EGO-NOISE REDUCTION FOR A HOSE-SHAPED RESCUE ROBOT USING DETERMINED RANK-1 MULTICHANNEL NONNEGATIVE MATRIX FACTORIZATION

[1] *Moe Takakusaki*, [2] *Daichi Kitamura*, [3,2] *Nobutaka Ono*, [1] *Takeshi Yamada*,
[1] *Shoji Makino*, [4] *Hiroshi Saruwatari*

[1]University of Tsukuba, Ibaraki, Japan
[2]SOKENDAI (The Graduate University for Advanced Studies), Kanagawa, Japan
[3]National Institute of Informatics, Tokyo, Japan
[4]The University of Tokyo, Tokyo, Japan

ABSTRACT

A hose-shaped rescue robot is one of the robots that have been developed for disaster response in times of large-scale disasters such as a massive earthquake. This robot is suitable for entering narrow and dark places covered with rubble in a disaster site and for finding victims inside it. It can transmit ambient sound captured by its built-in microphones to its operator. However, there is a serious problem, that is, the inherent noise of this robot, such as vibration sound or fricative sound, is mixed with the transmitted voice, thereby disturbing the operator's perception of a call for help from a disaster victim. In this paper, we apply the multichannel nonnegative matrix factorization (NMF) with the rank-1 spatial constraint (determined rank-1 MNMF), which was proposed by Kitamura *et al.*, to the reduction of the inherent noise.

Index Terms— determined rank-1 MNMF, hose-shaped rescue robot, ego-noise, noise reduction

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 robot cannot 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.

In the Tough Robotics Challenge, five types of remote-controlled and autonomous robots are developed. In this work, we deal with one of them, namely, a hose-shaped rescue robot [2], which is long and narrow like a snake. Using the microphones mounted on the robot, we develop its voice recording function to capture a disaster victim's voice in a disaster site. We examine the application of the determined rank-1 multichannel nonnegative matrix factorization (determined rank-1 MNMF) [3], [4] to the reduction in the inherent noise (ego-noise) of the robot, which is a particularly big problem for voice recording.

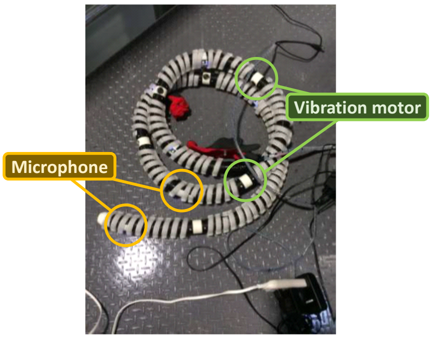

Fig. 1. Hose-shaped rescue robot

2. EGO-NOISE OF HOSE-SHAPED RESCUE ROBOT

2.1. Structure of robot and disposition of ego-noise

The hose-shaped rescue robot is suitable for entering narrow and dark places covered with rubble in a disaster site and finding victims inside it. Figure 1 shows a picture of the robot, and Fig. 2 shows its structure. The 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 (IMU), microphones, and speakers attached along the length of the robot.

According to the operation principle of the robot, very loud ego-noise is 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 person 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 sound.

978-1-5090-2007-2/16/$31.00 ©2016 IEEE

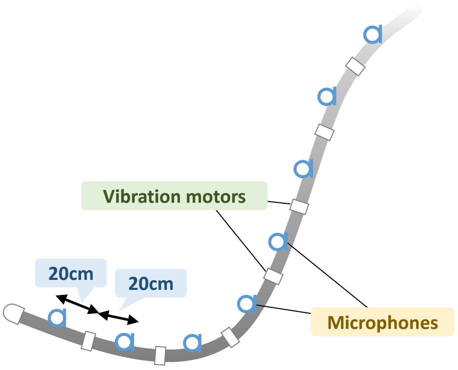

Fig. 2. Structure of hose-shaped rescue robot

2.2. Conventional noise reduction methods

In the conventional noise reduction methods for other robots, it is assumed that the acoustic characteristics of ego-noise do not change. However, it is considered that competent performance cannot be obtained by conventional methods, because the ego-noise characteristics are changed by the area a robot touches. Because a robot operates at the time of a disaster, a method that requires prior information is inappropriate. Therefore, in this work, we choose and apply a method that does not require prior information and can further reduce ego-noise compared with conventional methods.

3. APPLICATION OF BLIND SOURCE SEPARATION TO NOISE REDUCTION

3.1. Approach

Blind source separation (BSS) is a method that does not require prior information and it separates sound sources from only observed signals, and it is studied actively. It is considered that BSS, which uses multichannel signals, is effective because it utilizes the spatial information of sound sources obtained by the robot's many microphones. Thus, we consider an effective multichannel BSS method for the ego-noise reduction in the hose-shaped rescue robot.

It is considered that the main factors for the ego-noise are the vibration sound generated by vibration motors and fricative sound. We conjecture that ego-noise can be expressed effectively by nonnegative matrix factorization (NMF) [5] because it is considered that the time frequency structure is obtained by repeating several types of similar spectra. Because the hose-shaped rescue robot moves very slowly, the source separation by linear microphone array signal processing is effective when a linear time-varying mixture is assumed, which means that the positional relationship between the ego-noise sources and the microphones barely changes. In particular, the determined rank-1 MNMF [3], [4] method proposed by Kitamura *et al.* introduces the expression of NMF into the sound source model of the independent vector analysis (IVA) [6]. This is one of the sound source separation techniques by linear microphone array signal processing, and it realizes sound source separation more precisely than IVA. On the basis of the foregoing, we consider that the determined rank-1 MNMF is effective for the ego-noise reduction for the hose-shaped rescue robot and apply it.

3.2. Blind source separation

3.2.1. Formulation

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

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

where $1 \leq i \leq I \, (i \in \mathbb{N})$ describes the frequency index, $1 \leq j \leq J \, (j \in \mathbb{N})$ describes the time index, and ^{t} denotes the vector transpose, and all the entries of these vectors are complex values. We can approximately represent the observed signals as

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij}. \quad (4)$$

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

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}, \quad (5)$$

where $\boldsymbol{a}_{i,m}$ is the steering vector, $\boldsymbol{w}_{i,m}$ is the demixing filter, and ^{h} is the Hermitian transpose.

3.2.2. Determined rank-1 MNMF

The determined rank-1 MNMF [3], [4] is a method that adds the rank-1 spatial model limitation to the multichannel NMF (MNMF) [7]. We explain the formulation and algorithm [3], [4] derived by Kitamura *et al.* An observed signal is represented by the correlation matrix between the channels, X_{ij}, as

$$\mathsf{X}_{ij} = \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^h. \quad (6)$$

The separation model $\hat{\mathsf{X}}_{ij}$ that approximates X_{ij} is represented as

$$\mathsf{X}_{ij} \approx \hat{\mathsf{X}}_{ij} = \Sigma_k (\Sigma_m \mathsf{H}_{i,m} z_{mk}) t_{ik} v_{kj}, \quad (7)$$

where $m = 1, \cdots, M$ is the index of sound sources, and $k = 1, \cdots, K$ is the index of the spectral bases for NMF. $\mathsf{H}_{i,m}$ 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. $z_{mk} \in \mathbb{R}_{[0,1]}$ is a weight for distributing K NMF bases (frequent appearance spectrum) to each sound source. It shows that the kth base contributes to only the mth source. In addition, $t_{ik} \in \mathbb{R}_+$ and $v_{kj} \in \mathbb{R}_+$ are the elements of the basis matrix $\boldsymbol{T}$ and the activation matrix $\boldsymbol{V}$. MNMF obtains the separated signals $\boldsymbol{y}$ by assigning the spatial covariance matrices H and the source information $\boldsymbol{TV}$ with the partition function. However, the determined rank-1 MNMF separates the sound source by obtaining the demixing matrix $\boldsymbol{W}_i$ from the decomposed model described above. The update rules of the demixing matrix $\boldsymbol{W}_i$ to obtain the separated signal $\boldsymbol{y}$ is as follows:

$$r_{ij,m} = \Sigma_k z_{mk} t_{ik} v_{kj}, \quad (8)$$
$$V_{i,m} = \frac{1}{J} \Sigma_j \frac{1}{r_{ij,m}} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^h, \quad (9)$$
$$\boldsymbol{w}_{i,m} \leftarrow (\boldsymbol{W}_i V_{i,m})^{-1} \boldsymbol{e}_m, \quad (10)$$

where $\boldsymbol{e}_m$ is the unit vector and the only mth element equals to 1.

The partition function z_{mk} and the elements of the basis matrix t_{ik} and the activation matrix v_{kj} in the determined rank-1 MNMF are updated by two methods. One method involves the assignment of the bases to each separated sound source using the above partition function automatically, and the other method does not use the partition function and instead expresses all sound sources in the same number of bases. The method without the partition function does not have the update of z_{mk}. Therefore, it updates $t_{ik,m}$ and $v_{kj,m}$ by applying the update rules of NMF by the channels. The update rules are as follows:

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

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

On the other hand, in the method involving the assignment of the bases with the partition function z_{mk}, it is necessary to update z_{mk} as with MNMF.

$$z_{mk} \leftarrow z_{mk}\sqrt{\frac{\Sigma_{i,j} |y_{ij,m}|^2 t_{ik} v_{kj}\left(\Sigma_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-2}}{\Sigma_{i,j} t_{ik} v_{kj}\left(\Sigma_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-1}}} \tag{13}$$

$$t_{ik} \leftarrow t_{ik}\sqrt{\frac{\Sigma_{j,m} |y_{ij,m}|^2 z_{mk} v_{kj}\left(\Sigma_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-2}}{\Sigma_{j,m} z_{mk} v_{kj}\left(\Sigma_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-1}}} \tag{14}$$

$$v_{kj} \leftarrow v_{kj}\sqrt{\frac{\Sigma_{i,m} |y_{ij,m}|^2 z_{mk} t_{ik}\left(\Sigma_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-2}}{\Sigma_{i,m} z_{mk} t_{ik}\left(\Sigma_{k'} z_{mk'} t_{ik'} v_{k'j}\right)^{-1}}} \tag{15}$$

From the above, we obtain $\boldsymbol{W}_i$ to find separated signals by updating $\boldsymbol{W}_i$, z_{mk}, t_{ik}, and v_{kj} alternately and repeatedly. Finally, we restore the signal scale by applying a back-projection technique [8].

4. EXPERIMENT

4.1. Conditions

Using the actual ego-noise recorded by a hose-shaped rescue robot, we evaluated the ego-noise reduction performance. Specifically, we measured the impulse responses from a disaster victim to microphones using a robot with eight microphones, seven vibration motors, and a total length of 3 m, in the set simulating a disaster site. The distance of the robot from a sound is 1–3 m. We generated a mixed sound for simulation by convolving an impulse response with a speech and adding it to the ego-noise-adjusted SNR. We separated a mixed sound by the determined rank-1 MNMF and evaluated it. We used the signal-to-distortion ratio (SDR) [9] as an evaluation measure. Table 1 shows other experiment conditions.

In the experiment, we investigated each parameter of the determined rank-1 MNMF, which was appropriate for the ego-noise reduction. Moreover, on the basis of the quantity of SDR improvement, we confirmed whether the determined rank-1 MNMF was effective for the ego-noise reduction for the hose-shaped rescue robot.

4.2. Results

4.2.1. Results of experiment on fast Fourier fransform length

Figure 3 shows the experimental results for finding the number of bases and changing the analysis frame length. In this experiment, we uses the number of bases, which provides the best result with or without the partition function. Without the partition function, the number of bases assigned to each sound sources is fixed at 15. On the other hand, with the partition function, we fix the total number of bases of all sound sources to 40 and assign each base to each sound source by the partition function. Figure 3 shows the results in the case that the input SNR of the sound to the ego-noise was -5 dB; SDR was highest when the analysis frame length was 2048 samples. In the case of the determined rank-1 MNMF, the analysis frame length that provides good results was changed by the impluse response from a victim to eight microphones. Therefore, it was confirmed that SDR was markedly improved by setting the analysis frame length to 2048 samples.

Table 1. Experimental conditions

Sampling frequency		16 kHz
FFT length		1024, 2048, 4096, 8192 sample
Window shift length		FFT length/4
Number of bases	Without partition function	1, 5, 10, 15, 20
	With partition function	8, 40, 80, 120, 160
Number of iterations		200
Input SNR		0, -5, -10 dB

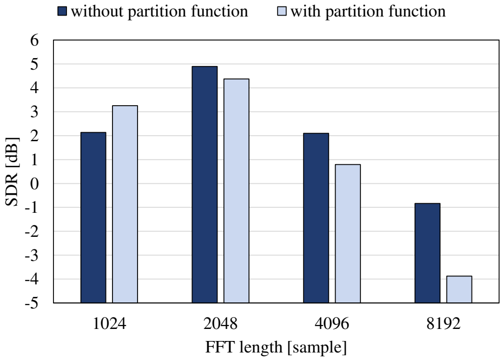

Fig. 3. SDRs of different analysis frame length (Input SNR: -5 dB)

4.2.2. Results of experiment on partitioning bases

We fixed the analysis frame length to 2048 samples and changed the partition function and number of bases. Figures 4 and 5 show the experimental results without and with the partition function, respectively. The input SNR of the sound to ego-noise was -5 dB. Figures 4 and 5 confirmed that SDR was highest when we did not use the partition function and the number of bases was 15. On the other hand, when we assigned 40 bases with the partition function, SDR was highest.

4.2.3. Results of experiment on effectiveness of determined rank-1 MNMF

Figure 6 shows SDR improvement achieved by IVA, determined rank-1 MNMF without the partition function, and determined rank-1 MNMF with the partition function, with various input SNRs. The analysis frame length was 2048 samples for 15 bases assigned to the sound sources for the determined rank-1 MNMF without the partition function and 40 bases in total for the determined rank-1 MNMF with the partition function. Figure 6 confirmed that the determined rank-1 MNMF has a higher ego-noise reduction performance than IVA. In addition, it is shown that a large number of bases is neces-

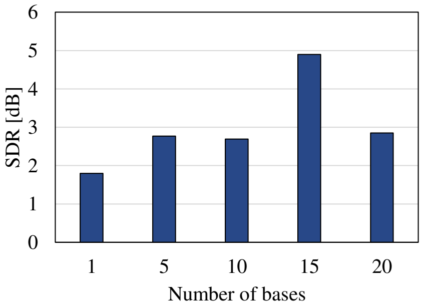

Fig. 4. SDRs of different number of bases (Without partition function, input SNR: −5 dB)

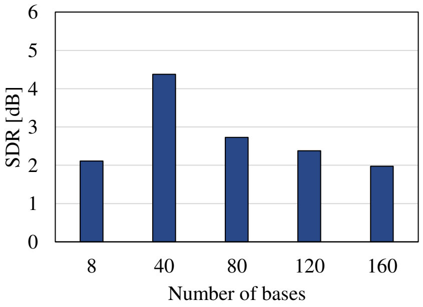

Fig. 5. SDRs of different number of bases (With partition function, input SNR: −5 dB)

sary to express each sound source, because IVA is equivalent in the case of 1 base. Furthermore, without the partition function, the SDR improvement was greater, which confirms that the partition function did not work effectively. By adding a restriction to the partition function to match the sound and ego-noise, we can make use of the flexibility of the partition function in ego-noise reduction.

5. CONCLUSIONS

In this paper, we applied the determined rank-1 MNMF to the ego-noise reduction for a hose-shaped rescue robot for response at a disaster site. First, we examined the analysis frame length and number of bases appropriate for ego-noise reduction in the determined rank-1 MNMF. Furthermore, we compared IVA and the determined rank-1 MNMF; the determined rank-1 MNMF had a higher SDR. We confirmed the effectiveness of the determined rank-1 MNMF for ego-noise reduction.

6. ACKNOWLEDGEMENTS

This work was supported by the Japan Science and Technology Agency and Impulsing Paradigm Change through Distributive Technologies Program (ImPACT) designed by the Council for Science, Technology and Innovation, and partly supported by SECOM Science and Technology Foundation.

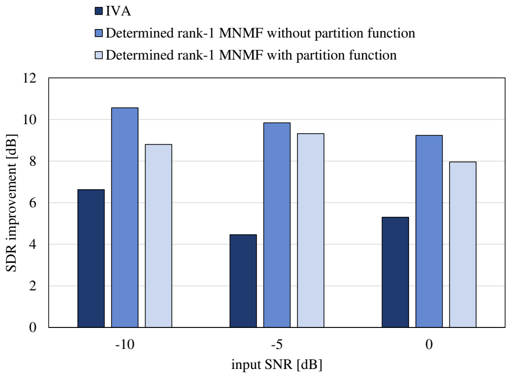

Fig. 6. SDR improvements according to method

7. 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] 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, pp. 276–280, 2015.

[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. Audio, Speech, and Language Processing, 16 pages, 2016 (in press).

[5] 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.

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

[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] N. Murata, S. Ikeda and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," Neurocomputing, vol. 41, no. 1–4, pp. 1–24, 2001.

[9] 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.