DEFICIENT BASIS ESTIMATION OF NOISE SPATIAL COVARIANCE MATRIX FOR RANK-CONSTRAINED SPATIAL COVARIANCE MATRIX ESTIMATION METHOD IN BLIND SPEECH EXTRACTION

Yuto Kondo[1], Yuki Kubo[1], Norihiro Takamune[1], Daichi Kitamura[2], Hiroshi Saruwatari[1]

[1]The University of Tokyo, Tokyo, Japan
[2]National Institute of Technology, Kagawa College, Kagawa, Japan

ABSTRACT

Rank-constrained spatial covariance matrix estimation (RCSCME) is a state-of-the-art blind speech extraction method applied to cases where one directional target speech and diffuse noise are mixed. In this paper, we proposed a new algorithmic extension of RCSCME. RCSCME complements a deficient one rank of the diffuse noise spatial covariance matrix, which cannot be estimated via preprocessing such as independent low-rank matrix analysis, and estimates the source model parameters simultaneously. In the conventional RC-SCME, a direction of the deficient basis is fixed in advance and only the scale is estimated; however, the candidate of this deficient basis is not unique in general. In the proposed RCSCM model, the deficient basis itself can be accurately estimated as a vector variable by solving a vector optimization problem. Also, we derive new update rules based on the EM algorithm. We confirm that the proposed method outperforms conventional methods under several noise conditions.

Index Terms — Blind speech extraction, diffuse noise, spatial covariance matrix, EM algorithm

1. INTRODUCTION

Blind speech extraction (BSE) is a technique for extracting a target speech signal from observed noisy mixture signals without any prior information, e.g., spatial locations of speech and noise sources and microphones. BSE can be interpreted as a special case of blind source separation (BSS) [1]; BSS aims to separate not only the target source but also the other sources. In this paper, we focus on the BSE problem for the observed noisy mixture that includes one directional target speech and diffuse background noise. Such BSE can be utilized for many applications including automatic speech recognition and hearing aid systems [2, 3].

In a determined or overdetermined situation (number of microphones $\geq$ number of sources), independent vector analysis [4, 5] and independent low-rank matrix analysis (ILRMA) [6, 7, 8, 9] provide better BSS performance. These methods assume that the frequency-wise acoustic path of each source can be modeled by a single time-invariant spatial basis, which is the so-called *steering vector*. In this model, the rank of a spatial covariance matrix (SCM) [10] becomes unity in all frequencies. Thus, hereafter, we call these BSS techniques *rank-1 methods*. Under diffuse noise conditions, rank-1 methods cannot separate directional sources in principle [2], and the estimated directional sources are always contaminated with a diffuse noise component remaining in the same direction. This is because the sources modeled by the rank-1 SCM (i.e., the steering vectors) cannot represent such spatially distributed noise components.

In contrast to rank-1 methods, multichannel nonnegative matrix factorization (MNMF) [11, 12] can represent the spatially spread sources and diffuse noise because MNMF utilizes a full-rank SCM for each source. However, the estimation of the full-rank SCM has a huge computational cost and lacks robustness against the parameter initialization [6]. FastMNMF [13, 14] is a model-constrained version of original MNMF and achieves an efficient SCM optimization with lower computational cost compared with that in MNMF, although its performance still depends on the initial values for the parameters.

To achieve fast and stable BSE under the diffuse noise condition, rank-constrained SCM estimation (RCSCME) [15, 16] was proposed, where the mixture of one directional speech and diffuse background noise is assumed. Fig. 1 illustrates a process flow of RC-SCME. In RCSCME, a rank-1 method such as ILRMA is utilized as a preprocess for BSE. From the rank-1 method, M estimated signals are obtained [2]; one includes target speech components contaminated with diffuse noise in the same direction and the other $M-1$ estimates consist of diffuse noise components in various directions, where M is the number of microphones. Then, the rank-$(M-1)$ SCM of the diffuse noise are calculated from the $M-1$ noise estimates. RCSCME estimates both the deficient rank of the noise SCM, which is used to compose the full-rank SCM for diffuse noise, and the source model parameters based on the expectation-maximization (EM) algorithm. The estimated target speech signal can be obtained via multichannel Wiener filtering using the full-rank SCM of diffuse noise. In [15], it was confirmed that RCSCME can outperform ILRMA, MNMF, and FastMNMF in terms of the speech extraction performance.

In this paper, we further improve the BSE performance by generalizing full-rank SCM estimation in conventional RCSCME. In the conventional method, the direction of the deficient basis in the rank-$(M-1)$ noise SCM is fixed to its eigenvector of the zero eigenvalue and only the scale of this basis is parameterized. However, the candidate of this deficient basis is not unique because any vector that is not included in the space spanned by column vectors of the rank-$(M-1)$ SCM can be used. In the proposed method, to estimate the optimal full-rank noise SCM, we parameterize the deficient vector itself and derive new update rules based on EM algorithm. Regarding its relation to prior works, the proposed RCSCME is interpreted as the world's first *spatial* model extension of the conventional RCSCME; this extension had been considered as a difficult *vector* optimization problem but this paper can successfully give the effective solution with a new mathematical claim. Also, since the advantage in accurate SCM estimation under the fixed rank-$(M-1)$ noise SCM is still

Fig. 1. Process flow of RCSCME.

valid, the proposed RCSCME always outperforms the conventional full-rank SCM methods like MNMF and FastMNMF.

2. CONVENTIONAL RCSCME [15]

2.1. Generative model

In this section, we explain the generative model in conventional RCSCME. Let $\boldsymbol{x}_{ij} = (x_{ij,1}, \cdots, x_{ij,m}, \cdots, x_{ij,M})^{\mathsf{T}}$ be the observed M-channel vector obtained by short-time Fourier transform (STFT), where $^{\mathsf{T}}$ denotes the transpose, and $i = 1, 2, \ldots, I, j = 1, 2, \ldots, J$, and , $m = 1, 2, \ldots, M$ are the indices of frequency bins, time frames, and channels (microphones), respectively. The observed signal is a mixture of the directional target speech and diffuse noise as

$$\boldsymbol{x}_{ij} = \boldsymbol{z}_{ij}^{(\mathrm{t})} + \boldsymbol{z}_{ij}^{(\mathrm{n})}, \tag{1}$$

where $\boldsymbol{z}_{ij}^{(\mathrm{t})} \in \mathbb{C}^M$ and $\boldsymbol{z}_{ij}^{(\mathrm{n})} \in \mathbb{C}^M$ are the source images of the target source and the diffuse noise, respectively.

The directional target source $\boldsymbol{z}_{ij}^{(\mathrm{t})}$ can be modeled as

$$\boldsymbol{z}_{ij}^{(\mathrm{t})} = \boldsymbol{a}_i^{(\mathrm{t})} s_{ij}^{(\mathrm{t})}, \tag{2}$$

$$s_{ij}^{(\mathrm{t})}|r_{ij}^{(\mathrm{t})} \sim \mathcal{N}_c(0, r_{ij}^{(\mathrm{t})}), \tag{3}$$

where $\boldsymbol{a}_i^{(\mathrm{t})} \in \mathbb{C}^M$, $s_{ij}^{(\mathrm{t})}, r_{ij}^{(\mathrm{t})} > 0$ are the steering vector, dry source component, and the time-frequency variance of the directional target source, respectively, and $\mathcal{N}_c(0, r)$ is the zero-mean circularly symmetric complex Gaussian distribution with a variance r. Note that M steering vectors, $\boldsymbol{a}_{i,1}, \boldsymbol{a}_{i,2}, \ldots, \boldsymbol{a}_{i,M}$, are estimated in advance via the rank-1 method, and we define the n_{t}th steering vector $\boldsymbol{a}_{i,n_t}$ corresponds to $\boldsymbol{a}_i^{(\mathrm{t})}$. This channel selection of the target source can be achieved based on the kurtosis values of each estimated signal [15]. The model (2) assumes that the source image of directional target source has time-invariant acoustic paths represented by the single spatial basis, steering vector $\boldsymbol{a}_i^{(\mathrm{t})}$, and the rank of SCM for $\boldsymbol{z}_{ij}^{(\mathrm{t})}$ becomes unity. In addition, since the power spectrogram of speech signals has sparsity property, we assume $r_{ij}^{(\mathrm{t})} \sim \mathcal{IG}(\alpha, \beta)$ as a prior distribution, where $\mathcal{IG}(\alpha, \beta)$ is the inverse gamma distribution with the shape parameter $\alpha > 0$ and the scale parameter $\beta > 0$.

In contrast to (2), the source image of diffuse noise, $\boldsymbol{z}_{ij}^{(\mathrm{n})}$, should have a full-rank SCM. The generative model of such sources is defined as [10]

$$\boldsymbol{z}_{ij}^{(\mathrm{n})} \sim \mathcal{N}_c^{(\mathrm{multi})}(\boldsymbol{0}, r_{ij}^{(\mathrm{n})}\mathbf{R}_i^{(\mathrm{n})}), \tag{4}$$

where $\mathcal{N}_c^{(\mathrm{multi})}(\boldsymbol{0}, \boldsymbol{R})$ is the zero-mean multivariate circularly symmetric complex Gaussian distribution with a covariance matrix $\boldsymbol{R}$, and $r_{ij}^{(\mathrm{n})} > 0$ and $\mathbf{R}_i^{(\mathrm{n})} \in \mathbb{C}^{M \times M}$ are the time-frequency variance and the time-invariant SCM for diffuse noise, respectively. The full-rank SCM of diffuse noise is modeled as

$$\mathbf{R}_i^{(\mathrm{n})} = \mathbf{R}_i'^{(\mathrm{n})} + \lambda_i \boldsymbol{b}_i \boldsymbol{b}_i^{\mathsf{H}}, \tag{5}$$

where $\mathbf{R}_i'^{(\mathrm{n})} \in \mathbb{C}^{M \times M}$ is the rank-$(M-1)$ SCM estimated by the rank-1 method in advance and is defined as

$$\mathbf{R}_i'^{(\mathrm{n})} = \frac{1}{J} \sum_j \hat{\boldsymbol{y}}_{ij}^{(\mathrm{n})}(\hat{\boldsymbol{y}}_{ij}^{(\mathrm{n})})^{\mathsf{H}}, \tag{6}$$

$$\hat{\boldsymbol{y}}_{ij}^{(\mathrm{n})} = \mathbf{W}_i^{-1}(\boldsymbol{w}_{i,1}^{\mathsf{H}}\boldsymbol{x}_{ij}, \ldots, \boldsymbol{w}_{i,n_t-1}^{\mathsf{H}}\boldsymbol{x}_{ij},$$
$$0, \boldsymbol{w}_{i,n_t+1}^{\mathsf{H}}\boldsymbol{x}_{ij}, \ldots, \boldsymbol{w}_{i,M}^{\mathsf{H}}\boldsymbol{x}_{ij})^{\mathsf{T}}, \tag{7}$$

$\boldsymbol{w}_{i,m}$ is the demixing filter estimated by the rank-1 method such as ILRMA, $\hat{\boldsymbol{y}}_{ij}^{(\mathrm{n})}$ is the sum of diffuse noise components whose scales are fixed by the back projection technique [17], and $^{\mathsf{H}}$ denotes the Hermitian transpose. Also, $\boldsymbol{b}_i \in \mathbb{C}^M$ is the deficient basis for the rank-$(M-1)$ SCM $\mathbf{R}_i'^{(\mathrm{n})}$, which makes the SCM $\mathbf{R}_i^{(\mathrm{n})}$ full-rank. Note that $\boldsymbol{b}_i$ is defined as a unit vector and its scale is defined by $\lambda_i > 0$. In conventional RCSCME, the direction of $\boldsymbol{b}_i$ is fixed to the eigenvector of the zero eigenvalue in $\mathbf{R}_i'^{(\mathrm{n})}$ and only the scale λ_i is estimated.

2.2. Update rules based on EM algorithm

The parameters $\Theta_{\mathrm{c}} = \{r_{ij}^{(\mathrm{t})}, r_{ij}^{(\mathrm{n})}, \lambda_i\}$ can be optimized by maximum a posteriori (MAP) estimation based on the EM algorithm with the latent variables $s_{ij}^{(\mathrm{t})}$ and $\boldsymbol{z}_{ij}^{(\mathrm{n})}$. Details of the derivation are described in [15]. The update rules are obtained as follows: In the E-step,

$$\tilde{\mathbf{R}}_i^{(\mathrm{n})} = \mathbf{R}_i'^{(\mathrm{n})} + \tilde{\lambda}_i \boldsymbol{b}_i \boldsymbol{b}_i^{\mathsf{H}}, \tag{8}$$

$$\tilde{\mathbf{R}}_{ij}^{(\mathrm{x})} = \tilde{r}_{ij}^{(\mathrm{t})}\boldsymbol{a}_i^{(\mathrm{t})}(\boldsymbol{a}_i^{(\mathrm{t})})^{\mathsf{H}} + \tilde{r}_{ij}^{(\mathrm{n})}\tilde{\mathbf{R}}_i^{(\mathrm{n})}, \tag{9}$$

$$\hat{r}_{ij}^{(\mathrm{t})} = \tilde{r}_{ij}^{(\mathrm{t})} - (\tilde{r}_{ij}^{(\mathrm{t})})^2(\boldsymbol{a}_i^{(\mathrm{t})})^{\mathsf{H}}(\tilde{\mathbf{R}}_{ij}^{(\mathrm{x})})^{-1}\boldsymbol{a}_i^{(\mathrm{t})}$$
$$+ |\tilde{r}_{ij}^{(\mathrm{t})}\boldsymbol{x}_{ij}^{\mathsf{H}}(\tilde{\mathbf{R}}_{ij}^{(\mathrm{x})})^{-1}\boldsymbol{a}_i^{(\mathrm{t})}|^2, \tag{10}$$

$$\hat{\mathbf{R}}_{ij}^{(\mathrm{n})} = \tilde{r}_{ij}^{(\mathrm{n})}\tilde{\mathbf{R}}_i^{(\mathrm{n})} - (\tilde{r}_{ij}^{(\mathrm{n})})^2\tilde{\mathbf{R}}_i^{(\mathrm{n})}(\tilde{\mathbf{R}}_{ij}^{(\mathrm{x})})^{-1}\tilde{\mathbf{R}}_i^{(\mathrm{n})}$$
$$+ (\tilde{r}_{ij}^{(\mathrm{n})})^2\tilde{\mathbf{R}}_i^{(\mathrm{n})}(\tilde{\mathbf{R}}_{ij}^{(\mathrm{x})})^{-1}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathsf{H}}(\tilde{\mathbf{R}}_{ij}^{(\mathrm{x})})^{-1}\tilde{\mathbf{R}}_i^{(\mathrm{n})}, \tag{11}$$

and in the M-step,

$$r_{ij}^{(\mathrm{t})} \leftarrow \frac{\hat{r}_{ij}^{(\mathrm{t})} + \beta}{\alpha + 2}, \tag{12}$$

$$\lambda_i \leftarrow \frac{1}{J} \sum_j \frac{1}{\tilde{r}_{ij}^{(\mathrm{n})}|\boldsymbol{b}_i^{\mathsf{H}}\boldsymbol{u}_i|^2}\boldsymbol{u}_i^{\mathsf{H}}\hat{\mathbf{R}}_{ij}^{(\mathrm{n})}\boldsymbol{u}_i, \tag{13}$$

$$\mathbf{R}_i^{(\mathrm{n})} \leftarrow \mathbf{R}_i'^{(\mathrm{n})} + \lambda_i \boldsymbol{b}_i \boldsymbol{b}_i^{\mathsf{H}}, \tag{14}$$

$$r_{ij}^{(\mathrm{n})} \leftarrow \frac{1}{M}\mathrm{tr}((\mathbf{R}_i^{(\mathrm{n})})^{-1}\hat{\mathbf{R}}_{ij}^{(\mathrm{n})}), \tag{15}$$

where $\tilde{\Theta}_{\mathrm{c}} = \{\tilde{r}_{ij}^{(\mathrm{t})}, \tilde{r}_{ij}^{(\mathrm{n})}, \tilde{\lambda}_i\}$ is the set of up-to-date parameters and $\boldsymbol{u}_i$ is the eigenvector of $\mathbf{R}_i'^{(\mathrm{n})}$ that corresponds to the zero eigenvalue.

3. PROPOSED RCSCME

3.1. Motivation

In conventional RCSCME, the direction of deficient basis $\boldsymbol{b}_i$ is fixed and only its scale λ_i is estimated. However, the candidate of the deficient basis is not unique because any complex vector that is not included in the space spanned by column vectors of $\mathbf{R}_i^{\prime(\mathrm{n})}$ can be used. In this paper, we propose a new vector optimization algorithm that estimates the deficient basis $\boldsymbol{b}_i$ itself, and derive new update rules based on the EM algorithm. This method can be interpreted as a generalization of conventional RCSCME.

3.2. New formulation of full-rank SCM and derivation of update rules

To simultaneously parameterize the direction and the scale of the deficient basis in $\mathbf{R}_i^{\prime(\mathrm{n})}$, we model the full-rank noise SCM $\mathbf{R}_i^{(\mathrm{n})}$ as follows:

$$\mathbf{R}_i^{(\mathrm{n})} = \mathbf{R}_i^{\prime(\mathrm{n})} + \boldsymbol{c}_i \boldsymbol{c}_i^{\mathsf{H}}, \tag{16}$$

where $\boldsymbol{c}_i \in \mathbb{C}^M$ is the deficient basis vector in $\mathbf{R}_i^{\prime(\mathrm{n})}$. We model all the other variables in the same way as conventional RCSCME.

We estimate the parameters $\Theta_{\mathrm{p}} = \{r_{ij}^{(\mathrm{t})}, r_{ij}^{(\mathrm{n})}, \boldsymbol{c}_i\}$ by MAP estimation based on the EM algorithm with the latent variables $s_{ij}^{(\mathrm{t})}$ and $\boldsymbol{z}_{ij}^{(\mathrm{n})}$. A Q function is defined by the expectation of the complete-data log-likelihood w.r.t. $p(s_{ij}^{(\mathrm{t})}, \boldsymbol{z}_{ij}^{(\mathrm{n})} | \boldsymbol{x}_{ij}; \tilde{\Theta}_{\mathrm{p}})$ as

$$Q(\Theta_{\mathrm{p}}; \tilde{\Theta}_{\mathrm{p}}) = \sum_{i,j} \left[-(\alpha+2)\log r_{ij}^{(\mathrm{t})} - M\log r_{ij}^{(\mathrm{n})} - \log \det \mathbf{R}_i^{(\mathrm{n})} \right.$$
$$\left. - \frac{\hat{r}_{ij}^{(\mathrm{t})} + \beta}{r_{ij}^{(\mathrm{t})}} - \frac{\mathrm{tr}\big((\mathbf{R}_i^{(\mathrm{n})})^{-1} \hat{\mathbf{R}}_{ij}^{(\mathrm{n})}\big)}{r_{ij}^{(\mathrm{n})}} \right] + \mathrm{const.}, \tag{17}$$

where $\tilde{\Theta}_{\mathrm{p}} = \{\tilde{r}_{ij}^{(\mathrm{t})}, \tilde{r}_{ij}^{(\mathrm{n})}, \tilde{\boldsymbol{c}}_i\}$ is the set of up-to-date parameters and const. are the constant terms that do not depend on Θ_{p}.

In the E-step, $\hat{r}_{ij}^{(\mathrm{t})}$ and $\hat{\mathbf{R}}_{ij}^{(\mathrm{n})}$ are obtained in the same way as the conventional RCSCME as follows:

$$\tilde{\mathbf{R}}_i^{(\mathrm{n})} = \mathbf{R}_i^{\prime(\mathrm{n})} + \tilde{\boldsymbol{c}}_i \tilde{\boldsymbol{c}}_i^{\mathsf{H}}, \tag{18}$$

$$\tilde{\mathbf{R}}_{ij}^{(\mathrm{x})} = \tilde{r}_{ij}^{(\mathrm{t})} \boldsymbol{a}_i^{(\mathrm{t})} (\boldsymbol{a}_i^{(\mathrm{t})})^{\mathsf{H}} + \tilde{r}_{ij}^{(\mathrm{n})} \tilde{\mathbf{R}}_i^{(\mathrm{n})}, \tag{19}$$

$$\hat{r}_{ij}^{(\mathrm{t})} = \tilde{r}_{ij}^{(\mathrm{t})} - (\tilde{r}_{ij}^{(\mathrm{t})})^2 (\boldsymbol{a}_i^{(\mathrm{t})})^{\mathsf{H}} (\tilde{\mathbf{R}}_{ij}^{(\mathrm{x})})^{-1} \boldsymbol{a}_i^{(\mathrm{t})}$$
$$+ |\tilde{r}_{ij}^{(\mathrm{t})} \boldsymbol{x}_{ij}^{\mathsf{H}} (\tilde{\mathbf{R}}_{ij}^{(\mathrm{x})})^{-1} \boldsymbol{a}_i^{(\mathrm{t})}|^2, \tag{20}$$

$$\hat{\mathbf{R}}_{ij}^{(\mathrm{n})} = \tilde{r}_{ij}^{(\mathrm{n})} \tilde{\mathbf{R}}_i^{(\mathrm{n})} - (\tilde{r}_{ij}^{(\mathrm{n})})^2 \tilde{\mathbf{R}}_i^{(\mathrm{n})} (\tilde{\mathbf{R}}_{ij}^{(\mathrm{x})})^{-1} \tilde{\mathbf{R}}_i^{(\mathrm{n})}$$
$$+ (\tilde{r}_{ij}^{(\mathrm{n})})^2 \tilde{\mathbf{R}}_i^{(\mathrm{n})} (\tilde{\mathbf{R}}_{ij}^{(\mathrm{x})})^{-1} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathsf{H}} (\tilde{\mathbf{R}}_{ij}^{(\mathrm{x})})^{-1} \tilde{\mathbf{R}}_i^{(\mathrm{n})}. \tag{21}$$

Among some update rules in the M-step, we especially describe the derivation of the update rule of $\boldsymbol{c}_i$. First, we differentiate the Q function with respect to $\boldsymbol{c}_i^*$ as follows:

$$\frac{\partial Q}{\partial \boldsymbol{c}_i^*} = -J(\mathbf{R}_i^{(\mathrm{n})})^{-1} \boldsymbol{c}_i + J(\mathbf{R}_i^{(\mathrm{n})})^{-1} \hat{\mathbf{T}}_i (\mathbf{R}_i^{(\mathrm{n})})^{-1} \boldsymbol{c}_i, \tag{22}$$

where * denotes the complex conjugate and $\hat{\mathbf{T}}_i$ is defined as follows:

$$\hat{\mathbf{T}}_i := \frac{1}{J} \sum_j \frac{1}{\tilde{r}_{ij}^{(\mathrm{n})}} \hat{\mathbf{R}}_{ij}^{(\mathrm{n})}. \tag{23}$$

From the equation $\partial Q / \partial \boldsymbol{c}_i^* = \boldsymbol{0}$, we obtain

$$\boldsymbol{c}_i = \hat{\mathbf{T}}_i (\mathbf{R}_i^{(\mathrm{n})})^{-1} \boldsymbol{c}_i. \tag{24}$$

It is diffult to directly solve (24) w.r.t. $\boldsymbol{c}_i$ since $\mathbf{R}_i^{(\mathrm{n})}$ itself includes $\boldsymbol{c}_i$. Hence, we use the following claim to resolve the equation.

Claim 1. *Let $\boldsymbol{u}_i \in \mathbb{C}^M$ be a vector that satisfies $\mathbf{R}_i^{\prime(\mathrm{n})} \boldsymbol{u}_i = \boldsymbol{0}$. Then, the following holds:*

$$(\mathbf{R}_i^{(\mathrm{n})})^{-1} \boldsymbol{c}_i = \frac{\boldsymbol{u}_i}{\boldsymbol{c}_i^{\mathsf{H}} \boldsymbol{u}_i}. \tag{25}$$

Proof. First, it holds that

$$\mathbf{R}_i^{(\mathrm{n})} \boldsymbol{u}_i = \mathbf{R}_i^{\prime(\mathrm{n})} \boldsymbol{u}_i + \boldsymbol{c}_i \boldsymbol{c}_i^{\mathsf{H}} \boldsymbol{u}_i$$
$$= (\boldsymbol{c}_i^{\mathsf{H}} \boldsymbol{u}_i) \boldsymbol{c}_i. \tag{26}$$

Then, by multiplying $(\mathbf{R}_i^{(\mathrm{n})})^{-1} / (\boldsymbol{c}_i^{\mathsf{H}} \boldsymbol{u}_i)$, we can obtain

$$(\mathbf{R}_i^{(\mathrm{n})})^{-1} \boldsymbol{c}_i = \frac{\boldsymbol{u}_i}{\boldsymbol{c}_i^{\mathsf{H}} \boldsymbol{u}_i}. \tag{27}$$

$\square$

By applying Claim 1 to (24), the equation can be deformed as

$$(\boldsymbol{c}_i^{\mathsf{H}} \boldsymbol{u}_i) \boldsymbol{c}_i = \hat{\mathbf{T}}_i \boldsymbol{u}_i. \tag{28}$$

By multiplying $\boldsymbol{u}_i^{\mathsf{H}}$, we can obtain

$$|\boldsymbol{c}_i^{\mathsf{H}} \boldsymbol{u}_i|^2 = \boldsymbol{u}_i^{\mathsf{H}} \hat{\mathbf{T}}_i \boldsymbol{u}_i, \tag{29}$$

and consequently $\boldsymbol{c}_i^{\mathsf{H}} \boldsymbol{u}_i$ can be represented as

$$\boldsymbol{c}_i^{\mathsf{H}} \boldsymbol{u}_i = \exp(-\mathrm{j}\phi) \sqrt{\boldsymbol{u}_i^{\mathsf{H}} \hat{\mathbf{T}}_i \boldsymbol{u}_i}, \tag{30}$$

where $\phi \in [0, 2\pi)$ is an arbitrary constant and j is an imaginary unit. Hence, the solution of the equation $\partial Q / \partial \boldsymbol{c}_i^* = \boldsymbol{0}$ is

$$\boldsymbol{c}_i = \frac{\exp(\mathrm{j}\phi)}{\sqrt{\boldsymbol{u}_i^{\mathsf{H}} \hat{\mathbf{T}}_i \boldsymbol{u}_i}} \hat{\mathbf{T}}_i \boldsymbol{u}_i. \tag{31}$$

We use $\phi = 0$ in our update rules since we model $\mathbf{R}_i^{(\mathrm{n})}$ by $\boldsymbol{c}_i \boldsymbol{c}_i^{\mathsf{H}}$, which does not depend on ϕ. We obtain the update rules of all other parameters than $\boldsymbol{c}_i$ in the same way as the conventional RCSCME. Finally, in the M-step, we update the parameters as follows:

$$r_{ij}^{(\mathrm{t})} \leftarrow \frac{\hat{r}_{ij}^{(\mathrm{t})} + \beta}{\alpha + 2}, \tag{32}$$

$$\boldsymbol{c}_i \leftarrow \frac{\hat{\mathbf{T}}_i \boldsymbol{u}_i}{\sqrt{\boldsymbol{u}_i^{\mathsf{H}} \hat{\mathbf{T}}_i \boldsymbol{u}_i}}, \tag{33}$$

$$\mathbf{R}_i^{(\mathrm{n})} \leftarrow \mathbf{R}_i^{\prime(\mathrm{n})} + \boldsymbol{c}_i \boldsymbol{c}_i^{\mathsf{H}}, \tag{34}$$

$$r_{ij}^{(\mathrm{n})} \leftarrow \frac{1}{M} \mathrm{tr}\big((\mathbf{R}_i^{(\mathrm{n})})^{-1} \hat{\mathbf{R}}_{ij}^{(\mathrm{n})}\big). \tag{35}$$

(33) can be interpreted as the *vector direction adaptation* by rotating $\boldsymbol{u}_i$ via $\hat{\mathbf{T}}_i$ composed of posterior.

Fig. 2. Recording conditions of impulse responses.

Fig. 3. Behavior of SDR improvement in babble noise case.

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

To confirm the efficacy of the proposed RCSCME, we conducted a BSE experiment using a simulated mixture of a target speech source and diffuse noise. To simulate the mixture, we convoluted dry sources with impulse responses from each position to four microphones as shown in Fig. 2. The diffuse noise was simulated by simultaneous reproduction from 19 positions and the target speech arrived from a closer position than each position of the diffuse noise. As the target speech source, we utilized six speech signals obtained from JNAS [18]. We imitated the babble, station, traffic, and cafe noises. We used 19 JNAS speech signals as the babble noise. The station, traffic, and cafe noise signals were obtained from DEMAND [19]. An STFT was performed by using a 64-ms-long Hamming window with a 32-ms-long shift. The speech-to-noise ratio was set to 0 dB.

We compared eight methods, namely, ILRMA [6], blind spatial subtraction array (BSSA) [2], multichannel Wiener filter with single-channel noise power estimation (MWF1) [20], multichannel Wiener filter with multichannel noise power estimation (MWF2) [21], the original FastMNMF [14], FastMNMF initialized by ILRMA (ILRMA +FastMNMF), the conventional RCSCME [15] and the proposed RCSCME. As for BSSA, ILRMA was used instead of frequency-domain independent component analysis in [2] and set the oversubtraction and flooring parameters to 1.4 and 0, respectively. For FastMNMF and ILRMA, the nonnegative matrix factorization (NMF) variables were initialized by nonnegative random values, and the demixing matrix was initialized by the identity matrix. As for ILRMA+FastMNMF, the NMF variables were handed over from ILRMA to FastMNMF. Also, the SCM was initialized by $\boldsymbol{a}_{i,n_t}\boldsymbol{a}_{i,n_t}^{\mathsf{H}} + \epsilon\mathbf{I}$ for ILRMA+MNMF and $\boldsymbol{a}_{i,n_t}\boldsymbol{a}_{i,n_t} + \epsilon\sum_{n\neq n_t}\boldsymbol{a}_{i,n}\boldsymbol{a}_{i,n}^{\mathsf{H}}$ for ILRMA+FastMNMF, where $\mathbf{I}$

Table 1. SDR improvements [dB] for each method and noise case. Each term represents "peak score / score after 200 iterations"

Method	babble noise	station noise	traffic noise	cafe noise
ILRMA	6.1 / -	6.2 / -	4.7 / -	6.4 / -
BSSA	6.8 / -	6.9 / -	5.7 / -	7.2 / -
MWF1	6.1 / -	6.9 / -	5.8 / -	7.0 / -
MWF2	6.9 / -	7.2 / -	5.6 / -	7.4 / -
FastMNMF	1.7 / 1.7	2.6 / 2.5	2.9 / 2.8	2.6 / 2.6
ILRMA +FastMNMF	6.5 / 6.2	6.6 / 6.6	5.4 / 5.4	7.3 / 7.3
Conventional RCSCME	8.4 / 7.5	9.5 / 8.9	7.5 / 7.2	9.6 / 8.9
Proposed RCSCME	**8.8 / 8.4**	**10.7 / 10.4**	**8.8 / 8.7**	**10.6 / 10.2**

is the identity matrix and ϵ was set to 10^{-5}. In ILRMA, which was used as the preprocessing for each method in the experiment, the number of bases was 10 and the number of iterations was 50. As for all methods, by selecting the channel whose kurtosis was the maximum in all channels, we blindly decided the index n_t of the target source in the demixed signals by ILRMA. In the conventional RCSCME, we utilized the minimum positive eigenvalue σ_i of $\mathbf{R}_i'^{(n)}$ as the initial value of λ_i. We used the inverse gamma distribution parameters $\alpha = 2.5$ and $\beta = 10^{-16}$ in the conventional RCSCME, which showed the best separation performance in the experiment of [16]. $\boldsymbol{c}_i$ was initialized by $\sqrt{\sigma_i}\boldsymbol{u}_i$ and we chose the parameters $\alpha = 0.1$ and $\beta = 10^{-16}$ in the proposed RCSCME experimentally. Source-to-distortion ratio (SDR) improvement [22] was used as a total evaluation score. The SDR improvement was averaged over 10 parameter-initialization random seeds, four target directions, and six target speech sources; thus each SDR improvement score is the average value over 240 trials.

4.2. Result

Figure 3 shows the behavior example of the averaged SDR improvement of the proposed and the conventional RCSCME for each iteration under a babble noise condition. Although the preprocessing, i.e., ILRMA, is an iterative method, we show the averaged SDR improvement of ILRMA after 50 iterations as a reference. From this figure, we can see that the SDR improvement of the conventional and proposed RCSCMEs have a peak and the proposed RCSCME outperforms the conventional RCSCME from the viewpoint of both the peak score and the score after 200 iterations.

Table 1 shows SDR improvements for each method under each noise condition. For FastMNMF, ILRMA+FastMNMF, conventional RCSCME, and proposed RCSCME, which are iterative methods, we display both the peak score and the score after 200 iterations. From this table, we reveal that the proposed RCSCME outperforms all the conventional methods under all the noise conditions.

5. CONCLUSION

In this paper, we proposed a new algorithmic extension of RCSCME to improve the BSE performance. In the conventional RCSCME, a direction of the deficient basis is fixed and only the scale is estimated. In the proposed RCSCM, we accurately estimated the deficient basis itself as a vector variable by solving a vector optimization problem. Also, we derived a new update rules based on the EM algorithm. We confirmed that the proposed method outperformed conventional methods under several noise conditions.

6. REFERENCES

[1] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: two converging routes to ILRMA originating from ICA and NMF," *APSIPA Trans. Signal Inf. Process.*, vol. 8, no. e12, pp. 1–14, 2019.

[2] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Trans. ASLP*, vol. 17, no. 4, pp. 650–664, 2009.

[3] Z. Koldovský, P. Tichavský, "Gradient algorithms for complex nongaussian independent component/vector extraction, question of convergence," *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 1050–1064, 2019.

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

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

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

[7] D. Kitamura, S. Mogami, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, Y. Takahashi, and K. Kondo, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation,," *EURASIP J. Adv. Signal Process.*, vol. 2018, no. 1, pp. 1–28, 2018.

[8] R. Ikeshita and Y.Kawaguchi, "Inpendent low-rank matrix analysis based on multivariate complex exponential power distribution," in *Proc. ICASSP*, 2018, pp. 741–745.

[9] S. Mogami, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, K. Kondo, and N. Ono, "Independent low-rank matrix analysis based on time-variant sub-gaussian source model for determined blind source separation," *IEEE/ACM Trans. ASLP*, vol. 28, pp. 503–518, 2019.

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

[11] 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.

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

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

[14] K. Sekiguchi, A. A. Nugraha, Y. Bando, and K. Yoshii, "Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices," in *Proc. EUSIPCO*, 2019, 5 pages.

[15] Y. Kubo, N. Takamune, D. Kitamura, and H. Saruwatari, "Efficient full-rank spatial covariance estimation using independent low-rank matrix analysis for blind source separation," in *Proc. EUSIPCO*, 2019, 5 pages.

[16] Y. Kubo, N. Takamune, D. Kitamura, and H. Saruwatari, "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, June 2020.

[17] 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.

[18] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *The Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[19] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," June 2013.

[20] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, 2001.

[21] R. Miyazaki, H. Saruwatari, R. Wakisaka, K. Shikano, and T. Takatani, "Theoretical analysis of parametric blind spatial subtraction array and its application to speech recognition performance prediction," in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays 2011*, 2011, pp. 19–24.

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