ICA 2016 BUENOS AIRES
Buenos Aires – 5 to 9 September 2016
Acoustics for the 21st Century...

PROCEEDINGS of the 22nd International Congress on Acoustics

Acoustics Array Systems: Paper ICA2016-312

Flexible microphone array based on multichannel nonnegative matrix factorization and statistical signal estimation

Hiroshi Saruwatari[(a)], **Kazuma Takata**[(a)], **Nobutaka Ono**[(b)], **Shoji Makino**[(c)]

[(a)]The University of Tokyo, Japan, hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp
[(b)]National Institute of Informatics, Japan, onono@nii.ac.jp
[(c)]University of Tsukuba, Japan, maki@tara.tsukuba.ac.jp

Abstract

In this paper, we propose a novel source separation method for the hose-shaped rescue robot based on multichannel nonnegative matrix factorization (MNMF) and statistical speech enhancement. The rescue robot is aimed to detect victims' speech in a disaster area, wearing multiple microphones around the body. Different from the common microphone array, the positions of microphones are unknown, and the conventional beamformer cannot be utilized. In addition, the vibration noise (ego-noise) is generated when the robot moves, yielding the serious contamination in the observed signals. Therefore, it is important to eliminate the ego-noise in this system. Blind source separation is a technique taken to separately estimate the sources without knowing the sensors' positions. Several methods, e.g., independent component analysis, independent vector analysis, and spatially rank-1 MNMF (Rank-1 MNMF) have been proposed so far, but their separation performance is not sufficient. To address this problem, in this study, first, supervised Rank-1 MNMF is proposed, thanks to the stationary properties of the ego-noise, where we train spectral bases of the ego-noise in advance. Secondly, to reduce the mismatch problem between the trained bases and the spectrogram in observed data, we propose an algorithm that an all-pole model is estimated to deform the bases using the reliable spectral components sampled by the statistical signal enhancement method. Thirdly, we propose to initialize Rank-1 MNMF by using the low-rank representation of the estimated speech spectrogram, and improve the convergence. Finally, we reveal that the proposed method outperforms the conventional methods in the source separation accuracy via experiments with actual sounds observed in the rescue robot.

Keywords: Microphone array, Source separation, NMF, Statistical signal estimation, Robot

Flexible microphone array based on multichannel nonnegative matrix factorization and statistical signal estimation

1 Introduction

In this paper, we propose a novel source separation method for the hose-shaped rescue robot based on multichannel nonnegative matrix factorization (MNMF) [1, 2] and statistical speech enhancement. The rescue robot is aimed to detect victims' speech in a disaster area, wearing multiple microphones around the body (see Fig. 1). Different from the common microphone array, the positions of microphones are unknown, and the conventional beamformer cannot be utilized. In addition, the vibration noise (ego-noise) is generated when the robot moves, yielding the serious contamination in the observed signals. Therefore, it is important to eliminate the ego-noise in this system.

Blind source separation is a technique taken to separately estimate the sources without knowing the sensors' positions. Several methods, e.g., independent component analysis (ICA) [3, 4, 5, 6], independent vector analysis (IVA) [7, 8], and spatially rank-1 MNMF (Rank-1 MNMF) [9, 10, 11] have been proposed so far (see Fig. 2 for their advantages and drawbacks). However, their separation performance is not sufficient, especially for the purpose of actual acoustic sound separation. To address this problem, in this study, first, supervised Rank-1 MNMF is proposed, thanks to the stationary properties of the ego-noise, where we train spectral bases of the ego-noise in advance.

Secondly, to reduce the mismatch problem between the trained bases and the spectrogram in observed data, we propose an algorithm that an all-pole model is estimated to deform the bases using the reliable spectral components sampled by the statistical signal enhancement method. Also, we propose to initialize Rank-1 MNMF by using the low-rank representation of the estimated speech spectrogram, and improve the convergence.

Finally, we reveal that the proposed method outperforms the conventional methods in the source separation accuracy via experiments with actual sounds observed in the rescue robot.

2 Preliminaries and related Works

2.1 Sound mixing model

The number of sources and that of microphones are assumed to be M. We represent multichannel sound source signals, observed signals, separated signals in each time-frequency slot as follows:

$$\boldsymbol{s}_{\omega,t} = [s_{\omega,t,1}, s_{\omega,t,2}, \cdots, s_{\omega,t,M}]^{\mathrm{T}}, \tag{1}$$

$$\boldsymbol{x}_{\omega,t} = [x_{\omega,t,1}, x_{\omega,t,2}, \cdots, x_{\omega,t,M}]^{\mathrm{T}}, \tag{2}$$

$$\boldsymbol{y}_{\omega,t} = [y_{\omega,t,1}, y_{\omega,t,2}, \cdots, y_{\omega,t,M}]^{\mathrm{T}}, \tag{3}$$

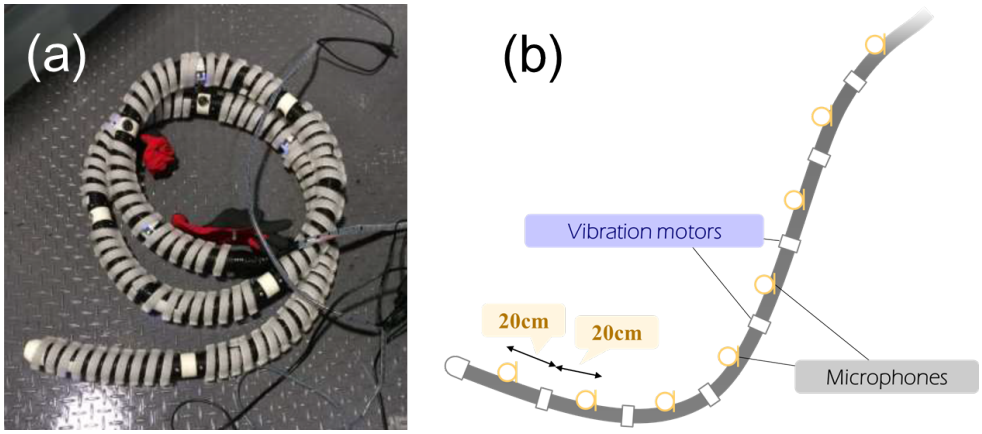

Figure 1: **(a) Overview of hose-shaped rescue robot, and (b) its location of microphones.**

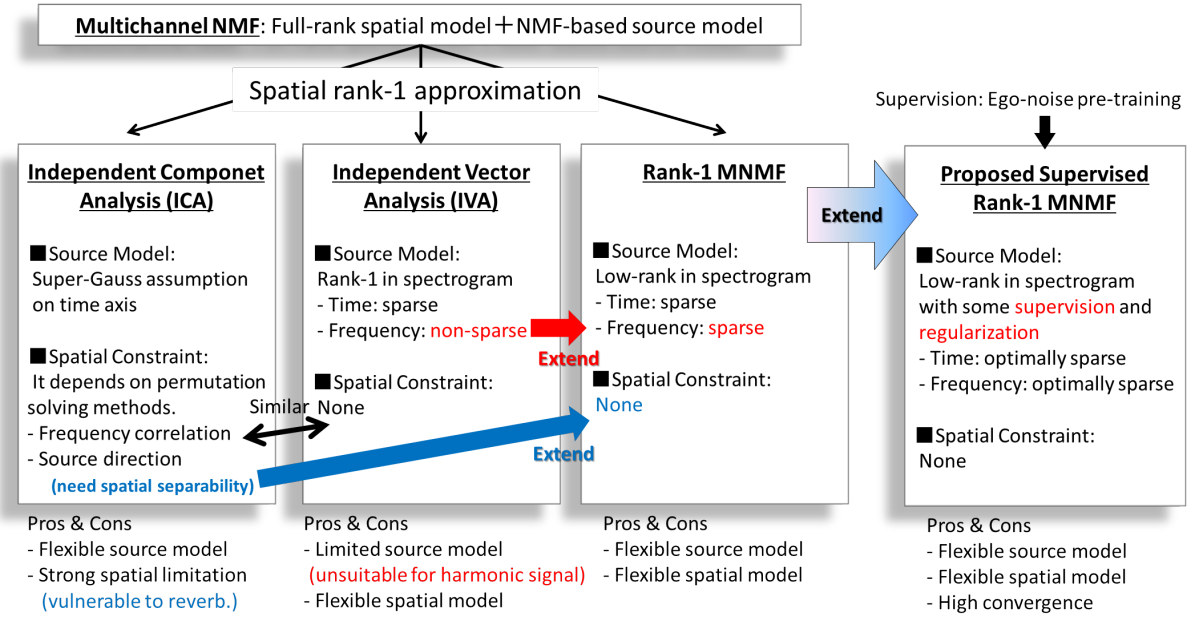

Figure 2: **Relationship between typical source separation algorithms.**

where $1 \leq \omega \leq \Omega$ and $1 \leq t \leq T$ denote the frequency and time indexes. Here we can express the observed signal as

$$\boldsymbol{x}_{\omega,t} = \boldsymbol{A}_{\omega}\boldsymbol{s}_{\omega,t}, \tag{4}$$

where $\boldsymbol{A}_{\omega}$ is called the *mixing matrix*.

2.2 Blind source separation

If we know the mixing matrix and its inverse, the separated signal is given by

$$\boldsymbol{y}_{\omega,t} = \boldsymbol{W}_{\omega}\boldsymbol{x}_{\omega,t}, \tag{5}$$

where $\boldsymbol{W}_{\omega} = \boldsymbol{A}_{\omega}^{-1}$ is referred to as the *demixing matrix*.

To blindly estimate the demixing matrix only from the observed signal, several methods have been proposed so far, e.g., ICA, IVA, and Rank-1 MNMF. In this study, we introduce Rank-1 MNMF, which models each sound source spectrogram as *low-rank nonnegative matrix* and decomposes the sources on the basis of their independence nature. Thus, this method can also be referred to as *independent low-rank matrix analysis*. For more detail algorithm, see [11].

2.3 Informed source separation

In the application of robot audition, we can often obtain the prototype of the ego-noise signal that can be used as training data in advance. This property is very suitable for embedding the supervision spectral bases into Rank-1 MNMF, yielding the rapid convergence of the algorithm. A priori ego-noise basis training is carried out via NMF, expressed as

$$\boldsymbol{S}_{\text{noise}} \simeq \boldsymbol{F}\boldsymbol{G}, \tag{6}$$

where $\boldsymbol{S}_{\text{noise}}$ is a nonnegative matrix that represents an amplitude spectrogram of the specific signal used for training, $\boldsymbol{F}$ is a nonnegative matrix that comprises the basis vectors of the ego-noise signal as column vectors, and $\boldsymbol{G}$ is a nonnegative matrix that corresponds to the activation of each basis vector of $\boldsymbol{F}$. Therefore, the basis matrix $\boldsymbol{F}$ is constructed by the supervision of the ego-noise signal, and embedded into Rank-1 MNMF as a part of ego-noise source model.

3 Proposed method

3.1 Overview of proposed method

One inherent problem of informed source separation is a mismatch between the trained basis $\boldsymbol{F}$ and real-world ego-noise confronted with the robot. Thus, it is necessary to adapt the supervised basis to the real ego-noise signal spectrogram to deal with real environmental sounds. However, it is difficult for Rank-1 MNMF to perform optimal basis deformation because it optimizes the deformation and separation simultaneously. In this paper, we propose a new method introducing the following schemes. (a) Apart from the source separation process, the basis deformation process is separately carried out with a linear time-invariant filter, namely an all-pole model, that consists of fewer parameters. (b) The parameters of the all-pole model can be optimized by utilizing “sampled convincing target components" obtained by a generalized minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator [12].

First, we perform Rank-1 MNMF with a current supervised basis $\boldsymbol{F}$. Second, using the generalized MMSE-STSA estimator with an estimated extra components of ego-noise signal, $\boldsymbol{Y}_{\text{mix}} -$

$\boldsymbol{FG}$, we obtain an estimated ego-noise signal $\boldsymbol{Y}$ and a binary mask $\boldsymbol{I}$ that extracts seldom overlapping components with the target speech signal from the estimated ego-noise signal $\boldsymbol{Y}$. Finally, we deform the original supervised basis $\boldsymbol{F}_{\text{org}}$ and update $\boldsymbol{F}$ as a deformed basis. After some iterations of the procedures, we conduct Rank-1 MNMF using the deformed basis and obtain the improved separation.

3.2 Convincing component sampler using statistical spectral amplitude estimator

The generalized MMSE-STSA estimator calculates the spectrum gain $\boldsymbol{J}$ that minimizes the average squared error between the true ego-noise signal and the estimated signal given the a priori probability distribution of the ego-noise signal. This process is expressed as follows:

$$\boldsymbol{Y} = \boldsymbol{J} \circ \boldsymbol{Y}_{\text{mix}}, \tag{7}$$

$$J_{\omega,t} = \frac{\sqrt{v_{\omega,t}}}{\tilde{\gamma}_{\omega,t}} \cdot \left(\frac{\Gamma(\rho+0.5)}{\Gamma(\rho)} \cdot \frac{\Phi(0.5-\rho,1,-v_{\omega,t})}{\Phi(1-\rho,1,-v_{\omega,t})} \right)^{1/\beta}, \tag{8}$$

where $\boldsymbol{Y}$ is the ego-noise signal estimated by the generalized MMSE-STSA estimator, $\circ$ is a Hadamard product, $J_{\omega,t}$ is an element of $\boldsymbol{J}$, $\Gamma(\cdot)$ is the gamma function, $\Phi(a,b;k) = F_1(a,b;k)$ is the confluent hypergeometric function, β is the amplitude compression parameter, and ρ is the shape parameter of the chi-squared distribution used as the prior distribution of the ego-noise signal. In addition, $v_{\omega,t}$ is defined using an a priori SNR $\tilde{\varepsilon}_{\omega,t}$ and a posteriori SNR $\tilde{\gamma}_{\omega,t}$ as

$$v_{\omega,t} = \tilde{\gamma}_{\omega,t} \tilde{\varepsilon}_{\omega,t} \left(1 + \tilde{\varepsilon}_{\omega,t}\right)^{-1}. \tag{9}$$

In the generalized MMSE-STSA estimator, it is necessary to obtain the power spectrum of the nontarget signal to calculate $\tilde{\gamma}_{\omega,t}$. In this study, we use $\boldsymbol{Y}_{\text{mix}} - \boldsymbol{FG}$ for this purpose. In addition, we use the method proposed in [13] to estimate ρ.

3.3 Basis deformation with all-pole model using generalized MMSE-STSA estimator

In this section, we propose basis deformation with an all-pole model controlled by the generalized MMSE-STSA estimator. Note that the basic idea has been introduced to describe a spectral mismatch in a music signal [14]. However, to the best of our knowledge, this method is the first approach to apply the model to the basis deformation problem for robot ego-noise.

In our method, we calculate the trained supervision and deform the basis $\boldsymbol{F}_{\text{org}}$ with reference to the estimated ego-noise signal $\boldsymbol{Y}$. Since the estimated ego-noise signal $\boldsymbol{Y}$ still has low accuracy, it is necessary to extract only a sufficient number of reliable components to deform the basis correctly. Otherwise, the basis deforms excessively and cannot accomplish the separation. Therefore, to avoid this, the thresholding of the spectrum gain $\boldsymbol{J}$ used to extract seldom overlapping components with the speech signal is introduced. In addition, although the few components are sampled by the thresholding that yields many blanks in the spectrogram, they are still sufficient to decide the all-pole model because the model has the time-invariant and frequency-interpolation properties. The above-mentioned concepts are described as

$$\boldsymbol{I} \circ \boldsymbol{Y} \simeq \boldsymbol{I} \circ (\boldsymbol{A}\boldsymbol{F}_{\text{org}}\boldsymbol{G}), \tag{10}$$

where $\boldsymbol{I}$ is an $\Omega \times T$ binary mask matrix with entries $i_{\omega,t}$, which was obtained from the spectrum gain matrix $\boldsymbol{J}$ of the generalized MMSE-STSA estimator, the entries of which were subjected to thresholding (e.g., if $J_{\omega,t} > 0.8$, then $i_{\omega,t} = 1$; otherwise $i_{\omega,t} = 0$). In addition, $\boldsymbol{A}$ is a diagonal matrix in which the diagonal elements are described using the all-pole model. The elements of $\boldsymbol{A}$ are described as

$$A_{\omega,\omega} = \frac{1}{|1 - \sum_{k=1}^{p} \alpha_k \exp(-\pi j k \frac{\omega}{\Omega})|}, \tag{11}$$

where p is the order and α_k are the coefficients of the all-pole model. In addition, we define $A_\omega = 1 - \sum_{k=1}^{p} \alpha_k \exp(-\pi j k \frac{\omega}{\Omega})$ to simplify the calculations.

3.4 Cost function and update rule

The cost function for (10) based on the generalized KL divergence is given by

$$\mathscr{J} = \sum_{\omega,t} i_{\omega,t} \Big\{ -y_{\omega,t} + \frac{\sum_k f_{\omega,k} g_{k,t}}{|A_\omega|} + y_{\omega,t} \log \frac{y_{\omega,t}}{\sum_k f_{\omega,k} g_{k,t} / |A_\omega|} \Big\}, \tag{12}$$

where $y_{\omega,t}$, $f_{\omega,k}$, and $g_{k,t}$ are the nonnegative elements of matrices $\boldsymbol{Y}$, $\boldsymbol{F}_{\text{org}}$, and $\boldsymbol{G}$, respectively. Since it is difficult to analytically derive the optimal $\boldsymbol{A}$ and $\boldsymbol{G}$, we define an auxiliary function that represents the upper bound of $\mathscr{J}$, as described below. First, applying Jensen's inequality to $\log \sum_k f_{\omega,k} g_{k,t}$ and the tangent inequality to $\log |A_\omega| = 1/2 \log |A_\omega|^2$, we have

$$\mathscr{J} \leq \sum_{\omega,t} i_{\omega,t} \Big\{ \frac{\sum_k f_{\omega,k} g_{k,t}}{|A_\omega|} + y_{\omega,t} \big(\frac{1}{2\rho_\omega} |A_\omega|^2 - \sum_k \zeta_{\omega,t,k} \log \frac{f_{\omega,k} g_{k,t}}{\zeta_{\omega,t,k}} \big) + C_{\omega,t} \Big\}, \tag{13}$$

where $C_{\omega,t}$ are unnecessary constants when calculating the update rules of the activation matrix $\boldsymbol{G}$ and the all-pole-model weight matrix $\boldsymbol{A}$, and ρ_ω and $\zeta_{\omega,t,k}$ are auxiliary variables. The equality in (13) holds if and only if the auxiliary variables are set to $\rho_\omega = |A_\omega|^2$ and $\zeta_{\omega,t,k} = f_{\omega,k} g_{k,t} / \sum_k f_{\omega,k} g_{k,t}$. Second, to make the auxiliary function a quadratic form of $|A_\omega|$, we conduct a Taylor expansion around τ_ω,

$$\begin{aligned} \mathscr{J} \leq & \sum_{\omega,t} i_{\omega,t} \Big\{ \sum_k f_{\omega,k} g_{k,t} \big(\frac{1}{\tau_\omega^3} |A_\omega|^2 - 3 \frac{1}{\tau_\omega^2} |A_\omega| + \frac{3}{\tau_\omega} \big) + \\ & y_{\omega,t} \big(\frac{1}{2\rho_\omega} |A_\omega|^2 - \sum_k \zeta_{\omega,t,k} \log \frac{f_{\omega,k} g_{k,t}}{\zeta_{\omega,t,k}} \big) + C_{\omega,t} \Big\}. \end{aligned} \tag{14}$$

The equality of (14) holds if and only if $\tau_\omega = |A_\omega|$. This approximation does not meet the condition of an auxiliary function, but if τ_ω is updated as $|A_\omega|$, this approximation is equivalent to Newton's method. Finally, using the inequality $\mathscr{R}e[\theta_\omega^* A_\omega] \leq |A_\omega|$, we can define the upper bound function $\mathscr{J}^+$ for $\mathscr{J}$ as

$$\begin{aligned} \mathscr{J} \leq & \sum_{\omega,t} i_{\omega,t} \Big\{ \sum_k f_{\omega,k} g_{k,t} \big(\frac{1}{\tau_\omega^3} |A_\omega|^2 - 3 \frac{1}{\tau_\omega^2} \mathscr{R}e[\theta_\omega^* A_\omega] + \frac{3}{\tau_\omega} \big) \\ & + y_{\omega,t} \big(\frac{1}{2\rho_\omega} |A_\omega|^2 - \sum_k \zeta_{\omega,t,k} \log \frac{f_{\omega,k} g_{k,t}}{\zeta_{\omega,t,k}} \big) + C_{\omega,t} \Big\}, \end{aligned} \tag{15}$$

where $\mathscr{R}e[\cdot]$ is a real part of $\cdot$ and $|\theta_\omega| = 1$. The equality of (15) holds if and only if $\theta_\omega = A_\omega / |A_\omega|$.

3.4.1 Multiplicative update rule for activation matrix G

The update rule for $\mathscr{J}^+$ with respect to the activation matrix $\boldsymbol{G}$ is determined by setting the gradient to zero. From $\partial \mathscr{J}^+ / \partial g_{k,t} = 0$, we obtain

$$\sum_{\omega} i_{\omega,t} \left\{ f_{\omega,k} \left(\frac{1}{\tau_{\omega}^3} |A_{\omega}|^2 - 3 \frac{1}{\tau_{\omega}^2} \mathscr{R}e[\theta_{\omega}^* A_{\omega}] + \frac{3}{\tau_{\omega}} \right) + y_{\omega,t} \left(-\zeta_{\omega,t,k} g_{k,t}^{-1} \right) \right\} = 0. \tag{16}$$

By substituting the auxiliary variables into (16) and simplifying it, we obtain the multiplicative update rule of $g_{k,t}$ as

$$g_{k,t} \leftarrow g_{k,t} \frac{\sum_{\omega} i_{\omega,t} y_{\omega,t} f_{\omega,k} / (\sum_{\kappa} f_{\omega,\kappa} g_{\kappa,t})}{\sum_{\omega} i_{\omega,t} f_{\omega,k} / |A_{\omega}|}. \tag{17}$$

3.4.2 Multiplicative update rule for all-pole-model weight matrix A

First, by differentiating $\mathscr{J}^+$ partially with respect to α_q and setting it to zero, we obtain

$$\begin{aligned}
&\sum_{k=1}^{p} \alpha_k \sum_{\omega,t} \left[i_{\omega,t} \left(\sum_k f_{\omega,k} g_{k,t} \frac{1}{\tau_{\omega}^3} + y_{\omega,t} \frac{1}{2\rho_{\omega}} \right) \left(\exp\left(-\pi j \frac{\omega}{\Omega} (k-q) \right) + \exp\left(\pi j \frac{\omega}{\Omega} (k-q) \right) \right) \right] \\
&- \sum_{\omega,t} i_{\omega,t} \left[\left(\sum_k f_{\omega,k} g_{k,t} \frac{1}{\tau_{\omega}^3} + y_{\omega,t} \frac{1}{2\rho_{\omega}} \right) \left(\exp(-\pi j \frac{\omega}{\Omega} q) + \exp(\pi j \frac{\omega}{\Omega} q) \right) - \frac{3}{\tau_{\omega}^2} \sum_k f_{\omega,k} g_{k,t} \mathscr{R}e[\theta_{\omega}^* \exp(-\pi j \frac{\omega}{\Omega} q))] \right] \\
&= 0,
\end{aligned} \tag{18}$$

where $1 \leq q \leq p$. Second, we define $\boldsymbol{R}$ and $\boldsymbol{r}$ as

$$R_{k,q} = \sum_{\omega,t} \left[i_{\omega,t} \left(\sum_k f_{\omega,k} g_{k,t} \frac{1}{\tau_{\omega}^3} + y_{\omega,t} \frac{1}{2\rho_{\omega}} \right) \left(\exp\left(-\pi j \frac{\omega}{\Omega} (k-q) \right) + \exp\left(\pi j \frac{\omega}{\Omega} (k-q) \right) \right) \right], \tag{19}$$

$$\begin{aligned}
r_q = \sum_{\omega,t} i_{\omega,t} \Bigg[&\left(\sum_k f_{\omega,k} g_{k,t} \frac{1}{\tau_{\omega}^3} + y_{\omega,t} \frac{1}{2\rho_{\omega}} \right) \left(\exp(-\pi j \frac{\omega}{\Omega} q) + \exp(\pi j \frac{\omega}{\Omega} q) \right) \\
&- \frac{3}{\tau_{\omega}^2} \sum_k f_{\omega,k} g_{k,t} \mathscr{R}e[\theta_{\omega}^* \exp(-2\pi j \frac{\omega}{\Omega} q))] \Bigg].
\end{aligned} \tag{20}$$

By substituting (19) and (20) into (18), we obtain

$$\boldsymbol{R}\boldsymbol{\alpha} = \boldsymbol{r}, \tag{21}$$

where $\boldsymbol{\alpha}$ is the vector of coefficients in the all-pole model. Since $\boldsymbol{R}$ is a Toeplitz matrix, we can derive $\boldsymbol{\alpha}$ using the Levinson–Durbin algorithm with a computationally efficient form.

3.5 Initialization of speech basis

In the previous subsections, we described the detail strategy of the deformation with regard to the ego-noise basis. In our method, we also propose to initialize Rank-1 MNMF by using the

low-rank representation of the estimated speech spectrogram, and improve the convergence. This can be accomplished by using the same methodology as the ego-noise enhancement, i.e., the generalized MMSE STSA estimator is applied to the speech signal candidate (not the ego-noise candidate) separated by Rank-1 MNMF, and we obtain more sparse representation. Then, we again set the sparse-aware speech basis into Rank-1 MNMF and restart the update of the demixing matrix.

4 Experimental evaluation

4.1 Experimental condition

To validate the efficacy of the proposed method, we conducted an experimental simulation based on the real apparatus with the hose-shaped robot shown in Fig. 1. The experimental conditions were set as follows.

The flexible robot had eight *location-unknown* microphones, which recorded an observed signals consisting of one speech signal and ego-noise. The target signal was imitated using clean male and female speech signals with real-recorded impulse responses from the source to each of the microphones. The multichannel ego-noise signals were independently recorded with the actual dynamics of the robot, and were added into the speech signals. The ego-noise signals were classified into two parts, i.e., (a) **matched**: this ego-noise signal was used for both initial basis training and separation test (for 2 patterns), and (b) **mismatched**: different ego-noise signals were independently used for basis training and separation test (for 3 patterns).

4.2 Results

The evaluation score of the separation performance is a signal-to-distortion ratio (SDR) via *BSSeval* [15], which indicates the total sound quality regarding separation accuracy and sound distortion. In this evaluation, we set the input SDRs of 0, -5, and -10 dB. As for the competitive methods, IVA [8], supervised NMF (SNMF) [16], simple Rank-1 MNMF are used.

Figure 3 shows the SDR scores for each of the methods, which are averaged over all experimental conditions. We can confirm that the proposed methods of both matched and mismatched cases outperform other conventional methods. The matched case is the best one because the same ego-noise signal can be used for basis training and separation, i.e., this corresponds to perfectly informed situation (but unrealistic). The mismatched case is more feasible situation and still gains certain SDR improvement, showing the proposed method's net efficacy.

5 Conclusions

In this paper, we proposed a new informed source separation method for the flexible microphone array system equipped in the hose-shaped rescue robot based on supervised Rank-1 MNMF and statistical speech enhancement. To reduce the mismatch problem between the

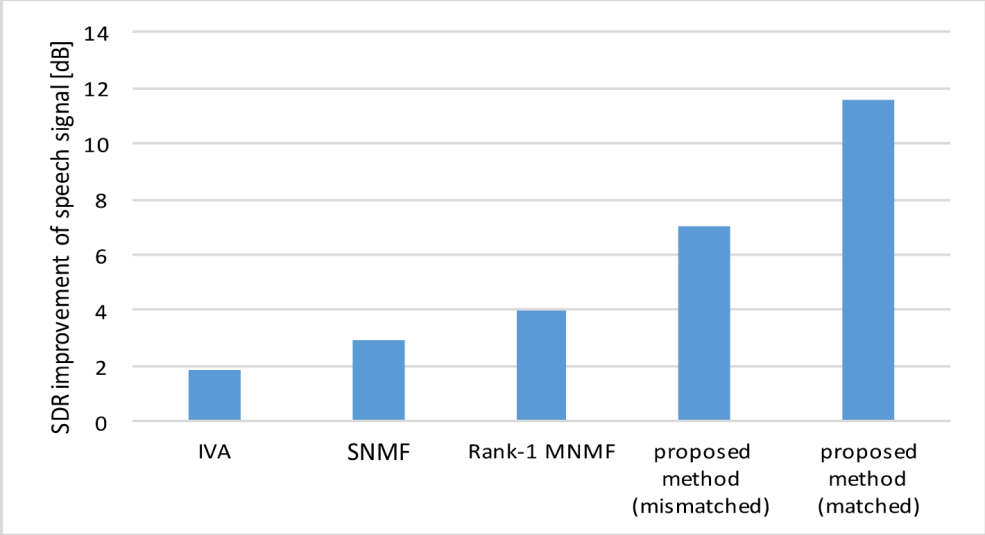

Figure 3: **SDR scores for each method, which are averaged with respect to each experimental condition.**

trained bases and the spectrogram in observed data, we proposed the algorithm that an all-pole model is estimated to deform the bases using the reliable spectral components sampled by the statistical signal enhancement method. We revealed that the proposed method outperforms the conventional methods via experiments with actual sounds in the rescue robot.

Acknowledgements The authors are grateful to Dr. Hiroshi G. Okuno of Waseda University, Dr. Katsutoshi Itoyama and Mr. Yoshiaki Bando of Kyoto University for their fruitful suggestions and discussions regarding this work. This work was supported by ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan), and SECOM Science and Technology Foundation.

References

[1] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550—563, 2010.

[2] D. Kitamura, H. Saruwatari, H. Kameoka, Y. Takahashi, K. Kondo, and S. Nakamura, "Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 654–669, 2015.

[3] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[4] 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. Speech and Audio Processing*, vol. 11, no. 2, pp. 109–116, 2003.

[5] H. Sawada, R. Mukai, S. Araki, S. Makino, “A robust and precise method for solving the permutation problem of frequency-domain blind source separation,” *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[6] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, K. Shikano, “Blind source separation based on a fast-convergence algorithm combining ICA and beamforming,” *IEEE Trans. Speech and Audio Processing*, vol. 14, no. 2, pp. 666-678, 2006.

[7] T. Kim, T. Eltoft and T.-W. Lee, “Independent vector analysis: an extension of ICA to multivariate components,” *Proc. International Conference on Independent Component Analysis and Blind Source Separation*, pp. 165–172, 2006.

[8] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 189–192, 2011.

[9] H. Kameoka, T. Yoshioka, M. Hamamura, J. Le Roux, K. Kashino, "Statistical model of speech signals based on composite autoregressive system with application to blind source separation," *Proc. 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2010)*, LNCS 6365, pp. 245–253, 2010.

[10] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model,” *Proc. ICASSP*, pp. 276–280, 2015.

[11] 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. Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016. (DOI: 10.1109/TASLP.2016.2577880).

[12] C. Breihaupt, M. Krawczyk, and R. Martin “Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," *Proc. ICASSP*, pp. 4037–4040, 2008.

[13] Y. Murota, D. Kitamura, S. Nakai, H. Saruwatari, S. Nakamura, K. Shikano, Y. Takahashi, K, Kondo, “Music signal separation based on bayesian spectral amplitude estimator with automatic target prior adaptation," *Proc. ICASSP*, pp. 7490–7494, 2014.

[14] H. Nakajima, D. Kitamura, N. Takamune, S. Koyama, H. Saruwatari, N. Ono, Y. Takahashi, and K. Kondo, “Music signal separation using supervised NMF with all-pole-model-based discriminative basis deformation,” *Proc. EUSIPCO*, 2016. (in printing)

[15] E. Vincent, R. Gribonval and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[16] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, 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.