

Speech Enhancement by Noise Self-Supervised Rank-Constrained Spatial Covariance Matrix Estimation via Independent Deeply Learned Matrix Analysis

Sota Misawa*, Norihiro Takamune*, Tomohiko Nakamura*, Daichi Kitamura†, Hiroshi Saruwatari*, Masakazu Une‡ and Shoji Makino‡§

* The University of Tokyo, Tokyo, Japan

† National Institute of Technology, Kagawa College, Kagawa, Japan

‡ University of Tsukuba, Ibaraki, Japan

§ Waseda University, Fukuoka, Japan

Abstract—Rank-constrained spatial covariance matrix estimation (RCSCME) is a method for the situation that the directional target speech and the diffuse noise are mixed. In conventional RCSCME, independent low-rank matrix analysis (ILRMA) is used as the preprocessing method. We propose RCSCME using independent deeply learned matrix analysis (IDLMA), which is a supervised extension of ILRMA. In this method, IDLMA requires deep neural networks (DNNs) to separate the target speech and the noise. We use Denoiser, which is a single-channel speech enhancement DNN, in IDLMA to estimate not only the target speech but also the noise. We also propose noise self-supervised RCSCME, in which we estimate the noise-only time intervals using the output of Denoiser and design the prior distribution of the noise spatial covariance matrix for RCSCME. We confirm that the proposed methods outperform the conventional methods under several noise conditions.

I. INTRODUCTION

In this study, we deal with speech enhancement. In particular, we deal with the problem of extracting directional target speech from diffuse noise. This situation arises, for example, when the speaker is close to the microphones in a noisy environment such as a crowded place or a train station.

From a practical perspective, it is effective to use blind source separation (BSS) [1], which does not require any spatial information or characteristics of each source. In a determined or overdetermined situation (number of microphones $M \geq$ number of sources N), independent vector analysis (IVA) [2–4] and independent low-rank matrix analysis (ILRMA) [5] have been proposed for BSS. ILRMA assumes that the power spectrogram for each source is modeled by nonnegative matrix factorization (NMF) [6]. Since these methods assume that the rank of the spatial covariance matrix (SCM) of each source is unity, we call these methods *rank-1 methods*. In a rank-1 method, the noise remains in the separated target signal when the noise has a full-rank SCM, such as diffuse noise [7–9].

Multichannel NMF (MNMF) [10, 11] and FastMNMF [12, 13] have been proposed as methods to model full-rank SCMs for each source. MNMF adopts the full-rank nature for the

source model [14] and can handle diffuse sources, but it is computationally expensive and highly dependent on initial values [5]. In FastMNMF, the computational complexity is reduced by assuming that the SCMs are jointly diagonalizable. However, the problem that the performance is not robust to changes in the initial value still remains.

Rank-constrained spatial covariance matrix estimation (RCSCME) [15, 16] has been proposed as a method of blind speech enhancement for mixed signals of directional target speech and diffuse noise. RCSCME uses a rank-1 method such as ILRMA in advance. Since a rank-1 method accurately obtains $M - 1$ noise signals, we can obtain the rank- $(M - 1)$ component of the noise SCM from these signals. RCSCME estimates the time-varying variance of the target speech and the diffuse noise, and complements the deficient rank-1 component of the noise SCM at the same time. As shown in [15], this method is robust and converges rapidly because there are fewer estimated parameters than in MNMF and FastMNMF.

For the case that training data are available, many supervised source separation methods using a deep neural network (DNN) have been proposed. For a determined or overdetermined situation, independent deeply learned matrix analysis (IDLMA) [17] has been proposed to improve source separation performance by changing the source model of ILRMA from NMF to DNNs. Owing to the appropriately trained DNNs, IDLMA achieves higher performance than ILRMA. Since IDLMA is a rank-1 method similarly to ILRMA, the noise remains in the separated target signal in a situation with diffuse noise.

In this paper, we propose the use of IDLMA, which is a rank-1 method, as the preprocessing method of RCSCME. Since IDLMA was originally proposed for musical source separation, we should prepare DNNs that estimate the power spectrogram of speech or noise. One of the DNN-based single-channel speech enhancement methods is Denoiser [18], which estimates the target speech from a mixed-signal input in the waveform domain. Whereas IDLMA requires DNNs for each source, Denoiser extracts only the target speech. Thus, we

also propose a scheme to estimate the power spectrogram of the noise using Denoiser. Furthermore, considering the high capability of Denoiser to reduce noise, we can find the noise-only time intervals of the observed signal using the output of Denoiser. The SCM of this noise-only signal is a good approximation of the noise SCM. Thus, we design the prior distribution of the noise using this approximated noise SCM for RCSCME, and we call this method *noise self-supervised RCSCME*. We derive the new update rule on the basis of the EM algorithm for noise self-supervised RCSCME. We conduct a simulated experiment to verify the effectiveness of the proposed methods.

II. CONVENTIONAL METHODS

A. ILRMA and IDLMA

Let $\mathbf{s}_{ij} = (s_{ij,1}, \dots, s_{ij,N})^\top$, $\mathbf{x}_{ij} = (x_{ij,1}, \dots, x_{ij,M})^\top$, and $\mathbf{y}_{ij} = (y_{ij,1}, \dots, y_{ij,N})^\top$ denote the short-time Fourier transform (STFT) of the source, observed, and separated signals, respectively, where $i = 1, \dots, I$ and $j = 1, \dots, J$ are the indices of the frequency bin and time frame, respectively. Here, \top denotes transpose. We also represent the spectrograms for each signal as $S_n, X_m, Y_n \in \mathbb{C}^{I \times J}$ whose (i, j) th elements are $s_{ij,n}, x_{ij,m},$ and $y_{ij,n}$, where $n = 1, \dots, N$ and $m = 1, \dots, M$ are the source and channel indices, respectively.

When the reverberation time is sufficiently shorter than the window length of the STFT and each source is a point source, we can denote the observed signal as

$$\mathbf{x}_{ij} = A_i \mathbf{s}_{ij}, \quad (1)$$

where $A_i = (\mathbf{a}_{i,1}, \dots, \mathbf{a}_{i,N}) \in \mathbb{C}^{M \times N}$ is a mixing matrix. Assuming that $N = M$ and A_i is invertible, we can obtain the separated signal by estimating its inverse matrix $W_i = (\mathbf{w}_{i,1}, \dots, \mathbf{w}_{i,N})^\text{H} \in \mathbb{C}^{N \times M}$, which is called the demixing matrix, such that

$$\mathbf{y}_{ij} = W_i \mathbf{x}_{ij}, \quad (2)$$

where H denotes the Hermitian transpose.

In ILRMA [5] and IDLMA [17], it is assumed that the separated signals are mutually independent and that the n th separated signal in each time-frequency frame is generated by the univariate complex Gaussian distribution as

$$y_{ij,n} \sim \mathcal{N}_c(0, \sigma_{ij,n}^2), \quad (3)$$

where 0 and $\sigma_{ij,n}^2$ are the mean and variance, respectively. We use the negative log-likelihood of the observed signal as the cost function:

$$\begin{aligned} \mathcal{L}_1 &= -\log \prod_{i,j,n} p(x_{ij,n}) \\ &= -\log \prod_{i,j,n} p(y_{ij,n}) - 2J \sum_i \log |\det W_i| \\ &= \sum_{i,j,n} \left(\frac{|w_{in}^\text{H} x_{ij}|^2}{\sigma_{ij,n}^2} + 2 \log \sigma_{ij,n} \right) \\ &\quad - 2J \sum_i \log |\det W_i| + \text{const.}, \end{aligned} \quad (4)$$

where const. includes the terms independent of the target variables. In ILRMA, $\sigma_{ij,n}^2$ is modeled by NMF [6] as

$$\sigma_{ij,n}^2 = \sum_k t_{ik,n} v_{kj,n}, \quad (5)$$

where $t_{ik,n} \geq 0$ and $v_{kj,n} \geq 0$ are the NMF variables. Instead, in IDLMA, we use a source separation DNN that estimates the power spectrogram:

$$\sigma_{ij,n}^2 = \max \left\{ \left| (\text{DNN}_n(Y_n))_{ij} \right|^2, \varepsilon \right\}, \quad (6)$$

where $\varepsilon > 0$ is a small number used to avoid numerical instability. Here, DNN_n is the trained DNN corresponding to the n th source and its input is the current separated signal.

Since diffuse noise is not a point source, ILRMA and IDLMA cannot separate noise in the same direction as the directional target speech in principle [7–9]. Therefore, we cannot directly apply ILRMA and IDLMA to diffuse noise reduction, which often arises in practical speech enhancement application.

B. RCSCME

RCSCME [15, 16] deals with the problem of speech enhancement when directional target speech and diffuse noise are mixed. In [15], it is assumed that the generative model of the observed signal is the multivariate complex Gaussian distribution

$$p(x_{ij}; 0, R_{ij}^{(o)}) \propto \frac{\exp(-x_{ij}^\text{H} R_{ij}^{(o)} x_{ij})}{\det R_{ij}^{(o)}}, \quad (7)$$

where $R_{ij}^{(o)}$ denotes the covariance matrix of the observed signal. $R_{ij}^{(o)}$ is modeled as

$$R_{ij}^{(o)} = r_{ij}^{(t)} R_i^{(t)} + r_{ij}^{(n)} R_i^{(n)}, \quad (8)$$

where $r_{ij}^{(t)} > 0$ and $r_{ij}^{(n)} > 0$ are the variances of the target speech and the noise, and $R_i^{(t)}$ and $R_i^{(n)}$ are the SCMs of the target speech and the noise, respectively. To induce the sparsity of the target speech, we assume the inverse gamma distribution for the prior distribution on $r_{ij}^{(t)}$,

$$p(r_{ij}^{(t)}; \alpha, \beta) \propto \left(r_{ij}^{(t)} \right)^{-\alpha-1} \exp\left(-\frac{\beta}{r_{ij}^{(t)}}\right), \quad (9)$$

where $\alpha > 0$ and $\beta > 0$ are the shape parameter and scale parameter, respectively. Using the estimated mixing matrix $W_i^{-1} = A_i$, we model the SCMs of the target source and noise as

$$R_i^{(t)} = \mathbf{a}_{i,n_t} \mathbf{a}_{i,n_t}^\text{H}, \quad (10)$$

$$R_i^{(n)} = R_i^{(n)} + \lambda_i \mathbf{v}_i \mathbf{v}_i^\text{H}, \quad (11)$$

$$R_i^{(n)} = \frac{1}{J} \sum_j \hat{y}_{ij}^{(n)} \left(\hat{y}_{ij}^{(n)} \right)^\text{H}, \quad (12)$$

$$\hat{y}_{ij}^{(n)} = A_i \left(w_{i,1}^\text{H} x_{ij}, \dots, w_{i,n_t-1}^\text{H} x_{ij}, 0, \right. \\ \left. w_{i,n_t+1}^\text{H} x_{ij}, \dots, w_{i,N}^\text{H} x_{ij} \right)^\top, \quad (13)$$

where n_t denotes the target index and $R_i^{(n)} \in \mathbb{C}^{M \times M}$ is the rank- $(M-1)$ SCM of the diffuse noise estimated by a rank-1 method. Since the rank of $R_i^{(n)}$ is $M-1$, \mathbf{v}_i must be linearly independent of any column vectors of $R_i^{(n)}$ in order that $R_i^{(n)}$ has full rank. For example, the unit eigenvector corresponding to the zero eigenvalue of $R_i^{(n)}$ satisfies this condition for \mathbf{v}_i . RCSCME simultaneously estimates λ_i , $r_{ij}^{(t)}$, and $r_{ij}^{(n)}$ by maximum a posteriori estimation. Using (7) and (9), we calculate the log-posterior of the observed signal as

$$\begin{aligned} \mathcal{L}_{\text{RC}}(\Theta_c) &= \log \prod_{i,j} p(\mathbf{x}_{ij} | \Theta_c) p(r_{ij}^{(t)}; \alpha, \beta) \\ &= - \sum_{i,j} \left(\mathbf{x}_{ij}^H R_{ij}^{(o)} \mathbf{x}_{ij} + \log \det R_{ij}^{(o)} \right. \\ &\quad \left. + (\alpha + 1) \log r_{ij}^{(t)} + \frac{\beta}{r_{ij}^{(t)}} \right) + \text{const.}, \quad (14) \end{aligned}$$

where $\Theta_c = \{\lambda_i, r_{ij}^{(t)}, r_{ij}^{(n)}\}$. We maximize the log-posterior function using the EM algorithm with the following Q function (see [15] for details):

$$\begin{aligned} Q_{\text{RC}}(\Theta_c; \tilde{\Theta}_c) &= - \sum_{i,j} \left((\alpha + 2) \log r_{ij}^{(t)} + \frac{\hat{r}_{ij}^{(t)} + \beta}{r_{ij}^{(t)}} \right. \\ &\quad \left. + M \log r_{ij}^{(n)} + \log \det R_i^{(n)} \right. \\ &\quad \left. + \frac{\text{tr} \left(\hat{R}_{ij}^{(n)} \left(R_i^{(n)} \right)^{-1} \right)}{r_{ij}^{(n)}} \right) + \text{const.}, \quad (15) \end{aligned}$$

where $\tilde{\Theta}_c = \{\tilde{\lambda}_i, \tilde{r}_{ij}^{(t)}, \tilde{r}_{ij}^{(n)}\}$ is the set of the up-to-date parameters, and $\hat{r}_{ij}^{(t)}$ and $\hat{R}_{ij}^{(n)}$ are calculated in the E-step as follows:

$$\tilde{R}_i^{(n)} = R_i^{(n)} + \tilde{\lambda}_i \mathbf{v}_i \mathbf{v}_i^H, \quad (16)$$

$$R_{ij}^{(o)} = \tilde{r}_{ij}^{(t)} R_i^{(t)} + \tilde{r}_{ij}^{(n)} \tilde{R}_i^{(n)}, \quad (17)$$

$$\begin{aligned} \hat{r}_{ij}^{(t)} &= \tilde{r}_{ij}^{(t)} - \left(\tilde{r}_{ij}^{(t)} \right)^2 \mathbf{a}_{i,n_t}^H \left(R_{ij}^{(o)} \right)^{-1} \mathbf{a}_{i,n_t} \\ &\quad + \left| \tilde{r}_{ij}^{(t)} \mathbf{x}_{ij}^H \left(R_{ij}^{(o)} \right)^{-1} \mathbf{a}_{i,n_t} \right|^2, \quad (18) \end{aligned}$$

$$\begin{aligned} \hat{R}_{ij}^{(n)} &= \tilde{r}_{ij}^{(n)} \tilde{R}_i^{(n)} - \left(\tilde{r}_{ij}^{(t)} \right)^2 \tilde{R}_i^{(n)} \left(R_{ij}^{(o)} \right)^{-1} \tilde{R}_i^{(n)} \\ &\quad + \left(\tilde{r}_{ij}^{(t)} \right)^2 \tilde{R}_i^{(n)} \left(R_{ij}^{(o)} \right)^{-1} \mathbf{x}_{ij} \mathbf{x}_{ij}^H \left(R_{ij}^{(o)} \right)^{-1} \tilde{R}_i^{(n)}. \quad (19) \end{aligned}$$

In the M-step, by finding the stationary point of Q_{RC} , we obtain the update rules as

$$r_{ij}^{(t)} \leftarrow \frac{\hat{r}_{ij}^{(t)} + \beta}{\alpha + 2}, \quad (20)$$

$$\lambda_i \leftarrow \mathbf{u}_i^H \left(\frac{1}{J} \sum_j \frac{1}{\tilde{r}_{ij}^{(n)}} \hat{R}_{ij}^{(n)} \right) \mathbf{u}_i, \quad (21)$$

$$R_i^{(n)} \leftarrow R_i^{(n)} + \lambda_i \mathbf{v}_i \mathbf{v}_i^H, \quad (22)$$

$$r_{ij}^{(n)} \leftarrow \frac{1}{M} \text{tr} \left(\hat{R}_{ij}^{(n)} \left(R_i^{(n)} \right)^{-1} \right), \quad (23)$$

where \mathbf{u}_i is the eigenvector corresponding to the zero eigenvalues of $R_i^{(n)}$ satisfying $\mathbf{u}_i^H \mathbf{v}_i = 1$. Here, we use the following property for $R_i^{(n)} = R_i^{(n)} + \lambda_i \mathbf{v}_i \mathbf{v}_i^H$, as proved in [16], in obtaining the update rule of λ_i :

$$\begin{aligned} \log \det R_i^{(n)} &= \log \lambda_i + \text{const.}, \quad (24) \\ \left(R_i^{(n)} \right)^{-1} &= (E - \mathbf{u}_i \mathbf{v}_i^H) \left(R_i^{(n)} \right)^+ (E - \mathbf{v}_i \mathbf{u}_i^H) + \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^H, \quad (25) \end{aligned}$$

where $E \in \mathbb{C}^{M \times M}$ is the identity matrix and $+$ denotes the Moore–Penrose inverse.

Finally, we obtain the extracted target speech using the multichannel Wiener filter as

$$\hat{s}_{ij}^{(t)} = r_{ij}^{(t)} R_i^{(t)} \left(R_{ij}^{(o)} \right)^{-1} \mathbf{x}_{ij}. \quad (26)$$

III. PROPOSED METHOD

A. Motivation and Framework

In recent years, research on DNNs has become more popular and DNN-based source separation methods have been proposed [17–21]. IDLMA [17] has been proposed for determined audio source separation. IDLMA has higher separation performance than ILRMA by introducing the supervised DNNs into the source models. Since IDLMA is also a rank-1 method similarly to ILRMA, we propose the use of IDLMA instead of ILRMA as the preprocessing method for RCSCME. We expect that RCSCME will achieve high separation performance using highly accurate parameters pre-estimated by IDLMA. However, since the original IDLMA dealt with musical source separation, we should prepare new DNNs for the speech and the noise. For the single-channel speech enhancement DNN, Denoiser [18] has been proposed. Here, we propose a scheme to estimate the power spectrogram not only of the target speech, but also of the noise using Denoiser.

As shown in Table I, Denoiser tends to have very high accuracy in terms of the source-to-inference ratio (SIR) [22] improvement, which is a metric of denoising, but poor results in terms of the sources-to-artifact ratio (SAR) [22], which is a metric of artifacts. In other words, Denoiser is very good at removing noise, but it also tends to partially remove the target speech. Considering the high SIR of Denoiser, we can extract the noise-only time intervals from the observed signal corresponding to the quiet time intervals of the output of Denoiser and approximate the noise SCM from these intervals.

TABLE I
SIR IMPROVEMENT AND SAR OF DENOISER

	Noise			
	Babble	Cafe	Station	Traffic
SIR improvement [dB]	17.45	19.23	18.33	20.64
SAR [dB]	8.89	9.94	7.54	10.83

Thus, we propose noise self-supervised RCSCME, in which the prior distribution of the noise SCM is introduced using the approximated noise SCM.

Finally, an overview of the proposed method is as follows:

- 1) Prepare the trained Denoiser.
- 2) Estimate the demixing matrix W_i by IDLMA using Denoiser.
- 3) Calculate the quiet time intervals of the output using Denoiser, and set the corresponding time intervals of the observed signal as the noise-only signal.
- 4) Determine the prior distribution of the noise from the noise-only signal calculated in 3) and perform RCSCME with the result of 2).
- 5) Construct the multichannel Wiener filter with the result of 4).

The process flow of the proposed method is shown in Fig. 1.

B. IDLMA for Speech Enhancement

In the separated signals Y_n obtained by IDLMA, the n_t th signal must be the target speech, and the other channels must be noise signals. Therefore, for the variance of the n_t th separated signal, we use the output of the DNN, and for the other channels, we subtract the outputs of the DNN from the separated signals to simulate the estimation of the noise signal:

$$\zeta_{ij,n} = \begin{cases} (\text{DNN}(Y_n))_{ij} & \text{if } n = n_t, \\ y_{ij,n} - (\text{DNN}(Y_n))_{ij} & \text{otherwise,} \end{cases} \quad (27)$$

$$\sigma_{ij,n}^2 = \max \left\{ |\zeta_{ij,n}|^2, \varepsilon \right\}. \quad (28)$$

The reason for using the DNN for each channel is that each separated signal and each output of the DNN need to be in phase when we obtain the noise signal by subtraction.

C. Self-Estimation of Prior Distribution of Noise

We consider the silent parts of the output of Denoiser to be the noise-only parts in the observed signal. By thresholding the denoised signal by Denoiser, we extract the noise-only signal $\{\mathbf{x}_{ij'}\}_{j' \in \mathcal{J}'}$, where

$$\mathcal{J}' = \left\{ j \in \{1, \dots, J\} \mid \sqrt{\sum_i |(\text{DNN}(X_{n_t}))_{ij}|^2} < \theta \right\}. \quad (29)$$

Here, $\theta > 0$ is a threshold parameter. The SCM of the noise-only time intervals is calculated as

$$\begin{aligned} \check{R}_i^{(n)} &= \mathbb{E} [\mathbf{x}_{ij'} \mathbf{x}_{ij'}^H] \\ &\approx \frac{1}{\#\mathcal{J}'} \sum_{j'} \mathbf{x}_{ij'} \mathbf{x}_{ij'}^H, \end{aligned} \quad (30)$$

where $\#$ denotes the cardinality of a finite set.

One of the conjugate prior distributions for the covariance matrix in the multivariate complex Gaussian distribution is the complex inverse matrix gamma distribution (CIMGD) [23], which is an extension of the inverse gamma distribution to a positive definite Hermitian matrix. We assume that $R_i^{(n)}$ is generated by the M -dimensional CIMGD:

$$\begin{aligned} p\left(R_i^{(n)}; \alpha', \beta', \check{R}_i^{(n)}\right) \\ \propto \left(\det R_i^{(n)}\right)^{-(\alpha'+M)} \exp\left(-\frac{1}{\beta'} \text{tr}\left(\check{R}_i^{(n)} \left(R_i^{(n)}\right)^{-1}\right)\right), \end{aligned} \quad (31)$$

where $\alpha' > M - 1$, $\beta' > 0$, and $\check{R}_i^{(n)}$ are the shape parameter, the scale parameter, and the scale matrix, respectively. If $\alpha' = \tau$ and $\beta' = 1$ in (31), this distribution coincides with the complex inverse Wishart distribution with degree of freedom τ .

D. Log-Posterior with Prior Distribution of Noise

The log-posterior likelihood given the prior distributions for the variance of target speech and the noise SCM is

$$\begin{aligned} \mathcal{L}(\Theta) &= \log \prod_{i,j} p(\mathbf{x}_{ij} | \Theta) p(r_{ij}^{(t)}; \alpha, \beta) p\left(R_i^{(n)}; \alpha', \beta', \check{R}_i^{(n)}\right) \\ &= - \sum_{i,j} \left(\mathbf{x}_{ij}^H R_{ij}^{(o)} \mathbf{x}_{ij} + \log \det R_{ij}^{(o)} \right. \\ &\quad \left. + (\alpha + 1) \log r_{ij}^{(t)} + \frac{\beta}{r_{ij}^{(t)}} \right) \\ &\quad - \sum_i \left((\alpha' + M) \log \det R_i^{(n)} \right. \\ &\quad \left. + \frac{1}{\beta'} \text{tr}\left(\check{R}_i^{(n)} \left(R_i^{(n)}\right)^{-1}\right) \right) + \text{const.} \end{aligned} \quad (32)$$

We calculate the Q function in the same manner as in [15].

$$\begin{aligned} Q(\Theta; \tilde{\Theta}) &= - \sum_{i,j} \left((\alpha + 2) \log r_{ij}^{(t)} + \frac{\hat{r}_{ij}^{(t)} + \beta}{r_{ij}^{(t)}} \right. \\ &\quad \left. + M \log r_{ij}^{(n)} + \log \det R_i^{(n)} \right. \\ &\quad \left. + \frac{\text{tr}\left(\hat{R}_{ij}^{(n)} \left(R_i^{(n)}\right)^{-1}\right)}{r_{ij}^{(n)}} \right) \\ &\quad - \sum_i \left((\alpha' + M) \log \det R_i^{(n)} \right. \\ &\quad \left. + \frac{1}{\beta'} \text{tr}\left(\check{R}_i^{(n)} \left(R_i^{(n)}\right)^{-1}\right) \right) + \text{const.} \end{aligned} \quad (33)$$

It is easily seen that we only need to change the update rule of λ_i in (21). Using (24) and (25), we focus on the term

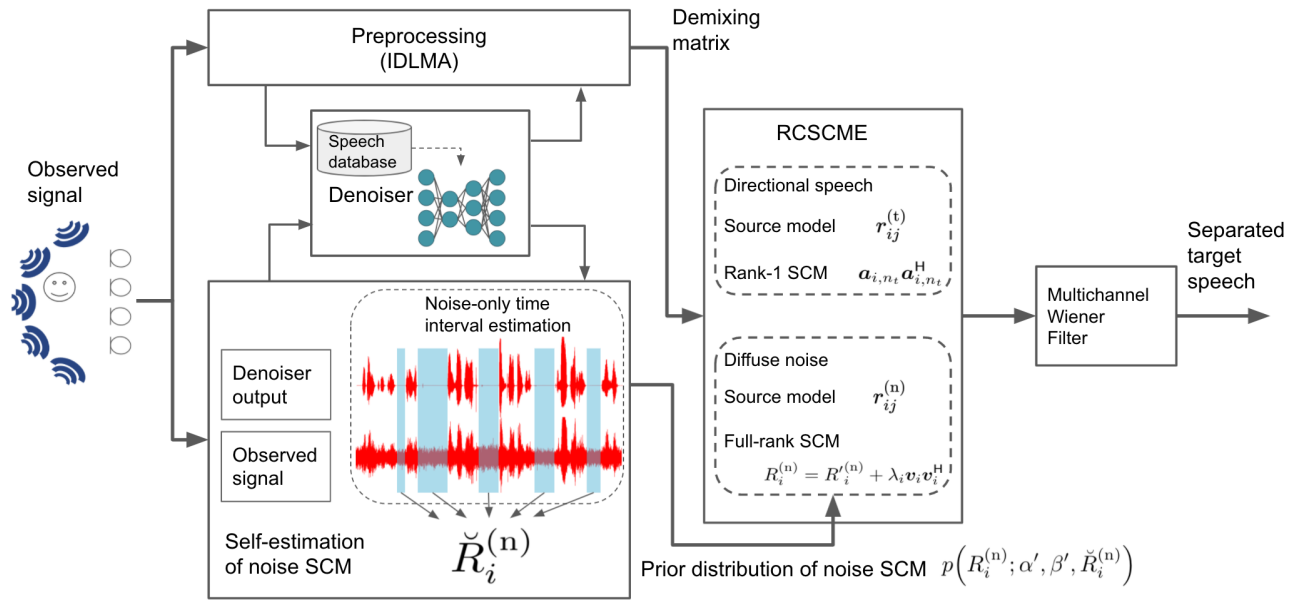


Fig. 1. Process flow of noise self-supervised RCSCME.

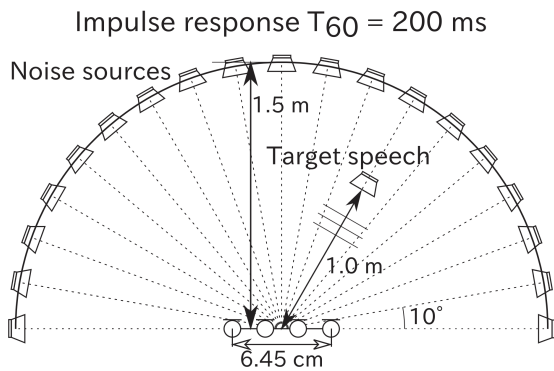


Fig. 2. Impulse response recording conditions in [16].

depending on λ_i and modify (33) to

$$\begin{aligned}
 Q(\Theta; \tilde{\Theta}) &= (\alpha' + M + J) \log \lambda_i \\
 &+ \frac{1}{\lambda_i} \mathbf{u}_i^H \left(\frac{1}{\beta'} \check{R}_i^{(n)} + \sum_j \frac{1}{r_{ij}^{(n)}} \hat{R}_{ij}^{(n)} \right) \mathbf{u}_i \\
 &+ \text{const.}
 \end{aligned} \quad (34)$$

We obtain the following update rule, instead of (21), by solving $\partial Q / \partial \lambda_i = 0$:

$$\lambda_i \leftarrow \frac{1}{\alpha' + M + J} \mathbf{u}_i^H \left(\frac{1}{\beta'} \check{R}_i^{(n)} + \sum_j \frac{1}{r_{ij}^{(n)}} \hat{R}_{ij}^{(n)} \right) \mathbf{u}_i. \quad (35)$$

IV. EXPERIMENT

A. Experimental Conditions

We conducted a simulated experiment to verify the effectiveness of the proposed method. The diffuse noise was simulated by convolution of the impulse response from 19 loudspeakers to four microphones. We used the four noise signals: babble, cafe, station, and traffic noise. For the babble noise, we used 19 speech signals from the JNAS speech corpus [24]. For the cafe, station, and traffic noise, we used 19 signals, which were obtained by dividing each noise signal from DEMAND [25]. The target speech signal was simulated by convolution of the impulse response from a closer location. For the target speech signal, we used one speech signal from the JNAS speech corpus [24]. We used the conditions in Fig. 2 to record the impulse response [16]. The sampling rate was 16 kHz. The input signal-to-noise ratio was set to 0 dB. The STFT was performed by using a 64-ms-long Hamming window and 32-ms-long shift.

We compared ILRMA, IDLMA, RCSCME using ILRMA (ILRMA+RCSCME), RCSCME using IDLMA with $\sigma_{ij,n}$ updated by (27) and (28) (proposed method 1), and noise self-supervised RCSCME using IDLMA with λ_i updated by (35) (proposed method 2). We used source-to-distortion ratio (SDR) [22] improvement as a measure. We used $(\alpha, \beta, \alpha', \beta') = (1.3, 10^{-16}, 8 \times 10^2, 10^4)$ for each method, which were selected experimentally. The flooring parameter in (6) and (28) was set to $\varepsilon = 0.1 \times \sum_{i,j} \sigma_{ij,n}^2 / (IJ)$ and the threshold parameter in (29) was set to $\theta = 10^{-3}$. The demixing matrix W_i in ILRMA and IDLMA was initialized by the identity matrix. The number of NMF bases of ILRMA was set to 10 and the NMF variables were initialized by random values taken from the uniform distribution on $[0, 1]$. We used the

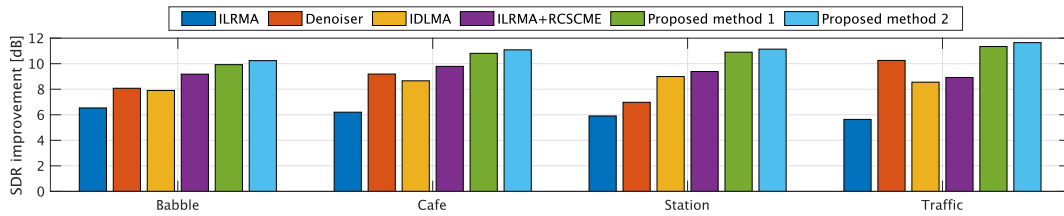


Fig. 3. Largest SDR improvement among the iterations under each noise.

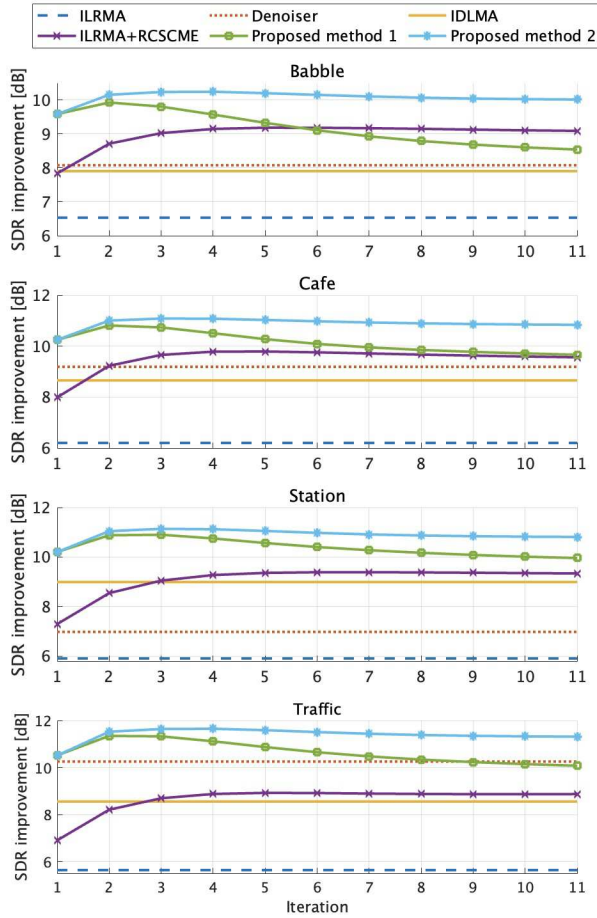


Fig. 4. Behavior of SDR improvements under each noise.

unit eigenvector corresponding to the zero eigenvalue of $R_i^{(n)}$ as v_i in RCSCME. For Denoiser, we used the trained model provided by [18] with $H = 48$ hidden channels, which was trained on the Deep Noise Suppression Challenge dataset [26]. Since Denoiser is the waveform-domain model, we used Denoiser as

$$\text{DNN}(Y_n) = \text{STFT}[\text{Denoiser}(\text{STFT}^{-1}[Y_n])]. \quad (36)$$

We set the number of iterations in ILRMA to 50, that in IDLMA to 90, and that in RCSCME for ILRMA+RCSCME, proposed method 1, and proposed method 2 to 10. In IDLMA, $\sigma_{ij,n}$ was updated by DNN for every 30 times W_i was updated.

These numbers of iterations were experimentally determined.

B. Results

We took the average of the SDR improvement results for 10 different sets of randomly initialized values of the NMF variables in ILRMA and ILRMA+RCSCME. The SDR improvements for each noise are shown in Fig. 3. For each RCSCME method, we show the largest SDR improvement among the iterations. As shown in Fig 3, Denoiser outperformed IDLMA for babble, cafe, and traffic noise. The reason is that Denoiser includes a nonlinear transformation that can provide stronger noise reduction than linear systems, whereas IDLMA estimates the output signals via a linear time-invariant spatial filter, i.e., demixing matrix W_i . However, proposed method 1 shows better performance than Denoiser and IDLMA for all the noises, implying the advantage of the time-variant properties in RCSCME. It is also shown that proposed method 2 achieves the best separation performance among all methods; this would be due to introduction of the self-estimated prior distribution of the noise SCM. The SDR improvement for each iteration is shown in Fig. 4. Fig. 4 shows that ILRMA+RCSCME requires more than five iterations to converge, while proposed method 1 and proposed method 2 take the maximum value in less than five iterations. Here in proposed method 1, the SDR improvement decreased after reaching the highest value. As discussed in [16], the phenomenon is caused by the sparsity of $r_{ij}^{(t)}$. On the other hand, even though proposed method 2 also imposes the sparsity on $r_{ij}^{(t)}$, the prior distribution of the noise SCM may contribute the stability of SDR improvement.

V. CONCLUSIONS

We proposed RCSCME using IDLMA, which is applied to speech enhancement by preparing a single-channel speech enhancement DNN for estimating the power spectrogram of the target speech and the noise. We used Denoiser as the single-channel speech enhancement DNN model for estimating the power spectrogram of the target speech. To estimate the power spectrogram of diffuse noise, we subtracted the speech extracted by Denoiser from the separated signal. We also proposed a method to estimate the noise-only time intervals using the output of Denoiser and design the prior distribution of the noise SCM for RCSCME. We confirmed that noise self-supervised RCSCME using IDLMA outperformed the other methods and took the maximum value in fewer iterations than the conventional methods.

ACKNOWLEDGMENT

This work was supported by the Japan–New Zealand Research Cooperative Program of JSPS and RSNZ (Grant Number JPJSBP120201002), JSPS KAKENHI Grant Numbers 19K20306, 19H01116, and 19H04131, and JST Moonshot R&D Grant Number JPMJPS2011.

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] A. Hiroe, “Solution of permutation problem in frequency domain ICA, using multivariate probability density functions,” in *Proc. ICA*, 2006, pp. 601–608.
- [3] 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, 2006.
- [4] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” in *Proc. WASPAA*, 2011, pp. 189–192.
- [5] 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.
- [6] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
- [7] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, “The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech,” *IEEE Trans. SAP*, vol. 11, no. 2, pp. 109–116, 3 2003.
- [8] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, “Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures,” *EURASIP J. Applied Signal Process.*, vol. 2003, no. 11, pp. 1157–1166, 2003.
- [9] 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, 5 2009.
- [10] 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, 2009.
- [11] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, “Multichannel extensions of non-negative matrix factorization with complex-valued data,” *IEEE Trans. ASLP*, vol. 21, no. 5, pp. 971–982, 2013.
- [12] N. Ito and T. Nakatani, “FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization,” in *Proc. ICASSP*, 2019, pp. 371–375.
- [13] 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.
- [14] N. Q. 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.
- [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, pp. 1–5.
- [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, 2020.
- [17] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari, and N. Ono, “Independent deeply learned matrix analysis for determined audio source separation,” *IEEE/ACM Trans. ASLP*, vol. 27, no. 10, pp. 1601–1615, 2019.
- [18] A. Defossez, G. Synnaeve, and Y. Adi, “Real time speech enhancement in the waveform domain,” in *Proc. Interspeech*, 2020.
- [19] E. M. Grais, M. U. Sen, and H. Erdogan, “Deep neural networks for single channel source separation,” in *Proc. ICASSP*, 2014, pp. 3734–3738.
- [20] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florencio, and M. Hasegawa-Johnson, “Deep learning based speech beamforming,” in *Proc. ICASSP*, 2018, pp. 5389–5393.
- [21] A. A. Nugraha, A. Liutkus, and E. Vincent, “Multichannel audio source separation with deep neural networks,” *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1652–1664, 2016.
- [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.
- [23] A. Iranmanesh, M. Arashi, D. Nagar, and S. Tabatabaey, “On inverted matrix variate gamma distribution,” *Commun. Stat. - Theory Methods*, vol. 42, no. 1, pp. 28–41, 2013.
- [24] 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,” *JASJ (E)*, vol. 20, no. 3, pp. 199–206, 1999.
- [25] J. Thiemann, N. Ito, and E. Vincent, “DEMAND: a collection of multichannel recordings of acoustic noise in diverse environments,” Zenodo, Jun. 2013. [Online]. Available: <https://doi.org/10.5281/zenodo.1227121>
- [26] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, “The INTERSPEECH 2020 Deep Noise Suppression Challenge: datasets, subjective testing framework, and challenge results,” *arXiv preprint arXiv:2005.13981*, 2020.