2012 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 7-12, 2012. Vilamoura, Algarve, Portugal

# Real-time Super-resolution Sound Source Localization for Robots

Keisuke Nakamura, Kazuhiro Nakadai, and Gökhan Ince

*Abstract*— **Sound Source Localization (SSL) is an essential function for robot audition and yields the location and number of sound sources, which are utilized for post-processes such as sound source separation. SSL for a robot in a real environment mainly requires noise-robustness, high resolution and real-time processing. A technique using microphone array processing, that is, Multiple Signal Classification based on Standard Eigen-Value Decomposition (SEVD-MUSIC) is commonly used for localization. We improved its robustness against noise with high power by incorporating Generalized EigenValue Decomposition (GEVD). However, GEVD-based MUSIC (GEVD-MUSIC) has mainly two issues: 1) the resolution of pre-measured Transfer Functions (TFs) determines the resolution of SSL, 2) its computational cost is expensive for real-time processing. For the first issue, we propose a TF interpolation method integrating time-domain-based and frequency-domain-based interpolation. The interpolation achieves super-resolution SSL, whose resolution is higher than that of the pre-measured TFs. For the second issue, we propose two methods, MUSIC based on Generalized Singular Value Decomposition (GSVD-MUSIC), and Hierarchical SSL (H-SSL). GSVD-MUSIC drastically reduces the computational cost while maintaining noise-robustness in localization. H-SSL also reduces the computational cost by introducing a hierarchical search algorithm instead of using greedy search in localization. These techniques are integrated into an SSL system using a robot embedded microphone array. The experimental result showed: the proposed interpolation achieved approximately 1 degree resolution although we have only TFs at 30 degree intervals, GSVD-MUSIC attained 46.4% and 40.6% of the computational cost compared to SEVD-MUSIC and GEVD-MUSIC, respectively, H-SSL reduces 59.2% computational cost in localization of a single sound source.**

## I. Introduction

Since a robot should work in a real environment, robot audition is essential for human-robot interaction. *Sound Source Localization* (*SSL*) in robot audition tells us the location and number of sound sources, which are utilized for other robot audition functions such as sound source separation. Since SSL affects the performance of a whole robot audition system drastically, it has been studied widely.

Since robots should work in real-time and localize sound sources in a noisy environment, SSL for robots mainly requires: noise-robustness, high-resolution, and real-time processing. To solve the first problem, we previously extended *Multiple Signal Classification (MUSIC)* based on *Standard EigenValue Decomposition* (*SEVD-MUSIC* [1]) by incorporating *Generalized EigenValue Decomposition* (*GEVD*), called *GEVD-MUSIC* [2]. The method successfully realized localization of target sources under noise with high power.

K. Nakamura, K. Nakadai, and Gökhan Ince are with the Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako, Saitama, 351-0114, Japan. {keisuke, nakadai, gokhan.ince}@jp.honda-ri.com

Although GEVD-MUSIC has such an advantage in noise-robustness, it has mainly two issues:

1) The resolution of pre-measured *Transfer Functions* (*TFs*) between a directional sound source and microphones decides that of SSL.
2) MUSIC algorithms are computationally expensive for subspace decomposition and high resolution SSL.

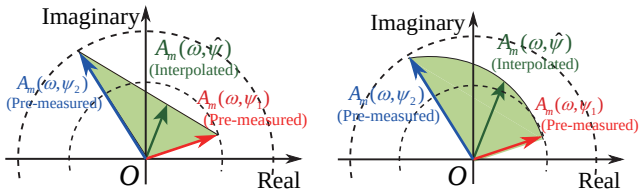
Issue 1) stems from the utilization of natural pre-measured TFs. The resolution of pre-measured TFs decides the resolution of SSL, and measurements of fine resolution are time consuming. To obtain fine resolution without fine measurements, one simple solution would be a numerical TF calculation using geometric information of a microphone array, known as a numerical method which implicitely assumes that microphones are in a free space. However, the accuracy of numerical methods is not sufficient because a robot-embedded microphone array is attached to a complex robot surface not in a free space, which includes high-order reflections at several different materials. Several numerical methods adaptable for a robot-embedded microphone array have high computational cost [3-4], which are not suitable for real-time SSL. The better solution would be interpolating pre-measured TFs to obtain TFs with any desired resolution, that is, an interpolation method. Since interpolation methods include a complex robot surface in pre-measured TFs, they are suitable for robots. Reported interpolation methods are divided into two categories: *all-points methods* and *adjacent-points methods*. The all-points methods [5-10] utilize all the known TFs to estimate a TF, which cause low estimation errors. However, their algorithm still has difficulties in real-time processing. On the other hand, the adjacent-points methods [11-13] utilize only the two closest pre-measured TFs from the estimated point. Therefore, they have advantages in terms of real-time processing, and are suitable for robot audition. The estimation error is not sufficiently small since the calculation relies on two adjacent TFs.

Issue 2) has two main points. First, the subspace decomposition in MUSIC algorithms, namely SEVD in SEVD-MUSIC and GEVD in GEVD-MUSIC, increases the computational cost dramatically and still has difficulties in real-time operation especially in a frame-by-frame manner. Secondly, SSL for higher resolution needs more time to calculate a spatial spectrum and search peaks of the spectrum. For robot audition, we need to achieve both high resolution and real-time processing simultaneously.

The purpose of this paper is to realize SSL solving these issues and apply the SSL to a robot in a real environment.

For Issue 1), we propose TF interpolation based on the integration of *Frequency- and Time-Domain Linear Inter-*

978-1-4673-1735-1/12/S31.00 ©2012 IEEE




(a) FDLI (b) TDLI

Fig. 1. Intuitive image of interpolation

*polation* (**FTDLI**), which is based on an adjacent-points method in consideration of real-time interpolation. This is the extension of a correlation matrix interpolation for TFs [14]. Namely, two interpolation methods for estimating amplitude and phase are integrated. The integration improves the interpolation accuracy for both amplitude and phase of TFs. By FTDLI, we can generate TFs with the desired resolution and achieve super-resolution SSL, where the super-resolution represents the resolution exceeding that of pre-measured TFs.

For Issue 2), we first extended GEVD-MUSIC [2] to utilize *Generalized SingularValue Decomposition* (*GSVD*), which is hereinafter called **GSVD-MUSIC**. The extension not only maintains the noise-robustness in GEVD-MUSIC but also reduces the computational cost enormously. Secondly, we introduced *Hierarchical SSL* (**H-SSL**) based on a coarse-to-fine approach [15]. It roughly localizes sound sources using the pre-measured TFs, and consequently it precisely localizes the sound source again around the estimated location using interpolated TFs. By this extension, H-SSL first assures the resolution of SSL with pre-measured TFs, and it can improve using the interpolated TFs afterwards. H-SSL allows performing super-resolution SSL in real-time.

The rest of the paper is organized as follows: Section II explains the details of TF interpolation by FTDLI for Issue 1). Section III gives a brief introduction of SEVD-MUSIC and GEVD-MUSIC, and describes GSVD-MUSIC and H-SSL to solve Issue 2). Section IV shows the system structure. Section V evaluates the techniques used in the system, and Section VI concludes this paper.

## II. Transfer Function Interpolation Using FTDLI

To solve the first issue discussed in Section I, FTDLI is described to obtain TFs with the desired resolution for super-resolution SSL.

### A. Related Work

Here, *Frequency Domain Linear Interpolation* (*FDLI*) [6], [11], [12] and *Time Domain Linear Interpolation* (*TDLI*) [13] are explained.

Let $\boldsymbol{A}(\omega,\psi_1) = [A_1(\omega,\psi_1),\ldots,A_M(\omega,\psi_1)]^T \in \mathbb{C}^M$ and $\boldsymbol{A}(\omega,\psi_2) = [A_1(\omega,\psi_2),\ldots,A_M(\omega,\psi_2)]^T \in \mathbb{C}^M$ denote pre-measured TFs between a microphone array and sound sources, in other words, steering vectors in SSL. $M$ is the number of microphones, $\psi_1$ and $\psi_2$ are directions of the pre-measured TFs, and $\omega$ represents frequency. Our objective is to estimate $\boldsymbol{A}(\omega,\hat{\psi})$ by interpolation, where $\hat{\psi}$ is the direction of an estimated point, which is $\psi_1 < \hat{\psi} < \psi_2$. **FDLI** [6] in the frequency domain interpolates a TF by:

$$\hat{A}_{m[\psi_1,\psi_2]}(\omega,\hat{\psi}) = D_A A_m(\omega,\psi_1) + (1-D_A)A_m(\omega,\psi_2)\ , \quad (1)$$

where $\hat{A}_{m[\psi_1,\psi_2]}(\omega,\hat{\psi})$ is an interpolated TF of the $m$-th microphone at $\hat{\psi}$ using $A_m(\omega,\psi_1)$ and $A_m(\omega,\psi_2)$. $D_A \in \mathbb{R}$ represents an interpolation factor, which is $0 \le D_A \le 1$.

In **TDLI** [13], $\hat{\boldsymbol{A}}_{[\psi_1,\psi_2]}(\omega,\hat{\psi})$ is obtained as follows:

$$\hat{A}_{m[\psi_1,\psi_2]}(\omega,\hat{\psi}) = A_m(\omega,\psi_2)(A_m(\omega,\psi_1)/A_m(\omega,\psi_2))^{D_A}\ . \quad (2)$$

Fig. 1 shows the intuitive image of the TF interpolation by FDLI and TDLI, respectively. Both FDLI and TDLI can interpolate transfer functions between $\boldsymbol{A}(\omega,\psi_1)$ and $\boldsymbol{A}(\omega,\psi_2)$ continuously in a short time by the factor $D_A$. Previously, we evaluated their interpolation errors based on amplitude and phase [14] and found that FDLI and TDLI have problems in estimating amplitude and phase, respectively.

### B. Frequency- and Time-Domain Linear Interpolation

With respect to the idea of the correlation matrix interpolation [14], we extended the method for TF interpolation. Namely, we integrated the phase interpolation result of FDLI and the amplitude interpolation result of TDLI in order to achieve the best interpolation performance. FDLI and TDLI are integrated by the following steps:

1) Take two interpolation results from Eq. (1) and Eq. (2). Here, $\hat{A}_{m[\psi_1,\psi_2]}(\omega,\hat{\psi})$ in Eq. (1) and Eq. (2) are redefined as $\hat{A}_{m[\mathrm{F}|\psi_1,\psi_2]}(\omega,\hat{\psi})$ and $\hat{A}_{m[\mathrm{T}|\psi_1,\psi_2]}(\omega,\hat{\psi})$, respectively.
2) Decompose $\hat{A}_{m[\mathrm{F}|\psi_1,\psi_2]}(\omega,\hat{\psi})$ and $\hat{A}_{m[\mathrm{T}|\psi_1,\psi_2]}(\omega,\hat{\psi})$ into phase and gain.

$$\hat{A}_{m[\mathrm{F}|\psi_1,\psi_2]}(\omega,\hat{\psi}) = \lambda_{m[\mathrm{F}]}\exp(-j\omega t_{m[\mathrm{F}]}) \quad (3)$$

$$\hat{A}_{m[\mathrm{T}|\psi_1,\psi_2]}(\omega,\hat{\psi}) = \lambda_{m[\mathrm{T}]}\exp(-j\omega t_{m[\mathrm{T}]}) \quad (4)$$

3) Calculate $\hat{A}_{m[\psi_1,\psi_2]}(\omega,\hat{\psi})$ as follows:

$$\hat{A}_{m[\psi_1,\psi_2]}(\omega,\hat{\psi}) = \lambda_{m[\mathrm{T}]}\exp(-j\omega t_{m[\mathrm{F}]})\ . \quad (5)$$

For SSL, $\hat{A}_{m[\psi_1,\psi_2]}(\omega,\hat{\psi})$ in Eq. (5) will be later utilized as $\boldsymbol{A}(\omega,\psi)$ in Eq. (7). Then, we can select a desired resolution in SSL depending on the resolution of $D_A$ in Eq. (1) and Eq. (2).

## III. SSL with GSVD-MUSIC and H-SSL

To solve the second issue, this section investigates computational cost reduction by using GSVD-MUSIC and H-SSL. This section first gives a brief introduction of SEVD-MUSIC [1] and GEVD-MUSIC [2]. Afterwards, we describe details of GSVD-MUSIC and H-SSL.

### A. SEVD-MUSIC and its Extension to GEVD-MUSIC

In advance of SSL, we need TFs (steering vectors), namely $\boldsymbol{A}(\omega,\psi)$. In Section II, we obtained $\boldsymbol{A}(\omega,\psi)$ for a desired interval of $\psi$ by measurements and interpolation.

In SSL, we first compute a correlation matrix of multi-channel input acoustic signals, denoted by $\boldsymbol{R}(\omega,f) \in \mathbb{C}^{M\times M}$. SEVD-MUSIC [1] performs SEVD of $\boldsymbol{R}(\omega,f)$ to decompose signal space into noise- and signal-subspaces as follows:

$$\boldsymbol{R}(\omega,f) = \boldsymbol{E}(\omega,f)\boldsymbol{\Lambda}(\omega,f)\boldsymbol{E}^{-1}(\omega,f) \quad (6)$$

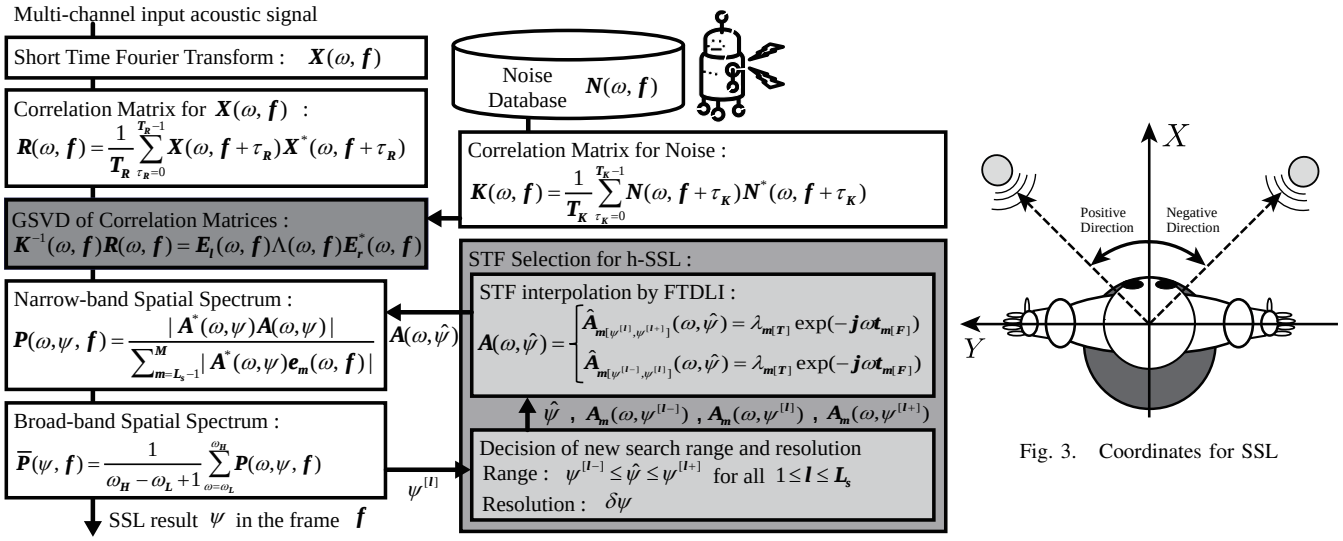



Fig. 2. Super-resolution SSL system working in real-time




Fig. 3. Coordinates for SSL

where $\Lambda(\omega,f) = \mathrm{diag}(\lambda_1(\omega,f),\ldots,\lambda_M(\omega,f))$ and $E(\omega,f) = [e_1(\omega,f),\ldots,e_M(\omega,f)]$ are eigenvalues and vectors, respectively. Here, $e_m(\omega,f)$ is sorted in order of $\lambda_m(\omega,f)$ $(1 \le m \le M)$.

The spatial spectrum for SSL is determined by

$$P(\omega,\psi,f) = \frac{|A^*(\omega,\psi)A(\omega,\psi)|}{\sum_{m=L_s+1}^{M}|A^*(\omega,\psi)e_m(\omega,f)|} , \quad (7)$$

where $L_s$ is the number of sound sources considered in SSL. For the *Direction of Arrival* (*DoA*) estimation, we accumulate $P(\omega,\psi,f)$ in Eq. (7) over $\omega$ as follows:

$$\bar{P}(\psi,f) = \frac{1}{k_h-k_l+1}\sum_{k=k_l}^{k_h} P(\omega_{[k]},\psi,f) , \quad (8)$$

where $k_h$ and $k_l$ are the frequency bin indices, which represent the maximum and minimum frequency for SSL, respectively.

GEVD-MUSIC [2] extends Eq. (6) to perform GEVD as:

$$K^{-1}(\omega,f)R(\omega,f) = E(\omega,f)\Lambda(\omega,f)E^*(\omega,f) , \quad (9)$$

where $K(\omega,f)$ is a freely designable correlation matrix and can be utilized for various purposes. For instance, when high power noise $N(\omega)$ exists, $K(\omega,f)$ is designed as $K(\omega,f) = N(\omega)N^*(\omega)$, which whitens the noise-related eigenvalues.

### B. GSVD-MUSIC

Theoretically, the GEVD of $K(\omega,f)$ and $R(\omega,f)$ can be equivalently described as the following SEVD:

$$K^{-\frac{1}{2}}(\omega,f)R(\omega,f)K^{-\frac{1}{2}}(\omega,f) = E(\omega,f)\Lambda(\omega,f)E^*(\omega,f) . \quad (10)$$

In [2], Eq. (9) is considered as the GEVD instead of Eq. (10) since it whitens noise without calculating $K^{-\frac{1}{2}}(\omega,f)$, which has a large calculation cost. However, Eq. (9) has the following problems:

- The SEVD still has a large calculation cost for frame-by-frame localization working in real-time.
- The eigenvectors are not mutually orthogonal since $K^{-1}(\omega,f)R(\omega,f)$ is not hermitian, which eventually degrades the SSL performance.

To solve these problems, this paper extends Eq. (9) to utilize the following GSVD:

$$K^{-1}(\omega,f)R(\omega,f) = E_l(\omega,f)\Lambda(\omega,f)E_r^*(\omega,f) , \quad (11)$$

where $E_l(\omega,f)$ and $E_r(\omega,f)$ are left- and right-singular vectors, respectively, which are unitary and are mutually orthogonal. $E_l(\omega,f)$ is used instead of $E(\omega,f)$ in Eq. (6).

The calculation cost of Eq. (11) is less than that of Eq. (9), which is evaluated in Section V.

We note the following aspects in GSVD-MUSIC. First, GSVD-MUSIC is equivalent to the following GEVD-MUSIC:

$$R^2 e_m = \lambda_m K^2 e_m \quad (12)$$

This indicates that the eigenvectors are mutually orthogonal, which improves SSL performance compared to GEVD-MUSIC[2].

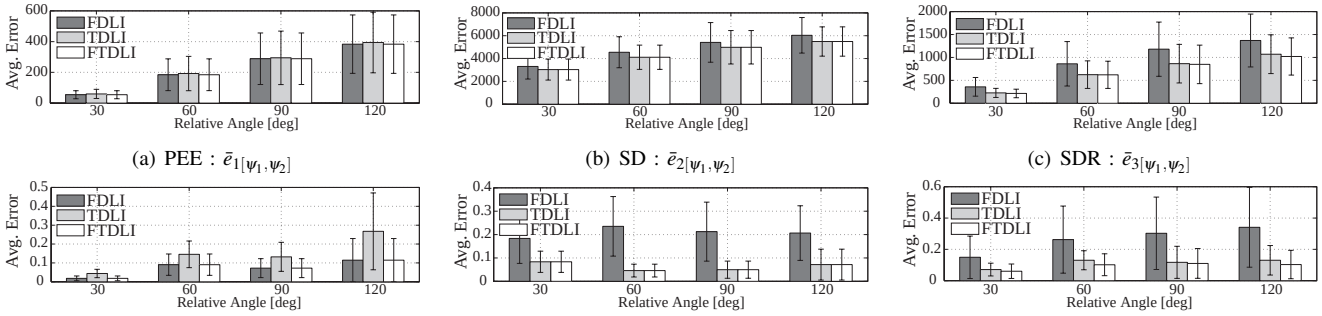
Secondly, GSVD-MUSIC is equivalent to SEVD-MUSIC when $K = I$, where $I \in \mathbb{C}^{M\times M}$ is an identity matrix. This means that we can reduce the calculation cost of SEVD-MUSIC with GSVD-MUSIC since GSVD is less computationally demanding than SEVD.

### C. H-SSL for Fast Super-resolution SSL

Although we can obtain fine TFs by using FTDLI, super-resolution SSL is computationally expensive. Therefore, it is not suitable for real-time processing for robot application. To solve the issue, we introduce H-SSL.

In GSVD-MUSIC, $P(\omega,\psi,f)$ as in Eq. (7) and $\bar{P}(\psi,f)$ as in Eq. (8) are $\psi$-dependent processes. Namely, in these processes, finer $\hat{A}_{m[\psi_1,\psi_2]}(\omega,\hat{\psi})$ increases the calculation cost linearly.

In H-SSL, we reduced the number of processed TFs while maintaining the resolution by the following steps:




(a) PEE : $\bar{e}_{1[\psi_1,\psi_2]}$ (b) SD : $\bar{e}_{2[\psi_1,\psi_2]}$ (c) SDR : $\bar{e}_{3[\psi_1,\psi_2]}$

(d) Linearity of $D_A$ in PEE : $\bar{d}_{1[\psi_1,\psi_2]}$ (e) Linearity of $D_A$ in SD : $\bar{d}_{2[\psi_1,\psi_2]}$ (f) Linearity of $D_A$ in SDR : $\bar{d}_{3[\psi_1,\psi_2]}$

Fig. 4. Interpolation error and the linearity of $D_A$ for PEE, SD, SDR using TF of a robot-embedded microphone array

1) Conduct SSL with rough pre-measured TFs and search peaks of spatial spectrum in Eq. (8). Let $\psi^{[l]}$ be the direction that has the $l$-th largest $\bar{P}(\psi, f)$, where $l$ is the index of sound sources ($1 \le l \le L_s$).
2) Take two closest $\psi$ from $\psi^{[l]}$, which are denoted as $\psi^{[l-]}$ and $\psi^{[l+]}$. Suppose $\psi^{[l-]} < \psi^{[l]} < \psi^{[l+]}$.
3) Generate $\hat{A}_{m[\psi^{[l-]},\psi^{[l]}]}(\omega,\hat{\psi})$ and $\hat{A}_{m[\psi^{[l]},\psi^{[l+]}]}(\omega,\hat{\psi})$ by Eq. (5) depending on a given finer resolution.
4) Conduct SSL again only with $\hat{A}_{m[\psi^{[l-]},\psi^{[l]}]}(\omega,\hat{\psi})$ and $\hat{A}_{m[\psi^{[l]},\psi^{[l+]}]}(\omega,\hat{\psi})$, and search peaks of Eq. (8).

The upper layer of the hierarchical process is for broad localization, while the lower layer provides finer localization. In this sense, SSL with H-SSL follows coarse-to-fine [15] localization, giving rough SSL results first, followed by fine SSL results.

## IV. System Structure

Fig. 2 shows the system structure to achieve super-resolution SSL working in real-time. All blocks explained above were implemented to a robot located in a normal room whose reverberation time was 0.2 seconds. Fig. 3 shows the coordinate system for $\psi$. We have utilized an 8-ch circular microphone array embedded in the robot's head and measured the TFs $\boldsymbol{A}(\omega,\psi)$ at every 1°, which were obtained by time-stretched pulse recording. The acoustic signal was sampled with 16 kHz and 16 bits. The window and shift length for frequency analysis were set to 512 and 160 samples, respectively. All the proposed functions were implemented as modules for robot audition software HARK [16]. The system worked in real-time with a laptop having a 2.0 GHz Intel Core i7 CPU and 8GB SDRAM running Linux.

## V. Experimental Validation

### A. Error of TF interpolation Using FTDLI

We evaluated the estimation errors of three interpolation methods, namely FDLI, TDLI, and FTDLI. We took the difference between estimated $\hat{\boldsymbol{A}}_{[\psi_1,\psi_2]}(\omega,\hat{\psi})$ and pre-measured $\boldsymbol{A}(\omega,\psi)$. $\psi_1$ was fixed at 0°, and $\psi_2 = \{30°, 60°, 90°, 120°\}$ was used. $\hat{\boldsymbol{A}}_{[\psi_1,\psi_2]}(\omega,\hat{\psi})$ was estimated at every 1°. We averaged the estimated errors for $\omega$ and $\psi$. The averaged error, $\bar{e}_{[\psi_1,\psi_2]}$, was calculated as:

$$\bar{e}_{[\psi_1,\psi_2]} = \frac{1}{i_\psi}\sum_{i=1}^{i_\psi}\frac{1}{k_h-k_l+1}\sum_{k=k_l}^{k_h} f_{[\psi_1,\psi_2]}(\omega_{[k]},\hat{\psi}_{[i]}) \ , \qquad (13)$$

where $f_{[\psi_1,\psi_2]}(\omega_{[k]},\hat{\psi}_{[i]})$ is the estimation error for specific $\hat{\psi}$ and $\omega$. $k_l$ and $k_h$ are defined as the same as Eq. (8) with the frequency band $500[\mathrm{Hz}] \le \omega \le 2800[\mathrm{Hz}]$. $i_\psi$ represents the number of $\hat{\psi}$ we conducted interpolation. For $\hat{\psi}$, we utilized all $\psi$ of the pre-measured $\boldsymbol{A}(\omega,\psi) = [A_1(\omega,\psi),\ldots,A_M(\omega,\psi)]^T$ in the range of $\psi_1 < \psi < \psi_2$.

We evaluated three kinds of $f_{[\psi_1,\psi_2]}(\omega_{[k]},\hat{\psi}_{[i]})$ in Eq. (13).

The first criterion is the summation of normalized inner-product of all channels:

$$f_{1[\psi_1,\psi_2]}(\omega,\hat{\psi}) = \sum_{m=1}^{M}\left|\frac{A_m(\omega,\hat{\psi})\cdot\hat{A}_{m[\psi_1,\psi_2]}(\omega,\hat{\psi})}{|A_m(\omega,\hat{\psi})||\hat{A}_{m[\psi_1,\psi_2]}(\omega,\hat{\psi})|}-1\right| \ , \qquad (14)$$

which represents the *Phase Estimation Error (PEE)*.

The second criterion is the summation of *Spectral Distortion (SD)* of all channels calculated as follows:

$$f_{2[\psi_1,\psi_2]}(\omega,\hat{\psi}) = \sum_{m=1}^{M}\left|20\log\frac{|\hat{A}_{m[\psi_1,\psi_2]}(\omega,\hat{\psi})|}{|A_m(\omega,\hat{\psi})|}\right| \ , \qquad (15)$$

which shows the amplitude estimation performance.

The third criterion is the summation of *Signal-to-Distortion Ratio (SDR)* of all channels calculated as follows [1]:

$$f_{3[\psi_1,\psi_2]}(\omega,\hat{\psi}) = \sum_{m=1}^{M}\frac{|A_m(\omega,\hat{\psi})-\hat{A}_{m[\psi_1,\psi_2]}(\omega,\hat{\psi})|^2}{|A_m(\omega,\hat{\psi})|^2} \ , \qquad (16)$$

which represents the total estimation performance. Let $\bar{e}_{1[\psi_1,\psi_2]}$, $\bar{e}_{2[\psi_1,\psi_2]}$, and $\bar{e}_{3[\psi_1,\psi_2]}$ denote $\bar{e}_{[\psi_1,\psi_2]}$ of PEE, SD, and SDR, respectively.

We also evaluated the linearity of $D_A$ calculated by

$$\bar{d}_{[\psi_1,\psi_2]} = \frac{1}{i_\psi}\sqrt{\sum_{i=1}^{i_\psi}\left(D_{A[i]}-\frac{\hat{\psi}_{[i]}-\psi_2}{\psi_1-\psi_2}\right)^2} \ , \qquad (17)$$

where $D_{A[i]}$ denotes $D_A$ having the smallest $f_{[\psi_1,\psi_2]}(\omega_{[k]},\hat{\psi}_{[i]})$ for each $\hat{\psi}_{[i]}$. A small $\bar{d}_{[\psi_1,\psi_2]}$ means $D_A$ is close to $\frac{\hat{\psi}-\psi_2}{\psi_1-\psi_2}$, being utilized for practical interpolation. Let $\bar{d}_{1[\psi_1,\psi_2]}$, $\bar{d}_{2[\psi_1,\psi_2]}$, and $\bar{d}_{3[\psi_1,\psi_2]}$ denote $\bar{d}_{[\psi_1,\psi_2]}$ of PEE, SD, and SDR, respectively.

[1] We have inverted the original SDR discussed in [6] since it shows the best estimation performance with $f_{[\psi_1,\psi_2]}(\omega,\hat{\psi}) = 0$.

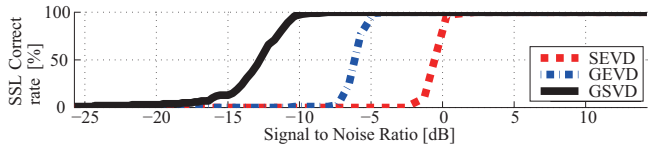



Fig. 5. SSL correct rate of SEVD-, GEVD-, and GSVD-MUSIC

Fig. 4 shows the comparison of the average errors, namely $\bar{e}_1$, $\bar{e}_2$, and $\bar{e}_3$, and their linearity of $D_A$, namely $\bar{d}_1$, $\bar{d}_2$, and $\bar{d}_3$, respectively. The horizontal axis in all the figures shows the difference between $\psi_1$ and $\psi_2$ for the interpolation.

FTDLI achieved as small $\bar{e}_1$ as FDLI and achieved as small $\bar{e}_2$ as TDLI thanks to the integration. As a result, it had the smallest $\bar{e}_3$. Also, FTDLI showed the smallest $\bar{d}_1$, $\bar{d}_2$, and $\bar{d}_3$. Thus, FTDLI has advantages in both high accuracy and intuitive parameter determination.

### *B. Noise-robustness and computational cost comparison among SEVD-, GEVD-, GSVD-MUSIC*

In the experiment, there were a target sound (white noise) and a noise source (robot's fan noise) 1m away from the microphone array in the directions of 60° and 180°, respectively. $\boldsymbol{A}(\omega,\psi)$ of $\psi = \{-175°,-170°,...,180°\}$ were used. Hence, the resolution of SSL was 5°.

We evaluated the relationship between *Signal-to-Noise Ratio* (*SNR*) and SSL correct rate. SNR is defined as follows[2]:

1) Calculate the average spectrum of $M$-ch input acoustic signals defined as $X_{s_a}(\omega)$. The *Power Spectrum Density* (*PSD*) of $X_{s_a}(\omega)$ is derived as $P_{s_a}(\omega) = \frac{1}{k_{wl}} X_{s_a}(\omega) X^*_{s_a}(\omega)$, where $k_{wl}$ is a window length.
2) Determine the PSD of the noise source $P_{n_a}(\omega)$ by using the same process as 1.
3) Calculate the SNR for each frequency bin, and we normalized the SNR of the bins between $k_h$ and $k_l$ as:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{1}{k_h - k_l + 1}\sum_{k=k_l}^{k_h}\frac{P_{s_a}(\omega_{[k]})}{P_{n_a}(\omega_{[k]})}\right), \quad (18)$$

where $k_h$ and $k_l$ are defined the same as those in Eq. (8). The frequency range was from 500Hz to 2800Hz.

The SSL correct rate was defined as the number of frames whose highest peaks in Eq. (8) were in the direction of the target sound source (60°) in 100 frames.

Fig. 5 shows the result. The horizontal axis shows SNR, and the vertical axis shows the SSL correct rate. The dotted-, chained-, and solid-line show the result of SEVD-MUSIC, GEVD-MUSIC, and GSVD-MUSIC, respectively.

GSVD-MUSIC shows better performance than both GEVD- and SEVD-MUSIC. This means that GSVD can obtain mutually-orthogonal signal and noise subspaces.

The computational cost of each method was also evaluated. We conducted SSL over 1000 frames for each method and measured the averaged processing time for only Eqs. (6)-(9), and (11). As a result, the average processing time for SEVD-, GEVD-, and GSVD-MUSIC methods was 11.9ms, 13.6ms, and 5.52ms, respectively. Clearly, GSVD-MUSIC showed considerable improvement in the computational cost compared to SEVD-MUSIC (approximately two times faster). Since the frame period was 10ms, only GSVD-MUSIC was able to work in real time in a frame-by-frame manner.

[2]This definition is the same as that in [17]. This paper includes the result of GSVD-MUSIC.

TABLE I

FRAME PROCESSING TIME WITH/WITHOUT H-SSL

| | 1 source | 2 sources | 3 sources | 4 sources |
|---|---|---|---|---|
| w/o H-SSL | 45.0 [ms] | 42.8 [ms] | 37.5 [ms] | 34.9 [ms] |
| w H-SSL | 18.4 [ms] | 19.0 [ms] | 20.2 [ms] | 20.8 [ms] |

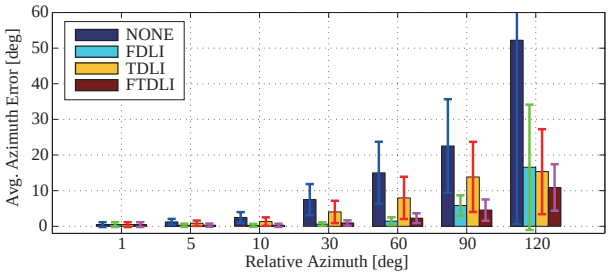



Fig. 6. DoA estimation errors

Finally, we confirmed that GSVD-MUSIC was the best method in terms of both noise-robustness and computational efficiency.

### *C. Computational cost of H-SSL*

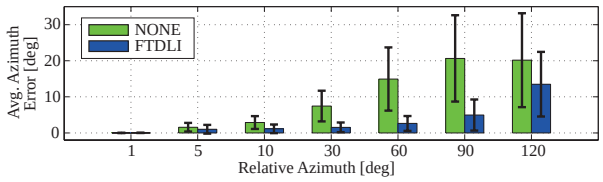
We compared the computational cost of SSL with and without H-SSL as explained in Section III-C. Same as Section V-B, we conducted SSL over 1000 frames for each method and measured the averaged processing time for only Eqs. (6)-(8). For SSL without H-SSL, $\boldsymbol{A}(\omega,\psi)$ of 1° intervals were used for higher resolution. For H-SSL, $\boldsymbol{A}(\omega,\psi)$ of 10° intervals were used as pre-measured TFs, and we conducted FTDLI to generate $\hat{A}_{m[\psi_1,\psi_2]}(\omega,\hat{\psi})$ of 1° intervals. Thus, these two SSLs have the same resolution.

Table I shows the average processing time for SSL with and without H-SSL. The table includes the change of $L_s$ in Eq. (7), which affects the hierarchical process. H-SSL reduced the processing time approximately half regardless of $L_s$, which shows the validity of H-SSL.

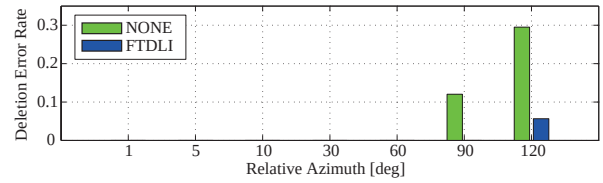
### *D. Applicability of the proposed methods to SSL for a robot*

We applied GSVD-MUSIC, FTDLI and H-SSL to SSL by our robot-embedded microphone array. We evaluated our methods by using two measures: 1) Error of DoA estimation towards a sound source in fixed position, 2) SSL performance towards a dynamic sound source.

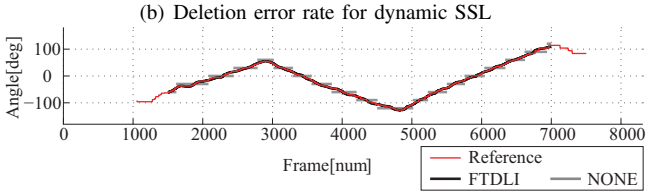
*1) SSL for a Sound Source in Fixed Position:* This section compares the error of DoA estimation towards a stationary sound source using different interpolation methods. We have recorded white noise by every 1° of $\psi$ ($-90° \le \psi \le 90°$). We localized the white noise and took the average error between estimated and recorded DoA of the white noise. We assume having $\boldsymbol{A}(\omega,\psi)$ at the intervals of $\psi = \{1°,5°,10°,30°,60°,90°,120°\}$. By interpolation, $\hat{\boldsymbol{A}}_{[\psi_1,\psi_2]}(\omega,\hat{\psi})$ of 1° intervals was estimated so that we can correctly localize the sound source.




(a) DoA estimation error for dynamic SSL




(b) Deletion error rate for dynamic SSL




(c) Example of dynamic SSL using $\boldsymbol{A}(\omega, \psi)$ at 30° intervals

Fig. 7. Dynamic SSL performance evaluation

Fig. 6 shows the error comparison of DoA estimation among the interpolation methods discussed in Section II. The horizontal axis shows the intervals of $\psi$ for $\boldsymbol{A}(\omega, \psi)$. The vertical axis shows the average error of DoA estimation. NONE means that no interpolation method was used.

We observed: 1) FTDLI had the smallest DoA estimation errors, 2) Up to the intervals of 30°, FTDLI maintained approximately the same performance as the case at 1° intervals, which showed the validity of FTDLI in SSL.

*2) SSL for a Dynamic Sound Source:* We used a moving white noise source. All other conditions are the same as those in Section V-D.1.

Similar to Fig. 6, Fig. 7(a) shows the error comparison between DoA estimations with and without FTDLI. FTDLI performed better than NONE and it maintained SSL performance until the intervals of 30°.

Fig. 7(b) shows the deletion error rate in SSL. The deletion error rate was defined as the rate of frames, which missed the DoA estimation result. FTDLI perfectly localized the sound source until the intervals of 90°.

Fig. 7(c) shows an SSL example using $\boldsymbol{A}(\omega, \psi)$ at 30° intervals. 30° intervals were selected in consideration of the evaluation results in Figs. 6 and 7(a). The horizontal axis shows the frame index, and the vertical axis shows the DoA estimation result $\psi$. In addition to the SSL results, we plotted the reference trajectory recorded by an ultrasonic positioning system [18]. The SSL with FTDLI showed a better and smoother trajectory than that with NONE.

As a result, we were able to confirm the validity of the SSL system in both static and dynamic environments.

## VI. Conclusion

This paper investigated super-resolution SSL working in real-time for robots in a real environment. To achieve a fast SSL with any desired resolution, we focused on two issues, pre-defined resolution based on the resolution of measurements, and high computational cost for conducting GEVD and high resolution SSL. To solve them, we proposed interpolation of TFs by FTDLI and computational cost reduction by GSVD-MUSIC and H-SSL.

All the proposed functions were integrated into a super-resolution SSL system and were evaluated. The evaluation showed: 1) FTDLI showed better interpolation performance compared to existing methods and realized super-resolution SSL, 2) GSVD-MUSIC reduced the calculation cost approximately by half and achieved better noise-robustness compared to GEVD-MUSIC, 3) SSL with H-SSL achieved approximately half the calculation cost than that without H-SSL, which successfully confirmed the validity of the whole system.

The future work will be the extension of FTDLI to 3-dimensional case and real-time 3-dimensional super-resolution SSL.

## References


[1] R. Schmidt, "Multiple emitter location and signal parameter estimation", *IEEE Trans. Ant. Prop.*, vol. 34, no. 3, pp. 276–280, 1986.

[2] K. Nakamura *et al.*, "Intelligent Sound Source Localization for Dynamic Environments", in *Proc. of IEEE/RSJ IROS*, pp. 664–669, 2009.

[3] K. Yamamoto *et al.*, "An acoustic simulation for speech interface of humanoid robot", in *Acoust. Soc. of Japan*, pp. 815–818, 2009.

[4] K. Nakadai *et al.*, "Applying scattering theory to robot audition system: robust sound source localization and extraction", in *Proc. of IEEE/RSJ IROS*, no. 2, pp. 1147–1152, 2003.

[5] L. Wang *et al.*, "Head-related transfer function interpolation through multivariate polynomial fitting of principal component weights", in *Acoust. Sci. & Tech.*, vol. 30, no. 6, pp. 395–403, 2009.

[6] T. Nishino *et al.*, "Interpolating Head Related Transfer Functions in the median plane", in *Proc. of IEEE WASPAA*, pp. 167–170, 1999.

[7] F. Keyrouz *et al.*, "A rational HRTF interpolation approach for fast synthesis of moving sound", in *the 12th DSP and 4th SPE Workshop*, pp. 222–226, 2006.

[8] T. Ajdler *et al.*, "Plenacoustic function on the circle with application to HRTF interpolation", in *Proc. of IEEE ICASSP*, pp. 273–276, 2005.

[9] C. I. Cheng *et al.*, "Spatial frequency response surfaces: an alternative visualizationtool for head-related transfer functions (HRTFs)", in *Proc. of IEEE ICASSP*, pp. 961–964, 1999.

[10] F. P. Freeland *et al.*, "HRTF interpolation through direct angular parameterization", in *IEEE Symp. Circ. Syst.*, pp. 1823–1826, 2007.

[11] K. Watanabe *et al.*, "Interpolation of Head-Related Transfer Functions based on the Common-Acoustical-Pole and Residue Model", in *Acoust. Sci. & Tech.*, vol. 24, no. 5, pp. 335–337, 2003.

[12] J. Zhang *et al.*, "A Piecewise Interpolation Method Based on Log-Least Square Error Criterion for HRTF", in *IEEE 7th Workshop on Multimedia Signal Processing*, pp. 1–4, 2005.

[13] M. Matsumoto *et al.*, "A method of interpolating binaural impulse responses for moving sound images", in *Acoust. Sci. & Tech.*, vol. 24, no. 5, pp. 284–292, 2003.

[14] K. Nakamura *et al.*, "Correlation Matrix Interpolation in Sound Source Localization for a Robot," in *Proc. of IEEE ICASSP*, Prague, Czech, pp. 4324-4327, 2011.

[15] D. Ringach, "Look at the big picture (details will follow)", *Nature Neuroscience*, vol. 6, no. 1, pp. 7–8, 2003.

[16] K. Nakadai *et al.*, "Design and Implementation of Robot Audition System HARK", *Advanced Robotics*, vol. 24, pp. 739–761, 2009.

[17] K. Nakamura *et al.*, "Intelligent Sound Source Localization and Its Application to Multimodal Human Tracking," in *Proc. of 2011 IEEE/RSJ IROS*, pp. 143-148, 2011.

[18] K. Nakadai *et al.*, "Sound source tracking with directivity pattern estimation using a 64 ch microphone array", in *Proc. of 2005 IEEE/RSJ IROS*, pp. 1690–1696, 2005.