Ego Noise Reduction for Hose-Shaped Rescue Robot Combining Independent Low-Rank Matrix Analysis and Noise Cancellation

Narumi Mae*, Daichi Kitamura†, Masaru Ishimura*, Takeshi Yamada*, and Shoji Makino*
* University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan
E-mail: {mae, ishimura}@mmlab.cs.tsukuba.ac.jp, takeshi@cs.tsukuba.ac.jp, maki@tara.tsukuba.ac.jp
† SOKENDAI (The Graduate University for Advanced Studies), Shonan Village, Hayama, Kanagawa 240-0193, Japan
E-mail: d-kitamura@nii.ac.jp

***Abstract*—In this paper, we present an ego noise reduction method for a hose-shaped rescue robot developed for search and rescue operations in large-scale disasters such as a massive earthquake. It can enter narrow and dark places covered with rubble in a disaster site and is used to search for disaster victims by capturing their voices with its microphone array. However, ego noises, such as vibration or fricative sounds, are mixed with the voices, and it is difficult to differentiate them from a call for help from a disaster victim. To solve this problem, we here propose a two-step noise reduction method as follows: (1) the estimation of both speech and ego noise signals from an observed multichannel signal by multichannel nonnegative matrix factorization (NMF) with the rank-1 spatial constraint, which was proposed by Kitamura *et al.*, and (2) the application of noise cancellation to the estimated speech signal using the noise reference. Our evaluations show that this approach is effective for suppressing ego noise.**

I. Introduction

It is important to develop robots for search and rescue operations in times of large-scale disasters such as earthquakes. Robots are required for emergency responses and for the restoration of disaster sites, which are difficult and dangerous tasks for humans. The "Tough Robotics Challenge" [1] is one of the research and development programs in the Impulsing Paradigm Change through Disruptive Technologies Program (ImPACT), whose aim is to develop five remote and autonomous robots. One of these robots is a hose-shaped rescue robot [2]. This robot is long and slim like a snake and makes it possible to investigate narrow spaces into which conventional remotely operable robots cannot enter. This robot searches for disaster victims by capturing their voices with a microphone array attached around itself at regular intervals. However, there is a serious problem of "ego noise". Ego noise is generated by the vibration motors used to move the robot by the vibrating cilia tape wrapped around the robot. Recently, many ego noise cancellation methods have been proposed [3]–[6]. This robot has reduced ego noise from the sound recorded by its microphone array compared with those in [7]–[8]. In addition, the many microphones on the hose-shaped rescue robot enable the application of the overdetermined source separation method. However, the microphone arrangement changes as the robot moves, making it difficult to control the microphone array geometry. Therefore, in our previous work [9], we applied independent vector analysis(IVA) [10]–[12] to the robot. However, IVA cannot capture the specific spectral structures of the sources. Thus, in this study, we apply determined rank-1 multichannel nonnegative matrix factorization [13]–[14] proposed by Kitamura *et al,* which can be interpreted as a method of independent low-rank matrix analysis (hereafter referred to as ILRMA) on the basis of the fact that the hose-shaped rescue robot has many microphones. We also apply a time-variant noise canceller to compensate for the time-invariant assumption of ILRMA. We examine the applicability of the proposed method for reducing ego noise.

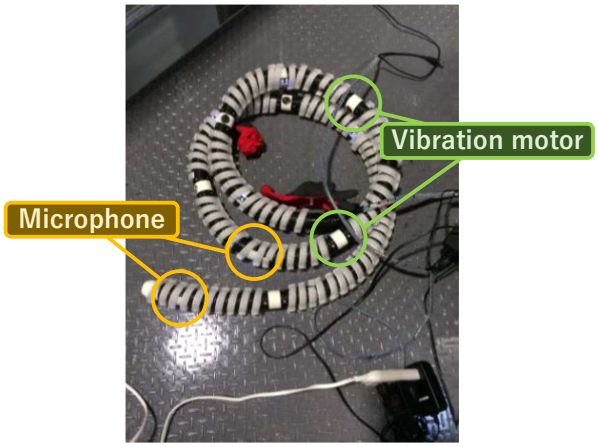

Fig. 1. Hose-shaped rescue robot.

II. Hose-Shaped Rescue Robot and Ego Noise

A. Hose-Shaped Rescue Robot

Figure 1 shows an image of the hose-shaped rescue robot and Fig. 2 shows its structure. The hose-shaped rescue robot basically consists of a hose as its axis with cilia tape wrapped around it; it moves forward slowly as a result of the friction between the cilia and floor through the vibration of the cilia

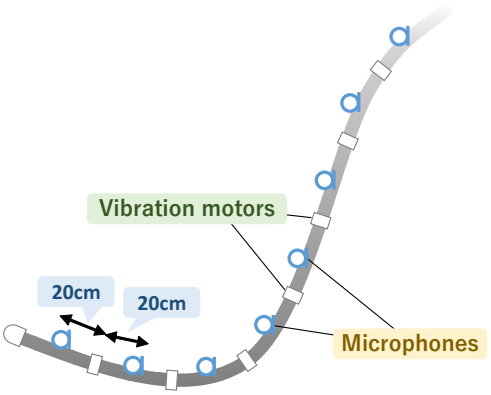

Fig. 2. Structure of hose-shaped rescue robot.

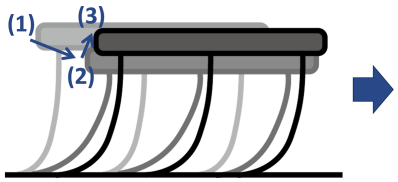

Fig. 3. Principle of movement of hose-shaped rescue robot [2].

tape induces by the vibration motors. Figure 3 schematically shows the principle of movement of the hose-shaped rescue robot. When the motors vibrate, state (1) changes to state (2) through the friction between the cilia and the floor, then state (2) changes into state (3) as a result the cilia slipping. The hose-shape rescue robot moves by repeating such changes in the state. It also performs various sensing functions using sensors such as microphones, cameras, an inertial measurement unit, and light sensors.

B. Problem in Recording Speech

Recording speech using the hose-shaped rescue robot has a serious problem. during the operation of the robot, very loud ego noise is mixed in the input to the microphones. The main sources of the ego noise are as follows:

- Driving sound of the vibration motors,
- Fricative sound generated between the cilia and floor,
- Noise generated by microphone vibration.

In an actual disaster site, the voice of a person seeking help may not be sufficiently loud to capture, and it may be smaller than the ego noise.

III. Conventional Method

Recently, many ego noise reduction methods have been proposed [3]–[5]. In [3], noise reduction based on the generalization of K-SVD was proposed, which can be used for an the underdetermined multichannel situation. Also, the authors of [4] and [5] have proposed a method of improving of the performances of ego noise reduction using an adaptive microphone array geometry. On the other hand, the many microphones on the rescue robot enable the application of an overdetermined source separation method. In a determined situation, IVA [10]–[12] is a commonly used method.

A. Formulation

We suppose that M sources are observed using M microphones (determined case). The multichannel source and the observed and separated signals in each time-frequency slot are as follows:

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

where $1 \leq i \leq I$ and $1 \leq j \leq J$ are indexes of frequency and time, and t denotes the vector transpose. All the entries of these vectors are complex values. When the window size in an STFT is sufficiently longer than the impulse response between source and microphone, we can approximately represent the observed signal as

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

Here, $\boldsymbol{A}_i = (\boldsymbol{a}_{i,1} \ \cdots \ \boldsymbol{a}_{i,M})$ is an $M \times M$ mixing matrix of the observed signals. When $\boldsymbol{W}_i = (\boldsymbol{w}_{i,1} \ \cdots \ \boldsymbol{w}_{i,M})^h$ denotes the demixing matrix, the separated signal $\boldsymbol{y}_{ij}$ is represented as

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

where ^{h} is the Hermitian transpose.

B. Independent Vector Analysis

IVA [10]–[12] is one of the techniques used to solve the permutation problem [15] and can be applied only in a determined situation. In this method, we define the source component as a vector that consists of all frequency bins, given as

$$\boldsymbol{y}_{j,m} = (y_{1j,m} \ \cdots \ y_{Ij,m}). \tag{6}$$

IVA can be used to estimate the demixing matrix $\boldsymbol{W}_i$ by assuming both independence between the sources (vectors) and a higher-order correlation between the frequency bins in each source. The cost function in IVA is defined as

$$Q(\boldsymbol{W}) = \sum_m \frac{1}{J} \sum_j G(\boldsymbol{y}_{j,m}) - \sum_i \log |\det \boldsymbol{W}_i|, \tag{7}$$

where J is the number of time frames and $G(\boldsymbol{y}_{j,m})$ is a contrast function. When $\boldsymbol{y}_{j,m}$ is a probability density function $p(\boldsymbol{y}_{j,m})$, the contrast function $G(\boldsymbol{y}_{j,m})$ is given as $-\log p(\boldsymbol{y}_{j,m})$. In IVA, $G(\boldsymbol{y}_{j,m}) = ||\boldsymbol{y}_{j,m}||_2$ is often used [12], where a spherical Laplace distribution is assumed for the source prior and $||\cdot||_2$ denotes the L_2 norm. For the minimization of (7), fast and stable update rules, which are derived by an auxiliary function technique, have been proposed [16].

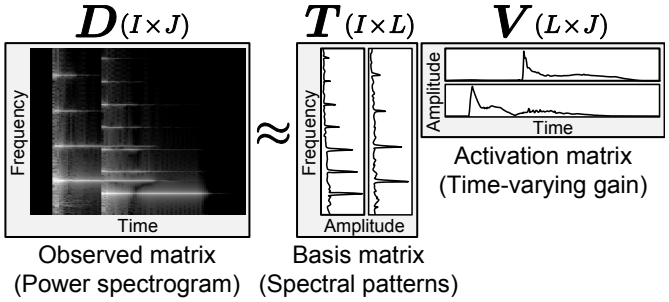

Fig. 4. Decomposition model of NMF.

IV. Proposed Method

IVA requires independence between the sources to estimate the demixing matrix. In general, in IVA, a spherical multivariate distribution is assumed as the source model to ensure a higher-order correlation between the frequency bins in all sources. However, this model does not include any particular information on sources, that is, IVA cannot capture specific spectral structures of the sources. Thus, the utilization of nonnegative matrix factorization (NMF) [17]–[19] as the source model has been proposed, which enables us to capture the spectral structures.

NMF decomposes a given spectrogram into several spectral bases $\boldsymbol{T}$ and temporal activations $\boldsymbol{V}$ as shown in Fig. 4, then the decomposed components are clustered into each separated source. Multichannel NMF (MNMF) [20]–[22] is one of the techniques for clustering the NMF bases and activations using a sourcewise spatial model. MNMF separately models the mixing system and the nonnegative power spectra of sources. However, this method is strongly dependent on its initial values because there are no constraints in the spatial models.

To solve the problem of MNMF, ILRMA [13], [14] was proposed, in which a rank-1 spatial model is introduces into MNMF [22]. This method estimates a demixing matrix while the represented source is determined using NMF bases, and can be optimized by the update rules of IVA and conventional single-channel NMF. Therefore, ILRMA is equivalent to a method that unifies IVA and NMF.

Since the hose-shaped rescue robot moves very slowly and the spatial locations of the sources and microphones barely change, we can assume a linear time-varying mixing system. In this case, IVA or ILRMA is effective for the separation because neither requires the locations of the sources and the microphones. In particular, ILRMA can efficiently capture the time-frequency structure of the ego noise because it repeatedly analyzes several types of similar spectra. The demixing filter in IVA or ILRMA is time-invariant over several seconds. To achieve time-variant noise reduction, in this study we apply a noise canceller for the post processing of ILRMA to reduce the remaining time-variant ego noise components. The noise canceller usually requires a reference microphone to observe only the noise signal. In this study, we utilize estimates of the noise estimates obtained by ILRMA as the noise reference signal.

A. Independent Low-Rank Matrix Analysis

We use ILRMA [13], [14] incorporate a rank-1 spatial model in MNMF [22]. Here, we explain the formulation and algorithm derived by Kitamura *et al.* MNMF is an extension of simple NMF for multichannel signals. The observed signals are represented as

$$\mathsf{X}_{ij} = \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^h, \tag{8}$$

where X_{ij} is the correlation matrix between the channels of size $M \times M$. The diagonal elements of X_{ij} represent real-valued powers detected by the microphones, and the non-diagonal elements represent the complex-valued correlations between the microphones. The separation model of MNMF $\hat{\mathsf{X}}_{ij}$ used to approximate X_{ij} is represented as

$$\mathsf{X}_{ij} \approx \hat{\mathsf{X}}_{ij} = \sum_m \mathsf{H}_{im} \sum_l t_{il,m} v_{lj,m}, \tag{9}$$

where $m = 1 \cdots M$ is the index of the sound sources. $\mathsf{H}_{i,m}$ is an $M \times M$ spatial covariance matrix for each frequency i and source m, and $\mathsf{H}_{i,m} = \boldsymbol{a}_{i,m}\boldsymbol{a}_{i,m}^h$ is limited to a rank-1 matrix. $t_{il,m} \in \mathbb{R}_+$ and $v_{lj,m} \in \mathbb{R}_+$ are the elements of the basis matrix $\boldsymbol{T}_m$ and activation matrix $\boldsymbol{V}_m$. In ILRMA, the spatial covariance matrix $\mathsf{H}_{i,m}$ is constrained to be a rank-1 matrix. This rank-1 spatial constraint leads to the following cost function:

$$\mathcal{Q} = \sum_{i,j} \left[\sum_m \frac{|y_{ij,m}|^2}{\sum_l t_{il,m} v_{lj,m}} - 2\log|\det \boldsymbol{W}_i| + \sum_m \log \sum_l t_{il,m} v_{lj,m} \right], \tag{10}$$

namely, the estimation of $\mathsf{H}_{i,m}$ can be transformed to the estimation of the demixing matrix $\boldsymbol{W}_i$. This cost function is equivalent to the Itakura–Saito divergence between X_{ij} and $\hat{\mathsf{X}}_{ij}$, and we can derive

$$t_{il,m} \leftarrow t_{il,m} \sqrt{\frac{\Sigma_j |y_{ij,m}|^2 v_{lj,m} \left(\Sigma_{l'} t_{il',m} v_{l'j,m}\right)^{-2}}{\Sigma_j v_{lj,m} \left(\Sigma_{l'} t_{il',m} v_{l'j,m}\right)^{-1}}}, \tag{11}$$

$$v_{lj,m} \leftarrow v_{il,m} \sqrt{\frac{\Sigma_i |y_{ij,m}|^2 t_{il,m} \left(\Sigma_{l'} t_{il',m} v_{l'j,m}\right)^{-2}}{\Sigma_i t_{il,m} \left(\Sigma_{l'} t_{il',m} v_{l'j,m}\right)^{-1}}}, \tag{12}$$

$$r_{ij,m} = \sum_l t_{il,m} v_{lj,m}, \tag{13}$$

$$V_{i,m} = \frac{1}{J} \sum_j \frac{1}{r_{ij,m}} \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^h, \tag{14}$$

$$\boldsymbol{w}_{i,m} \leftarrow \left(\boldsymbol{W}_i V_{i,m}\right)^{-1} \boldsymbol{e}_m, \tag{15}$$

where $\boldsymbol{e}_m$ is unit vector whose mth element is one. We can simultaneously estimate both the sourcewise time-frequency model $r_{ij,m}$ and the demixing matrix $\boldsymbol{W}_i$ by iterating (11)–(15) alternately. After the cost function converges, the separated signal $\boldsymbol{y}_{ij}$ can be obtained as (5). Note that since the

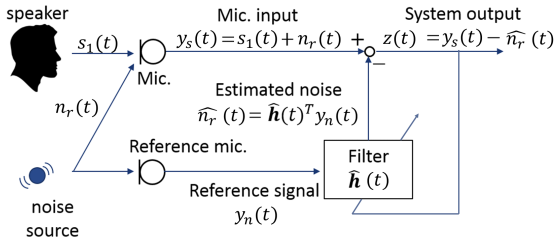

Fig. 5. Noise canceller.

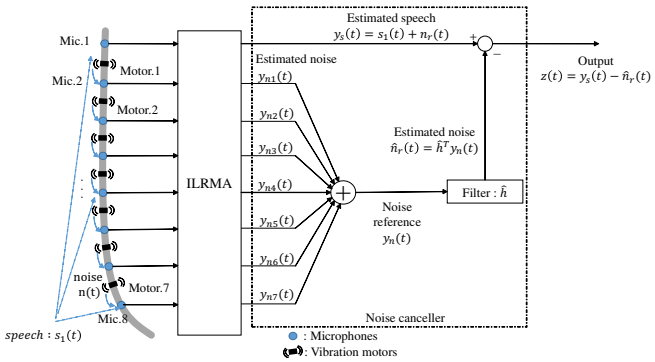

Fig. 6. Flow of proposed method.

signal scale of $\boldsymbol{y}_{ij}$ cannot be determined as well as by IVA, we apply a back-projection technique [12] to $\boldsymbol{y}_{ij}$ to restore the scale.

B. Noise Canceller

A noise canceller requires a reference microphone located near a noise source. The recorded noise reference signal $n_r(t)$ is utilized to reduce the noise in the observed speech signal $s_1(t)$ as shown in Fig. 5. We here assume that both $s_1(t)$ and $n_r(t)$ are simultaneously recorded. The observed signal contaminated with the noise source can be represented as

$$y_s(t) = s_1(t) + n_r(t). \tag{16}$$

We consider that the noise signal $n_r(t)$ is strongly correlated with the reference noise signal $y_n(t)$ and that $n_r(t)$ can be represented by a linear convolution model as

$$n_r(t) \simeq \hat{n}_r(t) = \hat{\boldsymbol{h}}(t)^t \boldsymbol{y}_n(t), \tag{17}$$

where $\boldsymbol{y}_n(t) = [y_n(t)\ y_n(t-1)\ \cdots\ y_n(t-N+1)]^t$ is the reference microphone input from the current time t to the past N samples, and $\hat{\boldsymbol{h}}(t) = [\hat{h}_1(t)\ \hat{h}_2(t)\ \cdots\ \hat{h}_N(t)]^t$ is the estimated impulse response. From (17), the speech signal $s_1(t)$ is extracted by subtracting the estimated noise $\hat{\boldsymbol{h}}(t)^t \boldsymbol{y}_n(t)$ from the observation as

$$z(t) = x(t) - \hat{\boldsymbol{h}}(t)^t \boldsymbol{y}_n(t), \tag{18}$$

where $z(t)$ is the estimated speech signal. The filter $\hat{\boldsymbol{h}}(t)$ can be obtained by a minimization of the mean square error. In this paper, we use the normalized least mean square (NLMS) algorithm [23] to estimate $\hat{\boldsymbol{h}}(t)$. From the NLMS algorithm, the update rule of the filter $\hat{\boldsymbol{h}}(t)$ is given as

$$\hat{\boldsymbol{h}}(t+1) = \hat{\boldsymbol{h}}(t) + \mu \frac{\boldsymbol{z}(t)}{||\boldsymbol{y}_n(t)||^2} \boldsymbol{y}_n(t). \tag{19}$$

C. Flow of Proposed Method

Figure 6 shows the flow of the proposed method. In Fig. 6, $y_s(t)$ is the speech signal estimated by ILRMA, $y_{n1(t)}$, …, $y_{n7(t)}$, are the residual outputs that correspond to the various components of ego noise, and y_n is the sum of such ego noise components. In the first step, the observed signals are separated into independent signals via IVA or ILRMA, where the number of separated signals is the same as the number of microphones ($M = 8$). Note that it is not fixed which channel includes most of the speech components owing to a permutation ambiguity in the outputs of ILRMA and IVA. Therefore, we must find an estimated signal that includes most of the speech components which we use as $y_s(t)$. In this paper, we manually choose the estimated signal from the output signals, although such a signal can be estimated by employing statistics or spectrograms of the output signals. Since a time-invariant demixing matrix (demixing filter) is applied for the separation in the first step, the ego noise, which does not follow the time-invariant assumption, remains in the separated speech signal $y_s(t)$. In the second step, we apply the noise canceller with the ego noise reference $y_n(t)$. In this step, we expect that the noise canceller will reduce the residual noise component in $y_s(t)$ because it models the time-variant noise as $\hat{\boldsymbol{h}}(t)^t \boldsymbol{y}_n(t)$, which can update the filter $\hat{\boldsymbol{h}}(t)$ at each iteration.

V. Experiment

A. Conditions

In our experiment, we produced an artificial observed signal using the hose-shaped rescue robot. This robot consists of eight microphones and seven vibration motors, and the total length of the robot is approximately 3m. The recorded speech signal was produced by convoluting a dry speech signal and the impulse response between a disaster victim and microphones on the robot. For the noise signal, we recorded actual ego noise by moving the robot in an area that simulated a disaster site. The observed multichannel signal was obtained as the sum of these speech and ego noise signals in each microphone, namely, it was a mixture of time-invariant speech and time-variant actual ego noise. In addition, we compared four methods; simple IVA, IVA with the noise canceller (IVA+NC), simple ILRMA, and the proposed method (ILRMA+NC), using the signal-to-distortion ratio (SDR) [24] to evaluate the separation performance. The other experimental conditions are shown in Table I. Note that we applied the back-projection method to the channel with the largest number of speech components for all the methods. However, we can apply the back-projection method to any of the channels without losing the advantages of the proposed method.

TABLE I
EXPERIMENTAL CONDITIONS

Sampling frequency	16 kHz
STFT length	1024 samples
Window shift length	STFT length/4
Number of bases	15
Number of iterations	200
Filter length of noise canceller	1600 taps
Step size of NLMS	0.1
Input SNR	-5, -10 dB

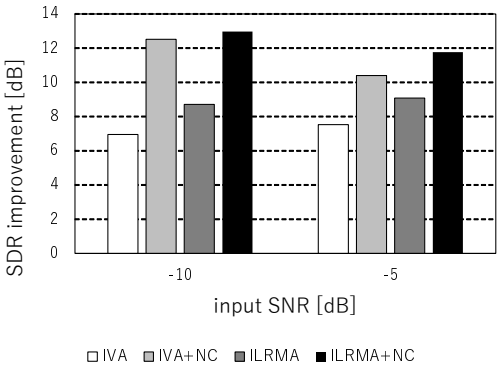

Fig. 7. SDR improvements for each method.

B. Results

Figure 7 shows the improvement in the SDR for each method. These results it is clearly show that the noise canceller greatly improves the separation performance for both IVA and ILRMA. This is because it efficiently reduces the time-variant ego noise components from the estimation result of IVA or ILRMA. In particular, ILRMA+NC outperforms the other methods for both values of the input SNR cases. ILRMA achieves the higher ego noise reduction than IVA. This difference is due to the existence of a source model, which is the bases decomposition using NMF, between IVA and ILRMA. The source model in ILRMA efficiently captures the spectral features of the speech and ego noise signals. However, as the input SNR increases, the difference between the improvements for ILRMA and ILRMA+NC becomes small. This results from the variation of the estimated filter coefficients in the noise canceller, which is caused by the remaining speech components in $\boldsymbol{y}_n$. The proposed method is more effective when the input SNR is low, namely, in a noisy environment.

VI. CONCLUSION

To enhance speech signals recorded by a hose-shaped rescue robot, we have proposed an ego noise suppression method using ILRMA and a noise canceller. We evaluated the proposed method by an experimental simulation and compared IVA, ILRMA, IVA with the noise canceller, and ILRMA with the noise canceller. It was found that the proposed method exhibited the performance in therms of the SDR under all conditions, thus confirmed the efficacy of combining ILRMA and the noise canceller.

VII. ACKNOWLEDGMENTS

This work was supported by the Japan Science and Technology Agency and the Impulsing Paradigm Change through Disruptive Technologies Program (ImPACT) commissioned by the Council for Science, Technology and Innovation, and partly supported by SECOM Science and Technology Foundation. We would also like to express our gratitude to Pf. Hiroshi Okuno and Mr. Yoshiaki Bando for providing experimental data.

REFERENCES

[1] "Impulsive Paradigm Change through Disruptive 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 (IROS)*, pp. 3679–3686, 2012.

[3] A. Deleforge and W. Kellerman, "Phase-optimized K-SVD for signal extraction from underdetermined multichannel sparse mixtures," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 355–359, 2015.

[4] H. Barfuss and W. Kellerman, "Improving blind source separation performance by adaptive array geometries for humanoid robots," *Proc. Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014.

[5] H. Barfuss and W. Kellerman, "An adaptive microphone array topology for target signal extraction with humanoid robots," *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 16–20, 2014.

[6] H. Sawada, J. Even, H. Saruwatari, K. Shikano, and T. Takatani, "Improvement of speech recognition performance for spoken-oriented robot dialog system using end-fire array," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS)*, 2010.

[7] Y. Bando, K. Itoyama, M. Konyo, T. Kazuhiro, K. Yoshii, and H. G. Okuno, "Human-voice enhancement based on online RPCA for a hose-shaped rescue robot with a microphone array," *IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, 2015.

[8] Y. Bando, K. Itoyama, M. Konyo, S. Tadokoro, K. Nakadai, K. Yoshii, and H. G. Okuno, "Variational bayesian multi-channel NMF for human-voice enhancement with a deformable and partially-occluded microphone array," *Proc. European Signal Processing Conference (EUSIPCO)*, 2016.

[9] M. Ishimura, S. Makino, T. Yamada, N. Ono, and H. Saruwatari, "Noise reduction using independent vector analysis and noise cancellation for a hose-shaped rescue robot," *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016.

[10] 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*, 2006.

[11] A. Hiroe, "Solution of permutation problem in frequency domain ica using multivariate probability density functions," *Proc. International Conference on Independent Component Analysis and Blind Source Separation*, 2006.

[12] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Speech and Audio Processing*, vol. 15, no. 1, 2007.

[13] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 276–280, 2015.

[14] 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. 10, 2016.

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

[16] 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 (WASPAA)*, pp. 189–192, 2011.

[17] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, 1999.
[18] 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.
[19] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*, John Wiley & Sons, New York, 2009.
[20] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 3, 2010.
[21] H. Kameoka, T. Yoshioka, M. Hamamura, J. Le Roux, and K. Kashino, "Statistical model of speech signals based on composite autoregressive system with application to blind source separation," *Proc. International Conference on Latent Variable Analysis and Signal Separation*, 2010.
[22] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.
[23] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, John Wiley&Sons, New York, 2004.
[24] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 1462–1469, 2006.