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

Ego Noise Reduction and Sound Localization Adapted to Human Ears Using Hose-shaped Rescue Robot

Narumi Mae[1], Koei Yamaoka[1], Yoshiki Mitui[2], Mitsuo Matsumoto[1], Shoji Makino[1]
Daichi Kitamura[2], Nobutaka Ono[3], Takeshi Yamda[1] and Hiroshi Saruwatari[2]

[1] University of Tsukuba, Japan
E-mail: mae@mmlab.cs.tsukuba.ac.jp,
yamaoka@mmlab.cs.tsukuba.ac.jp,
maki@tara.tsukuba.ac.jp,
takeshi@cs.tsukuba.ac.jp

[2] The University of Tokyo, Japan
E-mail: {yoshiki_mitsui, daichi_kitamura,
hiroshi_saruwatari}@ipc.i.u-tokyo.ac.jp,

[3] Tokyo Metropolitan University, Japan
E-mail: onono@tmu.ac.jp

Abstract

Rescue robots have been developed for search and rescue operations in times of large-scale disasters. Such a robot is used to search for survivors in disaster sites by capturing their voices with its microphone array. However, since the robot has many vibration motors, ego noise is mixed with voices, and it is difficult to differentiate the ego noise from a call for help from a disaster survivor. In our previous works, an ego noise reduction technique that combines a method of blind source separation called independent low-rank matrix analysis, noise cancellation and postfilter called MOSIE was proposed. Moreover, we experimentally confirm that the operator can perceive the direction of a survivor's location from the processed stereo sound. However, the stereo sound was not suitable for people to listen because it was recorded by robot's microphones. To solve this problem, in this study, we applied an extension of microphone spacing by virtual microphone technique. By performing in a simulated disaster site, we confirm that the operator can perceive the direction of a survivor's location by applying a speech enhancement technique combining independent low-rank matrix analysis, noise cancellation to the observed multichannel noisy signals and an extension of microphone spacing by virtual microphone technique.

1. Introduction

It is important to develop robots for search and rescue operations during large-scale disasters such as earthquakes. The Tough Robotics Challenge is one of the research and development programs in the Impulsing Paradigm Change through Disruptive Technologies Program (ImPACT) [1]. One of the robots developed in this program is a hose-shaped rescue robot. This robot is long and slim and it can investigate narrow spaces into which conventional remotely operable robots cannot enter. This robot searches for disaster survivors by capturing their voice with its microphones, which are attached around itself. However, there is a serious problem when recording speech using the robot. Because of the mechanism used to operate the robot, very loud ego noise is mixed in the microphones. In our previous works [2], an effective noise reduction technique for stereo signal was proposed. we also confirmed that an operator can perceive the direction of a survivor's location by applying our previous speech enhancement technique to observed multichannel noisy signals recorded by a hose-shaped rescue robot with a pairwise (stereo) microphone array. However, in our previous work, difference in spacing between microphones and the two ears of the operator was not taken into account. In this paper, we apply virtual microphone technique to extend the microphone spacing of the robot to match human ear spacing.

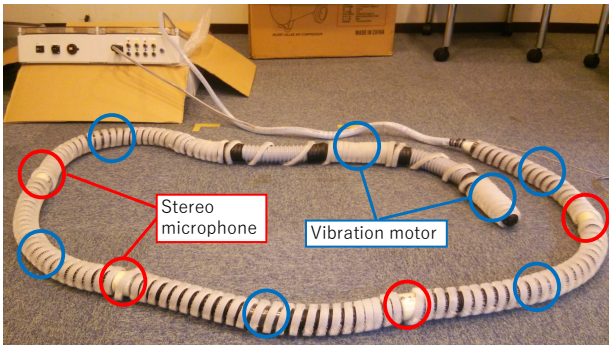

Figure 1: Hose-shaped rescue robot.

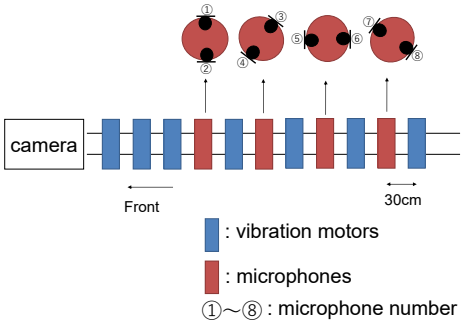

Figure 2: Structure of hose-shaped rescue robot.

2. Hose-shaped Rescue Robot and Ego Noise

2.1 Hose-Shaped Rescue Robot

Figure 1 shows an image of the hose-shaped rescue robot. The robot basically consists of a hose as its axis with cilia tape wrapped around it and has eight microphones, seven vibration motors, a camera, and lamps. Figure 2 shows the positions of its microphones and vibration motors. In the robot, two microphones are attached between each vibration motor, and the microphones are sequentially rotated by 45° with each edge. In other words, the robot has four stereo microphones that are rotated at 45° intervals. Furthermore, the robot moves forward slowly as a result of the reaction between the cilia and floor through the vibration of the cilia tape induced 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

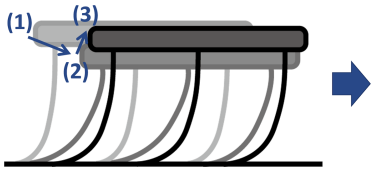

Figure 3: Principle of movement of hose-shaped rescue robot.

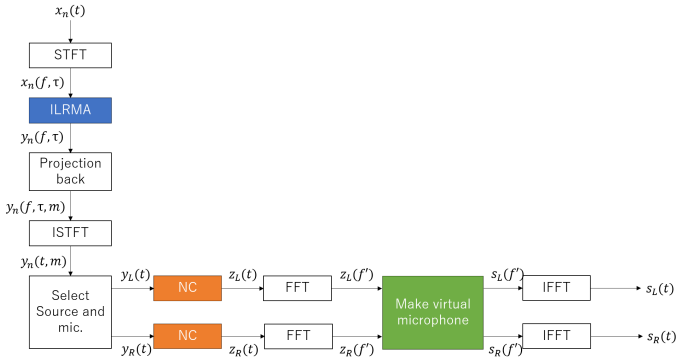

Figure 4: Flow of the proposed method.

the cilia and floor, then state (2) changes to state (3) as a result of the cilia slipping. The hose-shaped rescue robot moves by repeating such changes in its state.

2.2 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 the driving sound of the vibration motors, the fricative sound generated between the cilia and floor, and the noise generated by microphone vibration. In an actual disaster site, the voice of a person seeking help may be not sufficiently loud to capture and it may be smaller than the ego noise.

3. Proposed method

The method consists of three steps (Fig. 4). In the first step, we apply a BSS method called independent low-rank matrix analysis (ILRMA) that is used to estimate both the speech and ego noise for multichannel signal. In the second step, a noise cancellation process is applied to the resulting speech signal estimated by the BSS method. In the third step, we apply virtual microphone technique to extend the microphone spacing of the robot to match human ear spacing.

3.1 Independent Low-Rank Matrix Analysis

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

$$\boldsymbol{s}(f,\tau) = (s(f,\tau,1) \cdots s(f,\tau,M))^t, \quad (1)$$
$$\boldsymbol{x}(f,\tau) = (x(f,\tau,1) \cdots x(f,\tau,M))^t, \quad (2)$$
$$\boldsymbol{y}(f,\tau) = (y(f,\tau,1) \cdots y(f,\tau,M))^t, \quad (3)$$

where f and τ are indexes of frequency and time, respectively, and t denotes the vector transpose. All the entries of these vectors are complex values. When the window size in short-time Fourier transform (STFT) is sufficiently longer than the impulse response between a source and microphone, we can approximately represent the observed signal as

$$\boldsymbol{x}(f,\tau) = \boldsymbol{A}(f)\boldsymbol{s}(f,\tau). \quad (4)$$

Here, $\boldsymbol{A}(f) = (\boldsymbol{a}(f,1) \cdots \boldsymbol{a}(f,M))$ is an $M \times M$ mixing matrix of the observed signals. Denoting $\boldsymbol{W}(f) = (\boldsymbol{w}(f,1) \cdots \boldsymbol{w}(f,M))^h$ as the demixing matrix, the separated signal $\boldsymbol{y}(f,\tau)$ is represented as

$$\boldsymbol{y}(f,\tau) = \boldsymbol{W}(f)\boldsymbol{x}(f,\tau), \quad (5)$$

where ^{h} is the Hermitian transpose. We use ILRMA, which is a method unifying IVA and ISNMF. ILRMA allows us to model the statistical independence between sources and the source-wise time-frequency structure at the same time. We explain the formulation and algorithm derived by Kitamura *et al.* [3,4]. The observed signals are represented as

$$\mathsf{X}(f,\tau) = \boldsymbol{x}(f,\tau)\boldsymbol{x}(f,\tau)^h, \quad (6)$$

where $\mathsf{X}(f,\tau)$ is the correlation matrix between channels of size $M \times M$. The diagonal elements of $\mathsf{X}(f,\tau)$ represent real-valued powers detected by the microphones, and the nondiagonal elements represent the complex-valued correlations between the microphones. The separation model of MNMF that is an extension of simple NMF for multichannel signals, $\hat{\mathsf{X}}(f,\tau)$, used to approximate $\mathsf{X}(f,\tau)$ is represented as

$$\mathsf{X}(f,\tau) \approx \hat{\mathsf{X}}(f,\tau) = \sum_m \mathsf{H}(f,m) \sum_l t(f,l,m)v(l,\tau,m), \quad (7)$$

where $m = 1 \cdots M$ is the index of the sound sources, $\mathsf{H}(f,m)$ is an $M \times M$ spatial covariance matrix for each frequency i and source m, and $\mathsf{H}(f,m) = \boldsymbol{a}(f,m)\boldsymbol{a}(f,m)^h$ is limited to a rank-1 matrix. This assumption corresponds to $t(f,l,m) \in \mathbb{R}_+$ and $v(l.\tau,m) \in \mathbb{R}_+$ being the elements of the basis matrix $\boldsymbol{T}(m)$ and activation matrix $\boldsymbol{V}(m)$, respectively. This rank-1 spatial constraint leads to the following cost function:

$$\mathcal{Q} = \sum_{f,\tau} \left[\sum_m \frac{|y(f,\tau,m)|^2}{\sum_l t(f,l,m)v(l,\tau,m)} - 2\log|\det \boldsymbol{W}(i)| + \sum_m \log \sum_l t(f,l,m)v(l,\tau,m) \right], \quad (8)$$

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

$$t(f,l,m) \leftarrow t(f,l,m)\sqrt{\frac{\Sigma_j |y(f,\tau,m)|^2 v(l,\tau,m)\left(\Sigma_{l'} t(f,l',m)v(l',\tau,m)\right)^{-2}}{\Sigma_j v(l,\tau,m)\left(\Sigma_{l'} t(f,l',m)v(l',\tau,m)\right)^{-1}}}, \quad (9)$$

$$v(l,\tau,m) \leftarrow v(f,l,m)\sqrt{\frac{\Sigma_i |y(f,\tau,m)|^2 t(f,l,m)\left(\Sigma_{l'} t(f,l',m)v(l',\tau,m)\right)^{-2}}{\Sigma_i t(f,l,m)\left(\Sigma_{l'} t(f,l',m)v(l',\tau,m)\right)^{-1}}}, \quad (10)$$

$$r(f,\tau,m) = \sum_l t(f,l,m)v(l,\tau,m), \quad (11)$$

$$\boldsymbol{Z}(f,m) = \frac{1}{J}\sum_j \frac{1}{r(f,\tau,m)}\boldsymbol{x}(f,\tau)\boldsymbol{x}(f,\tau)^h, \quad (12)$$

$$\boldsymbol{w}(f,m) \leftarrow (\boldsymbol{W}(f)\boldsymbol{Z}(f,m))^{-1}\boldsymbol{e}(m), \quad (13)$$

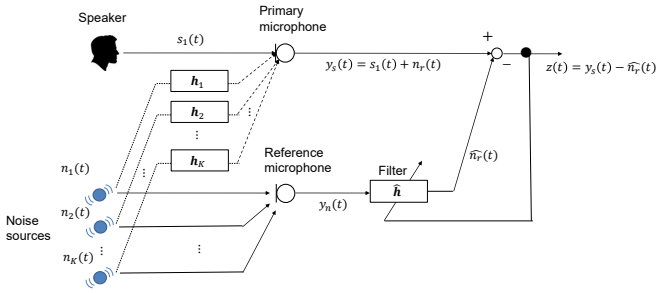

Figure 5: Single noise canceller.

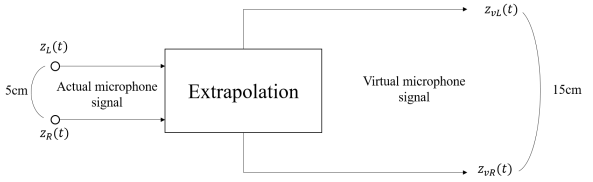

Figure 6: Block diagram of signal processing with virtual microphone array technique

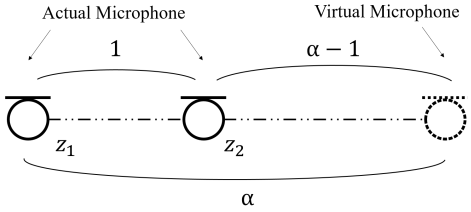

Figure 7: Arrangement of actual and virtual microphones

where $\boldsymbol{e}_m$ is a unit vector whose mth element is one. We can simultaneously estimate both the sourcewise time-frequency model $r(f,\tau,m)$ and the demixing matrix $\boldsymbol{W}(f)$ by iterating (9)–(13) alternately. After the cost function converges, the separated signal $\boldsymbol{y}(f,\tau)$ can be obtained as (5). Note that since the signal scale of $\boldsymbol{y}(f,\tau)$ cannot be determined, we apply a projection-back method [6] to $\boldsymbol{y}(f,\tau)$ to determine the scale.

The demixing filter in ILRMA is time-invariant over several seconds. To achieve time-variant noise reduction, we apply a noise canceller for the postprocessing of ILRMA to reduce the remaining time-variant ego noise components. An noise canceller usually requires a reference microphone to observe only the noise signal. Thus, we utilize the noise estimates obtained by ILRMA as the noise reference signals.

3.2 Noise Canceller

The noise canceller [7] 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

$$y_s(t) = s_1(t) + n_r(t). \tag{14}$$

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

$$n_r(t) \simeq \hat{n}_r(t) = \hat{\boldsymbol{h}}(t)^t \boldsymbol{y}_n(t), \tag{15}$$

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 (15), the speech signal $s_1(t)$ is extracted as follows by subtracting the estimated noise $\hat{\boldsymbol{h}}(t)^t \boldsymbol{y}_n(t)$ from the observation:

$$z(t) = x(t) - \hat{\boldsymbol{h}}(t)^t \boldsymbol{y}_n(t), \tag{16}$$

where $z(t)$ is the estimated speech signal. The filter $\hat{\boldsymbol{h}}(t)$ can be obtained by minimization of the mean square error. In this paper, we use the normalized least mean square (NLMS) algorithm [8] to estimate $\hat{\boldsymbol{h}}(t)$. From the NLMS algorithm, the update rule of the filter $\hat{\boldsymbol{h}}(t)$ is given as

$$\hat{\boldsymbol{h}}(t+1) = \hat{\boldsymbol{h}}(t) + \mu \frac{\boldsymbol{z}(t)}{||\boldsymbol{y}_n(t)||^2} \boldsymbol{y}_n(t). \tag{17}$$

3.3 Virtual Microphone Technique

F In this work, we used virtual microphone technique [9] to extend the microphone spacing of the robot to match human ear spacing. In this work, we generate a virtual microphone signal $z_v(f')$ by nonlinear interpolation or extrapolation of two real microphone signals in frequency domain as an estimate of a signal at a virtual microphone position where there is no real microphone (Fig. 6). In this work, we consider the virtual microphone at the point with a distance ratio of $1:(\alpha-1)$ $(\alpha > 1)$ from the two real microphone positions (Fig. 7).

Here, we formulate the extrapolation of the phase as follows. The phase and amplitude of the signal at microphone i are denoted by $A_i(f')$ and $\phi_i(f')$, and are respectively given as

$$A_i(f') = |z_i(f')|, \tag{18}$$

$$\phi_i(f') = \angle z_i(f') = \arctan \frac{\mathrm{Im}(z_i(f'))}{\mathrm{Re}(z_i(f'))}. \tag{19}$$

The phase of the virtual microphone signal is estimated as follows.

$$\phi_{\mathrm{v}}(f') = \alpha\phi_1(f') + (\alpha-1)\phi_2(f'). \tag{20}$$

Where, α is microphone interpolation parameter. Note that the observed phase has an aliasing ambiguity given by $\phi_i(f') \pm 2n_i\pi$ with integer n_i, therefore, we derived an unwrap phase from a phase angle for extrapolating phase difference between phases at Mic. 1 and Mic. 2. The virtual microphone signal is obtained as follows in terms of the extrapolated phase and amplitude:

$$z_v(f') = A_{\mathrm{v}}(f') \exp(j\phi_{\mathrm{v}}(f')). \tag{21}$$

In this paper, $A_v(f')$ is output signal of Noise canceller, and we obtained $z_i(f')$ by an FFT of the whole observed signal.

4. Experimental Evaluation

4.1 Condition

In the evaluation, we used signals recorded by the hose-shaped rescue robot. We recorded signals arriving from three different directions and evaluated processed sound without virtual microphone and proposed method. Figure 8 shows position of microphones and a speech source.

The processed stereo signal that the operator hears was recorded by microphones 1 and 2, which are attached to the front of the robot, to which the projection-back method is applied to adjust the scale to the observed signal at microphones 1 and 2.

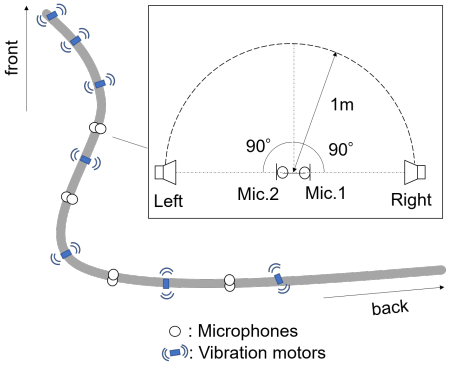

Figure 8: Position of microphones and a speech source.

Table 1: Experimental conditions

Sampling frequency	16 kHz
Window length of ILRMA	2048 samples
Window shift of ILRMA	STFT length/4
Number of bases	15
Number of iterations	50
Filter length of noise canceller	1600 taps
Step size of NLMS (μ)	0.1

We recorded signals arriving from left and right and evaluated each signal. First, we applied ILRMA to multichannel noisy signals that consist of a survivor's voice and ego noise. We found an estimated signal that includes most of the speech components, which was used as $y_L(t)$ and $y_R(t)$, by employing the spectrograms and microphones chosen in advance as reference microphones. To adjust the scale to the observed signal at microphones 1 and 2, we applied the projection-back method to an estimated signal.

Next, we applied the noise canceller for postprocessing of ILRMA to reduce the remaining time-variant ego noise. Then, we used the other microphone as a reference microphone, for example, we used microphone 2 as a reference microphone when applying the noise canceller to the estimated signal observed by microphone 1.

Next, we applied an extension of microphone spacing by virtual microphone technique. Finally, the operator hears the stereo sound. In this experiment, we use 4 signals: 2 signals that applied ILRMA and noise canceller and 2 signals that processed by proposed method.

To confirm that the location of the survivor can be determined by hearing the processed sound, the operator hears the sound with a headphone and chooses either left or right. Other experimental conditions are shown in Table 1.

4.2 Result

For each signal, directions of arrival which subjects answered are shown in Fig. 9. The subjects answered that sound images under the condition of " without virtual microphone " were localized in the azimuths of around 70-degree and 110-degrees (white circles in Fig. 9), close to the center (front) due to short inter-aural time difference. Oppositely, " with virtual microphones ", the sound images are perceived to be closer to the direction of the sound sources, in the azimuths of around 10-degree and 170-degree (black circles in Fig. 9). In this paper, since the amplitude is not extrapolated, the sound images

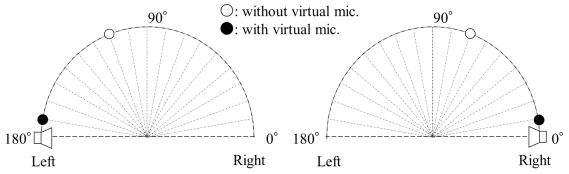

Figure 9: Direction of arrival which the subjects answered.

under the condition of " with virtual microphones " might not be localized in the azimuths of the actual sound sources.

5. Conclusion

In this paper, we have proposed noise reduction method adapted to human ears using hose-shaped rescue robot. The experimental results showed that the proposed method has better direction recognition accuracy than conventional method.

References

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

[2] N. Mae, Y. Mitsui, S. Makino, D. Kitamura, N. Ono, T. Yamada, and H. Saruwatari, "Sound source localization using binaural different for hose-shaped rescue robot," Proc. APSIPA 2017.

[3] 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.

[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 Trans. ASLP, vol. 24, no. 9, pp. 1626–1641, 2016.

[5] 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. IWAENC, 2016.

[6] 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.

[7] B.Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, Jr, E. Dong, and R. C. Goodlin, "Adaptive noise cancelling: Principles and applications," Proc. IEEE, vol. 63, pp. 1692–1716, 1975.

[8] E. Hansler and G. Schmidt, "Acoustic Echo and Noise Control: A Practical Approach", John Wiley & Sons, New York, 2004.

[9] H. Katahira, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Nonlinear speech enhancement by virtual increase of channels and maximum SNR beamformer, " EURASIP Journal on Advances in Signal Processing, vol. 2016, no. 1, pp. 1–8, Jan. 2016.

[10] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, " Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with β-divergence, " Proc. IEEE MLSP, pp. 283-288, 2010.