NOISE REDUCTION USING INDEPENDENT VECTOR ANALYSIS AND NOISE CANCELLATION FOR A HOSE-SHAPED RESCUE ROBOT

Masaru Ishimura[1], *Shoji Makino*[1], *Takeshi Yamada*[1], *Nobutaka Ono*[2,3], *Hiroshi Saruwatari*[4]

[1]University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
[2]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo 101-8430, Japan
[3]SOKENDAI (The Graduate University for Advanced Studies)
[4]The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-8654, Japan
ishimura@mmlab.cs.tsukuba.ac.jp, maki@tara.tsukuba.ac.jp, takeshi@cs.tsukuba.ac.jp
onono@nii.ac.jp, hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

ABSTRACT

In this paper, we present noise reduction for a hose-shaped rescue robot. The robot is used for searching for disaster victims by capturing their voice with its microphone array. However, the ego noise generated by its vibration motors makes it difficult to distinguish human voices. To solve this problem, we propose a noise reduction method using a blind source separation technique based on independent vector analysis (IVA) and noise cancellation. Our method consists of two steps: (1) estimating a speech signal and an ego noise signal from observed multichannel signals using the IVA-based blind source separation technique, and (2) applying noise cancellation to the estimated speech signal using the estimated ego noise signal as a noise reference. The experimental evaluations show that this approach is effective for suppressing the ego noise.

Index Terms— Rescue robot, Tough environment, Noise reduction, Independent vector analysis, Noise cancellation

1. INTRODUCTION

In recent years, the development of remotely operable robots for the efficient investigation of postdisaster situations in the event of natural disasters such as earthquakes has been promoted. The Impulsing Paradigm Change through Disruptive Technologies Program (ImPACT) "Tough Robotics Challenge"[1] is one such project. A hose-shaped rescue robot [2] is one of the robots developed under this project. The hose-shaped rescue robot is long and slim like a snake and makes it possible to investigate narrow spaces that conventional remotely operable robots cannot enter. A goal of the hose-shaped rescue robot is to search for disaster victims by capturing their voices with its microphones attached around itself at regular intervals. However, the hose-shaped rescue robot moves by vibrating cilia wrapped around itself using vibration motors, so the noise generated by itself is mixed with victims' voices that are captured by its microphones. This makes it difficult to distinguish human voices. In this paper, we refer to such noise as "ego noise".

Thus, in this study, we focus on reducing the ego noise from the recorded sound in order to search for victims by capturing their voices with the microphone array of the robot. Recently, many noise reduction methods for robots have been proposed, such as those described in [3, 4]. These methods improves the performance of noise reduction by adapting the microphone array geometry. However, microphone positions on the hose-shaped rescue robot changes as the robot moves, so we need a blind source separation (BSS) method, which does not need information about microphone and source positions. Hence, in this study, we apply a BSS method based on the statistical independence of each sound source, namely, the independent vector analysis (IVA) [5], because IVA has a remarkable advantage that it is not affected by permutation ambiguity. Furthermore, we apply the time-variant noise cancellation to compensate for the time-invariant assumption of IVA. We also evaluate the proposed method by an experiment in which we reproduce the sound captured by the hose-shaped rescue robot.

2. HOSE-SHAPED RESCUE ROBOT

2.1. Overview of hose-shaped rescue robot

Figure 1 shows an overview of the hose-shaped rescue robot. The hose-shaped rescue robot basically consists of a hose, cilia wrapped around the hose, and vibration motors that vibrate the hose, and performs various sensing functions using sensors such as microphones, cameras, and gas sensors.

Figure 2 shows the movement principle of the hose-shaped rescue robot. It schematically shows the contact area between the robot and the floor. When the motors vibrate, state (1) changes to state (2) by friction between the cilia and the floor, then state (2) changes to state (3) by cilia slipping. The hose-shaped rescue robot moves by repeating such changes in states.

978-1-5090-2007-2/16/$31.00

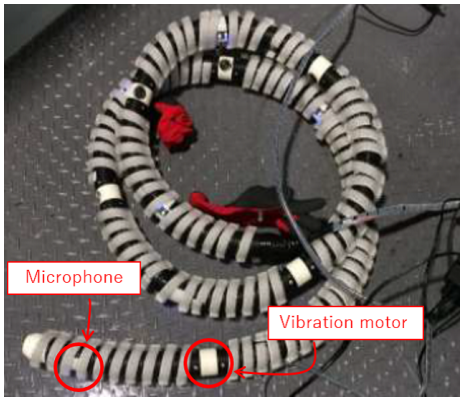

Fig. 1: Hose-shaped rescue robot

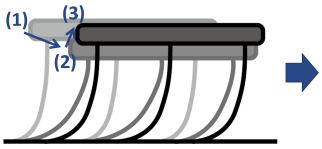

Fig. 2: Movement principle of the hose-shaped rescue robot [2]

2.2. Problems in recording speech

In recording speech, the hose-shaped rescue robot, as well as other robots, has problems caused by its movement principle, as indicated in sect. 2.1 . The problems are as follows:

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

In this paper, we refer to these noises as ego noise and propose the ego noise reduction method using IVA-based BSS and noise cancellation.

3. PROPOSED METHOD

The hose-shaped rescue robot has more than one microphone. Thus, we use a conventional multichannel BSS method, such as beam forming [6] and a method based on statistical independence. The hose-shaped rescue robot changes its shape while moving. Thus, it is difficult to reduce ego noise by techniques that assume the use of a position-fixed microphone array such as beam forming [6]. Thus, we use a method based on statistical independence, namely, the IVA-based BSS method.

However, the demixing filter of the IVA-based BSS method is time invariant for several seconds. Therefore, in this study, we apply the time-variant noise canceller for the postprocessing in IVA to reduce ego noise, which cannot be cancelled using a time-invariant filter. A noise canceller usually requires a reference microphone for observing only noise signals. However, in this study, we use IVA for estimating noise in place of the reference microphone.

In what follows, we first explain the IVA-based BSS and noise canceller, then explain the proposed method.

3.1. IVA-based Blind Source Separation

We use the IVA-based BSS method based on an auxiliary function technique [5]. We suppose that K sources are observed using K microphones. Let the short-time Fourier transform (STFT) representation of source signals and observed signals be $s_k(\tau,\omega)$ and $x_k(\tau,\omega)$, respectively. k is the index of sources and microphones, ω is the index of frequency bins, and τ is the index of the time frames.

Then, the mixing model is expressed as

$$\mathbf{x}(\tau,\omega) = \mathbf{A}(\omega)\mathbf{s}(\tau,\omega), \tag{1}$$

and the estimated source signals are expressed as

$$\mathbf{y}(\tau,\omega) = \mathbf{W}(\omega)\mathbf{x}(\tau,\omega), \tag{2}$$

where

$$\mathbf{A}(\omega) = (\mathbf{a}_1(\omega), \mathbf{a}_2(\omega), \cdots, \mathbf{a}_K(\omega))^h \tag{3}$$

is the mixing matrix,

$$\mathbf{W}(\omega) = (\mathbf{w}_1(\omega), \mathbf{w}_2(\omega), \cdots, \mathbf{w}_K(\omega))^h \tag{4}$$

is the demixing matrix, where ^{h} denotes the hermitian transpose, and $\mathbf{s}(\tau,\omega), \mathbf{x}(\tau,\omega)$, and $\mathbf{y}(\tau,\omega)$ are respectively defined as

$$\mathbf{s}(\tau,\omega) = (s_1(\tau,\omega), \cdots, s_K(\tau,\omega))^t, \tag{5}$$
$$\mathbf{x}(\tau,\omega) = (x_1(\tau,\omega), \cdots, x_K(\tau,\omega))^t, \tag{6}$$
$$\mathbf{y}(\tau,\omega) = (y_1(\tau,\omega), \cdots, y_K(\tau,\omega))^t, \tag{7}$$

where ^{t} denotes transposition.

On the basis of the above model, the IVA-based BSS is performed by finding the demixing matrix $\mathbf{W}(\omega)$, which maximizes the independence of $\mathbf{y}_k(\tau)$. Non-gaussianity, mutual information, and likelihood are used as measures of independence, and each case boils down to the minimization problem of the objective function as follows [7]:

$$J(\mathbf{W}_s) = \sum_{\tau=1}^{N_\tau} \frac{1}{N_\tau} \sum_{k=1}^{K} G(\mathbf{y}_k(\tau)) - \sum_{\omega=1}^{N_\omega} \log|\det \mathbf{W}(\omega)|, \tag{8}$$

where $\mathbf{W}_s$ denotes a $\mathbf{W}(\omega)$ set, N_ω is the number of frequency bins, N_τ is the number of time frames, $\mathbf{y}_k(\tau)$is the source-wise vector defined as

$$\mathbf{y}_k(\tau) = (y_k(1,\tau), \cdots, y_k(N_\omega,\tau))^t, \tag{9}$$

and $G(\mathbf{y}_k)$ is the contrast function. When $p(\mathbf{y}_k)$ is the probability density function that $\mathbf{y}_k(\tau)$ follows, we obtain $G(\mathbf{y}_k(\tau)) = -\log p(\mathbf{y}_k(\tau))$. In this paper, we assume the prior distribution of source signals as a multivariate super-Gaussian distribution [8]. Thus, we obtain $G(\mathbf{y}_k(\tau)) = ||\mathbf{y}_k(\tau)||_2$.

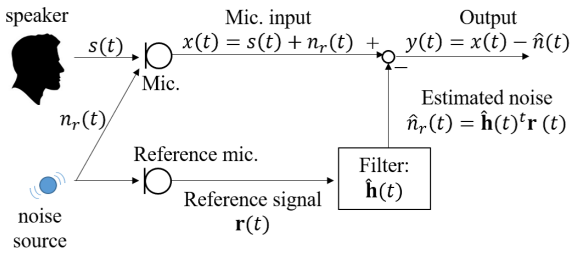

Fig. 3: Noise canceller

By the auxiliary function technique, we use the algorithm to obtain $\mathbf{W}$, which minimizes Eq.(8), as follows[5]:

- Update of auxiliary variable

$$r_k(\tau) = \sqrt{\sum_{\omega=1}^{N_\omega} |\mathbf{w}_k^h(\omega)\mathbf{x}(\tau,\omega)|^2}, \quad (10)$$

$$V_k(\omega) = \frac{1}{N_\tau}\sum_{\tau=1}^{N_\tau}\left[\frac{G'(r_k(\tau))}{r_k(\tau)}\mathbf{x}(\tau,\omega)\mathbf{x}^h(\tau,\omega)\right], \quad (11)$$

- Update of demixing matrix

$$\mathbf{w}_k(\omega) \leftarrow (W(\omega)V_k(\omega))^{-1}\mathbf{e}_k, \quad (12)$$

$$\mathbf{w}_k(\omega) \leftarrow \mathbf{w}_k(\omega)/\sqrt{\mathbf{w}_k^h(\omega)V_k(\omega)\mathbf{w}_k(\omega)}. \quad (13)$$

Here, $\mathbf{e}_k$ denotes the unit vector with the kth element unity.

3.2. Noise canceller

In addition to a microphone for recording speech, the noise canceller requires a reference microphone located near a noise source in order to record only a noise signal, as shown in Fig. 3. We assume the situation wherein we can record the noise reference signal at the same time as the speech signal.

Suppose that a speaker talks in a noisy environment, the input signal to the microphone for recording speech, $x(t)$, is the mixed signal of speech, $s(t)$, and the noise $n_r(t)$, and is as follows:

$$x(t) = s(t) + n_r(t), \quad (14)$$

where t is the index of time samples.

On the other hand, the noise signal is recorded by the reference microphone at the same time as the speech is recorded. Now, we can assume that the noise signal $n_r(t)$ mixed into the microphone for recording speech correlates closely with the reference microphone input signal $r(t)$. Thus, we assume that the relationship between the reference microphone input $r(t)$ and the noise signal mixed into the microphone for recording speech can be described by a linear convolution model as follows:

$$n_r(t) \approx \mathbf{h}(t)^t\mathbf{r}(t), \quad (15)$$

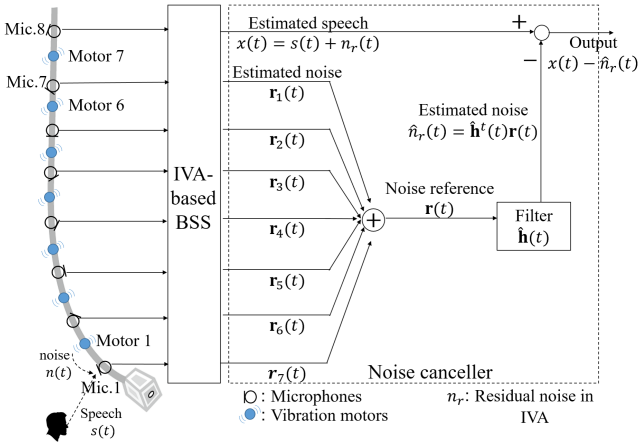

Fig. 4: Flow of the proposed method

where $\mathbf{r}(t) = [r(t), r(t-1), \cdots, r(t-N)]^t$ is the reference microphone input from the current time t to past N samples, and $\hat{\mathbf{h}}(t) = [\hat{h}(0), \hat{h}(1), \cdots, \hat{h}(N)]^t$ is the estimated impulse response.

On the basis of the above model, if the filter $\hat{\mathbf{h}}(t)$ is estimated for the noise signal $n_r(t)$ which mixed into the microphone for recording speech, the noise can be cancelled by subtracting the estimated noise $\hat{\mathbf{h}}(t)^t\mathbf{r}(t)$ from the speech microphone input as follows:

$$y(t) = x(t) - \hat{\mathbf{h}}(t)^t\mathbf{r}(t), \quad (16)$$

where $y(t)$ denotes the estimated speech signal.

As described above, we refer to the noise cancellation method carried out by estimating the noise from the reference signal highly correlated with the noise as the noise canceller. The filter $\mathbf{h}(t)$ can be estimated on the basis of the minimization of the mean square error.

The algorithms for estimating the filter are numerous, and in this study, we use the normalized least mean square (NLMS) algorithm [9]. From the NLMS algorithm, the update rule of the filter $\hat{\mathbf{h}}(t)$ is

$$\hat{\mathbf{h}}(t+1) = \hat{\mathbf{h}}(t) + \mu\frac{y(t)}{||\mathbf{r}(t)||^2}\mathbf{r}(t). \quad (17)$$

3.3. Flow of the proposed method

Figure 4 shows the flow of the proposed method, where s, n, n_r, and r denote the speech signal, the ego noise signal, the residual noise signal in the IVA-based BSS, and the reference signal of ego noise, respectively.

In the first step, the observed signals are separated into independent signals, the number of which is the same as the number of microphones. In this step, we use the IVA-based frequency-domain BSS method based on the auxiliary function technique [5]. The IVA-based BSS is based on independence, which is high-order statistics, so analyzing the statistics requires several-second signals and in this method, it is

Table 1: Experimental conditions

Sampling frequency	16 kHz
IVA iteration	100
Frame length of STFT	1024 samples
Shift length of STFT	256 samples
Filter length of noise canceller	1600 taps
Step size of NLMS	0.1
Input SNR	$-10, -5, 0$ dB

assumed that the demixing filter is time invariant. As a result, the ego noise, which does not follow the time-invariant assumption, remains.

In the second step, we choose the signal that contains speech from the separated signals, and apply the noise canceller using the sum of the other separated signals as the ego noise reference signal. Note that we manually choose the speech signal from the output signals. In this step, we expect that the noise canceller cancels the residual noise that does not follow the time-invariant assumption of IVA, because it can update the filter at each time sample.

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

To adjust the signal-to-noise ratio (SNR) of the input signal, we use the artificially mixed signals of ego noise and speech. As the ego noise signal, we use the sound recorded by moving the hose-shaped rescue robot that has 8 microphones and 7 vibration motors. As the speech signal, we use the convolved signal of a dry speech signal and impulse responses between the location of the speaker and those of the microphones. We use the source-to-distortion ratio (SDR) and source-to-interference ratio (SIR) [10] as the evaluation measure. SDR is a measure for evaluating the distortion of the output signal. SIR is a measure for evaluating the suppression of nontarget signals. SDR and SIR are calculated on the basis of the correct source signal and the estimated source signal. If the target sound is further enhanced, SDR and SIR increase. Other experimental conditions are shown in Table 1. Note that we here apply the back-projection to the channel that includes many speech components.

4.2. Results of evaluation experiment

Figure 5 shows the SDR and SIR improvements obtained at different input SNRs. (b) shows that under all the conditions, SDR improvement of the proposed method is the best. Comparison of the SDR improvement between SNR$=0$ and other SNRs shows that the difference between IVA and the proposed method is small when SNR$=0$. This indicates that the speech components in the reference signal distort the desired speech owing to the postprocessing with noise cancellation.

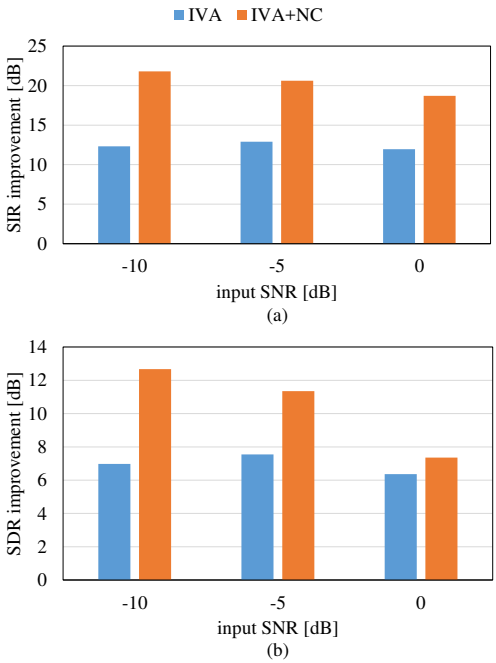

Fig. 5: (a) SIR and (b) SDR improvements for recording at SNRs$=-10, -5$, and 0 dB

However, (a) shows that SIR improvements under all conditions are approximately the same; therefore, the proposed method is effective for detecting the speech of disaster victims.

5. CONCLUSIONS

In this paper, in order to enhance speech on a recorded signal using a hose-shaped rescue robot, we proposed the noise suppression method based on IVA and noise cancellation, and evaluated the proposed method by an experimental simulation. As a result, using SDR, we obtained a 7 dB improvement when using IVA, and a 1-4 dB improvement when using postprocessing with the noise canceller.

6. ACKNOWLEDGEMENTS

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

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 (IROS)*, pp. 3679–3686, 2012.

[3] H. Barfuss and W. Kellerman, "Improving blind source separation performance by adaptive array geometries for humanoid robots," *Late-breaking poster presentation at the 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014.

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

[5] 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.

[6] H. L. Van Trees, *Optimum Array Processing*, Wiley, New York, 2002.

[7] A. Hyvärinen, J. Karhunerl, and E. Oja, *Independent Component Analysis*, Wiley, New York, 2001.

[8] T. Kim, H. T. Attias, S. Y. Lee, and T. W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. on Audio, Speech & Language Processing*, vol. 15, pp. 70–79, 2007.

[9] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, Wiley, New York, 2004.

[10] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech & Language Processing*, vol. 14, pp. 1462–1469, 2006.