MULTICHANNEL HEARING-AID SYSTEM BASED ON BASIS-SHARED SEMI-SUPERVISED INDEPENDENT LOW-RANK MATRIX ANALYSIS

Masakazu Une[1] **Yuki Kubo**[2] **Norihiro Takamune**[2]
Daichi Kitamura[3] **Hiroshi Saruwatari**[2] **Shoji Makino**[1]
[1] University of Tsukuba, Graduate School of Systems and Information Engineering, Japan
[2] The University of Tokyo, Graduate School of Information Science and Technology, Japan
[3] National Institute of Technology, Kagawa College, Japan

ABSTRACT

Independent low-rank matrix analysis (ILRMA) can achieve highest-quality separation performance among blind source separation methods, but it still fails to separate the target speech under diffuse noise conditions. To solve this problem, rank-constrained special covariance matrix (SCM) estimation has been proposed. This method estimates the full-rank SCM of diffuse noise using spatial parameters estimated by ILRMA as the preprocessing and suppresses the noise in the direction of the target source. On the other hand, a semi-supervised extension of simple ILRMA, basis-shared ILRMA (BS-ILRMA), has been proposed. This method employs two ILRMAs for noise training and separation and achieves more effective separation than simple ILRMA. BS-ILRMA, as well as ILRMA, is a linear separation filter using estimated spatial parameters. In this paper, we introduce BS-ILRMA to the spatial parameter estimation as the preprocessing of rank-constrained SCM estimation. We evaluate the practical performance of the proposed method using our developed hearing-aid system that consists of binaural microphone arrays and the user's smartphone microphones.

1. Introduction

When we use a binaural hearing-aid system in a noisy environment, target-speech extraction is necessary because speech is always contaminated by noise. In binaural hearing-aid systems, blind source separation (BSS) [1] and semi-supervised source separation are suitable because these work well without spatial information or partial source information as training data. Among many BSS methods, independent low-rank matrix analysis (ILRMA) [2] achieves effective and accurate separation. As a semi-supervised extension of simple ILRMA, basis-shared ILRMA (BS-ILRMA) has been proposed [3]. BS-ILRMA was originally proposed for a rescue robot and separates survivor's voice from the loud noise the robot generates (ego-noise). In this application, training data of the target speech cannot be obtained, but ego-noise can easily be acquired in advance. For hearing-aid systems, BS-ILRMA is applicable because noise information can be acquired as the observed signal before conversation. However, the efficacy of BS-ILRMA for hearing-aid systems has not yet been determined because it has some disadvantages compared with the ego-noise separation task; the available data length is small; the distance from microphones to noise sources is long; there are various interferers (e.g., human voice and footsteps).

Linear time-invariant filters such as ILRMA and BS-ILRMA cannot separate the target speech under diffuse noise conditions in principle [4]. To overcome this limitation, rank-constrained spatial covariance matrix (SCM) estimation [4] has been proposed as an effective method for a situation in which noise arrives from all directions, i.e., diffuse noise condition. Basically, this method estimates a full-rank SCM [5], which represents spatial characteristics of diffuse noise, just as multichannel nonnegative matrix factorization (NMF) [6, 7] does. Rank-constrained SCM estimation consists of the following two steps: estimation of spatial parameters using ILRMA as the first step and suppression of the noise in the direction of the target source using the estimated spatial parameters as the second step. Multichannel NMF is sensitive to parameter initialization and requires the estimation of an enormous number of parameters, leading to a high computational cost. In contrast to multichannel NMF, the rank-constrained SCM estimation reduces the number of parameters by using the highly accurate spatial parameters obtained by ILRMA and restores the lost spatial basis for diffuse noise.

We have developed a multichannel hearing-aid system including binaural ear-attached microphones and smartphone microphones [8]. The additional microphones from a smartphone increase the total number of microphones as well as provides spatial information apart from the user's ears, although these microphones are not synchronized with the ear-attached microphones. We confirmed that rank-constrained SCM estimation whose parameters are preestimated by ILRMA achieves a high-quality separation performance for our developed hearing-aid system [8]. If BS-ILRMA is more effective than ILRMA, rank-constrained SCM estimation is expected to achieve higher-quality separation performance by introducing BS-ILRMA to the first step. In this paper, first, we show the efficacy of BS-ILRMA for practical data recorded by our developed hearing-aid system. Next, we show that rank-constrained SCM estimation that employs BS-

ILRMA as the parameter initialization process can achieve high-quality separation performance.

2. Formulation and BSS Algorithms

2.1 Formulation

Let us consider separating a multichannel observed signal, which is obtained by M microphones capturing the signals arriving from N sources. The source, observed, and separated signals in each time-frequency slot are denoted as

$$\boldsymbol{s}_{ij} = (s_{ij,1}, \ldots, s_{ij,n}, \ldots, s_{ij,N})^\top, \tag{1}$$

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

$$\boldsymbol{y}_{ij} = (y_{ij,1}, \ldots, y_{ij,n}, \ldots, y_{ij,N})^\top, \tag{3}$$

where $i = 1, \ldots, I$, $j = 1, \ldots, J$, $n = 1, \ldots, N$, and $m = 1, \ldots, M$ indicate the indexes of the frequency bins, time frames, sources, and microphones, respectively. The operator $\cdot^\top$ indicates transpose. When each source is a directional source and the window length of short-time Fourier transform (STFT) is sufficiently larger than that of the impulse responses from each source to each microphone, the observed signal $\boldsymbol{x}_{ij}$ and the mixing matrix $\boldsymbol{A}_i = (\boldsymbol{a}_{i,1} \cdots \boldsymbol{a}_{i,N}) \in \mathbb{C}^{M \times N}$ in each frequency bin have the relation

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

where $\boldsymbol{a}_{i,n}$ is the steering vector for each source. If the number of microphones is equal to that of sources ($M = N$) and $\boldsymbol{A}_i$ is not a singular matrix, the separated signal $\boldsymbol{y}_{ij}$ can be obtained by estimating the demixing matrix $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1} = (\boldsymbol{w}_{i,1} \cdots \boldsymbol{w}_{i,N})^{\mathrm{H}} \in \mathbb{C}^{N \times M}$ as

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

where the operator $\cdot^{\mathrm{H}}$ denotes the Hermitian transpose.

2.2 ILRMA

In ILRMA [2], the component of the nth source in each time-frequency slot is assumed to be generated from a statistical model that follows the univariate complex Gaussian distribution as

$$s_{ij,n} \sim \mathcal{N}_c^{\text{uni}}\left(0, r_{ij,n}\right), \tag{6}$$

$$r_{ij,n} = \sum_l t_{il,n} v_{lj,n}, \tag{7}$$

where $\mathcal{N}_c^{\text{uni}}\left(\mu, \sigma^2\right)$ is a univariate complex Gaussian distribution with a mean μ and variance σ^2, $t_{il,n} \geq 0$ and $v_{lj,n} \geq 0$ are NMF variables of the basis matrix $\boldsymbol{T}_n \in \mathbb{R}^{I \times L}$ and the activation matrix $\boldsymbol{V}_n \in \mathbb{R}^{L \times J}$, respectively, $l = 1, \ldots, L$ is an index of the NMF basis, and L is the number of bases. $r_{ij,n}$ corresponds to the nth source model. Simultaneously, the observed signal $\boldsymbol{x}_{ij}$ follows the multivariate complex Gaussian distribution because of the reproductive property, i.e.,

$$\boldsymbol{x}_{ij} \sim \mathcal{N}_c^{\text{mul}}\left(\boldsymbol{0}, \sum_n r_{ij,n} \boldsymbol{a}_{i,n} \boldsymbol{a}_{i,n}^{\mathrm{H}}\right), \tag{8}$$

where $\mathcal{N}_c^{\text{mul}}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is a multivariate complex Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. The steering vector $\boldsymbol{a}_{i,n}$ corresponds to the spatial basis of the nth source, which constructs the rank-1 SCM as $\boldsymbol{a}_{i,n} \boldsymbol{a}_{i,n}^{\mathrm{H}}$. The cost function of ILRMA $\mathcal{J}_{\text{ILRMA}}$ is defined as the negative log likelihood function

$$\mathcal{J}_{\text{ILRMA}} = \sum_n \sum_{i,j} \left[\frac{|y_{ij,n}|^2}{r_{ij,n}} + \log r_{ij,n} \right] - 2J \sum_i \log |\det \boldsymbol{W}_i| + \text{const.}, \tag{9}$$

where the NMF variables $t_{il,n}$ and $v_{lj,n}$ and the demixing matrix $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1}$ are estimated by minimizing (9), which is the maximum likelihood estimation.

2.3 Rank-Constrained SCM Estimation

Linear time-invariant filters such as ILRMA cannot separate the target speech under diffuse noise conditions in principle. To solve this problem, rank-constrained SCM estimation has been proposed [4]. This estimation focuses on a situation where one directional target source and diffuse noise are mixed. This method consists of two processes. First, we estimate the linear time-invariant filter $\boldsymbol{W}_i$ by applying ILRMA to $\boldsymbol{x}_{ij}$. Next, the residual noise in the direction of the target source is suppressed by estimating the full-rank noise SCM. ILRMA outputs M separated signals which consist of one "noise-contaminated target speech" component and $M - 1$ noise-only components (see [9] for the physical mechanism of this phenomenon). The rank of SCM calculated from $M - 1$ "noise-only" signals is $M - 1$ [4]. The diffuse noise SCM should be a full rank (i.e., rank-M); however, the estimated SCM lacks one rank (one spatial basis) that corresponds to the target source direction. Since the full-rank noise SCM is required to suppress the noise in the direction of the target source, in rank-constrained SCM estimation, the noise SCM is modeled as an addition of rank-$(M-1)$ SCM (preestimated by ILRMA) and another rank-1 SCM whose eigenvalue is estimated.

The rank-constrained SCM estimation assumes the observed signal $\boldsymbol{x}_{ij}$ to be the sum of the target source image vector $\boldsymbol{h}_{ij} = (h_{ij,1}, \ldots, h_{ij,M})^\top$ and the diffuse noise image vector $\boldsymbol{u}_{ij} = (u_{ij,1}, \ldots, u_{ij,M})^\top$; i.e.,

$$\boldsymbol{x}_{ij} = \boldsymbol{h}_{ij} + \boldsymbol{u}_{ij}. \tag{10}$$

The source image vector $\boldsymbol{h}_{ij}$ is expressed using a vector corresponding to the target source, $\boldsymbol{a}_i^{(h)} =: \boldsymbol{a}_{i,n_h}$ out of the spatial bases $\boldsymbol{a}_{i,1}, \ldots, \boldsymbol{a}_{i,N}$ obtained by ILRMA, and the target source image $s_{ij}^{(h)}$ as follows:

$$\boldsymbol{h}_{ij} = \boldsymbol{a}_i^{(h)} s_{ij}^{(h)}, \tag{11}$$

$$s_{ij}^{(h)} \sim \mathcal{N}_c^{\text{uni}}\left(0, r_{ij}^{(h)}\right), \tag{12}$$

where n_h indicates the index corresponding to the target source and $r_{ij}^{(h)} =: r_{ij,n_h}$ is the variance of the target source (power spectrogram), where $r_{ij}^{(h)}$ is assumed to

have sparsity in the time-frequency domain by modeling its prior distribution with the inverse gamma distribution as

$$p(r_{ij}^{(h)};\alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}\left(r_{ij}^{(h)}\right)^{-\alpha-1}\exp\left(-\frac{\beta}{r_{ij}^{(h)}}\right), \quad (13)$$

where $\alpha>0$, $\beta>0$, and $\Gamma(\cdot)$ are the shape parameter, scale parameter, and gamma function, respectively.

The diffuse noise $\boldsymbol{u}_{ij}$ follows the following multivariate complex Gaussian distribution and is statistically independent of the target source $\boldsymbol{h}_{ij}$:

$$\boldsymbol{u}_{ij}\sim\mathcal{N}_c^{\mathrm{mul}}\left(\boldsymbol{0},r_{ij}^{(u)}\boldsymbol{R}_i^{(u)}\right), \quad (14)$$

where $r_{ij}^{(u)}$ and $\boldsymbol{R}_i^{(u)}$ are the variance and the full-rank SCM of diffuse noise, respectively. The SCM of the diffuse noise $\boldsymbol{R}_i^{(u)}$ is represented by the demixing filter $\boldsymbol{w}_{i,n}$ as

$$\boldsymbol{R}_i^{(u)}=\boldsymbol{R}'^{(u)}_i+\lambda_i\boldsymbol{b}_i\boldsymbol{b}_i^{\mathrm{H}}, \quad (15)$$

$$\boldsymbol{R}'^{(u)}_i=\frac{1}{J}\sum_j\boldsymbol{W}_i^{-1}\mathrm{diag}(|\boldsymbol{w}_{i,1}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2,\ldots,|\boldsymbol{w}_{i,n_h-1}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2, 0,|\boldsymbol{w}_{i,n_h+1}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2,\ldots,|\boldsymbol{w}_{i,N}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2)(\boldsymbol{W}_i^{-1})^{\mathrm{H}}, \quad (16)$$

where the operator $\mathrm{diag}(k_1,k_2,\ldots)$ generates the square diagonal matrix with the arguments on the main diagonal, $\boldsymbol{R}'^{(u)}_i$ is the noise SCM estimated by ILRMA whose rank is $M-1$, $\boldsymbol{b}_i$ is a unit eigenvector corresponding to the zero eigenvalue of $\boldsymbol{R}'^{(u)}_i$ and λ_i is the weight variable. Note that, in $\boldsymbol{R}_i^{(u)}$, only λ_i is the variable to be optimized because $\boldsymbol{R}'^{(u)}_i$ and $\boldsymbol{b}_i$ are given by ILRMA as fixed values in advance. By modeling the prior distribution of the target source variance in (13), we express the negative log posterior function $\mathcal{L}$ of the rank-constrained SCM estimation as

$$\mathcal{L}(r_{ij}^{(h)},r_{ij}^{(u)},\lambda_i)=\sum_{i,j}\left[\boldsymbol{x}_{ij}^{\mathrm{H}}(\boldsymbol{R}_{ij}^{(x)})^{-1}\boldsymbol{x}_{ij}+\log\det\boldsymbol{R}_{ij}^{(x)}+(\alpha+1)\log r_{ij}^{(h)}+\frac{\beta}{r_{ij}^{(h)}}\right]+\mathrm{const.}, \quad (17)$$

$$\boldsymbol{R}_{ij}^{(x)}=r_{ij}^{(h)}\boldsymbol{a}_i^{(h)}\boldsymbol{a}_i^{(h)\mathrm{H}}+r_{ij}^{(u)}\boldsymbol{R}_i^{(u)}. \quad (18)$$

The parameters in this negative log posterior function $\mathcal{L}$ are optimized by a maximum a posteriori estimation based on the expectation–maximization (EM) algorithm [4]. Finally, multichannel Wiener filtering is applied to suppress noise diffusing the observed signal using the estimated parameters.

3. BS-ILRMA and Application to Hearing-Aid System

BS-ILRMA has been proposed for a rescue robot that detects a survivor's voice in a disaster site [3]. BS-ILRMA simultaneously performs previously recorded ego-noise training and extraction of the survivor's voice. In the application to hearing-aid systems, the observed unvoiced period, which does not contain speech and conversation, can be obtained as the noise samples (training signal of diffuse noise) in advance. Therefore, BS-ILRMA is applicable for the hearing-aid system. Under the $M=N$ condition, BS-ILRMA assumes that $N'=M'=N-1$ noise sources and one target source exist. Let $\boldsymbol{x}_{ij'}^{(\mathrm{noise})}=(x_{ij',1}^{(\mathrm{noise})},\ldots,x_{ij',m'}^{(\mathrm{noise})},\ldots,x_{ij',M'}^{(\mathrm{noise})})^{\top}$ and $\boldsymbol{x}_{ij}^{(\mathrm{mix})}=(x_{ij,1}^{(\mathrm{mix})},\ldots,x_{ij,M}^{(\mathrm{mix})})^{\top}$ be the prepared M' ch noise samples and M mixture signals, respectively. $j'=1,\ldots,J'$ and $m'=1,\ldots,M'$ indicate the indexes of the time frame and sources for noise samples, respectively.

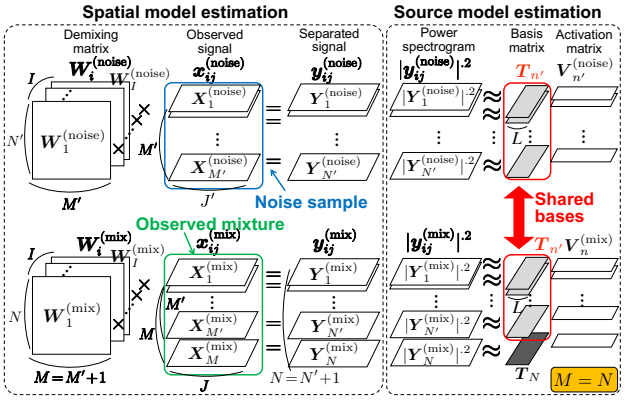

Figure 1. Overview of BS-ILRMA, where upper and lower models are *simultaneously* optimized.

We can consider a simple semi-supervised ILRMA by employing pretrained basis matrices for noise sources, which is a similar approach to the semi-supervised NMF [10]. We call this method semi-supervised ILRMA (SS-ILRMA) in this paper. SS-ILRMA trains the N' noise bases $\boldsymbol{T}_{n'}^{(\mathrm{noise})}\in\mathbb{R}_{\geq0}^{I\times L}$ by applying simple ILRMA to the noise samples $\boldsymbol{x}_{ij'}^{(\mathrm{noise})}$ in advance and separates the M-channel observed signals $\boldsymbol{x}_{ij}^{(\mathrm{mix})}$ by another ILRMA while fixing the trained noise basis matrices $\boldsymbol{T}_{n'}^{(\mathrm{noise})}$, where the index $n'=1,\ldots,N'$ indicates the source of the noise samples. However, this naive semi-supervised approach fails to fully receive the benefits of employing the noise sample. This is because the scale ambiguity in $\boldsymbol{W}_i$ among frequency bins can collapse the spectral structures in the supervised basis matrix $\boldsymbol{T}_{n'}^{(\mathrm{noise})}$ [3].

To solve this problem, BS-ILRMA has been proposed. The overview of BS-ILRMA is shown in Fig. 1, where $\boldsymbol{W}_i^{(\mathrm{noise})}\in\mathbb{C}^{N'\times M'}$ and $\boldsymbol{W}_i^{(\mathrm{mix})}\in\mathbb{C}^{N\times M}$ represent the demixing matrices for the noise samples $\boldsymbol{x}_{ij'}^{(\mathrm{noise})}$ and the observed signal $\boldsymbol{x}_{ij}^{(\mathrm{mix})}$, respectively. $\boldsymbol{X}_{m'}^{(\mathrm{noise})}\in\mathbb{C}^{I\times J'}$ and $\boldsymbol{Y}_{n'}^{(\mathrm{noise})}\in\mathbb{C}^{I\times J'}$ represent the spectrograms of the m'th channel in $\boldsymbol{x}_{ij'}^{(\mathrm{noise})}$ and the n'th channel in $\boldsymbol{y}_{ij'}^{(\mathrm{noise})}=(y_{ij',1}^{(\mathrm{noise})},\ldots,y_{ij',N'}^{(\mathrm{noise})})^{\top}$, respectively. $\boldsymbol{X}_{m}^{(\mathrm{mix})}\in\mathbb{C}^{I\times J}$ and $\boldsymbol{Y}_{n}^{(\mathrm{mix})}\in\mathbb{C}^{I\times J}$ represent the spectrograms of the mth channel in $\boldsymbol{x}_{ij}^{(\mathrm{mix})}$ and the nth channel in $\boldsymbol{y}_{ij}^{(\mathrm{mix})}=(y_{ij,1}^{(\mathrm{mix})},\ldots,y_{ij,N}^{(\mathrm{mix})})^{\top}$, respectively. $|\cdot|^{\cdot2}$ denotes entry-wise absolute and squaring operations, $\boldsymbol{T}_{n'}\in\mathbb{R}_{\geq0}^{I\times L}$ represents

the shared basis matrix for the noise sources, $\boldsymbol{T}_N \in \mathbb{R}_{\geq 0}^{I \times L}$ represents the unshared basis matrix for the target source, and $\boldsymbol{V}_{n'}^{(\text{noise})} \in \mathbb{R}_{\geq 0}^{L \times J'}$ and $\boldsymbol{V}_{n}^{(\text{mix})} \in \mathbb{R}_{\geq 0}^{L \times J}$ represent the activation matrices for approximating $\boldsymbol{Y}_{n'}^{(\text{noise})}$ and $\boldsymbol{Y}_{n}^{(\text{mix})}$, respectively. BS-ILRMA employs two ILRMAs; one is applied to the noise sample $\boldsymbol{x}_{ij'}^{(\text{noise})}$ to estimate $\boldsymbol{W}_i^{(\text{noise})}$ and $\boldsymbol{y}_{ij'}^{(\text{noise})}$, and the other one is applied to the noisy speech signal $\boldsymbol{x}_{ij}^{(\text{mix})}$ to estimate $\boldsymbol{W}_i^{(\text{mix})}$ and $\boldsymbol{y}_{ij}^{(\text{mix})}$. The important point is that the basis matrices for the noise sources, $\boldsymbol{T}_{n'}$, are *shared* between these two ILRMAs, and all the variables in these models are simultaneously optimized. Since the shared basis matrices $\boldsymbol{T}_{n'}$ must represent similar spectra in both $\boldsymbol{x}_{ij'}^{(\text{noise})}$ and $\boldsymbol{x}_{ij}^{(\text{mix})}$, the noise spectral patterns will be captured by $\boldsymbol{T}_{n'}$, and the other basis matrix $\boldsymbol{T}_N$ will consequently represent spectral patterns of the remaining source, i.e., the target source.

The cost function of BS-ILRMA is defined as the sum of two cost functions of ILRMA as follows:

$$\begin{aligned}\mathcal{J} =& \frac{1}{N'}\left\{\sum_{n'=1}^{N'}\sum_{i,j'}\left[\frac{|y_{i,j',n'}^{(\text{noise})}|^2}{\sum_l t_{il,n'} v_{lj',n'}^{(\text{noise})}} + \log\sum_l t_{il,n'} v_{lj',n'}^{(\text{noise})}\right]\right. \\ &\left. - 2J'\sum_i \log|\det \boldsymbol{W}_i^{(\text{noise})}|\right\} \\ &+ \frac{1}{N}\left\{\sum_{n=1}^{N'}\sum_{i,j}\left[\frac{|y_{i,j,n}^{(\text{mix})}|^2}{\sum_l t_{il,n} v_{lj,n}^{(\text{mix})}} + \log\sum_l t_{il,n} v_{lj,n}^{(\text{mix})}\right]\right. \\ &+ \sum_{i,j}\left[\frac{|y_{i,j,N}^{(\text{mix})}|^2}{\sum_l t_{il,N} v_{lj,N}^{(\text{mix})}} + \log\sum_l t_{il,N} v_{lj,N}^{(\text{mix})}\right] \\ &\left. - 2J\sum_i \log|\det \boldsymbol{W}_i^{(\text{mix})}|\right\}, \end{aligned} \tag{19}$$

where $t_{il,n'}$, $t_{il,N}$, $v_{lj,n'}^{(\text{noise})}$, and $v_{lj,n}^{(\text{mix})}$ indicate the elements of $\boldsymbol{T}_{n'}$, $\boldsymbol{T}_N$, $\boldsymbol{V}_{n'}^{(\text{noise})}$, and $\boldsymbol{V}_{n}^{(\text{mix})}$, respectively. The update rules of unshared parameters are the same as those in [2]. In contrast, since it is difficult to directly minimize Eq. (19) w.r.t. $t_{il,n'}$, we designed the auxiliary function and minimize it to obtain the local optimal solution [3].

BS-ILRMA is more effective than ILRMA for a task to separate the ego-noise and the target source [3]. However, applying BS-ILRMA to apply to the hearing-aid system is impractical because BS-ILRMA cannot be utilized in a situation in which diffuse noise exists. Thus, we introduce rank-constrained SCM estimation. In this paper, we propose the introduction of BS-ILRMA to the first step of rank-constrained SCM estimation. BS-ILRMA is a linear filter as well as ILRMA, and we can expect that the accuracy of the target speech extraction with the hearing-aid system is improved by employing BS-ILRMA.

4. Experimental Evaluation of Proposed System

The purpose of this experiment is to evaluate the separation performance of rank-constrained SCM estimation in which BS-ILRMA is applied as the first step for the recorded data by our developed hearing-aid system. In the preliminary experiment, we investigate the efficacy of BS-ILRMA for the hearing-aid system that we developed in Sect. 4.1. Next, we evaluate the separation performance of rank-constrained SCM estimation in which BS-ILRMA is applied as the first step in Sect. 4.2.

4.1 Performance of BS-ILRMA

The performance of BS-ILRMA for the hearing-aid system is unclear because the conditions are different from the task of separating ego-noise and target speech, e.g., small available data length, long-distance from microphones to the noise source, and various interference sources. Accordingly, we have to evaluate the performance of BS-ILRMA for the hearing-aid system we developed. We compared with the following three methods: ILRMA, SS-ILRMA, which separates a mixture using pretrained bases obtained by noise samples, and BS-ILRMA, which optimizes the parameters for noise training and separation by sharing bases. We recorded impulse response and diffuse noise in a room using the head-and-torso dummy we developed [8]. Figure 2 (a) shows the head-and-torso dummy, which simulates a person wearing a binaural hearing aid and holding a smartphone. The head-and-torso dummy wears eight microphones (three microphones are attached to each ear [see Figs. 2 (b) and (d)] and two microphones are attached to the smartphone [see Fig. 2 (c)]), which are synchronized with the same sampling rate. We numbered each microphone as shown in Figs. 2 (b)–(d). Figure 3 shows the shape of the recording room and the positions of loudspeakers. The reverberation time of the recording room is 300 ms. The depth, width, and height are 6.5, 5.3, and 2.6 m, respectively. The distance from the head-and-torso dummy to the loudspeaker was varied by 75, 100, and 150 cm, and the angle was varied by -20, 0, and $20°$, where $0°$ means the normal to the head-and-torso dummy and the negative angle means the left side. The height of the head-and-torso dummy was set to 170 cm, and the loudspeakers were set in front of the head-and-torso dummy to mimic the situation of conversation. The target speech is a female utterance from the JNAS database [11] convolved with the impulse response, which is down-sampled from 48 kHz to 16 kHz. We provided a two-second noise period before the utterance for the training of SS-ILRMA and BS-ILRMA. The hamming window was used (64 ms in length and 50% overlap) in STFT. The observed signal was generated by mixing the recorded diffuse noise and the target signal at the input SNRs of -10, -5, and 0 dB. The numbers of bases and iterations of ILRMA, SS-ILRMA, and BS-ILRMA were set to 10 and 50, respectively. As the noise pretraining for SS-ILRMA, the pretrained basis was optimized by 50 update iterations. The observed signal was preprocessed by sphering transformation by principal component analysis. The demixing matrix was initialized as the identity matrix, and the basis and activation matrices were initialized by nonnegative random values. We used source-to-distortion ratio (SDR) improvement [12] as the objective measurement criterion and averaged the scores of each direction and ten initialization trials using different

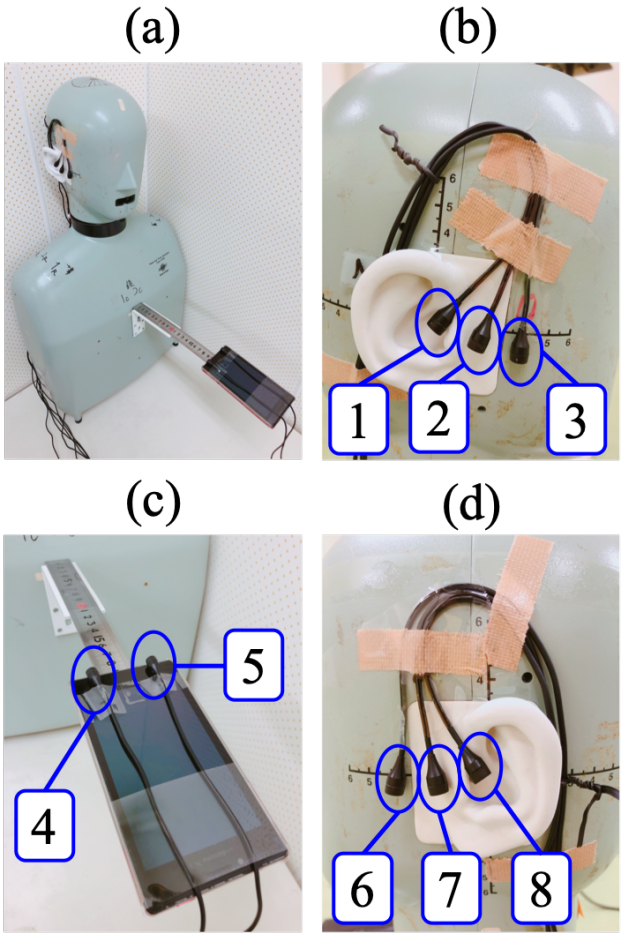

Figure 2. (a) Overall view of head-and-torso dummy, (b) right-ear microphone array, (c) smartphone microphones, and (d) left-ear microphone array.

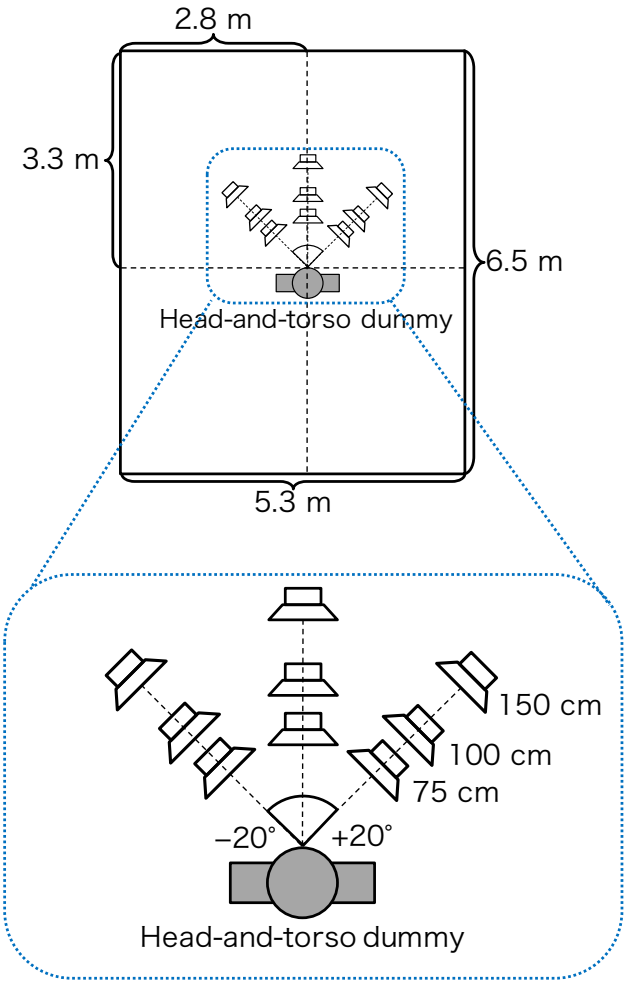

Figure 3. Shape of recording room and positions of loudspeakers (mouth position of conversation partner).

random values.

The results are shown in Fig. 4. The improvements of BS-ILRMA are greater than those of ILRMA and SS-ILRMA in almost all cases. From the results, BS-ILRMA is expected to be effective for the preprocessing of rank-constrained SCM estimation.

4.2 Performance of BS-ILRMA Applied to Rank-Constrained SCM Estimation

Next, we compare the performances when the rank-constrained SCM estimation is preprocessed by ILRMA, SS-ILRMA, and BS-ILRMA. The experimental condition was the same as that described in Sect. 4.1. The shape parameter α and the scale parameter β for the inverse gamma distribution, which is the prior distribution of rank-constrained SCM estimation, were set to 20 and 10^{-16}, respectively. Since we have determined that the rank-constrained SCM estimation can achieve high separation performance with few iterations, the parameters were optimized with only two iterations of the parameter update calculation [8].

Figure 5 shows the results of rank-constrained SCM estimation preprocessed by ILRMA, SS-ILRMA, and BS-ILRMA. We confirmed that rank-constrained SCM estimation clearly improves the SDR. In particular, rank-constrained SCM estimation preprocessed by BS-ILRMA achieves the greatest average SDR improvement among the methods analyzed in almost all cases. Furthermore, the scores of rank-constrained SCM estimation depend on the scores of the preprocessing method. From the results, we consider that BS-ILRMA is suitable for the parameter initialization of rank-constrained SCM estimation.

5. Conclusion

In this study, we investigated the efficacy of BS-ILRMA and rank-constrained SCM estimation preprocessed by BS-ILRMA for our developed hearing-aid system. In the first experimental evaluation, we confirmed that BS-ILRMA is effective in comparison with ILRMA and SS-ILRMA for the hearing-aid system in terms of separation performance. The second experimental evaluation showed that rank-constrained SCM estimation preprocessed by BS-ILRMA can achieve high-quality separation performance. As a future work, we extend rank-constrained SCM estimation to be a semi-supervised fashion.

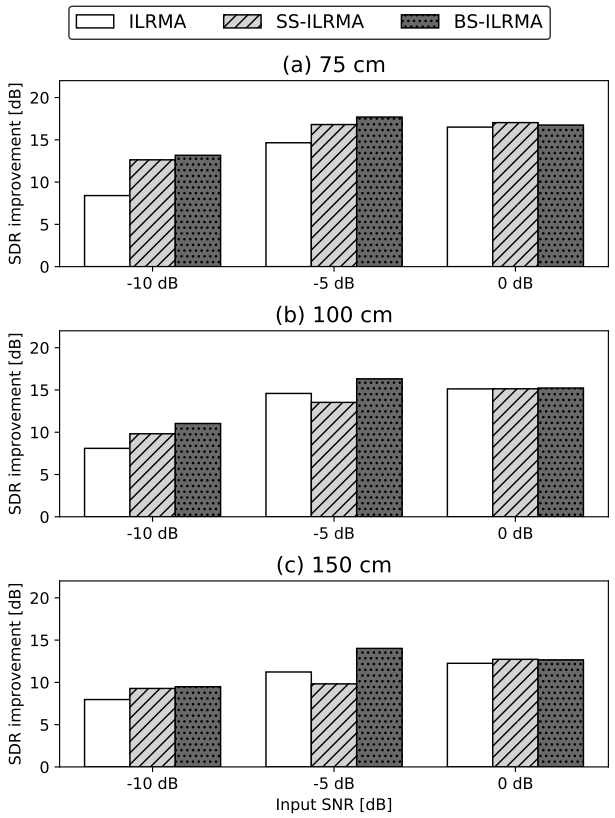

Figure 4. Average SDR improvements of ILRMA, SS-ILRMA, and BS-ILRMA under each input SNR condition. Three figures show results when distance from head-and-torso dummy to target source is set to (a) 75, (b) 100, and (c) 150 cm, respectively.

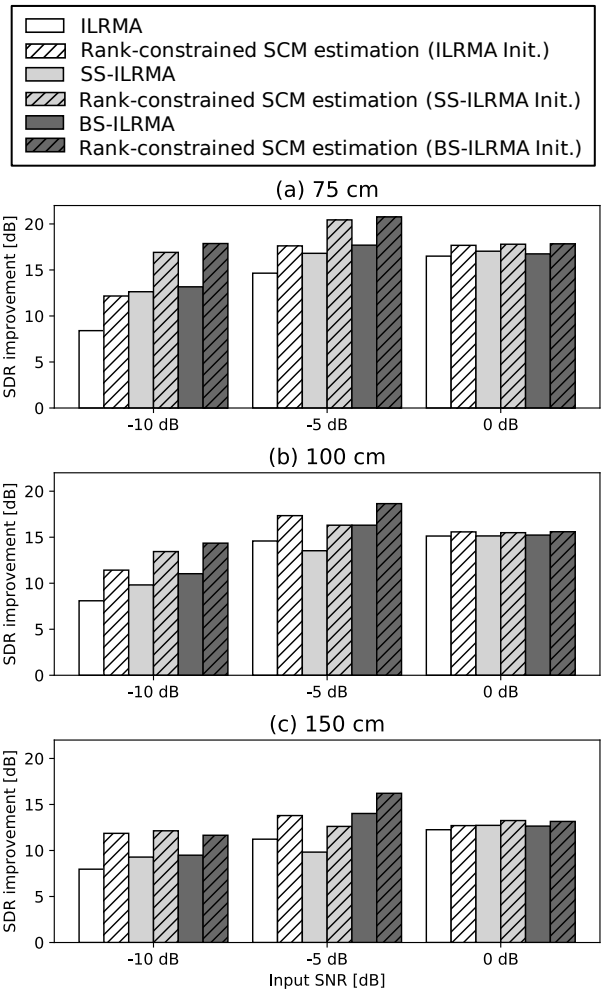

Figure 5. Average SDR improvements of ILRMA, rank-constrained SCM estimation preprocessed by ILRMA, SS-ILRMA, rank-constrained SCM estimation preprocessed by SS-ILRMA, BS-ILRMA, and rank-constrained SCM estimation preprocessed by BS-ILRMA, respectively. Three figures show results when distance from head-and-torso dummy to target source is set to (a) 75, (b) 100, and (c) 150 cm, respectively.

ACKNOWLEDGEMENT

This work was partly supported by SECOM Science and Technology Foundation and JSPS KAKENHI Grant Numbers 19H04131, 19H01116, and 19K20306.

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 and Information Processing*, vol. 8, no. e12, pp. 1–14, 2019.

[2] 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. on ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[3] M. Takakusaki, D. Kitamura, N. Ono, T. Yamada, S. Makino, and H. Saruwatari, "Ego-noise reduction for a hose-shaped rescue robot using basis shared semi-supervised independent low-rank matrix analysis," in *Proc. NCSP*, pp. 351–354, 2018.

[4] 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*, pp. 1814–1818, 2019.

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

[6] 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.

[7] 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.

[8] M. Une, Y. Kubo, N. Takamune, D. Kitamura,

H. Saruwatari, and S. Makino, "Evaluation of multichannel hearing aid system using rank-constrained spatial covariance matrix estimation," in *Proc. APSIPA*, pp. 1874–1879, 2019.

[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, 2009.

[10] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. ICA*, pp. 414–421, 2007.

[11] 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 Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[12] 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.