2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Real-time Super-resolution Three-dimensional Sound Source Localization for Robots

Keisuke Nakamura, Randy Gomez, and Kazuhiro Nakadai

Abstract— **This paper investigates Sound Source Localization (SSL) for a robot in a real world. Previously, we focused on one-dimensional SSL for azimuth and assumed that target sources are distributed close to a horizontal plane. Without this assumption, the SSL performance is drastically degraded. Thus, three-dimensional SSL is essential to improve the localization for sound sources distributed in a three-dimensional space. Compared to one-dimensional SSL, three-dimensional SSL mainly has the following problems: 1) a massive number of Transfer Function (TF) measurements for microphone array calibration are required for three dimensions to maintain the spatial resolution of SSL sufficiently-high, 2) the computational cost for searching for sound sources drastically increases in high-dimensional spaces. For the first issue, we extend the previously-proposed one-dimensional TF interpolation method, integrating time-domain-based and frequency-domain-based interpolation, to three dimensions. The interpolation achieves three-dimensional super-resolution SSL and reduction of the number of TF measurements while maintaining the spatial resolution of SSL. For the second issue, we propose optimal hierarchical SSL, which reduces computational cost for searching for sound sources by introducing a hierarchical search algorithm instead of using greedy search in localization. We previously proposed the concept of the algorithm. This paper additionally discusses theoretical optimality in hierarchization to minimize the total computational cost of SSL. The method determines the number of hierarchies and the resolution of each hierarchy depending on desired spatial resolution. These techniques are integrated into an SSL system using a robot. The experimental result showed: 1) the proposed interpolation method achieved super-resolution SSL working better than that with pre-measured TFs, 2) the optimal hierarchical SSL drastically reduced computational cost by approximately 97%.**

I. Introduction

Human-robot speech communication is essential for robots. In the communication by a robot-embedded microphone array, the distance between a speaker and a microphone array is usually further than that of using close-talk microphones in for example mobile-phone applications. Due to the long distance, we have to consider spatially-distributed sound/noise sources other than the target speech. These sound/noise sources make signal-to-noise ratio low and degrade the performance of automatic speech recognition. Moreover, the location, activity, and number of them are unknown. Robot audition[1] utilizes *Sound Source Localization (SSL)* and separation so that it can suppress those spatially-distributed sound/noise sources and enhance the target signal. Therefore, the performance improvement of SSL is essential to improve the overall performance.

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

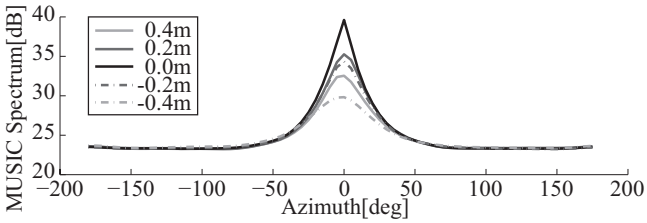


Fig. 1. 1D SSL Result with the Variation of Heights

In consideration of advantages and disadvantages in many conventional methods on SSL, robot audition mainly utilized *one-dimensional* (*1D*, namely azimuth) SSL by *MUltiple SIgnal Classification* (*MUSIC*[2]) which is well-known high-resolution SSL working robustly in a real environment [3], [4], [5]. The 1D SSL for azimuth worked when all the target sound sources are close to a horizontal plane such as simultaneous speech recognition of standing speakers[7].

Fig. 1 shows the 1D SSL result of MUSIC for a single source when the sound source was at 0° in azimuth with the variation of heights. The horizontal and vertical axes show azimuth and MUSIC spectrum, respectively. The black line (0.0m) means that the sound source is just on the horizontal plane, and others are apart from the plane. As seen in the figure, the resolution of MUSIC is degraded when the sound source is apart from the horizontal plane, which is an inevitable problem of 1D SSL. In a real environment, sound sources are usually distributed in a *three-dimensional (3D)* space such as a speaker sitting on a chair, foot step sounds, etc. Especially in SSL by a robot, robot's ego-noise such as motor- and fan-noise is not located on the horizontal plane. In order to achieve high-resolution SSL for such sounds, 3D SSL is essential.

Despite the demands on 3D SSL, only a few studied have been reported using a robot such as a binaural cue approach[11], classical beamformers[12], [13], MUSIC[15], [16]. The binaural cue approach [11] has an advantage in fast operation but assumes that there is only one source at a time. The classical beamformers [12], [13] achieved computationally efficient 3D SSL for multiple sources, but the resolution and noise-robustness is not as high as MUSIC. On the other hand, 3D SSL based on MUSIC [15], [16] is computationally restrictive for real-time processing.

Thus, the purpose of this paper is to develop a framework of 3D SSL by MUSIC achieving both high-resolution and real-time processing and apply it to a robot. Compared to 1D SSL, 3D SSL mainly has the following problems:

A1) A massive number of Transfer Function (TF) measure-

ments for microphone array calibration are required for 3D to maintain the spatial resolution sufficiently-high,

A2) The computational cost for searching for sound sources drastically increases in high-dimensional spaces.

For A1), we introduce trilinear interpolation to our previously reported *Frequency- and Time-Domain Linear Interpolation* (**FTDLI**)[6]. By this extension, we can generate TFs in a 3D space with desired resolution by a small number of pre-measured 3D TFs in rough resolution. The interpolated TFs achieve super-resolution SSL, where the super-resolution represents the resolution exceeding that of the pre-measured TFs. Also, interpolation works for reducing the number of 3D TF measurements, which is time consuming and normally not a simple task for non-professinals since it requires the spherical equipment used in common databases such as common databases [21], [22]. Moreover, FTDLI is algorithmically inexpensive compared to other state-of-the-art interpolation [8], [9] and computation [10], so that the interpolation for massive number of points in a 3D space can be computed in a sufficiently short time. Thus, FTDLI is conducted while operating SSL, which is important for the hierarchical source search described below.

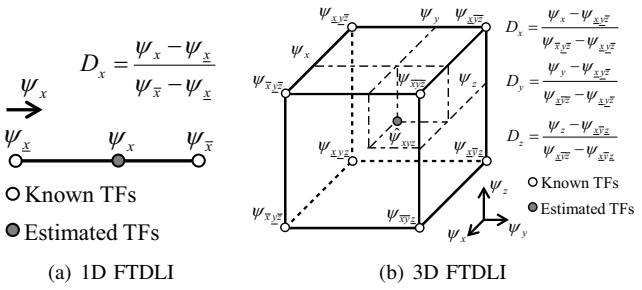
For A2), Ishii *et al.* proposed to use sufficiently-short frame size for fast Fourier transform[16]. However, the short frame size induces low frequency resolution, which has difficulty in dealing with general sounds. Valin *et al.* proposed a hierarchical source search with a spherical grid [13]. Hoang also proposed square grid search[14]. However, the resolution of each hierarchy in these methods is fixed even when having the variation of desired resolution. Thus, we propose 3D *Optimal Hierarchical SSL* (**OH-SSL**) based on a coarse-to-fine approach [17] as an extension of the hierarchical source search in [6]. It roughly localizes a sound source, and consequently it localizes the sound source again around the estimated location. In [6], we discussed only the hierarchical algorithm. Since this paper investigates 3D SSL gaining large computational cost, we additionally determine solutions for the optimal number of hierarchies and optimal resolution of each hierarchy that minimize the computational cost. Different from the conventional methods [13], [14], OH-SSL dynamically changes the number of hierarchies and the resolution of each hierarchy depending on the desired resolution. OH-SSL allows performing 3D super-resolution SSL in real-time.

II. Three-Dimensional Frequency- and Time-Domain Linear Interpolation (3D FTDLI)

First, we briefly explain FTDLI for 1D [6].

Let $\boldsymbol{A}(\omega,\psi)=[A_1(\omega,\psi),\ldots,A_M(\omega,\psi)]^T\in\mathbb{C}^M$ denote a TF between a microphone array and a sound source at ψ, in other words, a steering vector. M, ψ, and ω is the number of microphones, location of the sound source, and frequency, respectively. Our objective is to estimate an unknown TF at ψ_x, namely $\boldsymbol{A}(\omega,\psi_x)$, by interpolating two pre-measured TFs at $\psi_{\underline{x}}$ and $\psi_{\overline{x}}$, where $\psi_{\underline{x}}<\psi_x<\psi_{\overline{x}}$.

FTDLI is the integration between the phase interpolation of *Frequency Domain Linear Interpolation (FDLI)* [19] and the amplitude interpolation of *Time Domain Linear Interpolation (TDLI)* [20]. The integration procedure is as follows:



(a) 1D FTDLI (b) 3D FTDLI

Fig. 2. Difference of FTDLI between 1D and 3D

B1) Interpolate TFs by FDLI as follows:

$$\hat{A}_{m[\mathrm{F}]}(\omega,\psi_x)=(1-D_x)A_m(\omega,\psi_{\underline{x}})+D_xA_m(\omega,\psi_{\overline{x}})\ , \quad (1)$$

where $\hat{A}_{m[\mathrm{F}]}(\omega,\psi_x)$ is an interpolated TF of the m-th microphone at ψ_x using $A_m(\omega,\psi_{\underline{x}})$ and $A_m(\omega,\psi_{\overline{x}})$. $D_x\in\mathbb{R}$ represents an interpolation factor, $0\le D_x\le 1$.

B2) Interpolate TFs by TDLI as follows:

$$\hat{A}_{m[\mathrm{T}]}(\omega,\psi_x)=A_m^{1-D_x}(\omega,\psi_{\underline{x}})A_m^{D_x}(\omega,\psi_{\overline{x}})\ . \quad (2)$$

B3) Decompose interpolated TFs into phase and gain

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

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

B4) Calculate $\hat{A}_m(\omega,\psi_x)$ as follows:

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

This paper extends the 1D FTDLI to the 3D FTDLI. Let $\psi_{xyz}=[\psi_x,\psi_y,\psi_z]^T$ denote 3D location of a sound source. We estimate a TF at an unknown 3D location ψ_{xyz}, namely $\boldsymbol{A}(\omega,\psi_{xyz})$, by interpolating eight pre-measured TFs at $\psi_{\underline{x}\underline{y}\underline{z}}$, $\psi_{\underline{x}\underline{y}\overline{z}}$, $\psi_{\underline{x}\overline{y}\underline{z}}$, $\psi_{\underline{x}\overline{y}\overline{z}}$, $\psi_{\overline{x}\underline{y}\underline{z}}$, $\psi_{\overline{x}\underline{y}\overline{z}}$, $\psi_{\overline{x}\overline{y}\underline{z}}$, and $\psi_{\overline{x}\overline{y}\overline{z}}$, which are $\psi_{\underline{x}}<\psi_x<\psi_{\overline{x}}$, $\psi_{\underline{y}}<\psi_y<\psi_{\overline{y}}$, and $\psi_{\underline{z}}<\psi_z<\psi_{\overline{z}}$.

Instead of 1D FDLI, 3D FDLI extends Eq. (1) as follows using trilinear interpolation:

$$\begin{aligned}
\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xy\underline{z}})&=\begin{bmatrix}1-D_y & D_y\end{bmatrix}\begin{bmatrix}A_m(\omega,\psi_{\underline{x}\underline{y}\underline{z}}) & A_m(\omega,\psi_{\overline{x}\underline{y}\underline{z}})\\ A_m(\omega,\psi_{\underline{x}\overline{y}\underline{z}}) & A_m(\omega,\psi_{\overline{x}\overline{y}\underline{z}})\end{bmatrix}\begin{bmatrix}1-D_x\\ D_x\end{bmatrix}\\
\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xy\overline{z}})&=\begin{bmatrix}1-D_y & D_y\end{bmatrix}\begin{bmatrix}A_m(\omega,\psi_{\underline{x}\underline{y}\overline{z}}) & A_m(\omega,\psi_{\overline{x}\underline{y}\overline{z}})\\ A_m(\omega,\psi_{\underline{x}\overline{y}\overline{z}}) & A_m(\omega,\psi_{\overline{x}\overline{y}\overline{z}})\end{bmatrix}\begin{bmatrix}1-D_x\\ D_x\end{bmatrix}\\
\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xyz})&=(1-D_z)\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xy\underline{z}})+D_z\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xy\overline{z}})\ , \quad (6)
\end{aligned}$$

where D_x, D_y, and D_z are scalars in $0\le D_x,D_y,D_z\le 1$.

In the same manner, 3D TDLI extends Eq. (2) as follows:

$$\begin{aligned}
\hat{A}_{m[\mathrm{T}]}(\omega,\psi_{xyz})=&A_m^{(1-D_x)(1-D_y)(1-D_z)}(\omega,\psi_{\underline{x}\underline{y}\underline{z}})A_m^{(1-D_x)(1-D_y)D_z}(\omega,\psi_{\underline{x}\underline{y}\overline{z}})\\
&A_m^{(1-D_x)D_y(1-D_z)}(\omega,\psi_{\underline{x}\overline{y}\underline{z}})A_m^{(1-D_x)D_yD_z}(\omega,\psi_{\underline{x}\overline{y}\overline{z}})\\
&A_m^{D_x(1-D_y)(1-D_z)}(\omega,\psi_{\overline{x}\underline{y}\underline{z}})A_m^{D_x(1-D_y)D_z}(\omega,\psi_{\overline{x}\underline{y}\overline{z}})\\
&A_m^{D_xD_y(1-D_z)}(\omega,\psi_{\overline{x}\overline{y}\underline{z}})A_m^{D_xD_yD_z}(\omega,\psi_{\overline{x}\overline{y}\overline{z}}) \quad (7)
\end{aligned}$$

3D FTDLI integrates $\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xyz})$ in Eq. (6) and $\hat{A}_{m[\mathrm{T}]}(\omega,\psi_{xyz})$ in Eq. (7) by the steps B3) and B4) and obtains an interpolated TF, denoted as $\hat{A}_m(\omega,\psi_{xyz})$.

Although this is a straightforward extension, this computationally inexpensive method allows us to conduct the interpolation while operating SSL. This means that the points for interpolation can be decided dynamically, which is necessary for OH-SSL described below. Thus, this interpolation method is suitable for real-time 3D SSL.

III. Three-Dimensional Optimal Hierarchical Sound Source Localization (OH-SSL)

A. Algorithm of OH-SSL

Since fine TFs are obtained in a 3D space using FTDLI to achieve 3D super-resolution SSL, its computational cost becomes expensive. In MUSIC-based SSL algorithms [2], [18], the spatial spectrum for 3D SSL is determined by

$$P(\omega,\psi_{xyz},f)=\frac{|\hat{\boldsymbol{A}}^{*}(\omega,\psi_{xyz})\hat{\boldsymbol{A}}(\omega,\psi_{xyz})|}{\sum_{m=L_s+1}^{M}|\hat{\boldsymbol{A}}^{*}(\omega,\psi_{xyz})\boldsymbol{e}_m(\omega,f)|}\ , \quad (8)$$

where $()^{*}$, L_s and f are the conjugate transpose operator, dimension of subspace for targets and the frame index, respectively. $\boldsymbol{e}_m(\omega,f)$ is the m-th Eigen vector of a correlation matrix of multichannel signal. For SSL, we integrate $P(\omega,\psi_{xyz},f)$ over ω, denoted as $\bar{P}(\psi_{xyz},f)$, and search the L_sth-largest local maximum points of $\bar{P}(\psi_{xyz},f)$ with respect to ψ_{xyz}. Let $\psi_{xyz}^{[l]}$ be the direction that has the l-th largest $\bar{P}(\psi_{xyz},f)$, where $1\le l\le L_s$.

The computational cost for SSL is drastically affected by a resolution change of $\hat{\boldsymbol{A}}(\omega,\psi_{xyz})$ because:

C1) Eq. (8) is computed for all $\hat{\boldsymbol{A}}(\omega,\psi_{xyz})$.

C2) the number of interpolated TFs decides the number of operations for FTDLI.

Thus, we propose OH-SSL to reduce the computational cost for C1) and C2). As discussed above, thanks to the low complexity for FTDLI, we can assume that the computational cost for C2) is negligible compared to that of C1). Below we describe OH-SSL for 1D with respect to ψ_x, and it can be applied for ψ_y and ψ_z in the same manner. We previously proposed the algorithm [6], and this paper additionally discusses the *optimality* in Section III-B, which automatically determines optimal solutions for the number and resolution of hierarchies that minimize the total computational cost.

Let $\psi_x^{[l]}$, K, and $d_{x[k]}$ be ψ_x of $\psi_{xyz}^{[l]}$, the number of hierarchies, the resolution of the k-th hierarchy for ψ_x, respectively. The procedure of OH-SSL is as follows:

D1) Prepare TFs with the resolution $d_{x[k]}$. If necessary, interpolate TFs using pre-measured TFs.

D2) Conduct SSL and search peaks of $\bar{P}(\psi_{xyz},f)$. Let $\psi_{x[k]}^{[l]}$ be ψ_x having the l-th largest $\bar{P}(\psi_{xyz},f)$.

D3) Take the two closest measured/interpolated ψ_{xyz} from $\psi_{x[k]}^{[l]}$ in ψ_x, which are denoted as $\psi_{x[k]}^{[l-]}$ and $\psi_{x[k]}^{[l+]}$. Suppose $\psi_{x[k]}^{[l-]}<\psi_{x[k]}^{[l]}<\psi_{x[k]}^{[l+]}$.

D4) Using $\hat{A}_m(\omega,\psi_{x[k]}^{[l-]})$ and $\hat{A}_m(\omega,\psi_{x[k]}^{[l]})$, generate $\hat{A}_m(\omega,\psi_{xyz})$ between $\psi_{x[k]}^{[l-]}$ and $\psi_{x[k]}^{[l]}$ by Eq. (5) with the resolution of $d_{x[k+1]}$.

D5) Do D4) by using $\hat{A}_m(\omega,\psi_{x[k]}^{[l]})$ and $\hat{A}_m(\omega,\psi_{x[k]}^{[l+]})$.

D6) Conduct SSL again only with TFs in D4) and D5) and search a peak of $\bar{P}(\psi_{xyz},f)$.

D7) Repeat from D3) to D6) until k increments up to K.

The upper layer is for coarse localization, while the lower layer provides finer localization. In this sense, SSL with OH-SSL follows coarse-to-fine [17] localization, giving rough SSL results first, followed by fine SSL results.

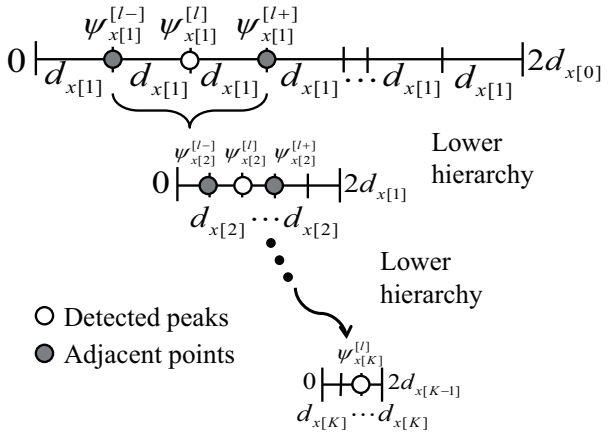


Fig. 3. Hierarchical Structure of OH-SSL

B. Optimality of OH-SSL

We determine an optimal solutions for K and $d_{x[k]}$ which minimize the computational cost of C1) and C2). Let c_S be the computational cost for one-point SSL in D6) and c_I be the computational cost for one-point interpolation in D4) and D5). Fig. 3 shows hierarchical structure for OH-SSL.

For C1), the number of points for SSL in the k-th hierarchy is $\frac{2d_{k-1}}{d_k}$, thus the computational cost is $\frac{2d_{k-1}}{d_k}c_S$. For C2), the number of points for interpolation in the k-th hierarchy is at most $\frac{2d_{k-1}}{d_k}$, thus the computational cost in the k-th hierarchy is at most $\frac{2d_{k-1}}{d_k}c_I$. As discussed above, since FTDLI is computationally inexpensive, this cost is assumed to be negligible, resulting the sum of both C1) and C2) as $g(k)=\frac{2d_{k-1}}{d_k}c_S$. Hereinafter, $c_S=1/2$ for the simplicity. Hence, the total computational cost is described as follows:

$$G(K)=\sum_{k=1}^{K}g(k)=\sum_{k=1}^{K}\frac{d_{k-1}}{d_k} \quad (9)$$

First, the optimality when $K=2$ is discussed. In this case, only d_1 is variable. $G(2)$ is minimized when $\frac{\partial G(2)}{\partial d_1}=0$, resulting $d_1=\sqrt{d_0d_2}$. The resulting computational cost in each hierarchy is $g(1)=g(2)=\sqrt{d_0/d_2}$. Thus, $G(2)$ is minimized when the computational cost of each hierarchy is equal. In the same manner, for $K\ge 3$, $G(K)$ is minimized when the computational cost of each hierarchy is equal. This can be proven by contradiction that if there is at least one pair of $g(k)$ and $g(k+1)$ that minimizes $G(K)$ and satisfies $g(k)\ne g(k+1)$, this contradicts to the fact that $\frac{\partial(g(k)+g(k+1))}{\partial d_k}=0$ is achieved when $g(k)=g(k+1)$.

Under the condition that $g(1)=g(2)=\cdots=g(K)$, $G(K)$ is obtained as follows:

$$\tilde{G}(K)=K\,(d_0/d_K)^{\frac{1}{K}} \quad (10)$$

Next, we obtain the optimal K that minimizes $\tilde{G}(K)$. For this, K satisfying $\frac{\partial\tilde{G}(K)}{\partial K}=0$ is determined as follows:

$$\frac{\partial\tilde{G}(K)}{\partial K}=\left(\frac{d_0}{d_K}\right)^{\frac{1}{K}}\left(1-\frac{1}{K}\log\left(\frac{d_0}{d_K}\right)\right)=0\Rightarrow K=\log\left(\frac{d_0}{d_K}\right)\ , \quad (11)$$

which is the *optimal* number of hierarchies to minimize $G(K)$. Finally, d_k with the optimal K is obtained as follows:

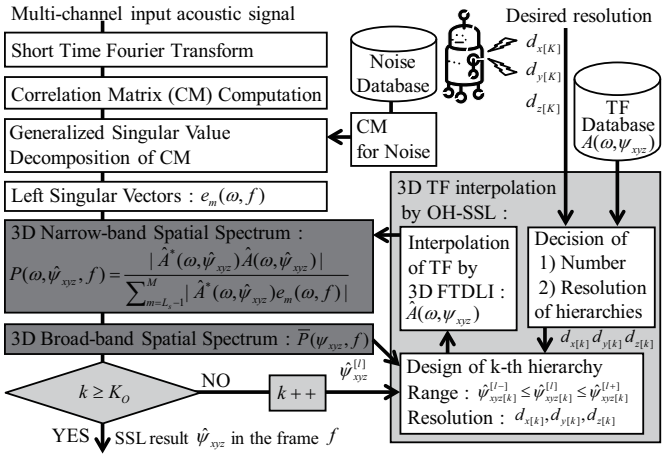


Fig. 4. Block Diagram of Real-time 3D Super-resolution SSL

$$d_k = d_0^{\frac{K-k}{K}} d_K^{\frac{k}{K}} \tag{12}$$

K and d_k in Eqs. (11) and (12) are used in D1)-D7). Hereinafter, K satisfying Eq. (11) is denoted as K_O. In the real application, K_O is rounded to unit.

IV. System Implementation

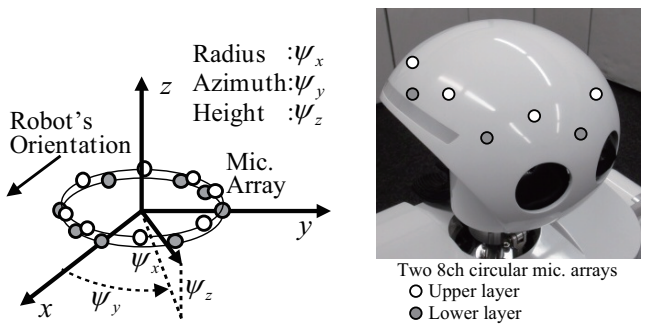
Fig. 4 shows the system structure for real-time 3D super-resolution SSL. As an extension of 1D super-resolution SSL (See [6] for details), the algorithm for 3D FTDLI and OH-SSL were added. We used a robot located in a 7m×4m room, where reverberation time was 0.2 seconds (RT_{20}). Fig. 5(a) shows the coordinate system for SSL. The robot faced in the same direction as the x-axis. We have utilized 16 microphones embedded in the robot's head shown in Fig. 5(b), which consist of two 8-channel circular microphone arrays with a variation in height. We pre-measured $\boldsymbol{A}(\omega,\psi_{xyz})$ cylindrically at every 5° in azimuth and every 0.2m in height from -0.4m to 0.4m, which were obtained by time-stretched pulse recording. With respect to the pre-measured $\boldsymbol{A}(\omega,\psi_{xyz})$, we defined ψ_{xyz} as a 3D location in a cylindrical coordinate, where ψ_x, ψ_y, and ψ_z denote radius, azimuth, and height, respectively (See Fig. 5(a)). The acoustic signal was sampled with 16kHz and 16bits. 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 [23]. The system worked with a laptop having a 2.0 GHz Intel Core i7 CPU and 8GB SDRAM.

V. Evaluation

This section shows the following evaluations:

1) Error of 3D TF interpolation
2) Computational cost of hierarchical source search
3) Application of the proposed SSL for a robot

In 1), we compared the errors of FTDLI, FDLI, and TDLI to see the validity of integration. In 2), we compared the computational cost of OH-SSL and several other conventional methods. Finally, we applied the SSL in Fig. 4 to a robot in a real environment in 3). The evaluation in 1) and 3) focused on 2D SSL (azimuth and elevation), and 2) was applied for



(a) Coordinates for SSL (b) Microphone Array

Fig. 5. Conditions for Experiment

1D hierarchical source search. Notice that the methods in Sections II and III are applicable for 3D SSL.

A. Error of 3D TF Interpolation Using FTDLI

We evaluated the 3D interpolation errors of FDLI, TDLI, and FTDLI. To see the general applicability of FTDLI (considering the robustness against the variations of number of microphones and the quality of pre-measured TFs), we evaluated the errors using TFs measured by Nagoya University[22] first and TFs of our robot afterward.

TFs measured by Nagoya University is widely known as a 2D standard binaural HRTF dataset, recorded by a spherical HRTF measurement equipment. Thus, we defined ψ_{xyz} in a spherical coordinate, where ψ_x, ψ_y, and ψ_z denote radius, azimuth, and elevation, respectively.

For the evaluation, we took the difference between $\hat{\boldsymbol{A}}(\omega,\psi_{xyz})$ and $\boldsymbol{A}(\omega,\psi_{xyz})$ with the variation of ψ_y and ψ_z because of the constant ψ_x=1.2m. We selected $\psi_{\underline{xyz}}$, $\psi_{\underline{xy}\bar{z}}$, $\psi_{\underline{x}\bar{y}\underline{z}}$, and $\psi_{\underline{x}\bar{y}\bar{z}}$ so that the center of the four locations becomes $[\psi_x,\psi_y,\psi_z]=[1.2\text{m},0°,0°]$. Namely, the four locations are described as $[\psi_x,\psi_y,\psi_z]=[1.2\text{m},\pm\delta_y,\pm\delta_z]$, where δ_y and δ_z are azimuth and elevation utilized as pre-measured TFs, respectively. The evaluation utilized $\delta_y=\{15°,30°,45°,60°\}$ and $\delta_z=\{15°,30°,45°\}$. $\hat{\boldsymbol{A}}(\omega,\psi_{xyz})$ was estimated at every 5° of ψ_y and ψ_z. We averaged the estimated errors for ω and ψ_{xyz}. The averaged error, denoted as $\bar{e}$, was calculated as:

$$\bar{e}=\frac{1}{i_\psi}\sum_{i=1}^{i_\psi}\frac{1}{k_h-k_l+1}\sum_{k=k_l}^{k_h} f(\omega_{[k]},\psi_{xyz[i]})\ , \tag{13}$$

where $f(\omega_{[k]},\psi_{xyz[i]})$ is the estimation error (defined later) for specific ψ_{xyz} and ω. k_l and k_h are defined so that the frequency band 500[Hz] $\leq \omega \leq$ 2800[Hz] is evaluated. i_ψ represents the number of ψ_{xyz} we conducted interpolation. For ψ_{xyz}, we utilized all ψ_{xyz} of the pre-measured $\boldsymbol{A}(\omega,\psi_{xyz})$ in the range of $-\delta_y<\psi_y<\delta_y$ and $-\delta_z<\psi_z<\delta_z$.

We evaluated three kinds of $f(\omega_{[k]},\psi_{xyz[i]})$ in Eq. (13), commonly used in interpolation error evaluation [19], [20].

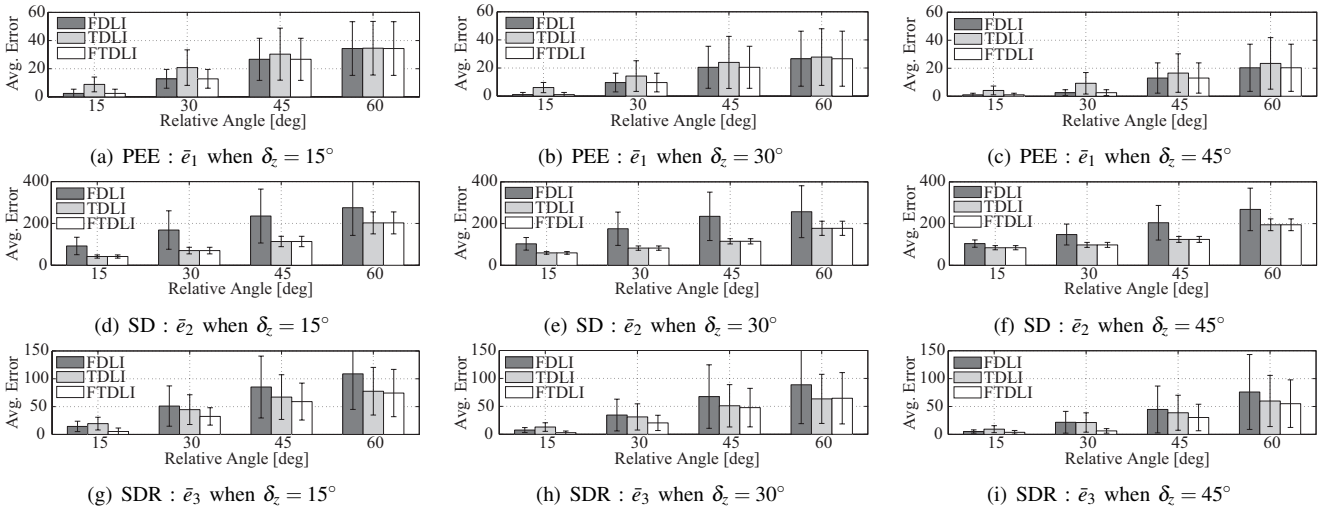
The first criterion is the normalized inner-product:

$$f_1(\omega,\psi_{xyz})=\sum_{m=1}^{M}\left|\frac{A_m(\omega,\psi_{xyz})\cdot\hat{A}_m(\omega,\psi_{xyz})}{|A_m(\omega,\psi_{xyz})||\hat{A}_m(\omega,\psi_{xyz})|}-1\right|\ , \tag{14}$$

which represents *Phase Estimation Error (PEE)*.

The second criterion is *Spectral Distortion (SD)*:

$$f_2(\omega,\psi_{xyz})=\sum_{m=1}^{M}\left|20\log\frac{|\hat{A}_m(\omega,\psi_{xyz})|}{|A_m(\omega,\psi_{xyz})|}\right|\ , \tag{15}$$

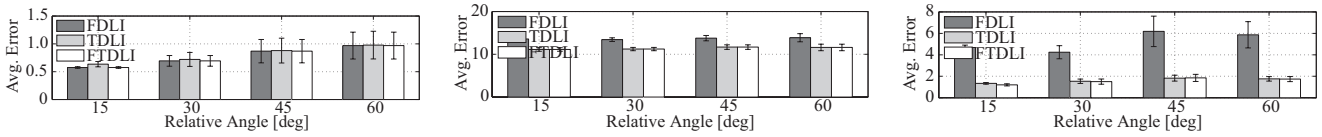


(a) PEE : $\bar{e}_1$ when $\delta_z = 15°$ (b) PEE : $\bar{e}_1$ when $\delta_z = 30°$ (c) PEE : $\bar{e}_1$ when $\delta_z = 45°$

(d) SD : $\bar{e}_2$ when $\delta_z = 15°$ (e) SD : $\bar{e}_2$ when $\delta_z = 30°$ (f) SD : $\bar{e}_2$ when $\delta_z = 45°$

(g) SDR : $\bar{e}_3$ when $\delta_z = 15°$ (h) SDR : $\bar{e}_3$ when $\delta_z = 30°$ (i) SDR : $\bar{e}_3$ when $\delta_z = 45°$

Fig. 6. Interpolation error of PEE, SD, SDR using Nagoya TF Database[22]



(a) PEE : $\bar{e}_1$ when δ_z = 0.2m (b) SD : $\bar{e}_2$ when δ_z = 0.2m (c) SDR : $\bar{e}_3$ when δ_z = 0.2m

Fig. 7. Interpolation error of PEE, SD, SDR using TFs of a robot-embedded microphone array

which shows the amplitude estimation performance.

The third criterion is *Signal-to-Distortion Ratio (SDR)* [1]:

$$f_3(\omega, \psi_{xyz}) = \sum_{m=1}^{M} \frac{|A_m(\omega, \psi_{xyz}) - \hat{A}_m(\omega, \psi_{xyz})|^2}{|A_m(\omega, \psi_{xyz})|^2} \quad , \qquad (16)$$

which represents the total estimation performance. Let $\bar{e}_1$, $\bar{e}_2$, and $\bar{e}_3$ denote $\bar{e}$ of PEE, SD, and SDR, respectively.

Fig. 6 shows the comparison of $\bar{e}_1$, $\bar{e}_2$, and $\bar{e}_3$, with the variation of δ_z. The horizontal axis shows δ_y.

As δ_y and δ_z increase, the resolution of pre-measured TFs becomes more coarse, resulting higher errors in interpolation. FDLI and FTDLI achieved the smallest $\bar{e}_1$, and TDLI and FTDLI achieved the smallest $\bar{e}_2$. Finally, FTDLI achieved the smallest $\bar{e}_3$ for the whole range of δ_y and δ_z.

In the same manner, we evaluated the pre-measured TFs of our robot-embedded microphone array. Fig. 7 shows the comparison of $\bar{e}_1$, $\bar{e}_2$, and $\bar{e}_3$, with δ_z=0.2m. The horizontal axis shows δ_y. The result also confirmed that FTDLI achieved the smallest interpolation error, confirming the high accuracy of the proposed method.

B. Computational Cost of OH-SSL

We compared the computational cost of OH-SSL and other conventional methods to see the efficiency of OH-SSL. As discussed in Section III, we calculated the processing time for D1)-D7) with the change of the desired resolution of SSL, $d_{x[K]}$. To see the validity of OH-SSL, we have measured the processing time of the following four methods and compared them towards the processing time of OH-SSL. The first method, described as **H1**, is the searching without hierarchical algorithm. The second and third methods, described as **H2** and **H3**, used two and three hierarchies every time, respectively. The resolution of each hierarchy in H2 and H3 was determined based on Eq. (12). The fourth method is the spherical grid[13], described as **SG**. The number of hierarchies in SG was determined so that the resolution of the grid can satisfy finer resolution than $d_{x[K]}$. In OH-SSL, denoted as **OH**, the resolution and number of hierarchies were determined based on Eqs. (11) and (12), respectively.

[1]We have inverted the original SDR discussed in [19] since it shows the best estimation performance with $f(\omega, \psi_{xyz}) = 0$.

TABLE I

COMPARISON OF COMPUTATIONAL COST

Condition		Comp. cost rate against OH-SSL			
$d_{x[K]}$	$d_{x[0]}$	H1/OS	H2/OS	H3/OS	SG/OS
10.0	360.0	4.5	1.4	1.1	1.4
1.0	360.0	30.0	3.0	1.8	1.4
0.1	360.0	225.0	7.5	2.8	1.6

Tables I shows the result. We evaluated the processing time ratio of H1, H2, H3, and SG towards OH, denoted as H1/OH, H2/OH, H3/OH, and SG/OH, respectively. Therefore, the value larger than one means that the computational cost is more expensive than that of OH. The result showed:

- All the ratios were larger than one, implying that the computational cost of OH is less than that of others,
- The ratios had larger values for finer $d_{x[K]}$, meaning that OH-SSL is efficient for super-resolution SSL,

which successfully confirmed the validity of OH-SSL.

C. Applicability of the proposed methods to SSL for a robot

The error in estimating location toward a moving sound source with/without interpolation was investigated. We focus on SSL of a single sound source by recording a moving white noise source. Three TF conditions were compared:

1) Pre-measured TFs at 5° azimuth intervals and 0.2m height intervals (Fine TFs),

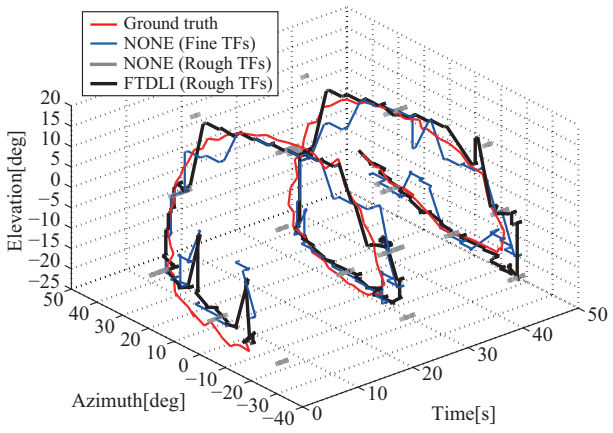


Fig. 8. Trajectory of the 3D Localization

2) Pre-measured TFs at 30° azimuth intervals and 0.4m height intervals (Rough TFs),
3) Interpolated TFs at 5° azimuth intervals and 0.01m height intervals using 2) (FTDLI).

The sound source was localized by an ultrasonic zone positioning system as a ground truth. Fig. 8 shows the comparison. The x-, y-, and z-axes show time, azimuth, and elevation of the sound source, respectively. The red-, blue-, gray-, and black-lines show ground truth, the result of 1), 2), and 3), respectively. Since the gray-line used roughly pre-measured TFs, the line shows many fragmentations. The line was also far from the ground truth because of the low resolution, and the averaged angle error between estimated location and ground truth, denoted as $\bar{e}_\psi$, was $\bar{e}_\psi = 10.5°$. The blue-line continuously localized the sound source using fine TFs. However, it had some variability because the resolution of height was not sufficient for the sound source motion, and $\bar{e}_\psi = 7.2°$. On the other hand, the black-line performed both continuous and stable localization using FTDLI, and $\bar{e}_\psi = 6.5°$. We also measured averaged processing time for SSL with/without OH-SSL when using TFs at 5° azimuth intervals and 0.01m height intervals. The processing time with and without OH-SSL was 0.028[sec] and 1.073[sec], respectively, which successfully showed the drastic computational cost reduction by approximately 97%. Finally, the result showed:

- 3D FTDLI achieved super-resolution SSL working better than pre-measured TFs,
- OH-SSL reduced computational cost drastically,

which confirmed the validity of the proposed techniques.

VI. Conclusion

This paper investigated a 3D SSL for a robot. Because the SSL should work in real-time with a sufficiently high accuracy, we focused on two issues: the resolution of SSL should be sufficiently high for a 3D space, the computational cost for 3D sound source search should be in real-time. For the first issue, we proposed 3D FTDLI which interpolates roughly-pre-measured TFs and generates fine TFs in a 3D space. For the second issue, we proposed OH-SSL which optimally minimizes the computational cost for searching for sound sources. The evaluation showed: 1) 3D FTDLI showed a better interpolation performance compared to the existing methods and provided super-resolution 3D SSL, 2) OH-SSL drastically reduced the computational cost by 97%.

Finally the proposed functions were integrated into real-time super-resolution 3D SSL. The application showed that 3D FTDLI and OH-SSL performed even better continuous and stable localization than that of finely-pre-measured TFs, which confirmed the validity of the proposed techniques.

Our future study will be the performance evaluation with multiple sound sources and the integration of 3D SSL with visually-obtained localization results and the construction of a robust localization system for robots in a real environment.

References

[1] K. Nakadai *et al.*, "Active Audition for Humanoid", in *Proc. of AAAI-2000*, pp. 832–839, 2000.
[2] R. Schmidt, "Multiple emitter location and signal parameter estimation", *IEEE Trans. Ant. Prop.*, vol. 34, no. 3, pp. 276–280, 1986.
[3] K. Nakadai *et al.*, "Robust Tracking of Multiple Sound Sources by Spatial Integration of Room and Robot Microphone Arrays", in *Proc. of IEEE ICASSP*, vol. IV, pp. 929–932, 2006.
[4] 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.
[5] S. Argentieri and P. Danés, "Broadband variations of the MUSIC high-resolution method for sound source localization in Robotics", in *IROS*, pp. 2009–2014, 2007.
[6] K. Nakamura *et al.*, "Real-time Super-resolution Sound Source Localization for Robots," in *IROS*, pp. 694–699, 2012.
[7] K. Nakadai *et al.*, "A robot referee for rock-paper-scissors sound games", in *ICRA*, pp. 3469–3474, 2008.
[8] L. Wang *et al.*, "Head-related transfer function interpolation through multivariate polynomial fitting of principal component weights", *Acoust. Sci. & Tech.*, vol. 30, no. 6, pp. 395–403, 2009.
[9] F. P. Freeland *et al.*, "HRTF interpolation through direct angular parameterization", in *Proc. of IEEE ISCAS*, pp. 1823–1826, 2007.
[10] M. Otani and S. Ise, "Fast calculation system specialized for head-related transfer function based on boundary element method", *J. Acoust. Soc. Am.*, vol. 119, no. 5, pp. 2589–2598, 2006.
[11] H. Nakashima *et al.*, "A Localization Method for Multiple Sound Sources by Using Coherence Function", in *Proc. of 18th EUSIPCO*, pp. 130–134, 2010.
[12] B. Rudzyn *et al.*, "Real time robot audition system incorporating both 3D sound source localization and voice characterization", in *ICRA*, pp. 4733–4738, 2007.
[13] J.-M. Valin *et al.*, "Localization of simultaneous moving sound sources for mobile robot using a frequency- domain steered beamformer approach", in *ICRA*, vol. 1, pp. 1033–1038, 2004.
[14] D. Hoang, "A Fast Microphone Array SRP-PHAT Source Location Implementation using Coarse-To-Fine Region Contraction(CFRC)", in *Proc. of IEEE WASPAA*, pp. 295–298, 2007.
[15] J.-S. Hu *et al.*, "Simultaneous localization of mobile robot and multiple sound sources using microphone array", in *ICRA*, pp. 29–34, 2009.
[16] C. T. Ishi *et al.*, "Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments", in *IROS*, pp. 2027–2032, 2009.
[17] D. Ringach, "Look at the big picture (details will follow)", *Nature Neuroscience*, vol. 6, no. 1, pp. 7–8, 2003.
[18] K. Nakamura *et al.*, "Intelligent Sound Source Localization for Dynamic Environments", in *IROS*, pp. 664–669, 2009.
[19] T. Nishino *et al.*, "Interpolating Head Related Transfer Functions in the median plane", in *Proc. of IEEE WASPAA*, pp. 167–170, 1999.
[20] 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.
[21] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR", *J. Acoust. Soc. Am.*, vol. 97, pp. 3907–3908, 1995.
[22] T. Nishino *et al.*, "Interpolating head related transfer functions," in *Proc. of 7th WESTPRAC VII*, 1A-1-3, pp.293-296, 2000.
[23] K. Nakadai *et al.*, "Design and Implementation of Robot Audition System HARK", *Advanced Robotics*, vol. 24, pp. 739–761, 2009.