Sound source localization using a single-point stereo microphone for robots

Futoshi Asano* and Mitsuharu Morisawa† and Kenji Kaneko† and Kazuhito Yokoi†
* Department of Computer Science, Kogakuin University, Tokyo, Japan
E-mail: asano@cc.kogakuin.ac.jp
† Intelligent Systems Research Institute, National Institute of Advanced Industrial Science and Technology, Tsukuba,Japan

***Abstract*—In this study, a secondary sound source localizer for robot applications is developed. The purpose of using the secondary sound localizer is to assist the camera and hand manipulation of disaster-response robots. A key requirement for this application is the compactness of the input device. In this study, a single-point 13-mm diameter stereo microphone mountable on the robot hand is employed, and its observation model is developed. The localizer must be usable while the hand scans to collect visual information. This is realized by the conversion of the parameter scan in conventional parametric modeling to the mechanical scan of the hand. Another requirement is robustness against the ego noise of the robot. Two localization methods for a single-point stereo microphone that have noise-whitening functions are proposed. Experimental results show that the proposed method achieves a performance comparable to that achieved by the conventional time delay estimation approach but with a much more compact configuration.**

I. Introduction

Sound source localization is an important function in robot applications and has been intensively researched for many years. In previous studies, such as [1], [2], [3], the main purpose of the sound localizer was as a human interface with which the locations of multiple human speakers could be estimated. For this purpose, the sound source localizer typically employs the frequency-domain approach with a large microphone array consisting of many microphones (e.g., 8 microphones) that are mounted on the head of the robot. In this study, the systems that require a large array aperture (including two-channel systems with a large inter-microphone spacing, e.g., [4]) are termed the *primary* sound localizers.

In recent years, the use of a mobile robot in disaster areas such as devastated nuclear plants has been attracting growing attention [5]. As described in detail in Section II-A, one of the important tasks for robots in this application is to investigate the damages to the facilities and repair them. For this purpose, the information from cameras has been typically used. On the other hand, damages that result in leakage of gases and liquids are often accompanied by the emission of sound. Therefore, the use of sound source localization techniques together with the camera is expected to be effective. As the damaged section is often located in narrow places or behind obstacles, the camera mounted on the hand of the robot is typically used for this task [6]. Thus, it is desired that the sound localizer is also mounted on the hand of the robot. A small localizer mountable on the robot hand is termed the *secondary* sound localizer, and is the topic of this study.

The requirements for this secondary localizer include the following:

- Compactness: As a secondary localizer, the input device must be sufficiently compact so that it is mountable on the hand of the robot.
- Scannability: As the localizer is assumed to be used while the hand scans over the region of interest to collect visual information with the camera, the localizer must be usable during the scanning.
- Robustness: The sound source localizer mounted on the robot often suffers from the ego noise of the robot. Therefore, the localizer must be robust against this noise.

Considering the compactness requirement, this study employs a small tie-clip type *single-point* stereo microphone. In this stereo microphone with a diameter of 13 mm, two directional microphone elements that have cardioid directivity patterns rotated to the left and right are embedded. As regards scannability, scanning of a single-point stereo microphone is realized by converting the parametric scan in the conventional parametric modeling approach into a mechanical scan of the robot hand based on the observation model for a single-point stereo microphone. To ensure robustness, the noise whitening technique is used to eliminate the effect of the ego noise of the robot.

II. Problem statement

A. Scenario

In this subsection, the problem addressed in this study is outlined. One of the important tasks for the robot utilized in disaster areas is to find damaged sections in the facilities and repair them.

As a strategy to find sound sources in the area of interest, the following two-stage sound source localization method is considered. In a large space such as a building or a warehouse, a primary sound source localizer can be used to estimate the location of sound sources. For the humanoid robot HRP-2 that is under consideration in this study, we developed an 8-element microphone array with an array aperture of approximately 20 cm mounted on the head of the robot [7]. Once the locations of the sound sources are estimated, the robot selects one of the sound sources as a target source to be investigated, and moves to the vicinity of the target sound source. This process is termed the first stage localization in this study.

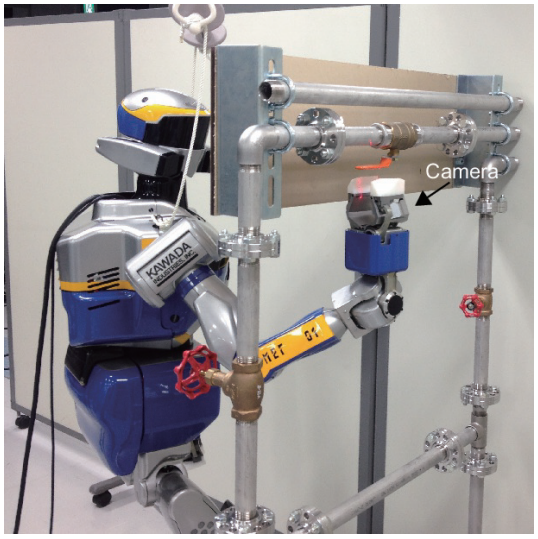

Fig. 1. An example of the tasks for a humanoid robot used in devastated areas. The configuration of the camera mounted on the palm (annotated by "Camera") is shown in Fig. 2. In this figure, microphones are not installed.

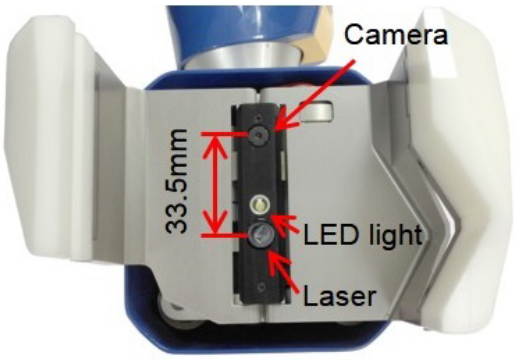

Fig. 2. The camera configuration on the palm of the humanoid robot HRP-2. In this figure, microphones are not installed.

In the second localization stage, the robot investigates the selected target sound source in more detail. Fig. 1 shows an example of the investigation of a damaged section currently conducted with a camera. As depicted in Fig. 2, the robot has a small camera on the palm to investigate the damaged section closely and assist the hand manipulation for repair. When the damaged section is in a narrow place such as those between the pipelines or behind obstacles as depicted in Fig. 1, the sound may reach to the microphone array mounted on the head with diffraction, and therefore the precise localization cannot be expected by using only the primary sound source localizer. Thus, a secondary sound localizer using a small input device that can be mounted on the robot hand is desired. If the sound localizer can be mounted on the hand, the sound information corresponding to the visual information from the camera is available while the hand scans, and therefore the efficiency of investigation is expected to become higher.

Fig. 3. The single-point stereo microphone AT-9901 used in this study.

B. Input device

As an input device for the secondary sound source localizer mountable on the hand of the robot, a *single-point* stereo microphone, the Audio-technica AT-9901, is used in this study and is shown in Fig. 3. Fig. 4 shows the position of AT-9901 tentatively attached to the edge of the robot palm in this study. Unlike a pair of microphones with a large spacing utilized in the conventional time difference of arrival (TDOA) estimation, two small directional microphone elements are embedded in the microphone's body with a diameter of ϕ =13 mm. The time difference τ_{12} between the stereo outputs is less than the propagation time through the microphone's body as

$$\tau_{12} < \frac{\phi}{c} = 38.2\ [\mu s]$$

where c is the velocity of sound. Note that the stereophonic effect based on the inter-channel time difference cannot be expected in the single-point stereo microphone.

To obtain the stereophonic effect, the two microphone elements in AT-9901 have cardioid directivity patterns rotated to the left and right as shown in Fig. 5. As the detailed data of AT-9901 is not available, this figure is used as the schematic for explaining the function of the single-point stereo microphone. These cardioid directivities yield gain difference between the stereo output that depends on the incident angle of sound. Fig. 6 shows the measured inter-channel gain difference as a function of the incident angle. In this study, the gain difference that depends on the incident angle is utilized for localizing sound sources.

C. Scanning

In the conventional sound source localization method, a parametric modeling approach is often used. In this approach, an observation model that includes the parameter of interest, such as the angle of the sound sources, is built. In the estimation process, the parameters in the model are scanned (varied) over the range of interest so that the model is fitted to the actual observation to find the optimal values of the parameters, while the location of the microphone array is fixed. This scan is called "parameter scan" in this study for the sake of convenience. In the frequency-domain approach,

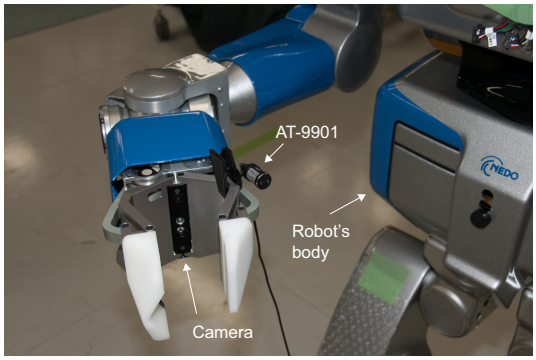

Fig. 4. Configuration of the single-point stereo microphone on the robot hand.

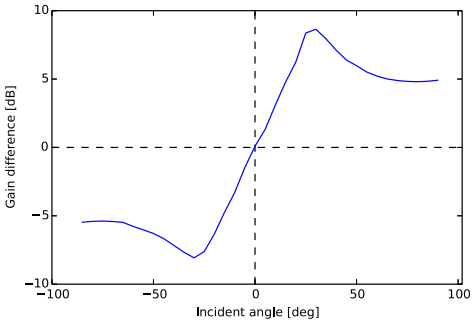

Fig. 6. Inter-channel gain difference of the single-point stereo microphone.

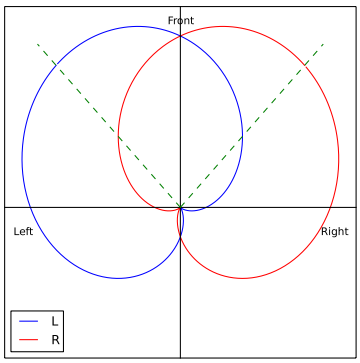

Fig. 5. Schematic diagram of the cardioid directivity of the single-point stereo microphone.

the steering vector (e.g. [8]) in the model is scanned over the possible incident angles.

In this study, by taking advantage of the mobility of the robot hand, the direction of the stereo microphone is scanned over the range of interest by rotating the hand around the wrist joint, while the parameter in the model is fixed. This is termed the "mechanical scan" in this study. It will be shown later in Sections IV and V that parameter scan can be easily converted to mechanical scan. The angle of the source can be estimated by obtaining the value of rotary encoder at the wrist joint when the fitness measure such as the likelihood of the model becomes maximum.

D. Characteristics of the ego noise

In this subsection, the characteristics of the ego noise of the robot are discussed.

Fig. 7 shows the configuration of the microphone. As can be seen from this figure, the hand is located in the vicinity of the body. Thus, the ego noise of the robot that mainly consists of the noise of the servo-motor emitted from the body is observed at the microphone position.

Fig. 8 (a) and (b) shows the short-time power of the observed signal at the right and left channel of the microphone during the scan of the hand. The range of the scan was $[+60°, -60°]$, and the duration of the scan was 15 s. For calculating the short-time signal power, the observation was sectioned into time blocks with a duration 0.5 s and an overlap of 0.25 s, and the averaged power in each block was calculated. In Fig. 8(a) and (b), the target source does not exist, and thus the observation mainly consists of the ego noise of the robot. The solid curve shows the average of the five trials. From this figure, the inter-channel level difference of approximately 7 dB is observed in a large part of the scan. This difference is attributed to the fact that the noise source (the body of the robot) is within the range of the left cardioid pattern shown in Fig. 5. Thus, it can be seen that the ego noise is directional. Moreover, the noise level is a function of the hand angle and becomes maximum when the hand is in the frontal direction for both channels. The reason for this observation is that when the hand is in the frontal direction, the robot's center of mass shifts forward and the servo motor functions intensively to maintain balance, causing the noise power level to become large. In the figure, the standard deviation for the five trials is also shown by the dotted line. It can be seen that the variance of the measurements is small, and thus the curve is reproducible.

Fig. 8 (c) and (d) shows the case when the target sound source exists. As a target source, a loudspeaker (Yamaha MS101-III) was placed at 0° (frontal direction of the robot) as depicted in Fig. 7. The distance of the target source from the center of the wrist joint was 1.5 m.

Fig. 9 shows the power spectrum of the observations for the case with the target signal (blue curve) and that without the target signal (red curve) at the time block when the robot hand was in the frontal direction. It can be seen that the noise mainly consists of low-frequency components, and thus the SNR in the frequency lower than around 800 Hz is low.

Based on these observations, the properties of the ego noise of the robot in this study can be summarized as follows:

- The noise is emitted from the body of the robot, and is directional.
- The SNR is low especially at the left channel.
- The noise level is a function of the robot hand angle.
- The curve of the noise level is reproducible, and thus can be measured in advance.

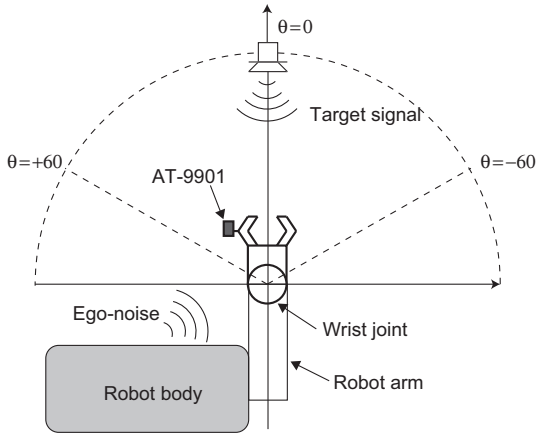

Fig. 7. Configuration of the robot, microphone, and target sound source.

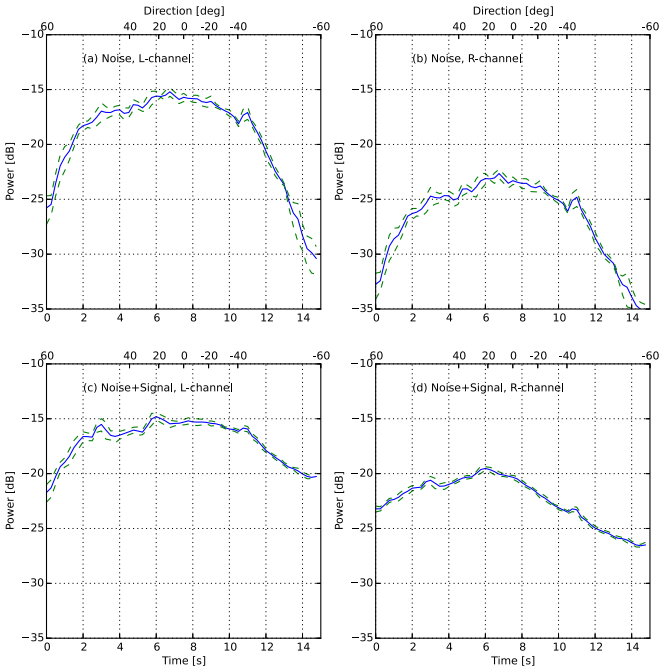

Fig. 8. Short-time power of the microphone observation. (a) and (b): Target source is switched off; (c) and (d): The target source is switched on. (a) and (c): Left channel; (b) and (d): Right channel.

- The noise mainly consists of low-frequency components.

As the ego noise of the robot is unavoidable in this application, the sound localization method developed needs to be robust against the ego noise.

III. Observation Model

In this section, the observation model for the estimation of source angle is discussed. For the sake of simplicity in notation, a static environment in which no hand scanning is conducted is assumed in Section III-A - Section III-B.

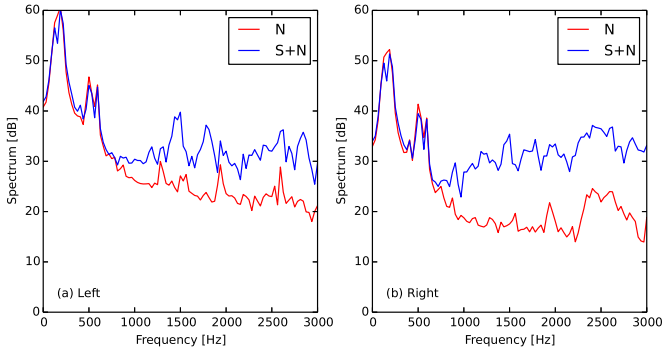

Fig. 9. Spectra of the observation for the left (a) and right (b) microphones. Red line: Target source is switched off (Noise); Blue line: The target source is switched on (Target signal + Noise).

In Section III-C, the time block-based estimation that is necessary for a dynamic environment involving hand scanning is introduced.

A. Gain difference model

In general, the observation at the microphone pair is modeled using the convolution as

$$\boldsymbol{z}_t = \begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix} = \begin{bmatrix} h_1(t) * s(t) + v_1(t) \\ h_2(t) * s(t) + v_2(t) \end{bmatrix} \tag{1}$$

where $z_m(t)$ denotes the observation at the mth microphone and the tth discrete time index. $s(t)$ is the source signal, and $h_m(t)$ denotes the impulse response from the sound source to the mth microphone. The symbol '$*$' denotes the convolution operator.

As described in Section II-B, the inter-channel time difference of the single-point stereo microphone AT-9901 is less than the sampling interval $T_s = 62.5$ μs corresponding to the sampling frequency $f_s = 16$ kHz employed in this study. On the other hand, it has the gain difference that is dependent on the incident angle θ as shown in Fig. 6. Therefore, (1) can be reduced to the following gain difference model:

$$\boldsymbol{z}_t = \begin{bmatrix} b_1(\theta)s(t) + v_1(t) \\ b_2(\theta)s(t) + v_2(t) \end{bmatrix} = \boldsymbol{b}(\theta)s(t) + \boldsymbol{v}_t \tag{2}$$

where $\boldsymbol{b}(\theta) = [b_1(\theta), b_2(\theta)]^T$ and $\boldsymbol{v}_t = [v_1(t), v_2(t)]^T$. The symbol $\cdot^T$ denotes the vector/matrix transpose. The common time delay is omitted. The vector $\boldsymbol{b}(\theta)$ represents the inter-channel gain difference

$$G_{12}(\theta) := \frac{b_2(\theta)}{b_1(\theta)} \tag{3}$$

that is dependent on the direction θ as shown in Fig. 6, and is termed the gain vector hereafter. This gain difference model was previously used in the blind separation of the acoustic sources together with the vertically-stacked two directional microphones to reduce the convolutive mixture to the instantaneous mixture [9], [10].

In the actual microphone device, the precision of the model (2) is degraded to some extent for $\theta \neq 0^\circ$ because the cardioid pattern shown in Fig. 5 differs at different frequencies. This causes some frequency distortion of the signal. As described in Section II-C and later in Section IV-C, the gain vector for the frontal direction $\boldsymbol{b}(\theta = 0^\circ)$ alone is used in the estimation by introducing the mechanical scan. Therefore, the effect of the signal distortion on sound source localization is considered to be small.

B. Noise whitening

From the property of the noise given in Section II-D, the noise $\boldsymbol{v}_t$ is spatially colored in the problem addressed in this study. In this subsection, a basic concept of how the effect of the colored noise is eliminated in the proposed sound source localizer is described.

Assuming that the signal $s(t)$ and the noise $\boldsymbol{v}_t$ are uncorrelated, the covariance matrix of the observation can be modeled using (2) as

$$\boldsymbol{R} := E[\boldsymbol{z}_t \boldsymbol{z}_t^T] = \gamma \boldsymbol{b}(\theta)\boldsymbol{b}^T(\theta) + \boldsymbol{Q} \tag{4}$$

where $\gamma = E[|s(t)|^2]$ and $\boldsymbol{Q} = E[\boldsymbol{v}_t \boldsymbol{v}_t^T]$. The symbol $E[\cdot]$ denotes the expectation operator. In this subsection and Section V, the signal $s(t)$ is modeled as a random variable (random signal model [11], see the discussion in Section IV-A.)

In the conventional parametric modeling approach, the noise $\boldsymbol{v}_t$ is often assumed to be spatially white. By this assumption, the noise covariance matrix becomes a diagonal matrix, i.e., $\boldsymbol{Q} = \sigma_v^2 \boldsymbol{I}$, where σ_v^2 represents the variance of the noise. This makes the derivation of the estimator much easier. In the case when the actual noise is not spatially white, the mismatch between the noise model and the actual noise causes estimation error [12].

A method to avoid this mismatch is the noise whitening. The whitening of $\boldsymbol{v}_t$ is defined as [13]:

$$\tilde{\boldsymbol{v}}_t := \boldsymbol{W}\boldsymbol{v}_t \tag{5}$$

where $\boldsymbol{W}$ is the whitening matrix that satisfies

$$\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{Q}^{-1} \tag{6}$$

Though the matrix $\boldsymbol{W}$ is not uniquely determined, a typical choice is $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{V}^T$ where $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ are the eigenvalue matrix and the eigenvector matrix of $\boldsymbol{Q}$, respectively. By applying the whitening to the observation, i.e., $\tilde{\boldsymbol{z}}_t = \boldsymbol{W}\boldsymbol{z}_t$, and taking its covariance matrix, we have

$$\tilde{\boldsymbol{R}} := E\left[\tilde{\boldsymbol{z}}_t \tilde{\boldsymbol{z}}_t^T\right] = \gamma \boldsymbol{W}\boldsymbol{b}(\theta)\boldsymbol{b}^T(\theta)\boldsymbol{W}^T + \boldsymbol{I} \tag{7}$$

Here we use the following relation obtained from (6).

$$\boldsymbol{W}\boldsymbol{Q}\boldsymbol{W}^T = \boldsymbol{I} \tag{8}$$

By comparing (7) and (4), it can be seen that the noise covariance $\boldsymbol{Q}$ is reduced to the identity matrix.

Instead of using the noise-whitened observation $\tilde{\boldsymbol{z}}_t$ explicitly as employed in [4], the noise-whitening process is embedded in the proposed method as described in Section IV-B and Section V-B.

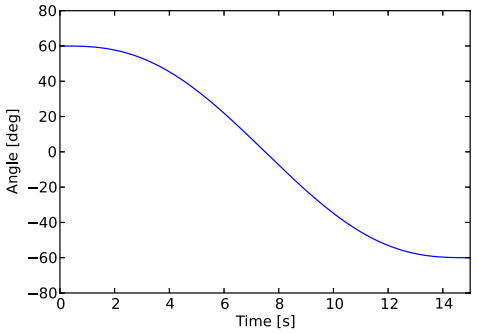

Fig. 10. Joint angle as a function of time.

C. Time block-based estimation

To estimate the source angle with the mechanical scan, time block-based estimation, typically employed in the estimation for dynamic systems, such as the tracking of moving targets [14], is introduced. The observation is sectioned into a time block with a block length of T samples, then the fitting measure between the model and the observation is evaluated in each time block. For time block-based estimation, the observation model (2) is expressed as

$$\boldsymbol{z}_{k,t} = \begin{bmatrix} b_1(\theta_k)s(k,t) + v_1(k,t) \\ b_2(\theta_k)s(k,t) + v_2(k,t) \end{bmatrix} = \boldsymbol{b}(\theta_k)s(k,t) + \boldsymbol{v}_{k,t} \tag{9}$$

where $z_m(k,t)$ denotes the microphone input at the mth channel, the kth time block, and the tth sample in the block.

In the same way as in the estimation for dynamic systems, it is assumed that the signal and noise are stationary and the location of the sound source θ_k is constant within a block. By assuming this, the methods developed for static cases can be applied in each time block. In this study, the robot hand scans in the range of $[-60^\circ, +60^\circ]$ to avoid contact with the body (see Fig. 7 for the configuration). The scanning speed and the block length T are determined on the basis of our previous studies on the tracking of moving human speakers using the primary sound localizer [15], [16]. The total scanning time for the range of $[+60^\circ, -60^\circ]$ was 15 s. Fig. 10 shows the value of the joint angle obtained from the rotary encoder as a function of time during the scan. The block length is 0.5 s (T=8000 samples in 16-kHz sampling) with a block overlap of 0.25 s.

IV. Maximum Likelihood Method

In this section, the maximum likelihood estimator based on the observation model (9) is developed. For the sake of simplicity, the "parameter scan" employed in the conventional parametric modeling approach is used in the derivation of the algorithm in Section IV-A and Section IV-B. The derived estimator is then converted to the "mechanical scan" in Section IV-C.

A. Estimator

In the observation model (9), there are three parameters, i.e., the angle of the sound source θ_k, the source signal $s(k,t)$, and the noise vector $\boldsymbol{v}_{k,t}$. Hereafter, the signal is denoted as $s_{k,t}$ for the sake of simplicity in notation. In this study, θ_k is the parameter to be estimated. Regarding the noise $\boldsymbol{v}_{k,t}$, its covariance $\boldsymbol{Q}_k = E[\boldsymbol{v}_{k,t}\boldsymbol{v}_{k,t}^T]$ is assumed to be known in advance. Regarding the source $s_{k,t}$, the deterministic signal model and the random signal model can be selected [11]. The difference between the two models is that in the deterministic model, $s_{k,t}$ is treated as a fixed but unknown parameter; in the random signal model, $s_{k,t}$ is treated as a random variable and the covariance model (4) is utilized in the fitting. The deterministic model, which is more complex in computation and derivation, is typically used for the estimation of the source signal. In this study, the estimation of the source $s_{k,t}$ is not required. Nevertheless, for the easiness in deriving the noise whitening in the ML approach described in Section IV-B, the deterministic signal model is selected in Section IV.

The likelihood for the signal parameters $\{\theta_k, s_{k,t}\}$ in the kth time block based on the deterministic signal model is given by

$$\begin{aligned} L(\theta_k, s_{k,t}) &= p(\boldsymbol{z}_{k,t}; \theta_k, s_{k,t}) \qquad (10)\\ &\propto \exp\left[-\left(\boldsymbol{z}_{k,t} - \boldsymbol{b}(\theta_k)s_{k,t}\right)^T \boldsymbol{Q}_k^{-1}\left(\boldsymbol{z}_{k,t} - \boldsymbol{b}(\theta_k)s_{k,t}\right)\right] \end{aligned}$$

The noise $\boldsymbol{v}_{k,t}$ in the kth block is assumed to have a Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_k)$. Expanding the likelihood (11) to that for the block data $\boldsymbol{Z}_k = [\boldsymbol{z}_{k,1}, \cdots, \boldsymbol{z}_{k,T}]$ yields

$$\begin{aligned} L(\theta_k, \boldsymbol{S}_k) &= \prod_{t=1}^{T} p(\boldsymbol{z}_{k,t}; \theta_k, s_{k,t}) \qquad (11)\\ &\propto \exp\left[-\sum_{t=1}^{T}(\boldsymbol{z}_{k,t} - \boldsymbol{b}(\theta_k)s_{k,t})^T \boldsymbol{Q}_k^{-1}(\boldsymbol{z}_{k,t} - \boldsymbol{b}(\theta_k)s_{k,t})\right] \end{aligned}$$

where $\boldsymbol{S}_k = [s_{k,1}, \cdots, s_{k,T}]$. The observations $\{\boldsymbol{z}_{k,1}, \cdots, \boldsymbol{z}_{k,T}\}$ are assumed to be mutually statistically independent. From the likelihood equation $\partial L/\partial s_{k,t} = 0$, the signal estimate is obtained as an intermediate estimate as follows:

$$\hat{s}_{k,t} = \frac{\boldsymbol{b}(\theta_k)^T \boldsymbol{Q}_k^{-1}}{\boldsymbol{b}^T(\theta_k)\boldsymbol{Q}_k^{-1}\boldsymbol{b}(\theta_k)}\boldsymbol{z}_{k,t} \qquad (12)$$

By substituting (12) into (11) and taking the logarithm, the log likelihood is written as

$$\begin{aligned} LL(\theta_k) &\propto -\sum_{t=1}^{T}(\boldsymbol{G}_k\boldsymbol{z}_{k,t})^T\boldsymbol{Q}_k^{-1}(\boldsymbol{G}_k\boldsymbol{z}_{k,t})\\ &= -\mathrm{tr}\left(\boldsymbol{Q}_k^{-1}\sum_{t=1}^{T}(\boldsymbol{G}_k\boldsymbol{z}_{k,t})(\boldsymbol{G}_k\boldsymbol{z}_{k,t})^T\right)\\ &= -\mathrm{tr}\left(\boldsymbol{Q}_k^{-1}\boldsymbol{C}_k\right) \qquad (13) \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{G}_k &:= \boldsymbol{I} - \frac{\boldsymbol{b}(\theta_k)\boldsymbol{b}^T(\theta_k)\boldsymbol{Q}_k^{-1}}{\boldsymbol{b}^T(\theta_k)\boldsymbol{Q}_k^{-1}\boldsymbol{b}(\theta_k)} \qquad (14)\\ \boldsymbol{C}_k &:= \boldsymbol{G}_k\bar{\boldsymbol{R}}_k\boldsymbol{G}_k^T \qquad (15)\\ \bar{\boldsymbol{R}}_k &:= \sum_{t=1}^{T}\boldsymbol{z}_{k,t}\boldsymbol{z}_{k,t}^T \qquad (16) \end{aligned}$$

The matrix $\bar{\boldsymbol{R}}_k$ is termed the sample covariance matrix (division by T is omitted.) Using the derived likelihood, the source angle can be estimated as

$$\hat{\theta}_k = \arg\max LL(\theta_k) \qquad (17)$$

B. Noise whitening

In this subsection, the means by which the effect of the noise $\boldsymbol{v}_{k,t}$ is removed in the ML approach is described.

By substituting (9) into (11), and assuming that the estimates of the parameters, $\{\hat{\theta}_k, \hat{s}_{k,t}\}$, are given, the value of the log likelihood is written as

$$LL(\hat{\theta}_k, \hat{\boldsymbol{S}}_k) \propto -\sum_{t=1}^{T}\boldsymbol{e}_{k,t}^T\boldsymbol{Q}_k^{-1}\boldsymbol{e}_{k,t} - \mathrm{tr}\left(\boldsymbol{Q}_k^{-1}\bar{\boldsymbol{Q}}_k\right) + \Omega \qquad (18)$$

where

$$\begin{aligned} \boldsymbol{e}_{k,t} &:= \boldsymbol{b}(\theta_k)s_{k,t} - \boldsymbol{b}(\hat{\theta}_k)\hat{s}_{k,t} \qquad (19)\\ \bar{\boldsymbol{Q}}_k &:= \sum_{t=1}^{T}\boldsymbol{v}_{k,t}\boldsymbol{v}_{k,t}^T \qquad (20)\\ \Omega &:= -\mathrm{tr}\left[\boldsymbol{Q}_k^{-1}\sum_{t=1}^{T}(\boldsymbol{e}_{k,t}\boldsymbol{v}_{k,t}^T + \boldsymbol{v}_{k,t}\boldsymbol{e}_{k,t}^T)\right] \qquad (21) \end{aligned}$$

$\{\theta_k, s_{k,t}\}$ denotes the true value of the parameters and the symbol $\boldsymbol{e}_{k,t}$ denotes the estimation error. When $T \to \infty$, $\frac{1}{T}\bar{\boldsymbol{Q}}_k \to \boldsymbol{Q}_k$, and the second term (divided by T) in (18) becomes

$$\mathrm{tr}\left(\boldsymbol{Q}_k^{-1}\frac{1}{T}\bar{\boldsymbol{Q}}_k\right) \to \dim(\boldsymbol{Q}_k) = 2 \qquad (22)$$

From this, it can be seen that the dependency on $\boldsymbol{v}_{k,t}$ is removed from the second term. By assuming that $\boldsymbol{e}_{k,t}$ and $\boldsymbol{v}_{k,t}$ are uncorrelated, $\Omega \to 0$. Thus, the $\boldsymbol{v}_{k,t}$-related term in (18) becomes constant, and the estimation error $\boldsymbol{e}_{k,t}$ is minimized by maximizing the log likelihood. The second term in (18) can be rewritten as

$$-\mathrm{tr}\left(\sum_{t=1}^{T}\boldsymbol{v}_{k,t}^T\boldsymbol{Q}_k^{-1}\boldsymbol{v}_{k,t}\right) = -\mathrm{tr}\left(\sum_{t=1}^{T}(\boldsymbol{W}\boldsymbol{v}_{k,t})^T(\boldsymbol{W}\boldsymbol{v}_{k,t})\right) \qquad (23)$$

where the matrix $\boldsymbol{W}$ that satisfies $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{Q}_k^{-1}$ is the whitening matrix defined in (6). Thus, it can be seen that the noise-whitening effect described in Section III-B is embedded in the ML approach.

In the actual estimation, (22) approximately holds as T is finite. The mismatch between $\boldsymbol{Q}_k$ and $\frac{1}{T}\bar{\boldsymbol{Q}}_k$ may yield imperfectness in the noise whitening, resulting in deterioration of the localization performance.

C. Conversion to the mechanical scan

The log likelihood for the parameter scan (13) - (16) can be rewritten as

$$LL(\theta_k^{(p)}) \propto -\mathrm{tr}\left(\boldsymbol{Q}_k^{-1}\boldsymbol{C}_k(\theta_k^{(p)})\right) \quad (24)$$

$$\boldsymbol{G}_k(\theta_k^{(p)}) = \boldsymbol{I} - \frac{\boldsymbol{b}(\theta_k^{(p)})\boldsymbol{b}^T(\theta_k^{(p)})\boldsymbol{Q}_k^{-1}}{\boldsymbol{b}^T(\theta_k^{(p)})\boldsymbol{Q}_k^{-1}\boldsymbol{b}(\theta_k^{(p)})} \quad (25)$$

$$\boldsymbol{C}_k(\theta_k^{(p)}) = \boldsymbol{G}_k(\theta_k^{(p)})\bar{\boldsymbol{R}}_k\boldsymbol{G}_k^T(\theta_k^{(p)}) \quad (26)$$

$$\bar{\boldsymbol{R}}_k = \sum_{t=1}^{T}\boldsymbol{z}_{k,t}\boldsymbol{z}_{k,t}^T \quad (27)$$

where the parameter $\theta_k^{(p)}$ to be scanned in the observation model is emphasized.

In the mechanical scan, by rotating the robot hand, an arbitrary angle θ in the world coordinate system is converted to that in the robot-hand coordinate system by

$$\tilde{\theta} = \theta - \theta_k^{(h)} \quad (28)$$

where $\theta_k^{(h)}$ is the angle of the robot hand in the world coordinate. The parameter $\tilde{\theta}_k^{(p)}$ in the observation model is fixed at $\tilde{\theta}_k^{(p)} = 0^\circ$. The corresponding gain vector is denoted as

$$\boldsymbol{b}_0 := \boldsymbol{b}(\tilde{\theta}_k^{(p)} = 0^\circ) \quad (29)$$

On the other hand, the covariance matrix becomes the function of the rotation angle $\theta_k^{(h)}$ as $\bar{\boldsymbol{R}}_k \to \bar{\boldsymbol{R}}_k(\theta_k^{(h)})$ and $\boldsymbol{Q}_k \to \boldsymbol{Q}_k(\theta_k^{(h)})$. By using these, the equations for the parameter scan (24) - (27) are converted to:

$$LL(\theta_k^{(h)}) \propto -\mathrm{tr}\left(\boldsymbol{Q}_k^{-1}(\theta_k^{(h)})\boldsymbol{C}_k(\theta_k^{(h)})\right) \quad (30)$$

$$\boldsymbol{G}_k(\theta_k^{(h)}) = \boldsymbol{I} - \frac{\boldsymbol{b}_0\boldsymbol{b}_0^T\boldsymbol{Q}_k^{-1}(\theta_k^{(h)})}{\boldsymbol{b}_0^T\boldsymbol{Q}_k^{-1}(\theta_k^{(h)})\boldsymbol{b}_0} \quad (31)$$

$$\boldsymbol{C}_k(\theta_k^{(h)}) = \boldsymbol{G}_k(\theta_k^{(h)})\bar{\boldsymbol{R}}_k(\theta_k^{(h)})\boldsymbol{G}_k^T(\theta_k^{(h)}) \quad (32)$$

$$\bar{\boldsymbol{R}}_k(\theta_k^{(h)}) = \sum_{t=1}^{T}\boldsymbol{z}_{k,t}(\theta_k^{(h)})\boldsymbol{z}_{k,t}^T(\theta_k^{(h)}) \quad (33)$$

The source angle is then estimated as:

$$\hat{\theta} = \arg\max_{\theta_k^{(h)}} LL(\theta_k^{(h)}) \quad (34)$$

When the direction of the hand $\theta_k^{(h)}$ matches the source angle, the sound wave arrivess from the frontal direction of the single-point stereo microphone, and the gain vector $\boldsymbol{b}_0$ in the model also matches the true gain vector in the observation $\boldsymbol{z}_{k,t}$, resulting in maximizing the likelihood function.

V. Subspace-based method

In this section, the method for estimating the spacial spectrum based on the subspace approach is developed. In the same manner as that in Section IV, the algorithm is derived using the "parameter scan" in Section V-A and Section V-B, and is then converted to the "mechanical scan" in Section V-C .

A. Estimator

In Section V, the random signal model is employed for the derivation (see the discussion in Section IV-A.) The covariance model (4) is extended to the time block-based estimation as:

$$\boldsymbol{R}_k := E[\boldsymbol{z}_{k,t}\boldsymbol{z}_{k,t}^T] = \gamma_k\boldsymbol{b}(\theta_k)\boldsymbol{b}^T(\theta_k) + \boldsymbol{Q}_k \quad (35)$$

In (35), $\mathrm{rank}\left(\gamma_k\boldsymbol{b}(\theta_k)\boldsymbol{b}^T(\theta_k)\right) = 1$, as this term consists of a single vector $\boldsymbol{b}(\theta_k)$. This is termed the one-rank model of the signal, and the vector $\boldsymbol{b}(\theta_k)$ spans the one-dimesional subspace termed the signal subspace [17], [8]. The orthogonal complement of the signal subspace is termed the noise subspace. Let us denote the signal subspace and the noise subspace as Ψ_S and Ψ_N, respectively. The dimension of the noise subspace is also one as the dimension of the entire vector space is two, i.e., $\dim(\boldsymbol{z}_{k,t}) = 2$. Let us denote the basis vector of the noise subspace as $\boldsymbol{d}$. As $\boldsymbol{b}(\theta) \in \Psi_S$, $\boldsymbol{d} \in \Psi_N$, and $(\Psi_S)^\perp = \Psi_N$, the two vectors $\boldsymbol{b}(\theta)$ and $\boldsymbol{d}$ are orthogonal, i.e.,

$$\boldsymbol{b}^T(\theta_k)\boldsymbol{d} = 0 \quad (36)$$

Based on the orthogonality given by (36), the spatial spectrum defined as

$$P(\theta_k^{(p)}) := \frac{\boldsymbol{b}^T(\theta_k^{(p)})\boldsymbol{b}(\theta_k^{(p)})}{|\boldsymbol{b}^T(\theta_k^{(p)})\boldsymbol{d}|^2} \quad (37)$$

is employed in this study. $\theta_k^{(p)}$ is an arbitrary direction for the parameter scanning. When $\theta_k^{(p)} = \theta_k$ where θ_k is the true direction of the sound source, the denominator in (37) becomes zero because of (36), resulting in $P(\theta_k^{(p)})$ having a peak in the true sound source direction. Therefore, the source angle can be estimated as

$$\hat{\theta}_k = \arg\max_{\theta_k^{(p)}} P(\theta_k^{(p)}) \quad (38)$$

The method of obtaining the vector $\boldsymbol{d}$ is described in Section V-B.

A similar spectrum estimator was employed in the well-known MUSIC (multiple signal classification) [18] or Minimum-norm [19] estimator. The difference between the proposed method and the conventional subspace-based spatial estimator is the observation model. The proposed method is based on the time-domain gain difference model shown in (9), whereas the conventional estimator is based on the frequency-domain observation model.

B. Noise whitening

In this section, the vector $\boldsymbol{d}$ that satisfies (36) is derived by using the GEVD approach. During the derivation, it is shown that the process of GEVD has the noise whitening function [18], [20].

The generalized eigenvalue problem for the subspace approach is given by

$$\boldsymbol{R}_k\boldsymbol{e} = \lambda\boldsymbol{Q}_k\boldsymbol{e} \quad (39)$$

where λ and $\boldsymbol{e}$ denote the eigenvalue and eigenvector, respectively. The eigenvectors that satisfy (39) jointly diagonalize the covariance matrices as

$$\boldsymbol{E}^T \boldsymbol{R}_k \boldsymbol{E} = \boldsymbol{\Lambda} \quad (40)$$

$$\boldsymbol{E}^T \boldsymbol{Q}_k \boldsymbol{E} = \boldsymbol{I} \quad (41)$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2)$ and $\boldsymbol{E} = [\boldsymbol{e}_1, \boldsymbol{e}_2]$ are the eigenvalue matrix and eigenvector matrix, respectively [21]. The eigenvalues and the corresponding eigenvectors are assumed to be sorted in a descending order with respect to the eigenvalues. Comparing (41) with (8), it can be seen that $\boldsymbol{E}^T$ is the whitening matrix for the noise $\boldsymbol{v}_{k,t}$. Thus, the GEVD approach has the noise whitening function that eliminates the effect of noise in the estimator (37).

The generalized eigenvalue problem given by (39) is equivalent to the following standard eigenvalue problem [21]:

$$\left(\boldsymbol{W}\boldsymbol{R}_k\boldsymbol{W}^T\right)\boldsymbol{f} = \lambda \boldsymbol{f} \quad (42)$$

where

$$\boldsymbol{f} = \boldsymbol{W}^{-T}\boldsymbol{e} \quad (43)$$

By substituting (35) into (42),

$$\left(\boldsymbol{W}\gamma_k \boldsymbol{b}\boldsymbol{b}^T\boldsymbol{W}^T + \boldsymbol{I}\right)\boldsymbol{f} = \lambda \boldsymbol{f} \quad (44)$$

Here, $\boldsymbol{b}(\theta_k^{(p)})$ is denoted as $\boldsymbol{b}$ for the sake of simplicity. As the rank of $\gamma_k \boldsymbol{b}\boldsymbol{b}^T$ is one, the rank of $\boldsymbol{W}\gamma_k \boldsymbol{b}\boldsymbol{b}^T\boldsymbol{W}^T$ is also one; thus, the smaller eigenvalue of $\boldsymbol{W}\gamma_k \boldsymbol{b}\boldsymbol{b}^T\boldsymbol{W}^T$ is zero. Let us denote the non-zero eigenvalue of $\boldsymbol{W}\gamma_k \boldsymbol{b}\boldsymbol{b}^T\boldsymbol{W}^T$ by μ. When the identity matrix $\boldsymbol{I}$ is added to $\boldsymbol{W}\gamma_k \boldsymbol{b}\boldsymbol{b}^T\boldsymbol{W}^T$, the eigenvectors do not change while the eigenvalues are added by one. Thus the eigenvalues $\{\lambda_i\}$ become

$$\lambda_i = \begin{cases} \mu + 1 & i = 1 \\ 1 & i = 2 \end{cases} \quad (45)$$

By multiplying the second eigenvector $\boldsymbol{f}_2^T$ corresponding to $\lambda_2 (= 1)$ from the left hand side of (44), and using (45),

$$\boldsymbol{f}_2^T \boldsymbol{W}\gamma_k \boldsymbol{b}\boldsymbol{b}^T\boldsymbol{W}^T\boldsymbol{f}_2 = 0 \quad (46)$$

By substituting (43) into (46),

$$\gamma_k |\boldsymbol{b}^T\boldsymbol{e}_2|^2 = 0 \quad (47)$$

Assuming $\gamma_k \neq 0$,

$$\boldsymbol{b}^T\boldsymbol{e}_2 = 0 \quad (48)$$

From (48), it can be seen that the vector $\boldsymbol{e}_2$ has the orthogonal property shown in (36). Thus, the eigenvector $\boldsymbol{e}_2$ can be used as $\boldsymbol{d}$.

C. Conversion to the mechanical scan

In the same manner as Section IV-C, the spatial spectrum estimator for the subspace method (37) can be converted to that for the mechanical scan as

$$P(\theta_k^{(h)}) := \frac{\boldsymbol{b}_0^T\boldsymbol{b}_0}{|\boldsymbol{b}_0^T\boldsymbol{e}_2(\theta_k^{(h)})|^2} \quad (49)$$

where $\boldsymbol{e}_2(\theta_k^{(h)})$ is the eigenvector for the smaller eigenvalue in the following GEVD problem:

$$\boldsymbol{R}_k(\theta_k^{(h)})\boldsymbol{e} = \lambda \boldsymbol{Q}_k(\theta_k^{(h)})\boldsymbol{e} \quad (50)$$

VI. Experiment

A. Conditions

The experiment was conducted in a large experiment room for robots with a reverberation time (RT_{60}) of approximately 0.3 s. A single sound source (loudspeaker, Yamaha MS101-III) was located on a circle of radius 1.5 m as depicted in Fig. 7. White Gaussian noise was emitted from this source as a target signal $s_{k,t}$ so that the frequency characteristics of the source signal does not affect the estimation performance. Another reason for this choice is that the white noise is considered to have characteristics similar to that of gas leakage sounds from pipelines. The right wrist joint of the robot, which is the center of the rotation of the right hand, was placed at the center of the circle. The direction of the sound source was selected from $\{+40°, +20°, 0°\}$. The conditions for the hand scanning are described in Section III-C

The ego noise of the robot was measured before the experiment was performed. Five trials were recorded and used to obtain the averaged noise covariance as

$$\bar{\boldsymbol{Q}}_k = \frac{1}{5}\sum_{i=1}^{5}\sum_{t=1}^{T}\boldsymbol{v}_{i,k,t}\boldsymbol{v}_{i,k,t}^T$$

where i is the index for the trial. Regarding the gain vector $\boldsymbol{b}_0$, the vector $\boldsymbol{b}_0 = [b_1(\theta = 0), b_2(\theta = 0)]^T$ measured in Section II-B was used.

For the experiment including the target sound source, five trials were recorded for each location of the sound source. As shown in Fig. 9, the ego noise is dominant below 800 Hz. Therefore, the following two cases: (a) the observation using the entire frequency range (denoted as"low SNR"), and (b) the observation processed by a high-pass filter with a cutoff frequency of 800 Hz (denoted as "high SNR"), were tested to compare the cases of different SNRs.

B. Results of the ML method

Fig. 11 shows the variation of likelihood as a function of time (lower axis) and joint angle (upper axis). The sound source was located at 20° as indicated by the vertical dotted line. For the noise covariance matrix, the previously measured covariance matrix $\bar{\boldsymbol{Q}}_k$, and the covariance matrix corresponding to the spatially white noise $\boldsymbol{I}$, were tested for the sake of comparison. For the high SNR case, the likelihood was at a maximum in the direction close to the true location in either case when $\bar{\boldsymbol{Q}}_k$ or $\boldsymbol{I}$ was employed (Fig. 11(a)-(b)). For the low SNR case with the noise covariance matrix $\boldsymbol{I}$, the peak of the likelihood was greatly shifted (Fig. 11(c)). When the measured noise covariance matrix $\bar{\boldsymbol{Q}}_k$ was employed, the maximum remained in the vicinity of the source location (Fig. 11(d)), though the peak was somewhat vague compared with the high SNR case.

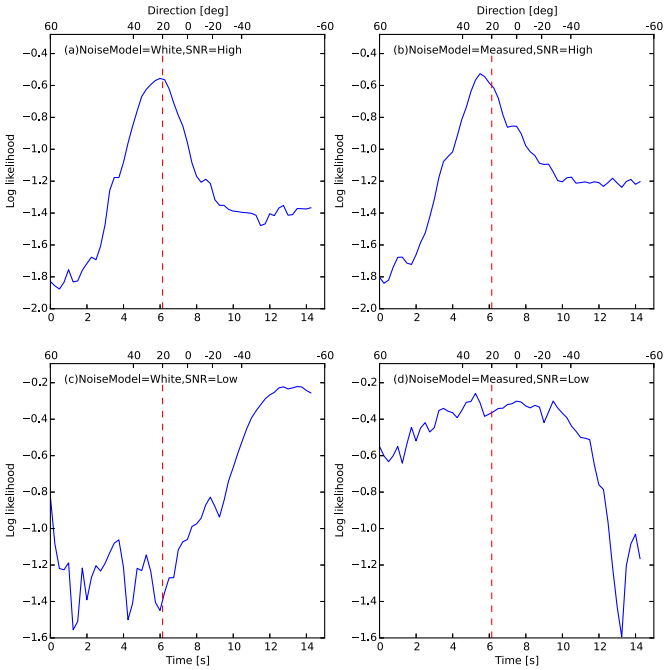

Fig. 11. Log likelihood of the ML-based method for the case of the true source angle $\theta = 20°$. (a) and (b): High SNR; (c) and (d): Low SNR. (a) and (c): White Noise Model; (b) and (d): Measured noise model.

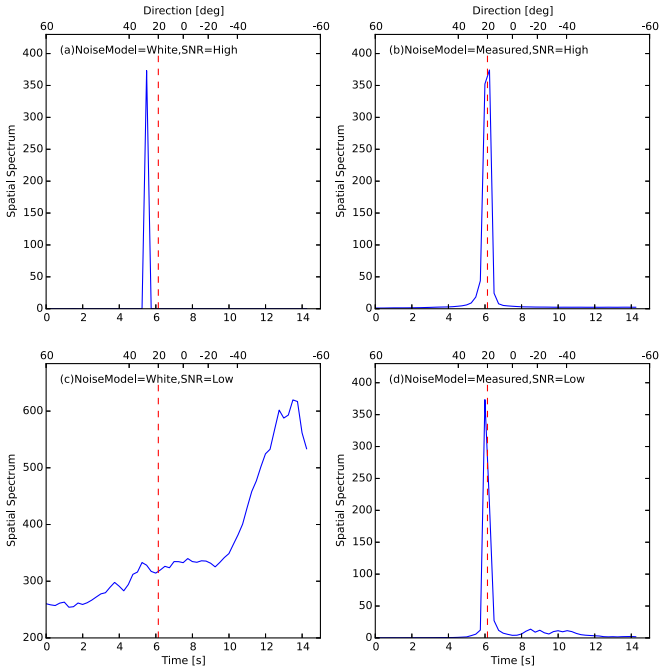

Fig. 12. Spatial spectrum of the subspace-based method for the case of the true source angle $\theta = 20°$. (a) and (b): High SNR; (c) and (d): Low SNR. (a) and (c): White Noise Model; (b) and (d): Measured noise model.

Table I shows the mean value and the standard deviation of the estimated angles for the five trials. For the high-SNR data, both the bias error (the difference between "Mean" and "True" angle) and the random error (standard deviation denoted as "SD") are small. For the low-SNR data, the results are strongly affected by the presence of the noise when the spatially white noise covariance matrix $\boldsymbol{I}$ was employed. When the measured noise covariance matrix $\bar{\boldsymbol{Q}}_k$ was employed, the estimation accuracy was recovered to some extent. Nevertheless, both the bias error and the random error were larger than those for the high-SNR data.

TABLE I
MEAN AND STANDARD DEVIATION (SD) OF THE ESTIMATED ANGLE FOR THE ML METHOD.

True angle	High SNR				Low SNR			
	White		Measured		White		Measured	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0	6.1	3.0	–0.7	1.5	–58.8	8.7	9.0	14.1
20	20.6	1.8	26.0	1.5	–59.1	7.6	16.3	9.8
40	36.6	1.8	37.8	2.4	–59.0	9.5	34.6	18.5

C. Results of the subspace method

Fig. 12 shows the spatial spectrum given by (49) for the subspace method. Compared with those using the ML method, the peak is much sharper. This is because a null is formed in the denominator of (49) when $\boldsymbol{b}_0$ in the model matches that in the observation. For the low-SNR data, in the same way as that with the ML method, a poor result was obtained when the spatially white noise covariance matrix $\boldsymbol{I}$ was employed. When the measured noise covariance matrix $\bar{\boldsymbol{Q}}_k$ was employed, the estimation performance was completely recovered. These results show that noise whitening is essential for the problem addressed in this study. From the results for the low-SNR data shown in Table II, the effect of noise whitening was confirmed.

TABLE II
MEAN AND STANDARD DEVIATION (SD) OF THE ESTIMATED ANGLE FOR THE SUBSPACE METHOD.

True angle	High SNR				Low SNR			
	White		Measured		White		Measured	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0	6.8	1.5	3.8	0.0	–57.9	7.7	–2.2	2.9
20	44.9	27.0	18.5	0.0	–58.8	8.6	18.5	2.4
40	47.2	1.5	37.2	1.5	–60.0	0.0	39.4	1.8

D. Comparison with the TDOA approach

In this subsection, the results of the proposed methods are compared with those of the conventional TDOA approach. For the TDOA approach, a pair of monaural omnidirectional microphones (Sony ECM-C115) that were placed on both sides of the robot hand with spacing $d = 12$ cm was used. As an estimator, the CSP method [22] was employed.

Table III shows the mean value and the standard deviation of the estimated angles for the TDOA approach. For both the high- and low-SNR data, some bias error with a small SD value, such as 0.0, was observed. This is attributed to the quantization effect in TDOA approach. The estimation precision of the TDOA approach is considered to be comparable to the proposed method for the currently discussed microphone configuration.

TABLE III
MEAN AND STANDARD DEVIATION (SD) OF THE ESTIMATED ANGLE FOR THE TDOA METHOD.

True angle	High SNR		Low SNR	
	Mean	SD	Mean	SD
0	-4.3	0.0	-4.3	0.0
20	16.4	0.0	16.4	0.0
40	30.4	5.2	30.4	5.2

VII. CONCLUSION AND DISCUSSION

In this study, a secondary sound source localizer for a humanoid robot that is assumed to be used in disaster areas was proposed. A requirement for this application is the compactness of the input device to ensure that it is mountable on the hand of the robot. The proposed method employs a single-point stereo microphone with a diameter of 13 mm as an input device, which can be easily mounted on the hand of the robot. The observation model for this microphone was developed and the source angle estimators based on this model were derived.

Furthermore, the secondary sound source localizer must be usable while the robot hand scans over the region of interest to collect visual information with the camera. This is accomplished by converting the parameter scan in the conventional parametric modeling approach into the mechanical scan of the hand based on the mathematical equivalence of the observation model between the parameter scan and the mechanical scan.

Another important issue involves the reduction of the effect of the robot ego noise. In the current application, the noise level is a function of the hand angle and is independently observable. Based on this noise characteristic, two methods, i.e., the ML-based method and the subspace-based method that incorporate a noise whitening function were proposed.

The results of the experiment show that the subspace method exhibited a higher spatial resolution compared to the ML method. This is attributed to the fact that the subspace method is a null-based approach as was described in Section V. As regards the noise whitening, the ML method is more robust against the modeling error of the noise covariance matrix compared to the subspace method. This is attributed to the fact that the subspace method is dependent on the one-rank model of the signal. The comparison with the conventional TDOA approach shows that the proposed method achieves a performance comparable to that achieved by the TDOA approach but with a much more compact configuration.

REFERENCES

[1] J.M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *Proc. IEEE ICRA 2004*, 2004, pp. 1033–1038.

[2] K. Nakadai, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and implementation of robot audition system HARK," *Advanced robotics*, vol. 24, pp. 739–761, 2009.

[3] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *Proc. IROS 2009*, 2009, pp. 664–669.

[4] Alban Portello, Patrick Danès, Sylvain Argentieri, and Sylvain Pledel, "Hrtf-based source azimuth estimation and activity detection from a binaural sensor," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, November 2013, pp. 2908–2913.

[5] DARPA, "Darpa robotics challenge," 2013.

[6] Kenji Kaneko, Toshio Ueshiba, Takashi Yoshimi, Yoshihiro Kawai, Mitsuharu Morisawa, Fumio Kanehiro, and Kazuhito Yokoi, "Development of sensor system built into a robot hand (in japanese)," in *Proc. of the 31st Annual Conference of the Robotics Society of Japan*, 2013, number 1C2-01.

[7] Isao Hara, Futoshi Asano, Hideki Asoh, Jun Ogata, Naoyuki Ichimura, Yoshihiro Kawai, Fumio Kanehiro, Hirohisa Hirukawa, and Kiyoshi Yamamoto, "Robust speech interface based on audio and video information fusion for humanoid hrp-2," in *Proc. IROS2004*, 2004.

[8] D. H. Johnson and D. E. Dudgeon, *Array signal processing*, Prentice Hall, Englewood Cliffs NJ, 1993.

[9] Masanori Ito, Yoshinori Takeuchi, Tetsuya Matsumoto, Hiroaki Kudo, Mitsuru Kawamoto, Toshiharu Mukai, and Noboru Ohnishi, "Moving-source separation using directional microphones," in *Proceedings of the 2nd IEEE International Symposium on Signal Processing and Information Technology (ISSPIT2002)*, December 2002, pp. pp.523–526.

[10] Masanori Ito, Mitsuru Kawamoto, Noboru Ohnishi, and Toshiharu Mukai, "A solution to blind separation of moving sources," *Transactions of the Society of Instrument and Control Engineers (in Japanese)*, vol. 41, no. 8, pp. 692–701, 2005.

[11] M. Miller and D. Fuhrmann, "Maximum-likelihood narrow-band direction finding and the EM algorithm," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 38, no. 9, pp. 1560–1577, 1990.

[12] F. Asano and H. Asoh, "Sound source localization using joint Bayesian estimation with a hierarchical noise model," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1953–1965, 2013.

[13] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, Wiley, New York, 2001.

[14] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman filter*, Artech house, Norwood, MA, 2004.

[15] H. Asoh, I. Hara, F. Asano, and K. Yamamoto, "Tracking human speech events using a particle filter," in *Proc. ICASSP 2005*, 2005, vol. II, pp. 1153–1156.

[16] A. Quinlan, M. Kawamoto, Y. Matsusaka, H. Asoh, and F. Asano, "Tracking intermittently speaking multiple speakers using a particle filter," *EURASIP journal on audio, speech and music processing*, vol. 2009, 2009, Article ID 67302.

[17] S. U. Pillai, *Array Signal Processing*, Springer-Verlag, New York, 1989.

[18] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagation*, vol. AP-34, no. 3, pp. 276–280, March 1986.

[19] R. Kumaresan and D. W. Tufts, "Estimating the angles of arrival of multiple plane waves," *IEEE Trans. Aerospace, Electro. System*, vol. AES-19, no. 1, pp. 134–139, January 1983.

[20] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 37, no. 7, pp. 984–995, July 1989.

[21] G. Strang, *Linear Algebra and Its Application*, Harcourt Brace Jovanovich Inc., Orlando, 1988.

[22] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 3, pp. 288–292, 1997.