The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Intelligent Sound Source Localization for Dynamic Environments

Keisuke Nakamura, Kazuhiro Nakadai, Futoshi Asano, Yuji Hasegawa, and Hiroshi Tsujino

***Abstract*— As robotic technology plays an increasing role in human lives, "robot audition", human-robot communication, is of great interest, and robot audition needs to be robust and adaptable for dynamic environments. This paper addresses sound source localization working in dynamic environments for robots. Previously, noise robustness and dynamic localized sound selection have been enormous issues for practical use. To correct the issues, a new localization system "Selective Attention System" is proposed. The system has four new functions: localization with Generalized EigenValue Decomposition of correlation matrices for noise robustness("Localization with GEVD"), sound source cancellation and focus ("Target Source Selection"), human-like dynamic Focus of Attention ("Dynamic FoA"), and correlation matrix estimation for robotic head rotation ("Correlation Matrix Estimation"). All are achieved by the dynamic design of correlation matrices. The system is implemented into a humanoid robot, and the experimental validation is successfully verified even when the robot microphones move dynamically.**

I. Introduction

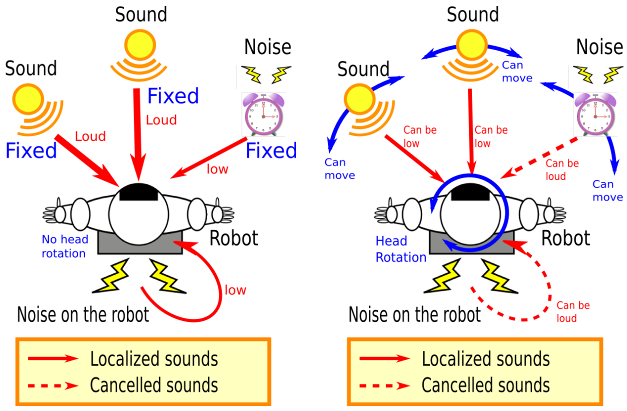
In recent years, robot technology has been rapidly developed, and more and more robots such as humanoid robots work with humans. For practical use of the robots, "Robot audition" proposed in [1], is now of great interest.

Sound source localization is one of the most fundamental processes for robot audition since results of localization induce post-signal processes such as sound source separation and speech recognition [2], [3].

Because of the importance, localization methods have been widely researched for several decades. Some outstanding methods are beamforming [4], [5], and MUSIC (Multiple SIgnal Classification) [6], [12] (See Sec. II-A for the detail of MUSIC), and MUSIC was applied to real world applications such as robotics since the peak of the spatial spectrum is more detectable than other reported methods.

However, these methods which used to focus on static conditions have difficulties in being applied to flexible dynamic environments. Most of them used to hold the following assumptions.

A1) Target sources have stronger power than noise sources
A2) No selective function of target sound sources
A3) Stationary sound sources
A4) Stationary microphone array



(a) Reported situation (b) Proposed situation

Fig. 1. Situation considered

Fig. 1(a) shows a schematic image of a reported situation from previous research.

As a matter of fact, a robot in a real environment is surrounded by many types of sounds such as its own fan noise. Therefore, A1 cannot be always satisfied. Moreover, a microphone array on the robot moves as the robot's head moves dynamically, so A3 and A4 are not always fulfilled. A2 is also not an ignorable problem since multiple signals and noises make the peak of the spectrum more undetectable. Therefore, the assumptions A1-A4 can be critical issues for real robot localization.

To solve the issues, a new sound source localization method is proposed. Fig. 1(b) shows the schematic image of the considered situation in the proposed method.

For A1, GEVD (Generalized EigenValue Decomposition) method is adopted [13]. The method is based on MUSIC, but a noise correlation matrix is additionally used in order to suppress environmental noises.

For A2, dynamic design of correlation matrices for GEVD is proposed in order to select specific sounds to be cancelled or focused on. Hereinafter, the function for cancelling and focusing on sounds is called "Target Source Selection".

As a further application of the Target Source Selection, the correlation matrices can be designed to specify the speed of cancelling and focusing on sounds with respect to their importance. It realizes human-like intelligent localization and is called "Dynamic FoA (Focus of Attention)" in this work.

For A3 and A4, dynamic correlation matrix transformation is proposed with regards to the robot's head orientation,

K. Nakamura is with the Department of Mechenical and Control Engineering, Tokyo Institute of Technology, 2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan. nakamura@sc.ctrl.titech.ac.jp

K. Nakadai is with the Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako, Saitama, 351-0114, Japan, and also with the Department of Information Science and Engineering, Tokyo Institute of Technology. nakadai@jp.honda-ri.com

F. Asano is with the Information Technology Research Institute, National Institute of Advanced Industrial Science and Technology, Tsukuba Central 2, 1-1-1 Umezono, Tsukuba 305-8568, Japan, and also with the Honda Research Institute Japan Co., Ltd. f.asano@aist.go.jp

Y. Hasegawa and H. Tsujino are with the Honda Research Institute Japan Co., Ltd. {yuji.hasegawa, tsujino}@jp.honda-ri.com

978-1-4244-3804-4/09/$25.00 ©2009 IEEE

TABLE I

PARAMETERS OF THE LOCALIZATION MODEL

M	Number of microphones
L	Number of sound sources
m	Index for the microphones ($1 \leq m \leq M$)
l	Index for the sound sources ($1 \leq l \leq L$)
θ_l	Direction of l-th sound source [deg]
ϕ	Orientation of the robot's head [deg]
ψ	Orientation of the steering vector [deg]
$S_l(\omega, \theta_l)$	Signal of the l-th sound source in frequency domain
$s_l(t, \theta_l)$	$S_l(\omega, \theta_l)$ in time domain
$N_m(\omega, \phi)$	Additive noise measured by m-th microphone with the head direction ϕ
$\boldsymbol{N}(\omega, \phi)$	$[N_1(\omega, \phi), N_2(\omega, \phi), ..., N_M(\omega, \phi)]^T$
$n_m(t, \phi)$	$N_m(\omega, \phi)$ in time domain
$\boldsymbol{n}(t, \phi)$	$[n_1(\omega, \phi), n_2(\omega, \phi), ..., n_M(\omega, \phi)]^T$
$A_{m,l}(\omega, \phi, \theta_l)$	Transfer function between l-th source and m-th microphone when the head orientation is ϕ
$\boldsymbol{A}_l(\omega, \phi, \theta_l)$	$[A_{1,l}(\omega, \phi, \theta_l), ..., A_{M,l}(\omega, \phi, \theta_l)]^T$
$a_{m,l}(t, \phi, \theta_l)$	$A_{m,l}(\omega, \phi, \theta_l)$ in time domain
$X_m(\omega)$	Signal measured by m-th microphone in frequency domain
$\boldsymbol{X}(\omega)$	$[X_1(\omega), X_2(\omega), ..., X_M(\omega)]^T$
$x_m(t)$	$X_m(\omega)$ in time domain
$\boldsymbol{x}(t)$	$[x_1(t), x_2(t), ..., x_M(t)]^T$
$\boldsymbol{G}(\omega, \psi)$	Steering vector toward the direction of ψ
$\boldsymbol{R}(\omega, \phi)$	Correlation matrix of measured signal with the head orientation ϕ ($\in \mathbb{C}^{M \times M}$)
$\boldsymbol{e}_m(\omega, \phi)$	Eigenvalue vectors of $\boldsymbol{R}(\omega, \phi)$
$\boldsymbol{E}(\omega, \phi)$	Eigenvalue vector matrix $[\boldsymbol{e}_1, ..., \boldsymbol{e}_M]$
λ_m	Eigenvalues of $\boldsymbol{R}(\omega, \phi)$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_M$
$\boldsymbol{\Lambda}$	diag$(\lambda_1, ..., \lambda_M)$

which is called "Correlation Matrix Estimation". It accomplishes localization over when rotating the robot's head.

The GEVD method, Target Source Selection, Dynamic FoA, and Correlation Matrix Estimation are integrated into a system called "Selective Attention System", and it is implemented for the Honda ASIMO robot, and the system was evaluated in simulated and real environments.

II. PRELIMINARIES

This section explains MUSIC for further discussion.

Here, the signal model commonly used in array processing is considered. The parameters are listed in Table I.

The model of signal measured by each microphone is regarded as following a linear receiving system.

$$x_m(t) = \sum_{l=1}^{L} \{a_{m,l}(t, \phi, \theta_l) s_l(t, \theta_l)\} + n_m(t, \phi) \ , \qquad (1)$$

where $n_m(t, \phi)$ is additive noise mainly in respect of environmental noises. The Fourier transform of $\boldsymbol{x}(t)$ is

$$\boldsymbol{X}(\omega) = \sum_{l=1}^{L} \{\boldsymbol{A}_l(\omega, \phi, \theta_l) S_l(\omega, \theta_l)\} + \boldsymbol{N}(\omega, \phi) \ . \qquad (2)$$

A. MUSIC Method [6]

In this section, the well known subspace-based method, MUSIC is briefly introduced for later discussion.

The first step is to derive the steering vector $\boldsymbol{G}(\omega, \psi)$ in advance of localization. Suppose $\phi = 0, L = 1$, and $n_m(t, \phi) = 0$. When $s_1(t, \theta_1)$ is an impulse signal, The Fourier transform of $x_m(t) = a_{m,1}(t, 0, \theta_1) s_1(t, \theta_1)$ is represented as

$$\boldsymbol{X}(\omega) = \boldsymbol{A}_1(\omega, 0, \theta_1) S_1(\omega, \theta_1) = \boldsymbol{A}_1(\omega, 0, \theta_1) \qquad (3)$$

since $S_1(\omega, \theta_1) = 1$. $\boldsymbol{X}(\omega)$ at arbitrary angle θ_1 is defined as a steering vector. Let $\boldsymbol{G}(\omega, \psi)$ be the pre-measured steering vector at each ψ represented as

$$\boldsymbol{G}(\omega, \psi) = \boldsymbol{A}_1(\omega, 0, \psi) \ . \qquad (4)$$

During localization, $\boldsymbol{x}(t)$ is measured at sampling time τ. (Notice that now $n_m(t, \phi)$ of (1) cannot be assumed to be zero.) Let $\boldsymbol{X}(\omega, \tau)$ be the Fourier transform of $\boldsymbol{x}(\tau)$. Then, the correlation matrices of $\boldsymbol{x}(\tau)$ are defined as

$$\boldsymbol{R}(\omega, \phi) = \boldsymbol{X}(\omega, \tau) \boldsymbol{X}^*(\omega, \tau) \ , \qquad (5)$$

where $()^*$ represents the complex conjugate transpose operator. Suppose $\phi = 0$ for simplicity. The eigenvalue decomposition of $\boldsymbol{R}(\omega, \phi)$ is obtained as

$$\boldsymbol{R}(\omega, \phi) = \boldsymbol{E}(\omega, \phi) \boldsymbol{\Lambda} \boldsymbol{E}^{-1}(\omega, \phi) \ . \qquad (6)$$

Since λ_m represents the power of each sound, λ_i and $\boldsymbol{e}_i$ when $1 \leq i \leq L$ are the eigenvalues and vectors in terms of the sound sources, and λ_i and $\boldsymbol{e}_i$ when $L + 1 \leq i \leq M$ are those of noises. The spatial spectrum is defined as

$$P(\omega, \psi) = \frac{|\boldsymbol{G}^*(\omega, \psi) \boldsymbol{G}(\omega, \psi)|}{\sum_{m=L+1}^{M} |\boldsymbol{G}^*(\omega, \psi) \boldsymbol{e}_m|} \ . \qquad (7)$$

Thus, when the direction of steering vector $\boldsymbol{G}(\omega, \psi)$ and that of a sound source is the same, $P(\omega, \psi)$ is theoretically infinity[1]. Therefore, MUSIC provides easy detectable and reliable peaks and has been used for robot localization. It is also easy to be implemented for robots because of its low computational cost.

However, the method only works when the assumption A1 is satisfied. In case of robot localization, the issue is not avoidable since $\boldsymbol{N}(\omega, \phi)$ of (2) is greater than the first term in many cases. Then, some of $\boldsymbol{e}_i (1 \leq i \leq L)$ are chosen from noises, and (7) returns undesired peaks.

III. SELECTIVE ATTENTION SYSTEM

In this section, the proposed localization system for achieving A1-A4 in Sec. I is introduced.

The parameters additionally defined for the discussion in this section are listed in Table II.

A. GEVD Method

In order to solve the problem of MUSIC, the GEVD method is utilized as a main localization method. The mathematical properties of GEVD are described in [9].

The problem is that when the power of noises is stronger than that of sounds to be localized, the eigenvectors for noises are mistakenly used as these for sounds.

The way to solve the problem is to determine the correlation matrix in terms of noises $\boldsymbol{N}(\omega, \phi)$. Let $\boldsymbol{K}(\omega, \phi)$ be

[1] In practical use, the noises cannot be assumed to be white, and the noises are actually cross correlated with the sound sources. Therefore, the peak is not going to be infinity.

TABLE II

ADDITIONAL PARAMETERS FOR SEC. III

$\boldsymbol{R}(\omega,\phi)$	Correlation matrix of measured signals (defined in the previous section)
$\boldsymbol{K}(\omega,\phi)$	Correlation matrix of pre-measured noises (defined in the previous section)
$\boldsymbol{C}_l(\omega,\phi)$	Correlation matrix for the l-th signal
$\boldsymbol{R}_l(\omega,\phi)$	Correlation matrix of measured signals when the l-th sound is detected.
$\boldsymbol{V}(\omega,\phi)$	Designed correlation matrix
D_l	Decay parameter for l-th sound($-1 \leq D_l \leq 1$)
$\boldsymbol{e}_{l,m}(\omega,\phi)$	Eigenvalue vectors of $\boldsymbol{C}_l(\omega,\phi)$
$\boldsymbol{E}_l(\omega,\phi)$	Eigenvalue vector matrix $[\boldsymbol{e}_{l,1},...,\boldsymbol{e}_{l,M}]$
$\lambda_{l,m}$	Eigenvalues of $\boldsymbol{C}_l(\omega,\phi)$, where $\lambda_{l,1} \geq \lambda_{l,2} \geq ... \geq \lambda_{l,M}$
$\boldsymbol{\Lambda}_l$	$\mathrm{diag}(\lambda_{l,1},...,\lambda_{l,M})$
$\hat{\boldsymbol{C}}_l(\omega,\phi,\phi')$	Estimated Correlation matrix of the l-th signal from ϕ to ϕ'
$\boldsymbol{I}$	M-th order identity matrix
$\boldsymbol{C}_l^{D_l}(\omega,\phi)$	Correlation matrix in the Dynamic FoA for the l-th signal

the correlation matrix derived by noise sources, which is described as

$$\boldsymbol{K}(\omega,\phi) = \boldsymbol{N}(\omega,\phi)\boldsymbol{N}^*(\omega,\phi) \ , \tag{8}$$

where $\boldsymbol{N}(\omega,\phi)$ can be measured when all $\boldsymbol{S}_l(\omega,\theta_l) = 0$ in (2). Then, the GEVD is described as

$$\boldsymbol{R}(\omega,\phi)\hat{\boldsymbol{e}}_m(\omega,\phi) = \hat{\lambda}_m \boldsymbol{K}(\omega,\phi)\hat{\boldsymbol{e}}_m(\omega,\phi) \ , \tag{9}$$

where $\hat{\lambda}_m$ and $\hat{\boldsymbol{e}}_m$ are new eigenvalues and vectors. This decomposition suppresses the noises.

If $\boldsymbol{K}$ is a regular matrix, the decomposition is simplified as a normal eigenvalue decomposition problem

$$\boldsymbol{K}^{-1}(\omega,\phi)\boldsymbol{R}(\omega,\phi)\hat{\boldsymbol{e}}_m(\omega,\phi) = \hat{\lambda}_m \hat{\boldsymbol{e}}_m(\omega,\phi) \ , \tag{10}$$

When the noises are uncorrelated to the sounds, $\boldsymbol{K}$ is not a regular matrix. However, in most cases of robot sound source localization, they are cross correlated, so (10) can be adopted as an equation of GEVD.

The new GEVD spatial spectrum is described as

$$\hat{P}(\omega,\psi) = \frac{|\boldsymbol{G}^*(\omega,\psi).\boldsymbol{G}(\omega,\psi)|}{\sum_{m=L+1}^{M} |\boldsymbol{G}^*(\omega,\psi).\hat{\boldsymbol{e}}_m|} \ . \tag{11}$$

Now, all noises are suppressed, so the eigenvectors of the noise sources are not chosen any more. This is a robust localization method for those noises.

B. Target Source Selection

In the GEVD method, noise correlation matrix $\boldsymbol{K}(\omega,\phi)$ is utilized to suppress noises $\boldsymbol{N}(\omega,\phi)$.

As an application of the method, it is shown that the method can select specific sounds to be localized and to be cancelled by the appropriate design of correlation matrices.

In Sec. III-A, (10) is used for cancelling the noises. In the decomposition, the inverse of $\boldsymbol{K}(\omega,\phi)$ can be regarded as a "cancel operator" of noises $\boldsymbol{N}(\omega,\phi)$ from the original correlation matrix $\boldsymbol{R}(\omega,\phi)$. Viceversa, $\boldsymbol{R}(\omega,\phi)$ can be regarded as a "focus operator" of all these sounds and noises.

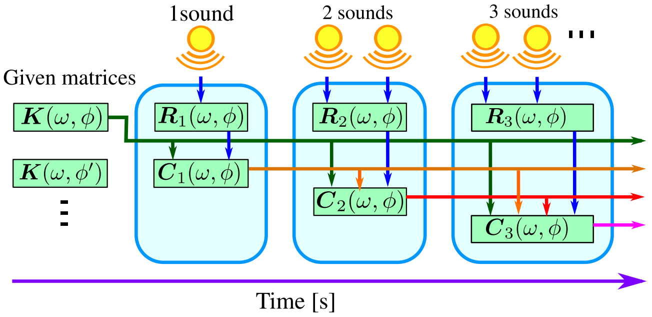

1sound
2 sounds
3 sounds
Given matrices
$\boldsymbol{K}(\omega,\phi)$
$\boldsymbol{K}(\omega,\phi')$
$\boldsymbol{R}_1(\omega,\phi)$
$\boldsymbol{C}_1(\omega,\phi)$
$\boldsymbol{R}_2(\omega,\phi)$
$\boldsymbol{C}_2(\omega,\phi)$
$\boldsymbol{R}_3(\omega,\phi)$
$\boldsymbol{C}_3(\omega,\phi)$
Time [s]

Fig. 2. Steps to derive each correlation matrix

The main idea of Target Source Selection is based on the cancel-and-focus operators. By the selection of those operators, we can design the desired localization environment.

The first step is to determine the operator for each sound. Assume that no sounds are detected coincidentally.

When no sound is in the environment, the microphone array senses only environmental noises $\boldsymbol{N}(\omega,\phi)$ such as a robot's own noise etc. which is pre-measured by ϕ in advance, and its correlation matrix is defined as $\boldsymbol{K}(\omega,\phi)$.

When the first sound is detected, a new correlation matrix of measured signal $\boldsymbol{R}_1(\omega,\phi)$ is obtained, which has information of both environmental noises and the first sound,

$$\boldsymbol{X}(\omega) = \boldsymbol{A}_1(\omega,\phi,\theta_1)S_1(\omega,\theta_1) + \boldsymbol{N}(\omega,\phi) \ . \tag{12}$$

Then, the correlation matrix of the first sound is

$$\boldsymbol{C}_1(\omega,\phi) = \boldsymbol{K}^{-1}(\omega,\phi)\boldsymbol{R}_1(\omega,\phi) \ , \tag{13}$$

since $\boldsymbol{K}^{-1}$ cancels the noise elements from $\boldsymbol{R}_1$. Intuitively, $\boldsymbol{C}_1(\omega,\phi)$ can be regarded as the correlation matrix of $\boldsymbol{A}_1(\omega,\phi,\theta_1)S_1(\omega,\theta_1)$. [2]

Same as the step, when the microphone array detects the second sound, $\boldsymbol{R}_2(\omega,\phi)$ is determined, and the correlation matrix for the second sound is obtained as

$$\boldsymbol{C}_2(\omega,\phi) = \boldsymbol{C}_1^{-1}(\omega,\phi)\boldsymbol{K}^{-1}(\omega,\phi)\boldsymbol{R}_2(\omega,\phi) \ . \tag{14}$$

Repeating the process up to the L-th sound, correlation matrices $\boldsymbol{C}_1,...,\boldsymbol{C}_L$ are determined. The general term of $\boldsymbol{C}_l$ is described as follows.

$$\boldsymbol{C}_l(\omega,\phi) = \prod_{i=1}^{l} \boldsymbol{C}_i^{-1}(\omega,\phi)\boldsymbol{K}^{-1}(\omega,\phi)\boldsymbol{R}_l(\omega,\phi) \ . \tag{15}$$

Thus, $\boldsymbol{C}_l$ is the focus operator for l-th sound, and $\boldsymbol{C}_l^{-1}$ is the cancel operator for l-th sound $\boldsymbol{A}_l(\omega,\phi,\theta_l)S_l(\omega,\theta_l)$.

The intuitive image of the steps to derive each correlation matrix is shown in Fig. 2. As seen in the hierarchical structure, it is easy to be implemented to robotic hardware which is one of the advantages of the function.

For cancelling and focusing on arbitrary sounds, $\boldsymbol{V}(\omega,\phi)$ is defined as a designed correlation matrix. The general form of $\boldsymbol{V}$ is described as

$$\boldsymbol{V}(\omega,\phi) = \prod_{i=1}^{L} \boldsymbol{C}_i^{p_i}(\omega,\phi)\boldsymbol{K}^{-1}(\omega,\phi) \ , \tag{16}$$

[2] Strictly speaking, there is an assumption that the noise and sounds are uncorrelated so that (12) does not have the cross term.

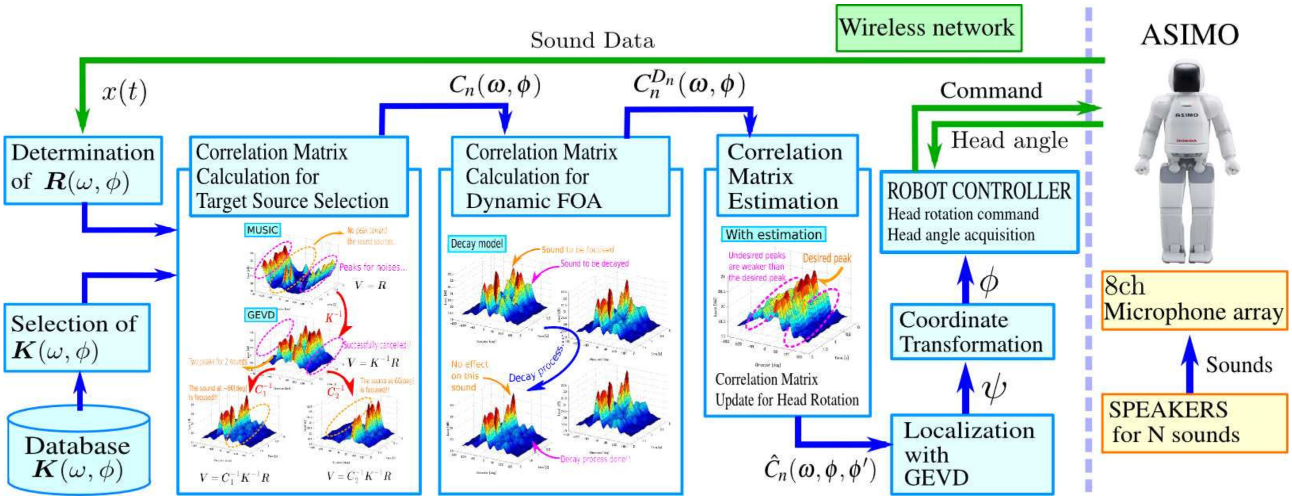


Fig. 3. Architecture of the proposed system

where p_i is integer $-1 \le p_i \le 1$. Then, (10) is rewritten as

$$\boldsymbol{V}(\omega,\phi)\boldsymbol{R}(\omega,\phi)\hat{\boldsymbol{e}}_m(\omega,\phi) = \hat{\lambda}_m\hat{\boldsymbol{e}}_m(\omega,\phi) \ , \qquad (17)$$

which enables the system to localize arbitrary sounds.

C. Dynamic FoA

In Sec. III-B, focusing on and cancelling sounds are discretely carried out. Namely, p_i in (16) can only be an integer. For localization, discrete focusing and cancelling are not significant problems. However, for "intelligent" localization, it should have a Dynamic FoA function.

Let us consider $\boldsymbol{C}_l$ again. The main idea is to transform each $\boldsymbol{C}_l$ continuously from I to $\boldsymbol{C}_l^{-1}$ and viceversa. (Here, only decaying is examined, so only $\boldsymbol{C}_l^{-1}$ is used. For focusing, $\boldsymbol{C}_l$ can be utilized with the same discussion.)

The eigenvalue decomposition of $\boldsymbol{C}_l^{-1}$ is described as

$$\boldsymbol{C}_l^{-1}(\omega,\phi) = \boldsymbol{E}_l(\omega,\phi)\boldsymbol{\Lambda}_l^{-1}\boldsymbol{E}^{-1}(\omega,\phi) \ , \qquad (18)$$

where $\boldsymbol{\Lambda}_l^{-1} = \mathrm{diag}(\lambda_{l,1}^{-1},...,\lambda_{l,M}^{-1})$.

For decaying,

$$\boldsymbol{C}_l^{D_l}(\omega,\phi) = \boldsymbol{E}_l(\omega,\phi)\mathrm{diag}(\lambda_{l,1}^{D_l},...,\lambda_{l,M}^{D_l})\boldsymbol{E}^{-1}(\omega,\phi) \qquad (19)$$

is defined, and D_l is continuously changed from 0 to -1, depending on the importance of each sound. In this paper, the time since the robot detects each sound is regarded as importance. Thus, the decay parameter is gradually getting smaller as time progresses.

When we hear a new sound, we notice the sound and check if it is the target sound. When we hear a non-target sound, we cancel the sound gradually. The Dynamic FoA achieves such human-like intelligent localization.

D. Correlation Matrix Estimation

In the above discussion, sound source localization without head rotation is investigated. In this section, localization with variable ϕ is considered.

The issue is that head rotation changes the orientation of the microphone array and all the sounds and noises. Even when $\boldsymbol{C}_l(\omega,\phi)$ is obtained by (15), head rotation from ϕ to ϕ' changes all correlation matrices into $\boldsymbol{C}_l(\omega,\phi')$. Thus, the estimation of $\boldsymbol{C}_l(\omega,\phi')$ from $\boldsymbol{C}_l(\omega,\phi)$ is required.

Here, the steering vector is used for estimation. As mentioned in Sec. III-B, $\boldsymbol{C}_l(\omega,\phi)$ can be regarded as correlation matrix of $\boldsymbol{A}_l(\omega,\phi,\theta_l)S_l(\omega,\theta_l)$, which is described as

$$\begin{aligned}\boldsymbol{C}_l(\omega,\phi) &= \{\boldsymbol{A}_l(\phi,\theta_l)S_l(\theta_l)\}\{\boldsymbol{A}_l(\phi,\theta_l)S_l(\theta_l)\}^* \\ &= \boldsymbol{A}_l(\phi,\theta_l)\boldsymbol{A}_l^*(\phi,\theta_l)S_l(\theta_l)S_l^*(\theta_l) \ . \qquad (20)\end{aligned}$$

$S_lS_l^*$ is not ϕ-dependent, so the transformation of $\boldsymbol{C}_l$ from ϕ to ϕ' is obtained as

$$\boldsymbol{T}_l(\omega,\phi,\phi') = \{\boldsymbol{A}_l(\phi)\boldsymbol{A}_l^*(\phi)\}^{-1}\{\boldsymbol{A}_l(\phi')\boldsymbol{A}_l^*(\phi')\} \ . \qquad (21)$$

Algorithmically, when the robot rotates its head, all $\boldsymbol{C}_l(\omega,\phi)$ is transformed as

$$\hat{\boldsymbol{C}}_l(\omega,\phi,\phi') = \boldsymbol{T}_l(\omega,\phi,\phi')\boldsymbol{C}_l(\omega,\phi) \ . \qquad (22)$$

IV. System Implementation

One of the remarkable points of these functions is that the functions proposed in Sec. III are "sound-independent operations", which algorithmically can be easily implemented.

Since the operation can deal with sounds one by one dynamically, it can be said that the integrated system is flexible for various environmental changes.

Fig. 3 shows the architecture of the proposed system.

On the PC side, the operation steps for localization are described as follows.

1) Determination of $R_l(\omega,\phi)$ $(1 \le l \le L)$
2) Determination of $\boldsymbol{C}_l(\omega,\phi)$ by (15)
3) Decision for Target Source Selection by p_i in (16)
4) Decision for the Dynamic FoA by defining D_l in (19) depending on the importance of the sounds
5) Transformation of $\boldsymbol{C}_l(\omega,\phi)$ with head rotation (22)

In this work, the Honda ASIMO robot with an embedded 8-ch microphone array is used as a robot for experimental validation. The position of the array is on its head, and it has an uniform-circular-array-like geometry.

Two speakers are located at 60[deg] and -60[deg] of the front side of ASIMO and emit different sounds (can be the same). The distance between ASIMO and the sounds is 1[m].

The robot has its own fan, which is regarded as a strong directional noise. The sounds lower than the fan noise are used for the validation of GEVD.

The architecture of the experimental setup is also shown in Fig. 3. The control PC and ASIMO are connected through a wireless network. ASIMO sends the data of the sounds recorded by the array, and the PC sends the angle for its head rotation. For real time processing, all the proposed functions are implemented as component blocks of HARK robot audition software [10] based on *FlowDesigner* [11], which is C++ based. All operations on the PC side are carried out by a laptop with a 2.5 GHz Intel Core 2 Duo CPU and 2 GB SDRAM running Linux.

The next section shows the results of simulations and experiments for validation.

V. Evaluation

In this section, the validity of the proposed method is evaluated by both simulations and experiments. In Sec. V-A, numerical comparisons between MUSIC and GEVD are shown in order to see how robust the GEVD method is for environmental noises. Sec. V-B gives experimental results of each function proposed in Sec. III.

The common conditions for validation are as follows.

- $K(\omega,\phi)$ is given by 5 [deg].
- There are 2 sounds at 60 [deg] and -60 [deg].
- Main robot noise is from an angle of 180[deg]
- The power of environmental noises $\boldsymbol{N}(\omega,\phi)$ is stronger than that of sounds to be localized.
- The sound from the angle of -60 [deg] is detected first, then that of 60 [deg] is detected afterwards.
- The steering vector $\boldsymbol{G}(\omega,\psi)$ is given by 5 [deg]. Namely, $\psi = \{-175,-170,...,180\}$[deg].
- The head rotation is done when the robot detects a sound, and it tries to face the sound.

A. Numerical Comparison Between MUSIC and GEVD

This section assumes that head orientation is fixed at 0[deg] for the evaluation.

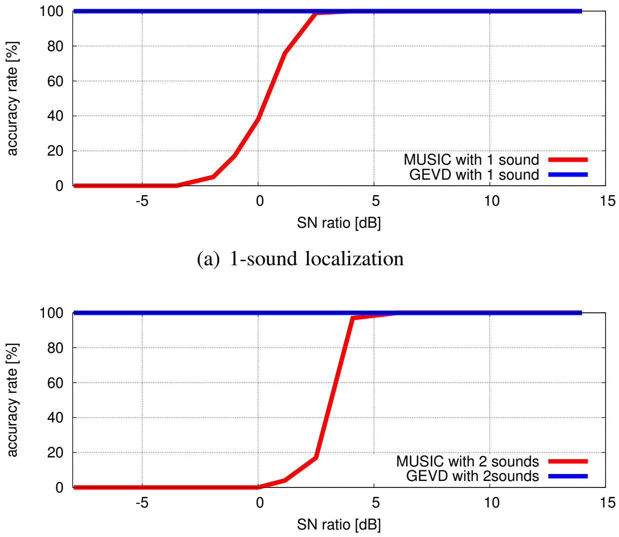
Fig. 4 shows the result. The horizontal axis represents the Signal-to-Noise (SN) ratio of the sources, and the vertical axis shows the localization accuracy. Specifically, the graph shows how many peaks are correctly detected in 100 frames.

Both methods were compared for 1- and 2-sound localization. In both cases, MUSIC performance was obviously degraded because of the assumption A1 in Sec. I. On the other hand, GEVD perfectly localized even when the SN ratio became negative.

In both cases, GEVD was more robust for noises than MUSIC. The validity of GEVD is successfully verified.

B. Experimental Validation of the Integrated System

Fig. 5 shows the snapshots of the experiment with the system proposed in Sec. IV.



(a) 1-sound localization

(b) 2-sound localization

Fig. 4. Comparison between MUSIC and GEVD

Fig. 5(a) shows the first phase of an experiment. ϕ is now 0[deg], and the speaker at -60[deg] starts emitting a low clock-alarm sound $s_1(t,-60)$. The right side of the figure shows the spatial spectrum of both MUSIC and GEVD when the robot notices the sound. As seen in the figure, MUSIC detects a high peak towards the fan noise, which is on the back side of the robot (180[deg] in the figure). Clearly, the A1 is not satisfied in this case. (In fact, the sound used in the experiment is much lower than the fan noise.) So, in the experiments with MUSIC, the robot cannot rotate its head. GEVD, however, can cancel the noise by using $\boldsymbol{K}(\omega,0)$, and the peak is apparently on the front side of the robot (-60[deg]). Head rotation is also successfully verified under the loud noise. Noise robustness of the GEVD method is clearly confirmed by the experimental result.

Fig. 5(b) represents the second phase of the experiment. Now, the robot faces $s_1(t,-60)$ and tries to cancel the sound using Dynamic FoA. First, $\boldsymbol{R}_1(\omega,-60)$ is calculated, and $\boldsymbol{C}_1(\omega,-60)$ is derived in order to hear a new sound. Dynamic FoA starts working when $\boldsymbol{C}_1(\omega,-60)$ is determined by

$$\boldsymbol{C}_1(\omega,-60) = \boldsymbol{K}^{-1}(\omega,-60)\boldsymbol{R}_1(\omega,-60) \ . \qquad (23)$$

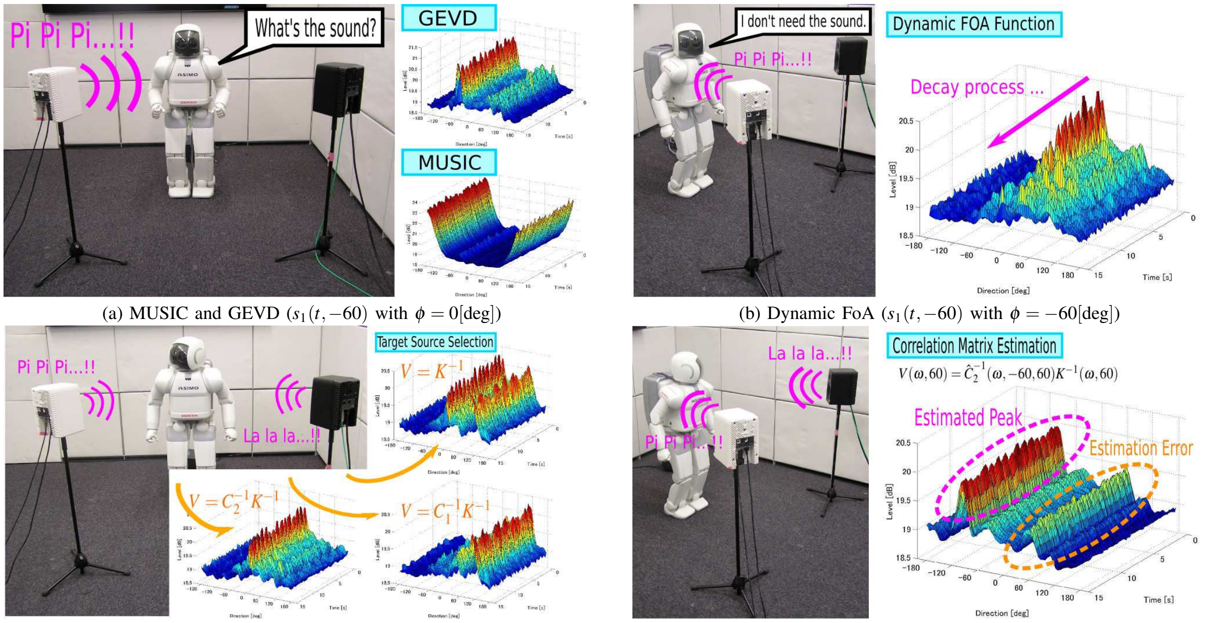
It is seen that the function is also working properly (See the right side of the figure). When it finishes cancelling the sound, it gets ready to hear new sounds.

In Fig. 5(c), the speaker at 60[deg] emits low music $s_2(t,60)$. The robot currently has correlation matrices $\boldsymbol{K}_1(\omega,-60)$,$\boldsymbol{R}_1(\omega,-60)$, and $\boldsymbol{C}_1(\omega,-60)$. When it hears the new sound, $\boldsymbol{C}_2(\omega,-60)$ is obtained from $\boldsymbol{R}_2(\omega,-60)$ as

$$\boldsymbol{C}_2(\omega,-60) = \boldsymbol{C}_1^{-1}(\omega,-60)\boldsymbol{K}^{-1}(\omega,-60)\boldsymbol{R}_2(\omega,-60) \ .$$

Therefore, it can decide to focus on or cancel each sound using Target Source Selection. The right side of the figure shows the result. The system successfully selects sounds by the appropriate design of the correlation matrices.

Now, the robot faces $s_2(t,60)$ in Fig. 5(d). Since, it does not have $\boldsymbol{C}_1(\omega,60)$ and $\boldsymbol{C}_2(\omega,60)$, Correlation Matrix



(a) MUSIC and GEVD ($s_1(t,-60)$ with $\phi = 0$[deg])

(b) Dynamic FoA ($s_1(t,-60)$ with $\phi = -60$[deg])

(c) Target Source Selection ($s_1(t,-60)$ and $s_2(t,60)$ with $\phi = -60$[deg])

(d) Correlation Matrix Estimation ($s_1(t,-60)$ and $s_2(t,60)$ with $\phi = 60$[deg])

Fig. 5. Snapshots of an experiment with the integrated system

Estimation is applied. The estimated matrices are derived as

$$\hat{\boldsymbol{C}}_1(\omega,-60,60) = \boldsymbol{T}_1(\omega,-60,60)\boldsymbol{C}_1(\omega,-60) \quad (24)$$

$$\hat{\boldsymbol{C}}_2(\omega,-60,60) = \boldsymbol{T}_2(\omega,-60,60)\boldsymbol{C}_2(\omega,-60)\ . \quad (25)$$

The right side of the figure shows the localization result when

$$\boldsymbol{V}(\omega,60) = \hat{\boldsymbol{C}}_2^{-1}(\omega,-60,60)\boldsymbol{K}^{-1}(\omega,60) \qquad (26)$$

is used. As seen in the figure, the peak for the sound is the highest, so estimation also works in a real environment. In fact, the estimation result included some undesired peaks. Even though it is not a big problem, an adaptive technique is necessary as future work to improve localization.

The validity of each function is successfully confirmed not only by numerical simulations but also by the experiments.

VI. Conclusion And Future Work

In this paper, the following functions were integrated in order to realize robust and intelligent localization, and the total system is proposed as Selective Attention System.

- **GEVD method** for cancelling environmental noises
- **Target Source Selection** for Selective Attention.
- **Dynamic FoA** for intelligent human-like localization.
- **Correlation Matrix Estimation** for head rotation

The validity of the system was confirmed not only by numerical simulations but also by experiments.

The following two points are considered as future work.

As mentioned in Sec. V, localization with the correlation matrix estimation has small undesired peaks in the spatial spectrum. Some adaptive subsystems are needed for more reliable localization.

In this paper, we need to consider localization "during" head rotation. During head rotation, the relative speed of sounds are extremely fast, so it is hard to achieve accurate localization now. Also, our localization method uses discrete noise correlation matrices which are digitized every 5 degrees. Therefore, continuous localization in respect of head angle is necessary. Some approximation methods for continuous head angles are to be examined.

References

[1] K. Nakadai *et al.*, "Active audition for humanoid", in *Proc. of 7th National Conf. on AAAI 2000*, pp. 832–839, 2000.

[2] K. Nakadai *et al.*, "A robot referee for rock-paper-scissors sound games", in *Proc. of ICRA 2008*, pp. 3469–3474, 2008.

[3] H. G. Okuno and K. Nakadai, "Computational Auditory Scene Analysis and its Application to Robot Audition", in *Proc. of the Hands-Free Speech Communication and Microphone Arrays, 2008 (HSCMA 2008)*, pp. 124–127, 2008.

[4] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering", in *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[5] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice-Hall, 1993.

[6] R. Schmidt, "Multiple emitter location and signal parameter estimation", in *IEEE Trans. on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[7] R. A. Kennedy *et al.*, "Broadband nearfield beamforming using a radial beampattern transformation", in *IEEE Trans. on Signal Processing*, vol. 46, no. 8, pp. 2147–2156, 1998.

[8] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach", in *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, 1996.

[9] G. Strang, *Linear Algebra and its Applications Third Edition*, Harcourt Brace Jovanovich, 1988.

[10] K. Nakadai *et al.*, "An Open Source Software System For Robot Audition HARK and Its Evaluation", in *Proc. of 8th IEEE-RAS Int. Conf. on Humanoid Robots*, 2008.

[11] C. Cote *et al.*, "Code reusability tools for programming mobile robots", in *Proc. of IROS 2004*, vol. 2, pp. 1820–1825, 2004.

[12] F. Asano *et al.*, "Real-time sound source localization and separation system and its application to automatic speech recognition", in *Proc. of EUROSPEECH-2001*, pp.1013–1016.

[13] F. Asano *et al.*, "Localization and extraction of brain activity using generalized eigenvalue decomposition", in *Proc. of ICASSP 2008*, pp.565–568.