain noise only [10]. By explicitly modeling this property, NoisyILRMA simultaneously estimates the separated target signal and the residual noise component in it. We focus on

the fact that the optimization problem of the demixing filters has the same form as that in independent vector extraction (IVE) [11]–[13], which is a method used to efficiently extract the target sources in overdetermined cases (the number of sources is less than that of microphones). For IVE, fast IVE (FIVE) [13], [14], in which the demixing filters are optimized by a fast algorithm, has been proposed. From these facts, we can use the same fast algorithm as FIVE to optimize the demixing filters in NoisyILRMA. In Sect. III-C, we show that NoisyILRMA can be viewed as an accelerated RC-FastMNMF used under the two-source assumption (for single point source extraction under diffuse noise). We also discuss OverILRMA [15], which is an extension of ILRMA to overdetermined cases, as another closely related method in Sect. III-C. Fig. 1 shows the relationships between NoisyILRMA and other methods.

In addition, to further improve the BSE performance of NoisyILRMA, we propose a source model switching mechanism. In NoisyILRMA, the limited expressiveness of the NMF model degrades the final BSE performance, although the NMF model is useful for estimating the demixing filters. To solve this problem, we focus on rank-constrained SCM estimation (RCSCME) [16]. By utilizing the spatial information estimated in the preprocessing BSS method such as ILRMA, RCSCME efficiently estimates the residual spatial information and the source models for the target source and noise. RCSCME achieves high-performance BSE owing to the use of a more expressive source model than the NMF model. We focus on this fact and propose a source model switching mechanism in which we first obtain the demixing filters accurately using the NMF model and then fix the demixing filters and switch to the same highly expressive source model as in RCSCME.

II. Background

A. ILRMA [1]

Let $\boldsymbol{x}_{ij} \in \mathbb{C}^M$ be the short-time Fourier transformation (STFT) of the multichannel observed signal, where i and j are the indices of frequency bins and time frames, respectively, and M is the number of microphones. When each source can be assumed to be a point source and the window in STFT is sufficiently longer than the reverberation, the following mixing system $\boldsymbol{x}_{ij} = \mathbf{A}_i \boldsymbol{s}_{ij}$ holds, where $\boldsymbol{s}_{ij} \in \mathbb{C}^N$ and $\mathbf{A}_i = (\boldsymbol{a}_{i1}, \ldots, \boldsymbol{a}_{iN})$ are the source signals and mixing matrix, respectively. Here, $\boldsymbol{a}_{in} \in \mathbb{C}^M$ denotes the steering vector, where $n \in \{1, \ldots, N\}$ is the index of the sources. If $N = M$ and $\mathbf{A}_i$ is invertible, the separated signal $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijM})^{\mathsf{T}}$ can be obtained as

$$\boldsymbol{y}_{ij} = \mathbf{W}_i^{\mathsf{H}} \boldsymbol{x}_{ij}, \tag{1}$$

where $\mathbf{W}_i = (\mathbf{A}_i^{\mathsf{H}})^{-1}$ is the demixing matrix, and $^{\mathsf{T}}$ and $^{\mathsf{H}}$ denote transpose and Hermitian transpose, respectively. We assume the following complex Gaussian distribution for y_{ijn}:

$$y_{ijn} \sim \mathcal{N}_{\mathbb{C}}\left(0, r_{ijn}\right), \tag{2}$$

where r_{ijn} is the time-variant variance of source n, which is modeled by NMF as $r_{ijn} = \sum_k t_{ikn} v_{kjn}$. Here, $t_{ikn}, v_{kjn} \geq 0$ are the NMF basis and activation, respectively, and $k \in \{1, \ldots, K\}$ is the index of the basis.

When there exist a single point source and diffuse noise, the separated signals satisfy the following property [10]: $y_{ijn_{\mathrm{s}}}$ contains both the target source and noise, while the other $M-1$ separated signals contain noise only, where n_{s} is the index of the separated signal corresponding to the target source.

B. RC-FastMNMF [7]

In MNMF, we assume the following multivariate complex Gaussian distribution for $\boldsymbol{x}_{ij}$:

$$\boldsymbol{x}_{ij} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}_M, \sum_{\tilde{n}=1}^{\tilde{N}} r_{ij\tilde{n}} \mathbf{R}_{i\tilde{n}}\right), \tag{3}$$

where $\mathbf{0}_M \in \mathbb{C}^M$ is a zero vector, $\mathbf{R}_{i\tilde{n}} \in \mathbb{C}^{M \times M}$ denotes the SCM of source $\tilde{n}$, and $\tilde{n} \in \{1, \ldots, \tilde{N}\}$ is the index of sources. Here, $r_{ij\tilde{n}}$ is also modeled by NMF as $r_{ij\tilde{n}} = \sum_k t_{ik\tilde{n}} v_{kj\tilde{n}}$. Note that $\tilde{N}$ is not necessarily equal to M, unlike in ILRMA. In FastMNMF, we assume the following joint diagonalizability for the SCMs of the $\tilde{N}$ sources to estimate them efficiently:

$$\tilde{\mathbf{W}}_i^{\mathsf{H}} \mathbf{R}_{i\tilde{n}} \tilde{\mathbf{W}}_i = \mathrm{diag}\left(\lambda_{i1\tilde{n}}, \ldots, \lambda_{iM\tilde{n}}\right), \tag{4}$$

where $\mathrm{diag}(q_1 \ldots, q_M) \in \mathbb{C}^{M \times M}$ is the diagonal matrix whose mth element is q_m, $\tilde{\mathbf{W}}_i = (\tilde{\boldsymbol{w}}_{i,1}, \ldots, \tilde{\boldsymbol{w}}_{iM}) \in \mathbb{C}^{M \times M}$ is the joint-diagonalization matrix, $\lambda_{im\tilde{n}} \geq 0$ is a diagonal element of diagonalized SCMs, and $m \in \{1, \ldots, M\}$ is the index of the column of the joint-diagonalization matrix. In RC-FastMNMF, from the fact that the SCM of a point source can be approximated as a rank-1 matrix, we introduce the rank-1 constraint for the point source $\tilde{n}'$ by setting $\lambda_{im\tilde{n}'} = 0$ for $m \in \{1, \ldots, M\} \backslash \{\tilde{n}'\}$.

The model parameters of RC-FastMNMF $\Theta^{\text{RC-FastMNMF}} = \{t_{ik\tilde{n}}, v_{kj\tilde{n}}, \tilde{\mathbf{W}}_i, \lambda_{im\tilde{n}}\}$ are estimated in the maximum likelihood sense. For $\tilde{\mathbf{W}}_i$, the IP algorithm [9] can be used [7], where each $\tilde{\mathbf{W}}_i$ column is alternately updated. A relationship with the proposed method is discussed in Sect. III-C.

C. RCSCME [16]

Using the property of ILRMA under diffuse noise environments, we can accurately obtain the steering vector of the target source $\boldsymbol{a}'_i \in \mathbb{C}^M$ and the rank-$(M-1)$ component of the noise SCM $\mathbf{R}'^{(\mathrm{n})}_i \in \mathbb{C}^{M \times M}$. Focusing on this fact, RCSCME uses ILRMA as a preprocess and utilizes the spatial information obtained in ILRMA for efficient estimation.

In RCSCME, we assume $\boldsymbol{x}_{ij}$ follows the following multivariate complex Gaussian distribution with the inverse-gamma prior distribution:

$$\begin{aligned} \boldsymbol{x}_{ij} \mid r^{(\mathrm{s})}_{ij} &\sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}_M, r^{(\mathrm{s})}_{ij} \boldsymbol{a}'_i (\boldsymbol{a}'_i)^{\mathsf{H}} + r^{(\mathrm{n})}_{ij} (\mathbf{R}'^{(\mathrm{n})}_i + \mu_i \boldsymbol{b}_i \boldsymbol{b}_i^{\mathsf{H}})\right), \\ r^{(\mathrm{s})}_{ij} &\sim \mathcal{IG}(\alpha, \beta), \end{aligned} \tag{5}$$

where $r^{(\mathrm{s})}_{ij}, r^{(\mathrm{n})}_{ij} > 0$ are time-variant variances of the target speech and noise, $\mu_i > 0$ and $\boldsymbol{b}_i \in \mathbb{C}^M$ are the weight and direction vector used to represent the deficient rank-1 component of the noise SCM, and $\alpha, \beta > 0$ are the shape and scale parameters of the inverse-gamma distribution, respectively. Here, $\boldsymbol{a}'_i = (\mathbf{W}'^{\mathsf{H}}_i)^{-1} \boldsymbol{e}_{n_{\mathrm{s}}}$ holds, where $\mathbf{W}'_i$ is the demixing matrix estimated in ILRMA and $\boldsymbol{e}_n$ is the unit vector whose nth element is one. We calculate $\mathbf{R}'^{(\mathrm{n})}_i$ as $\mathbf{R}'^{(\mathrm{n})}_i =$

$\sum_j \boldsymbol{x}'^{(\mathrm{n})}_{ij}(\boldsymbol{x}'^{(\mathrm{n})}_{ij})^{\mathsf{H}}/J$, where $\boldsymbol{x}'^{(\mathrm{n})}_{ij} = \boldsymbol{x}_{ij} - \boldsymbol{a}'_i(\mathbf{W}'_i\boldsymbol{e}_{n_\mathrm{s}})^{\mathsf{H}}\boldsymbol{x}_{ij}$ holds, and J is the number of time frames. $\boldsymbol{b}_i$ is a constant vector that is linearly independent of the column vectors of $\mathbf{R}'^{(\mathrm{n})}_i$. The inverse-gamma prior distribution in (5) is introduced for the sparsity of the target source in the time-frequency domain. Note that $r^{(\mathrm{s})}_{ij}$ and $r^{(\mathrm{n})}_{ij}$ are unconstrained parameters with the prior distribution in (5), which is more expressive than the NMF model. RCSCME can achieve high BSE performance owing to the more expressive source model.

III. Proposed methods

A. Motivation

In this section, we propose the diffuse-noise-aware ILRMA, which we call NoisyILRMA, for single-source extraction. When we use ILRMA in diffuse noise environments, the following properties hold [10]:

- y_{ijn_s} contains both speech and noise.
- The other $M-1$ separated signals contain noise only.

By explicitly modeling these properties, NoisyILRMA simultaneously estimates the demixing matrix $\mathbf{W}_i$ and the noise component in the separated target signal y_{ijn_s}. In NoisyILRMA, we also assume that the noise component in y_{ijn_s} has time synchronization with the other $M-1$ separated signals to enable a reasonable estimation. We can suppress the noise component in y_{ijn_s} by multichannel Wiener filter (MWF) using the estimated variances. As an additional advantage of such modeling in NoisyILRMA, we show that the optimization problem of the demixing matrix has the same form as that in IVE [11]–[13], which enables us to use the same considerably fast algorithm in FIVE [13], [14].

In addition, to further improve the BSE performance of NoisyILRMA, we propose a source model switching mechanism. The NMF model in NoisyILRMA is useful for estimating the demixing matrix $\mathbf{W}_i$ because it clusters the frequency bins corresponding to the same source. However, the NMF model may degrade the final BSE performance when using the MWF owing to NMF's limited expressiveness. Inspired by RCSCME, in the proposed switching mechanism, we first obtain $\mathbf{W}_i$ accurately using the NMF model and then switch to the same highly expressive source model as in RCSCME.

B. Method

In NoisyILRMA, we introduce the above-mentioned properties of the separated signal y_{ijn} and the assumption that the noise component in y_{ijn_s} has time synchronization with the other separated signals as follows:

$$y_{ij1} \sim \mathcal{N}_{\mathbb{C}}\left(0, r^{(\mathrm{s})}_{ij} + r^{(\mathrm{n})}_{ij}\lambda^{(\mathrm{n})}_i\right), \tag{6}$$

$$y_{ijn} \sim \mathcal{N}_{\mathbb{C}}\left(0, r^{(\mathrm{n})}_{ij}\right) \ (n \in \{2,\dots,M\}), \tag{7}$$

where $\lambda^{(\mathrm{n})}_i > 0$ denotes the weight of the noise component in y_{ij1}, $r^{(l)}_{ij}$ is modeled by NMF as $r^{(l)}_{ij} = \sum_k t^{(l)}_{ik}v^{(l)}_{kj}$, $l \in \{\mathrm{s},\mathrm{n}\}$ is the label used to distinguish the target source and noise, and $t^{(l)}_{ik}$, $v^{(l)}_{kj} \geq 0$ are the NMF basis and activation, respectively. Note that we assume $n_\mathrm{s} = 1$ without loss of generality and $y_{ij2},\dots,y_{ijM}$ to have the same variance by using the scale arbitrariness of $\boldsymbol{w}_{in}$.

In NoisyILRMA, the cost function is defined as the negative log-likelihood, which is obtained from (1), (6), and (7) as

$$\begin{aligned}\mathcal{L}(\Theta) = \sum_{i,j}\Bigg[&-2\log|\det\mathbf{W}_i| \\ &+ \log\left(r^{(\mathrm{s})}_{ij} + r^{(\mathrm{n})}_{ij}\lambda^{(\mathrm{n})}_i\right) + (M-1)\log r^{(\mathrm{n})}_{ij} \\ &+ \frac{|y_{ij1}|^2}{r^{(\mathrm{s})}_{ij} + r^{(\mathrm{n})}_{ij}\lambda^{(\mathrm{n})}_i} + \frac{\sum_{n=2}^{M}|y_{ijn}|^2}{r^{(\mathrm{n})}_{ij}}\Bigg] + \mathrm{const.},\end{aligned} \tag{8}$$

where $\Theta = \{t^{(l)}_{ik}, v^{(l)}_{kj}, \mathbf{W}_i, \lambda^{(\mathrm{n})}_i\}$ is the set of model parameters and const. includes the terms independent of Θ. $\mathbf{W}_i$ and the other parameters $\{t^{(l)}_{ik}, v^{(l)}_{kj}, \lambda^{(\mathrm{n})}_i\}$ are alternately updated to minimize (8). To derive the update rule for $\mathbf{W}_i$, we transform the cost function (8) with respect to $\mathbf{W}_i = (\boldsymbol{w}^{(\mathrm{s})}_i, \mathbf{W}^{(\mathrm{n})}_i)$ as

$$\begin{aligned}\mathcal{L}(\{\mathbf{W}_i\}) = J\sum_i\Bigg[&-2\log|\det\mathbf{W}_i| + (\boldsymbol{w}^{(\mathrm{s})}_i)^{\mathsf{H}}\mathbf{G}^{(\mathrm{s})}_i\boldsymbol{w}^{(\mathrm{s})}_i \\ &+ \mathrm{Tr}\Big((\mathbf{W}^{(\mathrm{n})}_i)^{\mathsf{H}}\mathbf{G}^{(\mathrm{n})}_i\mathbf{W}^{(\mathrm{n})}_i\Big)\Bigg] + \mathrm{const.},\end{aligned} \tag{9}$$

where we define $\mathbf{G}^{(\mathrm{s})}_i = \sum_j \boldsymbol{x}_{ij}\boldsymbol{x}^{\mathsf{H}}_{ij}/(r^{(\mathrm{s})}_{ij} + r^{(\mathrm{n})}_{ij}\lambda^{(\mathrm{n})}_i)/J$ and $\mathbf{G}^{(\mathrm{n})}_i = \sum_j \boldsymbol{x}_{ij}\boldsymbol{x}^{\mathsf{H}}_{ij}/r^{(\mathrm{n})}_{ij}/J$, and const. includes the terms independent of $\mathbf{W}_i$. Since this cost function (9) is the same form as that for IVE [11]–[13], the fast algorithm in FIVE [13], [14] can be used to optimize $\mathbf{W}_i$. By transforming the condition that the Wirtinger derivative of (9) with respect to $\mathbf{W}_i$ equals zero, we can obtain the following equations:

$$(\boldsymbol{w}^{(\mathrm{s})}_i)^{\mathsf{H}}\mathbf{G}^{(\mathrm{s})}_i\boldsymbol{w}^{(\mathrm{s})}_i = 1, \tag{10}$$

$$(\mathbf{W}^{(\mathrm{n})}_i)^{\mathsf{H}}\mathbf{G}^{(\mathrm{s})}_i\boldsymbol{w}^{(\mathrm{s})}_i = \mathbf{0}_{M-1}, \tag{11}$$

$$(\boldsymbol{w}^{(\mathrm{s})}_i)^{\mathsf{H}}\mathbf{G}^{(\mathrm{n})}_i\mathbf{W}^{(\mathrm{n})}_i = \mathbf{0}^{\mathsf{H}}_{M-1}, \tag{12}$$

$$(\mathbf{W}^{(\mathrm{n})}_i)^{\mathsf{H}}\mathbf{G}^{(\mathrm{n})}_i\mathbf{W}^{(\mathrm{n})}_i = \mathbf{E}_{M-1}, \tag{13}$$

where $\mathbf{E}_{M-1} \in \mathbb{C}^{(M-1)\times(M-1)}$ is the identity matrix. In [13], the update rule of $\boldsymbol{w}^{(\mathrm{s})}_i$ is derived as

$$\boldsymbol{w}^{(\mathrm{s})}_i \leftarrow \frac{\boldsymbol{h}_{i1}}{\sqrt{\boldsymbol{h}^{\mathsf{H}}_{i1}\mathbf{G}^{(\mathrm{s})}_i\boldsymbol{h}_{i1}}}, \tag{14}$$

where $\boldsymbol{h}_{i1} \in \mathbb{C}^M$ is the generalized eigenvector with the largest generalized eigenvalue in the following generalized eigenvalue problem:

$$\mathbf{G}^{(\mathrm{n})}_i\boldsymbol{v}_i = \kappa_i\ \mathbf{G}^{(\mathrm{s})}_i\boldsymbol{v}_i. \tag{15}$$

Here, $\kappa_i > 0$ and $\boldsymbol{v}_i \in \mathbb{C}^M$ are the generalized eigenvalue and the generalized eigenvector, respectively. We can update $\mathbf{W}^{(\mathrm{n})}_i$ to satisfy (11)–(13) as follows:

$$\mathbf{W}^{(\mathrm{n})}_i \leftarrow \left(\frac{\boldsymbol{h}_{i2}}{\sqrt{\boldsymbol{h}^{\mathsf{H}}_{i2}\mathbf{G}^{(\mathrm{n})}_i\boldsymbol{h}_{i2}}},\dots,\frac{\boldsymbol{h}_{iM}}{\sqrt{\boldsymbol{h}^{\mathsf{H}}_{iM}\mathbf{G}^{(\mathrm{n})}_i\boldsymbol{h}_{iM}}}\right), \tag{16}$$

where $\boldsymbol{h}_{i2},\dots,\boldsymbol{h}_{iM} \in \mathbb{C}^M$ are the other generalized eigenvectors of (15). By using (14) and (16), we can update all $\mathbf{W}_i$ columns simultaneously.

We can derive the update rules for the other parameters $t^{(l)}_{ik}$, $v^{(l)}_{kj}$, and $\lambda^{(\mathrm{n})}_i$ in the same manner as in FastMNMF [7]

by using the majorization-minimization (MM) algorithm [17]:

$$t_{ik}^{(\mathrm{s})} \leftarrow t_{ik}^{(\mathrm{s})} \sqrt{\frac{\sum_j \frac{|y_{ij1}|^2}{(r_{ij}^{(\mathrm{s})}+r_{ij}^{(\mathrm{n})}\lambda_i^{(\mathrm{n})})^2} v_{kj}^{(\mathrm{s})}}{\sum_j \frac{1}{r_{ij}^{(\mathrm{s})}+r_{ij}^{(\mathrm{n})}\lambda_i^{(\mathrm{n})}} v_{kj}^{(\mathrm{s})}}}, \tag{17}$$

$$v_{kj}^{(\mathrm{s})} \leftarrow v_{kj}^{(\mathrm{s})} \sqrt{\frac{\sum_i \frac{|y_{ij1}|^2}{(r_{ij}^{(\mathrm{s})}+r_{ij}^{(\mathrm{n})}\lambda_i^{(\mathrm{n})})^2} t_{ik}^{(\mathrm{s})}}{\sum_i \frac{1}{r_{ij}^{(\mathrm{s})}+r_{ij}^{(\mathrm{n})}\lambda_i^{(\mathrm{n})}} t_{ik}^{(\mathrm{s})}}}, \tag{18}$$

$$t_{ik}^{(\mathrm{n})} \leftarrow t_{ik}^{(\mathrm{n})} \sqrt{\frac{\sum_j \left(\frac{\lambda_i |y_{ij1}|^2}{(r_{ij}^{(\mathrm{s})}+r_{ij}^{(\mathrm{n})}\lambda_i^{(\mathrm{n})})^2} + \frac{\sum_{n=2}^{M} |y_{ijn}|^2}{(r_{ij}^{(\mathrm{n})})^2} \right) v_{kj}^{(\mathrm{n})}}{\sum_j \left(\frac{\lambda_i}{r_{ij}^{(\mathrm{s})}+r_{ij}^{(\mathrm{n})}\lambda_i^{(\mathrm{n})}} + \frac{M-1}{r_{ij}^{(\mathrm{n})}} \right) v_{kj}^{(\mathrm{n})}}}, \tag{19}$$

$$v_{kj}^{(\mathrm{n})} \leftarrow v_{kj}^{(\mathrm{n})} \sqrt{\frac{\sum_i \left(\frac{\lambda_i |y_{ij1}|^2}{(r_{ij}^{(\mathrm{s})}+r_{ij}^{(\mathrm{n})}\lambda_i^{(\mathrm{n})})^2} + \frac{\sum_{n=2}^{M} |y_{ijn}|^2}{(r_{ij}^{(\mathrm{n})})^2} \right) t_{ik}^{(\mathrm{n})}}{\sum_i \left(\frac{\lambda_i}{r_{ij}^{(\mathrm{s})}+r_{ij}^{(\mathrm{n})}\lambda_i^{(\mathrm{n})}} + \frac{M-1}{r_{ij}^{(\mathrm{n})}} \right) t_{ik}^{(\mathrm{n})}}}, \tag{20}$$

$$\lambda_i^{(\mathrm{n})} \leftarrow \lambda_i^{(\mathrm{n})} \sqrt{\frac{\sum_j \frac{r_{ij}^{(\mathrm{n})} |y_{ij1}|^2}{(r_{ij}^{(\mathrm{s})}+r_{ij}^{(\mathrm{n})}\lambda_i^{(\mathrm{n})})^2}}{\sum_j \frac{r_{ij}^{(\mathrm{n})}}{r_{ij}^{(\mathrm{s})}+r_{ij}^{(\mathrm{n})}\lambda_i^{(\mathrm{n})}}}}. \tag{21}$$

We obtain the extracted target source signal $\hat{\boldsymbol{s}}_{ij}$ by using the following MWF [8]:

$$\hat{\boldsymbol{s}}_{ij} = \underbrace{\boldsymbol{a}_i^{(\mathrm{s})}}_{\text{projection back [18]}} \underbrace{\frac{r_{ij}^{(\mathrm{s})}}{r_{ij}^{(\mathrm{s})}+r_{ij}^{(\mathrm{n})}\lambda_i^{(\mathrm{n})}}}_{\text{postfiltering}} \underbrace{(\boldsymbol{w}_i^{(\mathrm{s})})^{\mathsf{H}} \boldsymbol{x}_{ij}}_{\text{linear filtering}}, \tag{22}$$

where $\boldsymbol{a}_i^{(\mathrm{s})} = (\mathbf{W}_i^{\mathsf{H}})^{-1}\boldsymbol{e}_1$.

C. Relationship with other methods

In this section, we describe the relationship between NoisyILRMA and other methods. When we assume the two-source situation (the target source and noise), RC-FastMNMF becomes equivalent to NoisyILRMA in terms of modeling. From (3) and (4), $\tilde{y}_{ijm} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_{\tilde{n}=1}^{\tilde{N}} r_{ij\tilde{n}}\lambda_{im\tilde{n}}\right)$ holds, where $\tilde{\boldsymbol{y}}_{ij} = (\tilde{y}_{ij1}, \ldots, \tilde{y}_{ijM})^{\mathsf{T}} = \tilde{\mathbf{W}}_i^{\mathsf{H}}\boldsymbol{x}_{ij}$ is the decorrelated signal. By substituting $\tilde{N} = 2$, $\lambda_{i11} = 1$, $\lambda_{im1} = 0$ ($m \in \{2, \ldots, M\}$), $\lambda_{i12} = \lambda_i^{(\mathrm{n})}$, and $\lambda_{im2} = 1$ ($m \in \{2, \ldots, M\}$), we can confirm that $\tilde{\boldsymbol{y}}_{ij}$ satisfies (6)–(7) when we assume $\tilde{\mathbf{W}}_i = \mathbf{W}_i$. One major advantage of NoisyILRMA over RC-FastMNMF is that by setting $\tilde{N} = 2$, we can use a fast algorithm in which all $\mathbf{W}_i$ columns are updated simultaneously, whereas each $\tilde{\mathbf{W}}_i$ column is alternately updated in RC-FastMNMF. This makes NoisyILRMA considerably faster than RC-FastMNMF.

OverILRMA [15] is a method closely related to NoisyILRMA. The main difference is that $r_{ij}^{(\mathrm{n})}$ is a time-invariant parameter in OverILRMA, while $r_{ij}^{(\mathrm{n})}$ is modeled as a time-variant parameter using the NMF model in NoisyILRMA. There are two advantages of time-variant modeling. Firstly, we can model time-variant diffuse noise. Secondly, the noise component in y_{ij1} can be estimated using the time synchronization with the other separated signals in time-variant modeling, whereas it is estimated independently of the other separated signals in time-invariant modeling. Because of these advantages of time-variant modeling, NoisyILRMA is expected to achieve higher performance than OverILRMA.

D. Source model switching in NoisyILRMA

In this section, we propose a source model switching mechanism. In this method, we first use NoisyILRMA for several iterations to obtain $\mathbf{W}_i$ accurately by the NMF model. After that, we fix $\mathbf{W}_i$ and switch to the same source model as in RCSCME, i.e., $r_{ij}^{(\mathrm{s})}$ and $r_{ij}^{(\mathrm{n})}$ are the unconstrained parameters with the inverse-gamma prior distribution in (5). This enables a finer estimation of $r_{ij}^{(l)}$ and we expect higher BSE performance using MWF (22).

We can derive the update rules based on maximum a posteriori estimation in a manner similar to that in RCSCME [16] by using the MM algorithm as follows (the update rule for $\lambda_i^{(\mathrm{n})}$ is the same form as (21)):

$$r_{ij}^{(\mathrm{s})} \leftarrow r_{ij}^{(\mathrm{s})} \sqrt{\frac{\frac{|y_{ij1}|^2}{(r_{ij}^{(\mathrm{s})}+r_{ij}^{(\mathrm{n})}\lambda_i^{(\mathrm{n})})^2} + \frac{\beta}{(r_{ij}^{(\mathrm{s})})^2}}{\frac{1}{r_{ij}^{(\mathrm{s})}+r_{ij}^{(\mathrm{n})}\lambda_i^{(\mathrm{n})}} + \frac{\alpha+1}{r_{ij}^{(\mathrm{s})}}}}, \tag{23}$$

$$r_{ij}^{(\mathrm{n})} \leftarrow r_{ij}^{(\mathrm{n})} \sqrt{\frac{\frac{\lambda_i |y_{ij1}|^2}{(r_{ij}^{(\mathrm{s})}+r_{ij}^{(\mathrm{n})}\lambda_i^{(\mathrm{n})})^2} + \frac{\sum_{n=2}^{M} |y_{ijn}|^2}{(r_{ij}^{(\mathrm{n})})^2}}{\frac{\lambda_i}{r_{ij}^{(\mathrm{s})}+r_{ij}^{(\mathrm{n})}\lambda_i^{(\mathrm{n})}} + \frac{M-1}{r_{ij}^{(\mathrm{n})}}}}. \tag{24}$$

IV. Experimental analysis

A. Experimental conditions

We conducted a BSE experiment using simulated mixtures of a target source and diffuse noise in $M = 4$. For the target source, we used four different directions: 0, 10, 20, and 30 degrees clockwise from the normal to the microphone array. As the target source signals, we used six speech signals from JNAS [19]. For diffuse noise, we used four types of noise: babble, cafe, station, and traffic. We simulated diffuse noise by different signals arriving from 19 directions. The babble noise was prepared from the speech of 19 speakers in JNAS. For cafe, station, and traffic noises, we obtained noise signals from the DEMAND [20] dataset and split them into 19 signals. Each signal was convoluted with the impulse response shown in Fig. 2. The signal length was 8.8 s and the sampling frequency was 16 kHz. The input SNR was set to 0 dB. For STFT, a 64-ms-long Hamming window was used, and the frameshift was 32 ms. The source-to-distortion ratio (SDR) [21] improvement was used to evaluate the BSE performance.

The compared methods were RC-FastMNMF [7], OverILRMA [15], RCSCME [16], proposed NoisyILRMA, and proposed NoisyILRMA with switching. In RCSCME, we attempted 20 and 50 iterations for the preprocessing ILRMA, which are labeled "ILRMA 20" and "ILRMA 50", respectively. For all the methods using the NMF model, the numbers of NMF bases for the target sound and noise were set to three. All NMF variables were initialized by uniform

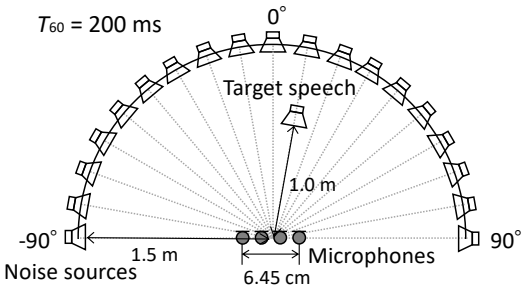

Fig. 2: Recording conditions of impulse responses.

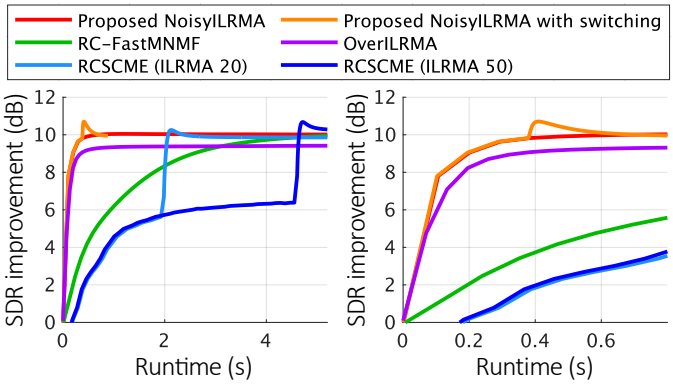

Fig. 3: (a) SDR improvement behaviors with respect to runtime averaged over 960 trials. (b) Enlarged view of (a).

random numbers on (0,1) intervals. $\mathbf{W}_i$ was initialized as $\boldsymbol{w}_{im} = \boldsymbol{u}_{im}/\sqrt{d_{im}}$, where $d_{im} > 0$ and $\boldsymbol{u}_{im} \in \mathbb{C}^M$ are the eigenvalue and eigenvector of $\sum_j \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathsf{H}}/J$, respectively, and d_{i1} is the largest eigenvalue. $\tilde{\mathbf{W}}_i$ was initialized in the same manner as $\mathbf{W}_i$. $\lambda_{im\tilde{n}}$, $\lambda_i^{(\mathrm{n})}$, and μ_i were initialized as one. In RCSCME, $\boldsymbol{b}_i$ was fixed to $\boldsymbol{a}_i'$ to achieve good performance. In OverILRMA and NoisyILRMA, the proportion between the number of $\mathbf{W}_i$ updates and the others was experimentally determined to be 1:10. In NoisyILRMA with switching, the number of iterations before switching was experimentally determined to be four. All methods were implemented in MATLAB (R2022a) and the calculation was performed on Intel Core i9-12900K.

B. Experimental results

To compare the methods in terms of performance and optimization speed, SDR improvement versus runtime is measured (Fig. 3). The runtime and SDR improvement were averaged over 960 trials (10 random initializations, four target speech directions, six speech signals, and four noise types). Fig. 3 shows that the proposed NoisyILRMA was the fastest among all the methods while maintaining the convergence performance of RC-FastMNMF. In addition, NoisyILRMA achieved approximately 1 dB higher performance than OverILRMA. This may be due to the time-variant noise modeling in NoisyLRMA. Fig. 3 also shows that RCSCME achieves good maximum performance with 50 ILRMA iterations and that NoisyILRMA with switching achieved comparable performance about 10 times faster.

V. Conclusion

In this paper, we proposed NoisyILRMA, a diffuse-noise-aware ILRMA, which can be optimized with the same fast algorithm as proposed in FIVE. We also proposed a switching mechanism of NoisyILRMA to further improve the BSE performance. The experimental result showed that NoisyILRMA ran faster than the conventional methods while maintaining the BSE performance. We also confirmed that the switching mechanism improves the BSE performance of NoisyILRMA to become comparable to that of RCSCME at an approximately 10 times faster speed.

References

[1] 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.

[2] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[3] 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.

[4] 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.

[5] 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.

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

[7] K. Sekiguchi *et al.*, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM Trans. ASLP*, vol. 28, pp. 2610–2625, 2020.

[8] N. Ito *et al.*, "A joint diagonalization based efficient approach to underdetermined blind audio source separation using the multichannel Wiener filter," *IEEE/ACM Trans. ASLP*, vol. 29, pp. 1950–1965, 2021.

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

[10] 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.

[11] Z. Koldovský and P. Tichavský, "Gradient algorithms for complex non-Gaussian independent component/vector extraction, question of convergence," *IEEE Trans. SP*, vol. 67, no. 4, pp. 1050–1064, 2019.

[12] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *Proc. WASPAA*, 2019, pp. 185–189.

[13] R. Ikeshita *et al.*, "Block coordinate descent algorithms for auxiliary-function-based independent vector extraction," *IEEE Trans. SP*, vol. 69, pp. 3252–3267, 2021.

[14] R. Scheibler and N. Ono, "Fast independent vector extraction by iterative SINR maximization," in *Proc. ICASSP*, 2020, pp. 601–605.

[15] M. Togami and R. Scheibler, "Over-determined speech source separation and dereverberation," in *Proc. APSIPA*, 2020, pp. 705–710.

[16] Y. Kubo *et al.*, "Blind speech extraction based on rank-constrained spatial covariance matrix estimation with multivariate generalized Gaussian distribution," *IEEE/ACM Trans. ASLP*, vol. 28, pp. 1948–1963, 2020.

[17] D. R. Hunter and K. Lange, "Quantile regression via an MM algorithm," *J. Comput. Graph. Stat.*, vol. 9, no. 1, pp. 60–77, 2000.

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

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

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

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