2018 RISP International Workshop on Nonlinear Circuits,
Communications and Signal Processing (NCSP2018)
Honolulu, Hawaii, USA, March 4-7, 2018

Ego-Noise Reduction for Hose-Shaped Rescue Robot Using Basis-Shared Semi-Supervised Independent Low-Rank Matrix Analysis

Moe Takakusaki[1], Daichi Kitamura[2], Nobutaka Ono[3], Shoji Makino[1],
Takeshi Yamada[1], Hiroshi Saruwatari[2]

[1] University of Tsukuba, 1–1–1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan
[2] The University of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[3] Tokyo Metropolitan University, 6-6 Asahigaoka, Hino, Tokyo 191–0065, Japan

1. Introduction

It is an important task to develop robots for coping with large-scale disasters. Robots working in a disaster site are required in emergency response and in the restoration of the disaster site, which are difficult or dangerous tasks for humans. There are tasks for modern-day robots in disaster response, but their functions are insufficient outdoors and their ability to respond to unexpected situations is unsatisfactory. For example, a typical robot cannot smoothly move in a disaster site and evaluate the situation, as well as act in an environment that does not fit its working condition. To support the development of robots in a disaster site, which can overcome some of the problems of conventional robots, the Council for Science, Technology and Innovation promoted the ImPACT Tough Robotics Challenge [1]. In this research and development program, we aim to realize remote-controlled and autonomous robots that are effective in extreme situations, and develop technologies that will become the basis for the development of outdoor robots.

The hose-shaped rescue robot (Fig. 1) [2] was developed for entering narrow and dark places covered with rubble in a disaster site and finding survivors inside it. This robot vibrates its body by the motors attached at each node to move forward without any control devices or operators. The purpose of the robot is to detect voice of survivors buried in rubbles using the microphones, which are also attached around the body of the robot. However, since the vibration causes loud noise (ego-noise), the observed voice and ego-noise must be separated for robustly detecting the survivors. In [3], a novel multichannel source separation technique called independent low-rank matrix analysis (ILRMA) [4, 5, 6] was applied to reduce ego-noise of the robot, where ILRMA does not require any prior information, e.g., array geometry of microphones, direction of sources, or sample sounds of sources. To improve its performance, the utilization of a sample sound of ego-noise for learning its timbers (spectral patterns) can be considered because the sample sound of ego-noise can easily be obtained by operating the robot in advance. For this reason, in this paper, we propose a new effective semi-supervised algorithm based on ILRMA. The experimental analysis shows that the proposed method can outperform the conventional methods of ego-noise reduction.

2. Ego-noise of hose-shaped rescue robot

2.1 Structure of robot and disposition of ego-noise

The hose-shaped rescue robot consists of a hose for the axis and ciliary tape wrapped around it; it moves forward slowly against the direction of the cilia by vibrating the ciliary tape with vibration motors. A camera and a lighting are attached to the tip of the robot, in addition to an inertial measurement unit, microphones, and speakers attached along the length of the robot. According to the operation principle of the robot, very loud ego-noise is always mixed into the microphones. The main factors for the ego-noise are considered to be the vibration sound generated by the vibration motors and fricative sound. In an actual disaster site, the voice of a survivor seeking help is not loud enough to capture, and it is fainter than the ego-noise. To capture the voice in such a situation, it is necessary to separate the voice from the recorded noisy sound.

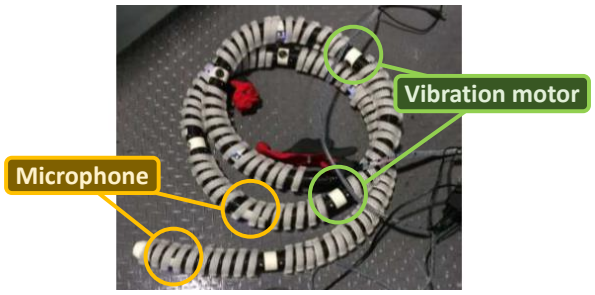

Figure 1: Hose-shaped rescue robot.

2.2 Conventional method for ego-noise reduction

2.2.1 Ego-noise reduction based on blind source separation

The most popular and reliable approach for audio source separation using microphone array is a blind source separation (BSS) technique [7, 8, 9], which does not require any information about the locations of microphones and sources. The assumption in BSS is valid for the hose-shape rescue robot because relative positions of the microphones (array geometry) can be changed depending on posture of the robot and the location of an observed voice is also unknown. In [3], a novel BSS technique, ILRMA [4, 5, 6], is applied to the observed multichannel signal. ILRMA assumes both statistical independence between sources and low-rank structure of each source in time-frequency domain and separates the sources in a fully blind manner, which will be introduced in the following sections.

2.2.2 Formulation in BSS

Let both the numbers of sources and microphones be M. The complex-valued source, observed, and separated signals in each time-frequency slot are defined as follows:

$$\boldsymbol{s}_{ij} = (s_{ij,1}, \cdots, s_{ij,m}, \cdots, s_{ij,M})^t, \quad (1)$$
$$\boldsymbol{x}_{ij} = (x_{ij,1}, \cdots, x_{ij,m}, \cdots, x_{ij,M})^t, \quad (2)$$
$$\boldsymbol{y}_{ij} = (y_{ij,1}, \cdots, y_{ij,m}, \cdots, y_{ij,M})^t, \quad (3)$$

where $1 \leq i \leq I\,(i \in \mathbb{N})$ is the index of frequency bins, $1 \leq j \leq J\,(j \in \mathbb{N})$ is the index of time frames, $1 \leq m \leq M\,(m \in \mathbb{N})$ is the index of sources and channels, and ^{t} denotes

the vector or matrix transpose. When the mixing system is time-invariant and the reverberation time is much shorter than the window length used in short-time Fourier transform (STFT), the following approximated mixture becomes valid:

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

where $\boldsymbol{A}_i = (\boldsymbol{a}_{i,1} \cdots \boldsymbol{a}_{i,M}) \in \mathbb{C}^{M \times M}$ is a frequency-wise mixing matrix and $\boldsymbol{a}_{i,m}$ is the steering vector for each source. When $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1}$ can be defined, the separated signal $\boldsymbol{y}_{ij}$ is represented as

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

where $\boldsymbol{W}_i = (\boldsymbol{w}_{i,1} \cdots \boldsymbol{w}_{i,M})^h \in \mathbb{C}^{M \times M}$ is the frequency-wise demixing matrix, $\boldsymbol{w}_{i,m}$ is the demixing filter for each source, and ^{h} denotes the Hermitian transpose. Note that there exists scale ambiguity in $\boldsymbol{W}_i$ among the frequency bins. To fix the scale of $\boldsymbol{y}_{ij}$, a back-projection technique [10] is often applied after the estimation of $\boldsymbol{W}_i$.

The typical BSS algorithms including independent vector analysis (IVA) [8, 9] and ILRMA assume (4) and (5) and estimate the frequency-wise demixing matrix $\boldsymbol{W}_i$. Strictly speaking, the mixing assumption (4), which is called linear time-invariant mixing, is not valid for an observed multichannel signal by the hose-shaped rescue robot. This is because the robot always moves very slowly, and relative locations of microphones and sources are barely changing with time. However, it is reported that BSS with assumption (4) can achieve high separation performance for a short-time (batch) observed signal [3], which means that the batch-wise algorithm can be used for a long-time observation.

2.2.3 ILRMA

ILRMA is a method unifying IVA and nonnegative matrix factorization (NMF) [11] with Itakura–Saito divergence (IS-NMF) [12], which allows us to simultaneously model the statistical independence between sources and the low-rank source-wise time-frequency structure. Whereas IVA employs source-wise frequency vectors, ILRMA estimates sourcewise power spectrograms that are approximately modeled by the NMF variables. The demixing matrix $\boldsymbol{W}_i$ and the separated signal $\boldsymbol{y}_{ij}$ are optimized so that the spectrogram of each source tends to be a low-rank matrix. This separation mechanism of ILRMA is shown in Fig. 2, where $\boldsymbol{T}_m \in \mathbb{R}_{\geq 0}^{I \times L}$ and $\boldsymbol{V}_m \in \mathbb{R}_{\geq 0}^{L \times J}$ are the basis and activation matrices (NMF variables) for the mth estimated source, respectively, and $1 \leq l \leq L\,(l \in \mathbb{N})$ is the index of the bases. $\boldsymbol{W}_i$, $\boldsymbol{T}_m$, and $\boldsymbol{V}_m$ can consistently be estimated in a fully blind manner.

The cost function in ILRMA is defined as follows:

$$\begin{aligned} \mathcal{J} = \sum_{i,j,m} \left[\frac{|\boldsymbol{w}_{i,m}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2}{\sum_l t_{il,m} v_{lj,m}} + \log \sum_l t_{il,m} v_{lj,m} \right] \\ - 2J \sum_i \log |\det \boldsymbol{W}_i|, \end{aligned} \tag{6}$$

where $t_{il,m}$ and $v_{lj,m}$ are the nonnegative elements of $\boldsymbol{T}_m$ and $\boldsymbol{V}_m$, respectively. The rank-L matrix $\boldsymbol{T}_m \boldsymbol{V}_m$ corresponds to the NMF decomposition and represents an estimated model spectrogram of the mth source. An efficient update algorithm for all $\boldsymbol{W}_i$, $\boldsymbol{T}_m$, and $\boldsymbol{V}_m$ is derived [4, 6] for ILRMA.

For the ego-noise reduction of the hose-shaped rescue robot, we can assume that ego-noise sources contain similar spectral patterns with repetitions, and the power spectrogram of ego-noise sources tends to be a low-rank matrix. For this reason, ILRMA is suitable for this task and can effectively capture the ego-noise components to separate them [3].

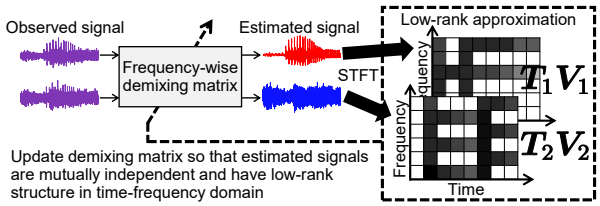

Figure 2: Separation mechanism of ILRMA.

3. Proposed method

3.1 Motivation

In this paper, similar to [3], we assume that the number of ego-noise sources is $M' = M - 1$ and only one speech source (survivor's voice) is mixed in the observed noisy speech signal $\boldsymbol{x}_{ij}^{(\mathrm{mix})}$. Since the sample sound of ego-noise, $\boldsymbol{x}_{ij}^{(\mathrm{noise})}$, can easily be obtained in advance, a semi-supervised approach for ego-noise reduction can be considered to improve the separation performance. The simplest way to utilize such an ego-noise sample sound $\boldsymbol{x}_{ij}^{(\mathrm{noise})}$ is to combine a semi-supervised NMF algorithm [13, 14] and ILRMA as the following steps: (a) the ego-noise sample sound $\boldsymbol{x}_{ij}^{(\mathrm{noise})}$ is separated by simple ILRMA in advance, (b) the supervised (pre-trained) basis matrix for ego-noise, $\boldsymbol{T}_{m'}^{(\mathrm{noise})}$, is obtained from the optimization result of (a), where $1 \leq m' \leq M' \in \mathbb{N}$ is the index of the ego-noise sources, and (c) we apply another ILRMA to the noisy speech signal $\boldsymbol{x}_{ij}^{(\mathrm{mix})}$ using the supervised basis matrix $\boldsymbol{T}_{m'}^{(\mathrm{noise})}$, where $\boldsymbol{T}_{m'}^{(\mathrm{noise})}$ is fixed and the other variables (activation matrix for the supervised ego-noise sources, basis and activation matrices for the unknown speech source, and $\boldsymbol{W}_i$) are optimized, as well as the semi-supervised NMF algorithm [13, 14]. However, this naive semi-supervised approach may fail to fully receive benefits from the sample sound of ego-noise because the scale ambiguity in $\boldsymbol{W}_i$ among frequency bins can collapse the spectral structures in the supervised basis matrix $\boldsymbol{T}_{m'}^{(\mathrm{noise})}$.

To cope with this problem, in this paper, we propose a new semi-supervised algorithm based on ILRMA, which is called *basis-shared ILRMA* (BS-ILRMA). The overview of the proposed method is depicted in Fig. 3, where $\boldsymbol{W}_i^{(\mathrm{noise})} \in \mathbb{C}^{M' \times M'}$ and $\boldsymbol{W}_i^{(\mathrm{mix})} \in \mathbb{C}^{M \times M}$ are the demixing matrices for the ego-noise sample signal $\boldsymbol{x}_{ij}^{(\mathrm{noise})}$ and the observed noisy speech signal $\boldsymbol{x}_{ij}^{(\mathrm{mix})}$, respectively, $\boldsymbol{X}_{m'}^{(\mathrm{noise})} \in \mathbb{C}^{I \times J}$ and $\boldsymbol{Y}_{m'}^{(\mathrm{noise})} \in \mathbb{C}^{I \times J}$ show the spectrograms of m'th channel in $\boldsymbol{x}_{ij}^{(\mathrm{noise})}$ and $\boldsymbol{y}_{ij}^{(\mathrm{noise})}$, respectively, $\boldsymbol{X}_m^{(\mathrm{mix})} \in \mathbb{C}^{I \times J}$ and $\boldsymbol{Y}_m^{(\mathrm{mix})} \in \mathbb{C}^{I \times J}$ show the spectrograms of mth channel in $\boldsymbol{x}_{ij}^{(\mathrm{mix})}$ and $\boldsymbol{y}_{ij}^{(\mathrm{mix})}$, respectively, an absolute symbol with dotted exponent for matrices denotes entry-wise absolute and exponent operations, $\boldsymbol{T}_{m'} \in \mathbb{R}_{\geq 0}^{I \times L}$ is the shared basis matrix for ego-noise sources, $\boldsymbol{T}_M \in \mathbb{R}_{\geq 0}^{I \times L}$ is the unshared basis matrix for the speech source, and $\boldsymbol{V}_{m'}^{(\mathrm{noise})} \in \mathbb{R}_{\geq 0}^{L \times J}$ and $\boldsymbol{V}_m^{(\mathrm{mix})} \in \mathbb{R}_{\geq 0}^{L \times J}$ are the activation matrices for $\boldsymbol{Y}_{m'}^{(\mathrm{noise})}$ and $\boldsymbol{Y}_m^{(\mathrm{mix})}$, respectively. In this method, we employ two ILRMAs; one is applied to the ego-noise sample sound $\boldsymbol{x}_{ij}^{(\mathrm{noise})}$ to estimate $\boldsymbol{W}_i^{(\mathrm{noise})}$ and $\boldsymbol{y}_{ij}^{(\mathrm{noise})}$, and the other one is applied to the noisy speech signal $\boldsymbol{x}_{ij}^{(\mathrm{mix})}$ to estimate $\boldsymbol{W}_i^{(\mathrm{mix})}$ and $\boldsymbol{y}_{ij}^{(\mathrm{mix})}$. The important point is that the basis matrices for ego-noise sources, $\boldsymbol{T}_{m'}$, are *shared* between these two ILRMAs, and all the variables in these models are *simultaneously* optimized. Since the shared basis matrices $\boldsymbol{T}_{m'}$ must represent the similar spectra in both $\boldsymbol{x}_{ij}^{(\mathrm{noise})}$ and $\boldsymbol{x}_{ij}^{(\mathrm{mix})}$, the ego-noise spectral patterns will be captured by $\boldsymbol{T}_{m'}$, and the other basis matrix $\boldsymbol{T}_M$ will consequently represent spectral patterns of the rest

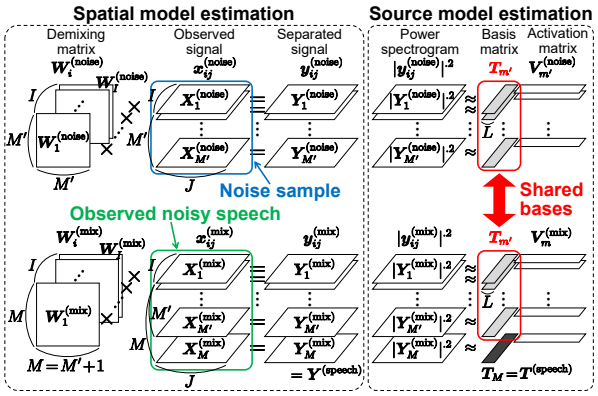

Figure 3: Overview of BS-ILRMA, where upper and lower models are *simultaneously* optimized.

source, i.e., the target speech. Thus, we expect that the separated signals $\boldsymbol{y}_1^{(\mathrm{mix})}, \cdots, \boldsymbol{y}_{M'}^{(\mathrm{mix})}$ are corresponding to the ego-noise sources in $\boldsymbol{x}_{ij}^{(\mathrm{mix})}$, and the other separated signal $\boldsymbol{y}_M^{(\mathrm{mix})}$ is corresponding to the target speech source in $\boldsymbol{x}_{ij}^{(\mathrm{mix})}$. Even if the scale ambiguities in $\boldsymbol{W}_i^{(\mathrm{noise})}$ and $\boldsymbol{W}_i^{(\mathrm{mix})}$ exist, this basis sharing indirectly improves the estimation accuracy of spectral patterns of ego-noise and speech sources, resulting in more accurate source separation.

3.2 Cost function and update rules in BS-ILRMA

We assume that the ego-noise sample sound $\boldsymbol{x}_{ij}^{(\mathrm{noise})}$ is obtained as an M'-channel signal and the observed noisy speech $\boldsymbol{x}_{ij}^{(\mathrm{mix})}$ is an M-channel signal, which contains one speech source. The cost function of BS-ILRMA is defined as the sum of two costs of ILRMA as follows:

$$\begin{aligned}\mathcal{L} =& \frac{1}{M'}\left\{\sum_{m'=1}^{M'}\sum_{i,j}\left[\frac{|y_{ij,m'}^{(\mathrm{noise})}|^2}{\sum_l t_{il,m'}v_{lj,m'}^{(\mathrm{noise})}} + \log\sum_l t_{il,m'}v_{lj,m'}^{(\mathrm{noise})}\right]\right.\\ &\left. -2J\sum_i \log|\det \boldsymbol{W}_i^{(\mathrm{noise})}|\right\}\\ &+\frac{1}{M}\left\{\sum_{m'=1}^{M'}\sum_{i,j}\left[\frac{|y_{ij,m'}^{(\mathrm{mix})}|^2}{\sum_l t_{il,m'}v_{lj,m'}^{(\mathrm{mix})}} + \log\sum_l t_{il,m'}v_{lj,m'}^{(\mathrm{mix})}\right]\right.\\ &+\sum_{i,j}\left[\frac{|y_{ij,M}^{(\mathrm{mix})}|^2}{\sum_l t_{il,M}v_{lj,M}^{(\mathrm{mix})}} + \log\sum_l t_{il,M}v_{lj,M}^{(\mathrm{mix})}\right]\\ &\left. -2J\sum_i \log|\det \boldsymbol{W}_i^{(\mathrm{mix})}|\right\},\end{aligned} \tag{7}$$

where $y_{ij,m'}^{(\mathrm{noise})} = \boldsymbol{w}_{i,m'}^{(\mathrm{noise})\mathsf{H}}\boldsymbol{x}_{i,j}^{(\mathrm{noise})}$ and $y_{ij,m'}^{(\mathrm{mix})} = \boldsymbol{w}_{i,m'}^{(\mathrm{mix})\mathsf{H}}\boldsymbol{x}_{i,j}^{(\mathrm{mix})}$ are the elements of $\boldsymbol{Y}_{m'}^{(\mathrm{noise})}$ and $\boldsymbol{Y}_{m'}^{(\mathrm{mix})}$, $t_{il,m'}$ and $t_{il,M}$ are the elements of $\boldsymbol{T}_{m'}$ and $\boldsymbol{T}_M$, $v_{lj,m'}^{(\mathrm{noise})}$ and $v_{lj,m}^{(\mathrm{mix})}$ are the elements of $\boldsymbol{V}_{m'}^{(\mathrm{noise})}$ and $\boldsymbol{V}_m^{(\mathrm{mix})}$, and $\boldsymbol{w}_{i,m'}^{(\mathrm{noise})}$ and $\boldsymbol{w}_{i,m}^{(\mathrm{mix})}$ are the row vectors of $\boldsymbol{W}_i^{(\mathrm{noise})}$ and $\boldsymbol{W}_i^{(\mathrm{mix})}$.

The update rules of unshared parameters, namely, $\boldsymbol{W}_i^{(\mathrm{noise})}$, $\boldsymbol{W}_i^{(\mathrm{mix})}$, $\boldsymbol{T}_M$, $\boldsymbol{V}_{m'}^{(\mathrm{noise})}$, and $\boldsymbol{V}_m^{(\mathrm{mix})}$ are the same as those in [4, 6]. In this paper, we only derive the update rule of the shared bases $t_{il,m'}$. Since it is difficult to directly minimize (7) w.r.t. $t_{il,m'}$, we design a majorization function of (7) and minimize it. Since the first and fourth terms in (7) are convex functions for $t_{il,m'}$, we apply Jensen's inequality to them for obtaining their upper bound functions using an auxiliary variable $\alpha_{ij,m'l}^{(*)} \geq 0$ that satisfies $\sum_l \alpha_{ij,m'l}^{(*)} = 1$:

$$\frac{1}{\sum_l t_{il,m'}v_{lj,m'}^{(*)}} \leq \sum_l \frac{{\alpha_{ij,m'l}^{(*)}}^2}{t_{il,m'}v_{lj,m'}^{(*)}}, \tag{8}$$

where $* = \{\mathrm{noise}, \mathrm{mix}\}$. Also, the second and fifth terms in (7) are concave functions, and we apply the tangent line inequality to them using an auxiliary variable $c_{ij,m'}^{(*)} \geq 0$ as

$$\begin{aligned}&\log\sum_l t_{il,m'}v_{lj,m'}^{(*)}\\ &\leq \frac{1}{c_{ij,m'}^{(*)}}\left(\sum_l t_{il,m'}v_{lj,m'}^{(*)} - c_{ij,m'}^{(*)}\right) + \log c_{ij,m'}^{(*)}.\end{aligned} \tag{9}$$

The equality of (9) and (10) holds if and only if the auxiliary variables are set as follows:

$$\alpha_{ij,m'l}^{(*)} = \frac{t_{il,m'}v_{lj,m'}^{(*)}}{\sum_{l'} t_{il',m'}v_{l'j,m'}^{(*)}}, \tag{10}$$

$$c_{ij,m'}^{(*)} = \sum_l t_{il,m'}v_{lj,m'}^{(*)}. \tag{11}$$

We can obtain the majorization function of (7) using (8) and (9) as

$$\begin{aligned}\mathcal{L} \leq& \mathcal{L}^+\\ \overset{\mathrm{c}}{=}& \sum_{i,j,m'}\left[\frac{1}{M'}\left(\sum_l \frac{|y_{ij,m'}^{(\mathrm{noise})}|^2\alpha_{ij,m'l}^{(\mathrm{noise})}}{t_{il,m'}v_{lj,m'}^{(\mathrm{noise})}} + \frac{\sum_l t_{il,m'}v_{lj,m'}^{(\mathrm{noise})}}{c_{ij,m'}^{(\mathrm{noise})}}\right)\right.\\ &\left.+\frac{1}{M}\left(\sum_l \frac{|y_{ij,m'}^{(\mathrm{mix})}|^2\alpha_{ij,m'l}^{(\mathrm{mix})}}{t_{il,m'}v_{lj,m'}^{(\mathrm{mix})}} + \frac{\sum_l t_{il,m'}v_{lj,m'}^{(\mathrm{mix})}}{c_{ij,m'}^{(\mathrm{mix})}}\right)\right],\end{aligned} \tag{12}$$

where $\overset{\mathrm{c}}{=}$ denotes equality up to a constant term. The update rule of $t_{il,m'}$ is derived by setting the gradient of $\mathcal{L}^+$ to zero and substituting (10) and (11), as

$$\boldsymbol{T}_{m'} \leftarrow \boldsymbol{T}_{m'} \otimes \sqrt{\frac{\frac{1}{M'}\left(\frac{|\boldsymbol{Y}_{m'}^{(\mathrm{noise})}|^{\cdot 2}}{\boldsymbol{G}_{m'}^{(\mathrm{noise})\cdot 2}}\right)\boldsymbol{V}_{m'}^{(\mathrm{noise})\mathsf{T}} + \frac{1}{M}\left(\frac{|\boldsymbol{Y}_{m'}^{(\mathrm{mix})}|^{\cdot 2}}{\boldsymbol{G}_{m'}^{(\mathrm{mix})\cdot 2}}\right)\boldsymbol{V}_{m'}^{(\mathrm{mix})\mathsf{T}}}{\frac{1}{M'}\boldsymbol{G}_{m'}^{(\mathrm{noise})\cdot -1}\boldsymbol{V}_{m'}^{(\mathrm{noise})\mathsf{T}} + \frac{1}{M}\boldsymbol{G}_{m'}^{(\mathrm{mix})\cdot -1}\boldsymbol{V}_{m'}^{(\mathrm{mix})\mathsf{T}}}}, \tag{13}$$

where $\otimes$ and the quotient symbol for matrices denote element-wise multiplication and division, and $\boldsymbol{G}_{m'}^{(*)} = \boldsymbol{T}_{m'}\boldsymbol{V}_{m'}^{(*)}$.

4. Experimental evaluation

4.1 Conditions

We compared the ego-noise reduction performance of simple unsupervised ILRMA, naive semi-supervised ILRMA (SS-ILRMA) described in Sect. 3.1, and proposed semi-supervised BS-ILRMA using the actual recorded signals by the hose-shaped rescue robot. The number of microphones used in the experiment was $M = 8$, namely, we assumed that $M' = M - 1 = 7$ ego-noise sources are mixed in the observed noisy speech signal $\boldsymbol{x}_{ij}^{(\mathrm{mix})}$. To produce the ego-noise sample and the observed mixture signals, the robot was driven in a simulated disaster site, and the speech signal was convolved with impulse responses from a survivor to each microphone with $\mathrm{SNR} = -20$ dB. We used the signal-to-distortion ratio (SDR) [15] as an evaluation measure, which indicates total quality of

Table 1: Experimental conditions

Sampling frequency	16 kHz
Window length in STFT	2048 samples
Window shift length in STFT	512 samples
Number of bases in each bases matrix	L=5,10 ,15

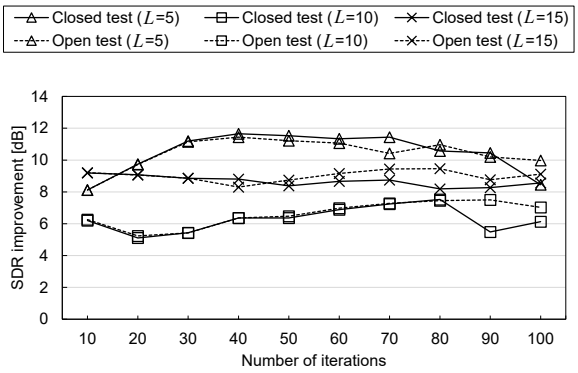

Figure 4: SDR improvements of different number of bases.

source separation. Table 1 shows the other experimental conditions.

In this experiment, we used two semi-supervised situations related to a sample sound of ego-noise: one is a *closed test* using the same ego-noise signal in both the sample sound and the observed signals, and the other one is an *open test* using the different ego-noise sample recorded by a different posture of the same robot. The signal length of the ego-noise sample and observed signals was five seconds.

4.2 Results

4.2.1 Optimal number of NMF bases for BS-ILRMA

We first investigate the optimal number of bases L for the ego-noise reduction with BS-ILRMA. Figure 4 shows the result of BS-ILRMA with various numbers of bases, where the horizontal axis indicates the number of iterations in optimization. We can confirm that BS-ILRMA with $L = 5$ achieves the best separation performance in open test. This fact means that when we increase the number of bases, shared ego-noise basis matrices have a risk to represent not only the ego-noise components but also the speech components.

4.2.2 Comparison with conventional and proposed methods

Figure 5 shows the comparison of SDR improvements achieved by ILRMA, SS-ILRMA, and BS-ILRMA, where the number of bases L was set to five for all the methods. From this result, the proposed method outperforms the other methods for both closed and open tests. As described in Sect. 3.1, SS-ILRMA cannot achieve better separation performance after several iterations because the scale ambiguity collapses the structures of supervised bases during the iterative optimization. In contrast, the separation result of BS-ILRMA becomes higher along with the iterative optimization.

5. Conclusions

In this paper, we proposed a novel semi-supervised extension of ILRMA that shares the NMF basis between sample and observed signals. This method is applied to the ego-noise reduction task for the hose-shaped rescue robot. The experimental results show the efficacy of the proposed method.

Acknowledgment

This work was supported by the Japan Science and Technology Agency and Impulsing Paradigm Change through Distributive

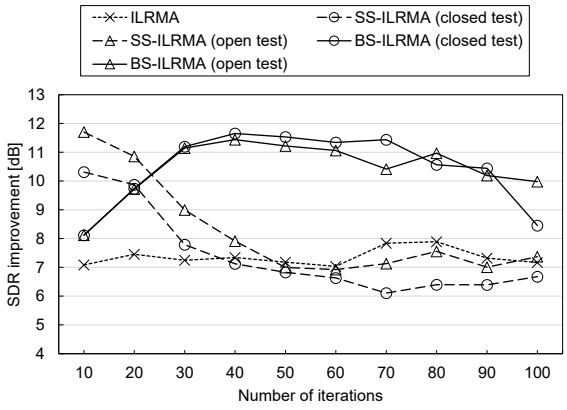

Figure 5: SDR improvements according to methods.

Technologies Program (ImPACT) designed by the Council for Science, Technology and Innovation, and partly supported by SECOM Science and Technology Foundation and JSPS KAKENHI Grant Number JP17H06572.

References

[1] Impulsive Paradigm Change through Distributed Technologies Program (ImPACT), http://www.jst.go.jp/impact/program07.html.

[2] H. Namari, K. Wakana, M. Ishikura, M. Konyo and S. Tadokoro, "Tube-type active scope camera with high mobility and practical functionality," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3679–3686, 2012.

[3] M. Takakusaki, D. Kitamura, N. Ono, T. Yamada, S. Makino, and H. Saruwatari, "Ego-noise reduction for a hose-shaped rescue robot using determined rank-1 multichannel nonnegative matrix factorization," *Proc. IWAENC*, 2016.

[4] 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. 23, no. 4, pp. 654–669, 2015.

[5] D. Kitamura, N. Ono, and H. Saruwatari, "Experimental analysis of optimal window length for independent low-rank matrix analysis," *Proc. EUSIPCO*, pp. 1210–1214, 2017.

[6] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," *Audio Source Separation*, Shoji Makino, Ed. Springer, 31 pages (to appear).

[7] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[8] 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.

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

[10] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," *Proc. ICA*, pp. 722–727, 2001.

[11] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Proc. Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[12] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

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

[14] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, and K. Kondo, "Music signal separation based on supervised nonnegative matrix factorization with orthogonality and maximum-divergence penalties," *IEICE Trans. Fundamentals*, vol. E97-A, no. 5, pp. 1113–1118, 2014.

[15] 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.