Sound Source Localization Using Binaural Difference for Hose-Shaped Rescue Robot

Narumi Mae*, Yoshiki Mitsui†, Shoji Makino*, Daichi Kitamura†,
Nobutaka Ono‡, Takeshi Yamada*, and Hiroshi Saruwatari†
* University of Tsukuba, Japan, E-mail: mae@mmlab.cs.tsukuba.ac.jp, takeshi@cs.tsukuba.ac.jp, maki@tara.tsukuba.ac.jp
† The University of Tokyo, Japan, E-mail: {yoshiki_mitsui, daichi_kitamura, hiroshi_saruwatari}@ipc.i.u-tokyo.ac.jp,
‡ National Institute of Informatics (NII), Japan, E-mail: onono@nii.ac.jp
§ SOKENDAI (The Graduate University for Advanced Studies), Japan

***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 and postprocessing for noise cancellation was proposed. In the practical use of this robot, to determine the precise location of survivors, the direction of the observed voice should be estimated after the ego noise reduction process. To achieve this objective, in this study, a new hose-shaped rescue robot with microphone arrays was developed. Moreover, we adapt postfilter called MOSIE to our previous noise reduction method to listen to stereo sound because this robot can record stereo sound. 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, and postfiltering to the observed multichannel noisy signals.**

I. 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]–[4], an effective technique of ego noise reduction that combines a method of blind source separation called independent low-rank matrix analysis (ILRMA) [5], [6] and postprocessing for noise cancellation [2] was proposed. In this paper, we add postfiltering to a noise reduction method called the minimum mean-square error (MMSE) estimation with an optimizable speech model and inhomogeneous error criterion (MOSIE) to obtain better sound. This method is a spatial-cue-aware binaural signal separation algorithm and can estimate binaural signals using a statistical model based on the chi distribution. A method of posture estimation for the hose-shaped rescue robot has also been recently proposed [7].

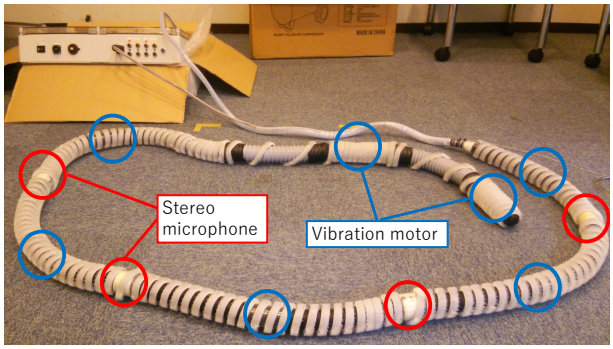

Fig. 1: Hose-shaped rescue robot.

In the practical use of this robot, to determine the precise location of survivors, the direction of the observed voice should be estimated after the ego noise reduction process. To achieve this objective, we have developed a new hose-shaped rescue robot that has a pairwise (stereo) microphone array. As reported in this paper, we experimentally confirm that the operator can perceive the direction of a survivor's location by listening to stereo sound after noise reduction by a combining ILRMA, noise cancellation and MOSIE.

II. Hose-shaped Rescue Robot and Ego Noise

A. 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

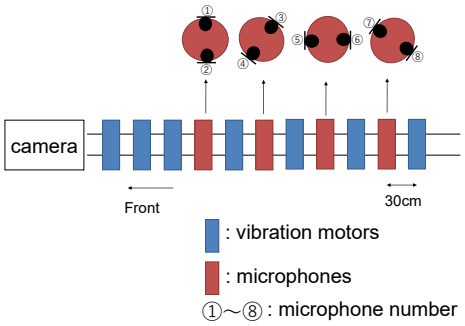

Fig. 2: Structure of hose-shaped rescue robot.

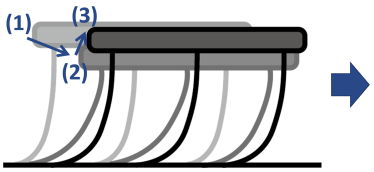

Fig. 3: Principle of movement of hose-shaped rescue robot [8].

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 [8]. When the motors vibrate, state (1) changes to state (2) through the friction between 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.

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 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.

III. Ego Noise Reduction Method

Recently, many ego noise reduction methods have been proposed [9]–[13]. In our case, the target and noise source locations are unknown. For this reason, we can consider the use of a blind source separation (BSS) method. However, using only a BSS method, time-varying components remain because the robot moves. To solve this problem, in our previous work [2]–[4], we proposed an ego noise reduction method for a hose-shaped rescue robot combining the BSS method and noise cancellation. Figure 4 shows the flow of the ego noise reduction method. The method consists of three steps. In the first step, a BSS method such as independent vector analysis (IVA) or ILRMA[1] is used to estimate both the speech and

[1]Note that the authors have renamed the method. In [5], [6], ILRMA was called *rank-1 MNMF*.

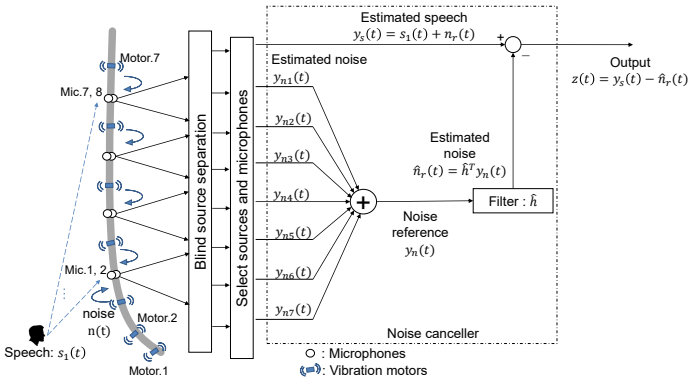

Fig. 4: Flow of the ego noise reduction method.

ego noise. 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 adapt MOSIE [14]. In this work, we use ILRMA for the BSS method because the time-frequency structure of the ego noise consists of several repeated spectral patterns, enabling it to be expressed effectively by nonnegative matrix factorization (NMF), which is used to estimate the source model in ILRMA.

A. 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. Here we use ILRMA, which is a method unifying IVA and ISNMF. ILRMA allows us to model the statistical independence between sources and the sourcewise time-frequency structure at the same time. We explain the formulation and algorithm derived by Kitamura *et al.* [5], [6]. The observed signals are represented as

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

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 non-diagonal elements represent the complex-valued correlations between the microphones. The separation model of MNMF, $\hat{\mathsf{X}}(f,\tau)$, 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:

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

namely, the estimation of $\mathsf{H}(f,m)$ can be transformed to the estimation of the demixing matrix $\boldsymbol{W}(f)$. 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

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

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

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

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

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

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 [16] 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 applied a noise canceller (NC) for the postprocessing of ILRMA to reduce the remaining time-variant ego noise components. An NC usually requires a reference microphone to observe only the noise signal. Thus, we utilized the noise estimates obtained by ILRMA as the noise reference signals.

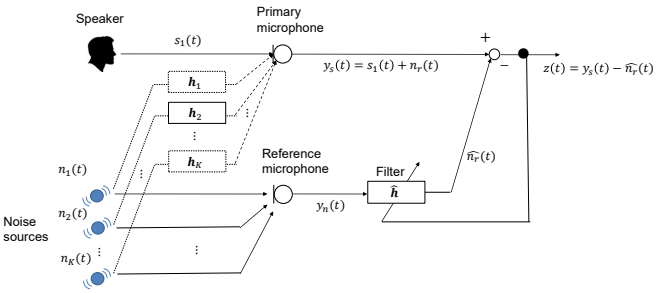

Fig. 5: Single noise canceller.

B. Noise Canceller

The NC [17] 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

$$y_s(t) = s_1(t) + n_r(t). \quad (14)$$

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

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

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:

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

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 [18] to estimate $\hat{\boldsymbol{h}}(t)$. From the NLMS algorithm, the update rule of the filter $\hat{\boldsymbol{h}}(t)$ is given as

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

IV. Postfiltering Based on MOSIE

We adopt MOSIE to the postprocessing stage for achieving further noise reduction. In this work, we process the stereo sound so that operator can listen in binaural fashion. The stereo sound is recorded and reconstructed at the arbitrarily selected pairwise microphone, and hereafter we denote the stereo signals as $z_L(f',\tau')$ and $z_R(f',\tau')$ for the operator's left and right ears, respectively, which are the STFT outputs of the NC. This method includes two steps. In the first

step, we estimate the amplitude spectrum of the target signal on the basis of the MMSE criterion under a certain target prior for each signal. Next, we derive the optimal spectral gain that minimizes the residual interference power in terms of the MMSE under the condition that the spectral gains obtained from the formulation derived by Murota *et al.* [19] are equivalent in both ears. This is because simple MOSIE only provides the statistically fluctuating spectral gains for each of ears independently, and the fluctuation of the gain function in interaural level differences at the left and right ears cause the deterioration in sound localization. Hereafter, we call this gain the equi-binaural optimal spectral gain.

A. Single-Channel MOSIE

At first, we explain the case of single channel signal. This method is based on the generalized Bayesian estimator with automatic target prior adaptation [14]. For MOSIE, we apply STFT to the mixture signal as the output signal of the NC. The stereo sound is a mixture of target and interference signals and can be expressed as

$$z(f',\tau')=s(f',\tau')+n(f',\tau'), \tag{18}$$

where $z(f',\tau')$ is the mixture signal of the output of the NC, $s(f',\tau')$ is the target signal, $n(f',\tau')$ is the interference signal recorded, f is the frequency bin number, and τ is the time-frame index. For MOSIE, the amplitude spectrum of the target signal is estimated on the basis of the MMSE criterion under a certain target prior. The processed signal $\tilde{s}(f',\tau')$ obtained via MOSIE is given by

$$\tilde{s}(f',\tau') = G(f',\tau')z(f',\tau'), \tag{19}$$

$$G(f',\tau') = \frac{\sqrt{\nu(f,\tau)}}{\gamma(f',\tau')}\cdot\left(\frac{\Gamma(\rho(f)+\beta/2)}{\Gamma(\rho(f))}\cdot\frac{\Phi(1-\beta/2-\rho(f),1,-\nu(f,\tau))}{\Phi(1-\rho(f),1,-\nu(f,\tau))}\right)^{1/\beta}, \tag{20}$$

where $\Gamma(\cdot)$ 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

$$\nu(f',\tau')=\tilde{\gamma}(f,\tau)\tilde{\xi}(f',\tau')\left(1+\tilde{\xi}(f',\tau')\right)^{-1}. \tag{21}$$

Here, $\tilde{\xi}(f',\tau')$ and $\tilde{\gamma}(f',\tau')$ are the estimated a priori and a posteriori SNRs, respectively, which are defined as

$$\tilde{\xi}(f',\tau') = \alpha\tilde{\gamma}(f,\tau-1)G^2(f',\tau') + (1-\alpha)\max[\tilde{\gamma}(f',\tau')-1,0], \tag{22}$$

$$\tilde{\gamma}(f',\tau') = |z(f',\tau')|^2/P_{\tilde{n}}(f), \tag{23}$$

where $P_{\tilde{n}}(f)$ is the estimated interference power spectral density and α is the forgetting factor. In MOSIE, the a priori statistical model of the target signal amplitude spectrum is set to the chi distribution

$$p(z)=2\phi^{\rho(f')}\Gamma(\rho(f))^{-1}z^{2\rho(f')-1}\exp(-\phi z^2), \tag{24}$$

where $p(z)$ is the probability density function (p.d.f.) of signal z in the amplitude domain, $\phi=\rho(f')/E\{|z|^2\}$, and $\rho(f')$ is the shape parameter with respect to the frequency bin number f'. Here, $\rho(f')=1$ gives a Rayleigh distribution that corresponds to a Gaussian distribution in the time domain, and a smaller value of $\rho(f')$ corresponds to a super-Gaussian distribution signal. In this paper, we employ the adaptive estimation method for these parameters proposed in [20] by assuming the stationarity of background noise.

B. Derivation of Equi-Binaural Optimal Spectral Gain

We consider a mixing model with two inputs that is recorded by specific pairwise microphone and assume that the observed signal contains the target signal and an interference signal. Hereafter, the observed signal vector (outputs of the NC) in the time-frequency domain, $\boldsymbol{z}(f',\tau')=[z_{\mathrm{L}}(f',\tau'),z_{\mathrm{R}}(f',\tau')]^t$, is given by

$$\boldsymbol{z}(f',\tau')=\boldsymbol{h}(f)s(f',\tau')+\boldsymbol{n}(f',\tau'), \tag{25}$$

where $\boldsymbol{h}(f)=[h_{\mathrm{L}}(f),h_{\mathrm{R}}(f)]^t$ is the column vector of the transfer functions between the target source and each selected microphone, $s(f',\tau')$ is the target signal component, and $\boldsymbol{n}(f',\tau')=[n_{\mathrm{L}}(f',\tau'),n_{\mathrm{R}}(f',\tau')]^t$ is the column vector of the interference signal that represents the residual noise component in the output of the NC. The derivation of the equi-binaural optimal spectral gain is described below. This is an extended version of [21] for a *generalized* cost function and can be formulated as the minimization problem of the following error e:

$$e=\mathrm{E}\Big[\left\{|h_{\mathrm{L}}(f)s(f',\tau')|^\beta-(G(f',\tau')|z_{\mathrm{L}}(f',\tau')|)^\beta\right\}^2 + \left\{|h_{\mathrm{R}}(f)s(f',\tau')|^\beta-(G(f',\tau')|z_{\mathrm{R}}(f',\tau')|)^\beta\right\}^2\Big], \tag{26}$$

where $G(f',\tau')$ is the equi-binaural spectral gain, which is considered as a variable. The optimization problem based on (26) is given by

$$\begin{aligned}
&G_{\mathrm{opt}}(f',\tau')\\
&=\arg\min_{G(f',\tau')}\mathrm{E}\Big[\left\{|h_{\mathrm{L}}(f)s(f',\tau')|^\beta\right.\\
&\left.-\{(G(f',\tau')-G_{\mathrm{L}}(f',\tau'))+G_{\mathrm{L}}(f',\tau')\}^\beta|z_{\mathrm{L}}(f',\tau')|^\beta\right\}^2\\
&+\left\{(|h_{\mathrm{R}}(f)s(f',\tau'))^\beta|-\{(G(f',\tau')\right.\\
&\left.-G_{\mathrm{R}}(f',\tau'))+G_{\mathrm{R}}(f',\tau')\}^\beta|z_{\mathrm{R}}(f',\tau')|^\beta\right\}^2\Big]\\
&=\arg\min_{G(f',\tau')}\mathrm{E}\left[\{|h_{\mathrm{L}}(f)s(f',\tau')|^\beta-(G_{\mathrm{L}}(f',\tau')|z_{\mathrm{L}}(f',\tau')|)^\beta\}^2\right.\\
&+\{|h_{\mathrm{R}}(f)s(f',\tau')|^\beta-(G_{\mathrm{R}}(f',\tau')|z_{\mathrm{R}}(f',\tau')|)^\beta\}^2\\
&+\{(G^\beta(f',\tau')-G_{\mathrm{L}}^\beta(f',\tau'))|z_{\mathrm{L}}(f',\tau')|^\beta\}^2\\
&\left.+\{(G^\beta(f',\tau')-G_{\mathrm{R}}^\beta(f',\tau'))|z_{\mathrm{R}}(f',\tau')|^\beta\}^2+2C\right],
\end{aligned} \tag{27}$$

where $G_{\mathrm{opt}}(f',\tau')$ is the equi-binaural optimal spectral gain to be estimated, and $G_{\mathrm{L}}(f',\tau')$ and $G_{\mathrm{R}}(f',\tau')$ are individual spectral gains for the L and R ears, respectively, which are auxiliary parameters for calculating an approximate solution of $G_{\mathrm{opt}}(f',\tau')$ because the direct Bayesian estimation of

$G_{\mathrm{opt}}(f',\tau')$ is difficult. In addition, C is related to the correlation between the estimation error and the observed signal in each channel when we estimate the target speech signals in the L and R ears using the parameters $G_{\mathrm{L}}(f',\tau')$ and $G_{\mathrm{R}}(f',\tau')$, and is defined by

$$\begin{aligned} C = &\{G^{\beta}(f',\tau') - G_{\mathrm{L}}^{\beta}(f',\tau')\} \cdot \{(G_{\mathrm{L}}(f',\tau')|z_{\mathrm{L}}(f',\tau')|)^{\beta} \\ &-|h_{\mathrm{L}}(f',\tau')s(f',\tau')|^{\beta}\}|z_{\mathrm{L}}(f',\tau')|^{\beta} \\ &+\{G^{\beta}(f',\tau') - G_{\mathrm{R}}^{\beta}(f',\tau')\} \cdot \{(G_{\mathrm{R}}(f',\tau')|z_{\mathrm{R}}(f',\tau')|)^{\beta} \\ &- |h_{\mathrm{R}}(f',\tau')s(f',\tau')|^{\beta}\}|z_{\mathrm{R}}(f',\tau')|^{\beta}. \end{aligned} \tag{28}$$

We discuss the minimization of (27). First, the first and second terms on the right-hand side correspond to the problem of target signal estimation in each ear. These terms can be minimized if we obtain the optimal values of $G_{\mathrm{L}}(f',\tau')$ and $G_{\mathrm{R}}(f',\tau')$ using MOSIE. Next, C in the fifth term on the right-hand side can be disregarded if the parameters $G_{\mathrm{L}}(f',\tau')$ and $G_{\mathrm{R}}(f',\tau')$ provide an accurate estimate of the target signals by approximately considering C to be negligible. Hence, the remaining third and fourth terms, i.e., $\{(G^{\beta}(f',\tau') - G_{\mathrm{L}}^{\beta}(f',\tau'))|z_{\mathrm{L}}(f',\tau')|^{\beta}\}^2 + \{(G^{\beta}(f',\tau') - G_{\mathrm{R}}^{\beta}(f',\tau'))|z_{\mathrm{R}}(f',\tau')|^{\beta}\}^2$, should be minimized. This problem can be formulated as

$$\begin{aligned} &G_{\mathrm{opt}}(f',\tau') \\ &\doteqdot \arg\min_{G(f',\tau')} \mathrm{E}\left[\{(G^{\beta}(f',\tau') - G_{\mathrm{L_{opt}}}^{\beta}(f',\tau'))|z_{\mathrm{L}}(f',\tau')|^{\beta}\}^2\right. \\ &\left.+\{(G^{\beta}(f',\tau') - G_{\mathrm{R_{opt}}}^{\beta}(f',\tau'))|z_{\mathrm{R}}(f',\tau')|^{\beta}\}^2\right], \end{aligned} \tag{29}$$

subject to

$$\begin{aligned} &G_{\mathrm{L_{opt}}}(f',\tau') \\ &=\arg\min_{G_{\mathrm{L}}(f',\tau')} \mathrm{E}\left[\{|h_{\mathrm{L}}(f)s(f',\tau')|^{\beta} - (G_{\mathrm{L}}(f',\tau')|z_{\mathrm{L}}(f',\tau')|)^{\beta}\}^2\right], \end{aligned} \tag{30}$$

$$\begin{aligned} &G_{\mathrm{R_{opt}}}(f',\tau') \\ &=\arg\min_{G_{\mathrm{R}}(f',\tau')} \mathrm{E}\left[\{|h_{\mathrm{R}}(f)s(f',\tau')|^{\beta} - (G_{\mathrm{R}}(f',\tau')|z_{\mathrm{R}}(f',\tau')|)^{\beta}\}^2\right], \end{aligned} \tag{31}$$

where $G_{\mathrm{Lopt}}(f',\tau')$ and $G_{\mathrm{Ropt}}(f',\tau')$ are the L- and R-ear optimal spectral gains, respectively. To solve (29), we first obtain $G_{\mathrm{Lopt}}(f',\tau')$ and $G_{\mathrm{Ropt}}(f',\tau')$ from MOSIE in (30) and (31). Then by substituting them into (29), we solve the following equation in $G(f',\tau')$:

$$\begin{aligned} \frac{\partial e}{\partial G(f',\tau')} = &G^{\beta}(f',\tau')|z_{\mathrm{L}}(f',\tau')|^{2\beta} \\ &- G_{\mathrm{Lopt}}^{\beta}(f',\tau')|z_{\mathrm{L}}(f',\tau')|^{2\beta} \\ &+ G^{\beta}(f',\tau')|z_{\mathrm{R}}(f',\tau')|^{2\beta} \\ &- G_{\mathrm{Ropt}}^{\beta}(f',\tau')|z_{\mathrm{R}}(f',\tau')|^{2\beta} \\ &= 0. \end{aligned} \tag{32}$$

The solution of (32) is given by

$$\begin{aligned} &G_{\mathrm{opt}}(f',\tau') \\ &= \left(\frac{G_{\mathrm{Lopt}}^{\beta}(f',\tau')|z_{\mathrm{L}}(f',\tau')|^{2\beta} + G_{\mathrm{Ropt}}^{\beta}(f',\tau')|z_{\mathrm{R}}(f',\tau')|^{2\beta}}{|z_{\mathrm{L}}(f',\tau')|^{2\beta} + |z_{\mathrm{R}}(f',\tau')|^{2\beta}}\right)^{1/\beta}. \end{aligned} \tag{33}$$

V. Estimation of Survivor's Location

In the practical use of this robot, it is necessary to determine the precise location of survivors. To achieve this objective, a method of posture estimation for the hose-shaped rescue robot was proposed in [7]. In contrast, to estimate the direction of the observed voice, we have developed a new hose-shaped rescue robot that has a pairwise (stereo) microphone array, allowing it to use stereo sound. Using the new robot, we can hear stereo sound processed by the speech enhancement technique using the observed multichannel noisy signals. Therefore, we experimentally confirm that the operator can perceive the direction of a survivor's location from the processed stereo sound. In the evaluation experiment, we first apply a noise reduction method combining ILRMA and the NC to the multichannel noisy sound including the survivor's voice arriving from a predetermined direction and ego noise. Next, the operator hears the processed sound and is asked to state the direction of the survivor's location. The correct answer rate of the operator is taken as the evaluation value.

A. Selection of Estimated Speech Signal

In the ego noise reduction process, ILRMA cannot determine the permutation of the estimated signal and their signals scales.

The scale ambiguity can easily be resolved by applying a projection-back technique to the estimated signals, namely, all the signal scales are projected onto the observation of the reference microphone. However, we still do not know which estimated signal mostly contains the speech components. Although the selection of the estimated speech signal may be automated by using another criterion, such as kurtosis, in this paper, we assume that the operator can manually select the estimated speech signal $y_s(t)$ from the output signals of ILRMA.

B. Evaluation Method

In the evaluation, we used signals recorded by the hose-shaped rescue robot. Figure 6 shows the positions of the microphones and a speech source and Fig. 7 shows the recording environment that simulated a disaster site. The stereo processed 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. We recorded signals arriving from three different directions 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 $y_s(t)$, by employing the spectrograms and microphones chosen in advance

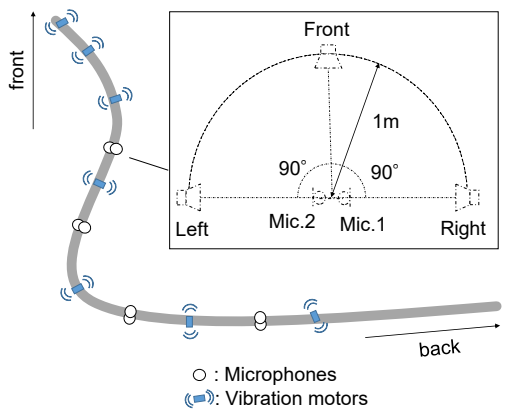

Fig. 6: Position of microphones and a speech source.

Fig. 7: Recording environment.

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 twice. Next, we applied the NC 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 NC to the estimated signal observed by microphone 1. Next, we applied postfiltering using the equi-binaural spectral gain estimated by MOSIE. Finally, the operator hears the stereo sound, which consists of an estimated speech signal observed by microphone 1 as the left channel source and an estimated speech signal observed by microphone 2 as the right channel source. In this experiment, we use 12 signals: three unprocessed signals that include the voice from either the left, right or front and nine processed signals that include the voice from either the left, right or front. 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, right or front. Other experimental conditions are shown in Table I.

TABLE I: Experimental conditions

Sampling frequency	16 kHz
Window length of ILRMA	2048 samples
Window shift of ILRMA	STFT length/4
Number of bases	10
Number of iterations	50
Filter length of noise canceller	1600 taps
Window length of MOSIE	1024 samples
Window shift of MOSIE	STFT length/4
Forgetting factor	0.96
Amplitude compression parameter	0.005
Sound source direction	0/90/180 degrees
Number of subjects	10 person

TABLE II: Accuracy rate

Direction	Unprocessed	ILRMA	ILRMA+NC	proposed method
Left	10	80	80	90
Front	70	70	90	80
Right	10	70	80	80

C. Results

We had listening test to evaluate performance of our new proposed method. Ten subjects joined the test. Each subject was required to answer the direction of the voice as a sound source, left, front or right for a total 12 stimuli. Table II shows propotion of correct answer. According to the result, the subjects perceived the direction of the sound source correctly by listening to the sound processed by our new method. In particular, the proposed method can increase the accuracy rate by 10%, compared with that of conventional ILRMA.

VI. Conclusion

We confirmed that an operator can perceive the direction of a survivor's location by applying our speech enhancement technique to observed multichannel noisy signals recorded by a hose-shaped rescue robot with a pairwise (stereo) microphone array. According to our experimental result, the operator could perceive the direction of a survivor's location by hearing the sound subjected to a noise reduction method combining ILRMA, the NC, and MOSIE.

Acknowledgments

This work was supported by Japan Science and Technology Agency and the 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.

References

[1] "Impulsive Paradigm Change through Disruptive Technologies Program (ImPACT)," http://www.jst.go.jp/impact/program07.html.
[2] 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.
[3] N. Mae, D. Kitamura, M. Ishimura, T. Yamada, and S. Makino, "Ego noise reduction for hose-shaped rescue robot combining independent low-rank matrix analysis and noise cancellation," Proc. APSIPA, 2016.

[4] N. Mae, M. Ishimura, S. Makino, D. Kitamura, N. Ono, T. Yamada, and H. Saruwatari, "Ego noise reduction for hose-shaped rescue robot combining independent low-rank matrix analysis and multichannel noise cancellation," Proc. LVA/ICA, 2017.
[5] 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.
[6] 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.
[7] Y. Bando, T. Otsuka, T. Mizumoto, K. Itoyama, M. Konyo, S. Tadokoro, K. Nakadai and H. G. Okuno, "Posture estimation of hose-shaped robot by using active microphone array", Adv. Rob. vol. 29, no. 1, pp. 35–49, 2015.
[8] 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 IROS, pp. 3679–3686, 2012.
[9] A. Deleforge and W. Kellerman, "Phase-optimized K-SVD for signal extraction from underdetermined multichannel sparse mixtures," Proc. ICASSP, pp. 355–359, 2015.
[10] Y. Bando, K. Itoyama, M. Konyo, S. Tadokoro, K. Nakadai, K. Yoshii, and H. G. Okuno, "Human-voice enhancement based on online RPCA for a hose-shaped rescue robot with a microphone array," Proc. SSRR, 2015.
[11] H. Barfuss and W. Kellerman, "Improving blind source separation performance by adaptive array geometries for humanoid robots," Proc. HSCMA, 2014.
[12] H. Barfuss and W. Kellerman, "An adaptive microphone array topology for target signal extraction with humanoid robots," Proc. IWAENC, pp. 16–20, 2014.
[13] R. Aichner, M. Zourub, H. Buchner, and W. Kellerman, "Post-processing for convolutive blind source separation," Proc. ICASSP, vol. 5, pp. 37–40, 2006.
[14] C. Breithaupt and R. Martin, "Analysis of the decisiondirected SNR estimator for speech enhancement with respect to low-SNR and transient conditions," IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 2, pp. 277–289, 2011.
[15] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of nonnegative matrix factorization with complex-valued data," IEEE Trans. ASLP, vol. 21, no. 5, pp. 971–982, 2013.
[16] 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.
[17] 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.
[18] E. Hansler and G. Schmidt, "Acoustic Echo and Noise Control: A Practical Approach", John Wiley & Sons, New York, 2004.
[19] Y. Murota, D. Kitamura, S. Koyama, H. Saruwatari and S. Nakamura, "Statistical modeling of binaural signal and its application to binaural source separation," Proc. ICASSP, pp. 494–498, 2015.
[20] Y. Murota, D. Kitamura, H. Saruwatari, S. Nakamura, Y. Takahashi and K. Kondo, "Music signal separation based on Bayesian spectral amplitude estimator with automatic target prior adaptation," Proc. ICASSP, pp. 7540–7544, 2014.
[21] H. Saruwatari, M. Go, R. Okamoto, K. Shikano and H. Hosoi, "Binaural hearing aid using sound-localization-preserved MMSE STSA estimator with ICA-based noise estimation," Proc. IWAENC, 2010.