人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202-2

ロボットのための実環境ロバストな実時間超解像三次元音源定位

Real-time Noise-robust Super-resolution Sound Source Localization in Three-dimension for Robots

中村圭佑 [1,2]，中臺一博 [1]，奥乃博 [2]

Keisuke NAKAMURA[1,2], Kazuhiro NAKADAI[1], Hiroshi OKUNO[2]

1.（株）ホンダ・リサーチ・インスティチュート・ジャパン，2. 京都大学大学院

1. Honda Research Institute Japan Co., Ltd., 2. Kyoto University

keisuke@jp.honda-ri.com, nakadai@jp.honda-ri.com, okuno@kuis.kyoto-u.ac.jp

Abstract

This paper investigates Three-dimensional Sound Source Localization (3D-SSL) for a robot. 3D-SSL by a robot mainly requires: 1) robustness against high power noise such as robot's ego-noise, 2) sufficiently-high resolution for a 3D space, 3) real-time operation for searching for sound sources in a 3D space. For these, we propose: 1) multiple signal classification based on generalized singular value decomposition (GSVD-MUSIC), 2) transfer function interpolation based on integration of linear interpolation in frequency- and time-domain (FT-DLI), 3) optimal hierarchical sound source localization (OH-SSL). These techniques are integrated into an SSL system using a robot, and the experimenral result showed 3D-SSL in real-time.

1 序論

人とロボットの会話によるインタラクションは重要である．携帯電話などの接話マイクを用いた音声認識に比べ，ロボットに搭載されたマイクを使った音声認識は，発話者からマイクの距離が遠く信号対雑音比が低い，発話者数は単独であると仮定できないという特徴がある．ロボット聴覚[1]では，マイクロホンアレイを用いて空間的に複数の音源を定位・分離することでこれらの問題に対処している．従って，音源定位・分離の性能向上はロボット聴覚システム全体の性能向上に必要不可欠である．

これまで，ロボット聴覚のための音源定位において，解像度が高いという利点のある Multiple Signal Classification (MUSIC[2]) を使用して，主に一次元（方位角）のみの定位が使用されてきた[3; 4; 5]．方位角に対する一次元の音源定位は，複数同時発話[7]に代表されるように，全ての音源がある水平面に近く，高さが近い場合に高い性能を維持する．Figure 1 は，一音源の距離と方位角をそれぞれ 1.0m と 0° に固定し，音源とマイクアレイとの相対高さを変化させたときの MUSIC による方位角推定結果である．図の横軸と縦軸はそれぞれ，方位角と MUSIC スペクトルを示す．図中の黒実線 (0.0m) は音源が水平面上にある時の方位角推定結果，それ以外は水平面から離れている時の方位角推定結果を示す．図のように，音源が高さを持たない場合は 0° に鋭角なピークが見られるが，高さが変化してしまうと解像度が低下し，音源定位性能が劣化してしまう．従って，高さの異なる複数の音源や雑音の定位性能を向上するために三次元の音源定位は必要不可欠である．

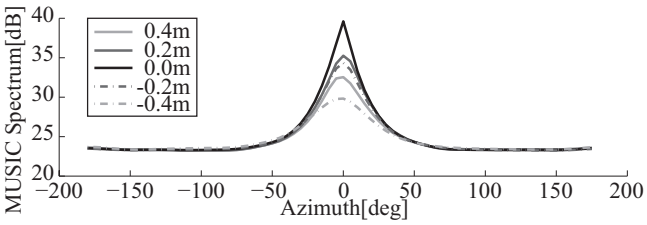


Figure 1: 1D SSL Result with the Variation of Heights

三次元音源定位の必要性に関わらず，ロボットに適用された報告は少なく，これまで，両耳聴音源定位による手法[8]，ビームフォーミングによる手法[9; 10]，MUSIC による手法[11; 12]等が報告されている．両耳聴音源定位による手法[8]は高速に定位できるが，フレーム内に一音源以上存在しないという制約が存在する．ビームフォーミングによる手法[9; 10] は複数音源に対応した高速な音源定位が実現できるが，空間解像度や雑音ロバスト性が MUSIC と比べて十分でない．一方，MUSIC による手法[11; 12]は，計算コストが高く，実時間性が保証されない．

本稿の目的は，高い雑音ロバスト性，高い空間解像度，実時間処理を実現する MUSIC による三次元音源定位の枠組を構築することにある．このため，本稿では次の問題に取り組む．

A1) ロボットの自己雑音等の目的音より大きなパワーを持つ雑音下で定位性能が劣化する．

A2) 三次元空間で高い空間解像度が求められる場合，マイクロホンアレイのキャリブレーションのための伝達関数の計測回数が増大する．

A3) 一次元音源定位に比べ，三次元音源定位では，音源探索にかかる計算コストが増大する．

A1) に対し，一般化特異値展開 (Generalized Singular Value Decomposition, GSVD) を用いた MUltiple SIgnal Classi cation(GSVD-MUSIC) を提案する．この手法は，MUSIC で一般に用いられる標準固有値展開を一般化特異値展開に拡張することにより，マイクロホン間の入力信号の相関行列に加え，自由に設計が可能な相関行列を導入することができる．この相関行列を既知または動的に取得した雑音信号から生成すれば，たとえその雑音が目的音よりも大きなパワーを持っていたとしても，雑音の影響を吸収する効果をこの相関行列が持つため，ロバストに目的音の定位ができる．

A2) に対し，一次元音源定位に対して提案されたハイブリッド伝達関数補間 (Frequency- and Time-Domain Linear Interpolation, FTDLI[6]) にトリリニア補間を導入する．この拡張によって，粗い解像度で計測された少数の三次元空間内伝達関数から所望の解像度の伝達関数を得ることができる．補間された伝達関数を音源定位に応用することで，超解像三次元音源定位を実現する．

A3) に対し，Coarse-to-fine 認識[13]に基づく最適階層的音源定位 (Optimal Hierarchical Sound Source Localization (OH-SSL)) を提案する．本手法により，空間解像度の粗い定位から細かい定位へと階層的に定位を行うことで，定位の精度を制御して精度を維持しつつ，探索にかかる計算コストを削減し，実時間性を向上する．本稿では，探索にかかるコストを最小化する最適な階層数と各階層の粒度について述べる．

2 GSVD-MUSIC による雑音ロバストな三次元音源定位

2.1 MUSIC による三次元音源定位

まずは，MUSIC [2]について概説する．音源の三次元位置を $\psi_{xyz} = [\psi_x, \psi_y, \psi_z]^T$ と表記する．ここで，座標系は円筒座標系，球座標系，直交座標系等，任意のものとする．MUSIC では ψ_{xyz} の音源のインパルス応答の計測もしくは，計算で得られる伝達関数（ステアリングベクトル）$\boldsymbol{A}(\omega, \psi_{xyz})$ を既知情報として用いる．ここで，ω は周波数である．定位では，f フレーム目の入力音響信号を短時間フーリエ変換して得られる $\boldsymbol{X}(\omega, f)$ から，以下のように入力信号の相関行列 $\boldsymbol{R}(\omega, f)$ を計算する．

$$\boldsymbol{R}(\omega, f) = \frac{1}{\mathrm{T_R}} \sum_{\tau=0}^{\mathrm{T_R}-1} \boldsymbol{X}(\omega, f+\tau)\boldsymbol{X}^*(\omega, f+\tau) \quad (1)$$

ここで，$()^*$ は複素共役転置演算子を表す．また，雑音に対するロバスト性向上のため，$\boldsymbol{R}(\omega, f)$ は $\mathrm{T_R}$ フレームで平滑化されている．

MUSIC では $\boldsymbol{R}(\omega, f)$ が張る空間を，以下の標準固有値展開 (SEVD) により，目的音と雑音の部分空間に分解する．

$$\boldsymbol{R}(\omega, f) = \boldsymbol{E}(\omega, f)\boldsymbol{\Lambda}(\omega, f)\boldsymbol{E}^{-1}(\omega, f) \quad (2)$$

以降，提案法と区別するため，式 (2) を用いた MUSIC を SEVD-MUSIC と呼ぶ．空間スペクトルは以下で求められる．

$$P(\omega, \psi_{xyz}, f) = \frac{|\boldsymbol{A}^*(\omega, \psi_{xyz})\boldsymbol{A}(\omega, \psi_{xyz})|}{\sum_{m=L_s+1}^{M} |\boldsymbol{A}^*(\omega, \psi_{xyz})\boldsymbol{e}_m(\omega, f)|} \quad (3)$$

ここで，L_s は音源数を，$\boldsymbol{e}_m(\omega, f)$ は式 (2) の $\boldsymbol{E}(\omega, f)$ の m 番目の固有ベクトルを表す．音源方向を推定するため，$P(\omega, \psi_{xyz}, f)$ を以下のように ω 方向に平均する．

$$\bar{P}(\psi_{xyz}, f) = \frac{1}{j_h - j_l + 1} \sum_{j=j_l}^{j_h} P(\omega_{[j]}, \psi_{xyz}, f) \quad (4)$$

ここで，j_l, j_h は音源定位で用いる最低・最高周波数に相当する周波数ビン番号を表す．音源探索では，$\bar{P}(\psi_{xyz}, f)$ がなす三次元超平面の極大点を探索し，その極大点の集合から極大値が大きいものから L_s 個の ψ_{xyz} を選択し，音源位置を決定する．以降，この L_s 個の ψ_{xyz} を $\psi_{xyz}^{[l]}$ と表記することとする ($1 \le l \le L_s$).

2.2 GSVD-MUSIC への拡張

式 (3) では，$\boldsymbol{A}(\omega, \psi_{xyz})$ は事前情報である為，入力音響信号に関わる項は $\boldsymbol{e}_m(\omega, f)$ のみであり，$\boldsymbol{e}_m(\omega, f)$ の選択が定位の性能に大きく影響する．MUSIC では式 (2) で求まる固有値が音源のパワーと相関があることを利用し，固有値の小さなものが雑音である（雑音のパワーは必ず目的音のパワーより小さい）という仮定のもとで $\boldsymbol{e}_m(\omega, f)$ を選択する．しかし，実環境では，雑音が目的音より大きなパワーを持つ場合が存在し，固有値の大きさに逆転が生じるため性能が劣化することが知られている．

これまで我々はこの問題の解決のため，一般化固有値分解を用いた MUSIC (GEVD-MUSIC) を導入しており，その対雑音ロバスト性を確認した[14]．この手法では，自由に設計が可能な相関行列 $\boldsymbol{V}$ を導入し，$\boldsymbol{V}$ を非発話区間の入力信号から式 (1) を用いて生成することで，非発話区間に存在するロボットの自己雑音等を白色化し，定位のロバスト性を向上させた．しかし，この手法は，計算量が大きい，もしくは，固有ベクトルの直交性が保証できないといった問題をかかえていた．

そこで，こうした問題を解決するため，本稿では一般化特異値分解による MUSIC(GSVD-MUSIC) を導入する．本手法は式 (2) を次のように拡張する．

$$\boldsymbol{V}^{-1}\boldsymbol{R}(\omega,f)=\boldsymbol{E}_l(\omega,f)\boldsymbol{\Lambda}(\omega,f)\boldsymbol{E}_r^*(\omega,f) \tag{5}$$

ここで，$\boldsymbol{E}_l(\omega,f),\boldsymbol{E}_r(\omega,f)$ は，それぞれ左特異ベクトル行列，右特異ベクトル行列を表し，いずれもユニタリ行列で互いの直交性が保証される．音源方向推定において，$\boldsymbol{E}_l(\omega,f)$ を式 (2) の $\boldsymbol{E}(\omega,f)$ として用いることで，MUSIC の対雑音ロバスト性を向上することができる．

3 FTDLI による三次元伝達関数の補間

まず，FTDLI による一次元伝達関数の補間[6]について概説する．本章では，ψ 方向の音源とマイクロホンアレイ間の伝達関数を $\boldsymbol{A}(\omega,\psi)=[A_1(\omega,\psi),\ldots,A_M(\omega,\psi)]^T$ と，M 個のマイクで別に表記する．2 つの $\psi_{\underline{x}}$ 方向と $\psi_{\overline{x}}$ 方向の事前計測伝達関数から，補間によって ψ_x 方向 ($\psi_{\underline{x}}<\psi_x<\psi_{\overline{x}}$) の未知伝達関数 $\boldsymbol{A}(\omega,\psi_x)$ を推定する．

FTDLI は周波数領域での線形補間法 (Frequency Domain Linear Interpolation, FDLI[15]) による位相情報と，時間領域での線形補間法 (Time Domain Linear Interpolation, TDLI[16]) による振幅情報を統合し，高い精度の補間を実現する．二つの線形補間法は以下のように統合される．

B1) FDLI による補間を行う．

$$\hat{A}_{m[\mathrm{F}]}(\omega,\psi_x)=(1-D_x)A_m(\omega,\psi_{\underline{x}})+D_xA_m(\omega,\psi_{\overline{x}}) \tag{6}$$

ここで，$\hat{A}_{m[\mathrm{F}]}(\omega,\psi_x)$ は，$A_m(\omega,\psi_{\underline{x}})$ と $A_m(\omega,\psi_{\overline{x}})$ の事前計測伝達関数を用いて推定された ψ_x 方向の音源と m 番目のマイクロホン間の伝達関数である．また，D_x は補間係数である ($0\le D_x\le 1$).

B2) TDLI による補間を行う．

$$\hat{A}_{m[\mathrm{T}]}(\omega,\psi_x)=A_m^{1-D_x}(\omega,\psi_{\underline{x}})A_m^{D_x}(\omega,\psi_{\overline{x}})\,. \tag{7}$$

B3) B1) と B2) によって補間された伝達関数を以下のように位相と振幅に分解する．

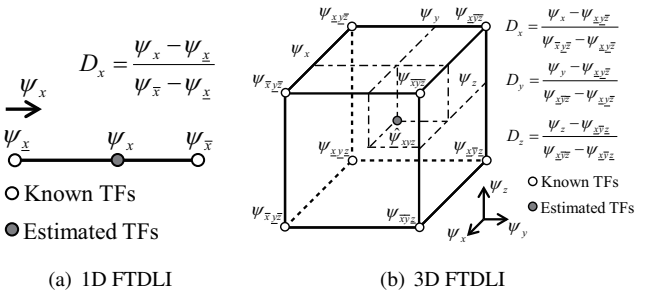
$$\hat{A}_{m[\mathrm{F}]}(\omega,\psi_x)=\lambda_{m[\mathrm{F}]}\exp(-j\omega t_{m[\mathrm{F}]}) \tag{8}$$

$$\hat{A}_{m[\mathrm{T}]}(\omega,\psi_x)=\lambda_{m[\mathrm{T}]}\exp(-j\omega t_{m[\mathrm{T}]}) \tag{9}$$

B4) 補間伝達関数 $\hat{A}_m(\omega,\psi_x)$ は以下で求まる．

$$\hat{A}_m(\omega,\psi_x)=\lambda_{m[\mathrm{T}]}\exp(-j\omega t_{m[\mathrm{F}]})\,. \tag{10}$$

本稿は一次元伝達関数に対する FTDLI を三次元伝達関数 $\boldsymbol{A}(\omega,\psi_{xyz})$ の補間に拡張する．8 つの三次元位置 ($\psi_{\underline{x}\underline{y}\underline{z}}$, $\psi_{\underline{x}\underline{y}\overline{z}}$, $\psi_{\underline{x}\overline{y}\underline{z}}$, $\psi_{\underline{x}\overline{y}\overline{z}}$, $\psi_{\overline{x}\underline{y}\underline{z}}$, $\psi_{\overline{x}\underline{y}\overline{z}}$, $\psi_{\overline{x}\overline{y}\underline{z}}$, $\psi_{\overline{x}\overline{y}\overline{z}}$) の事前計測伝達関数から，補間によって三次元空間の未知伝達関数 $\boldsymbol{A}(\omega,\psi_{xyz})$ を推定する．ただし，$\psi_{\underline{x}}<\psi_x<\psi_{\overline{x}}$, $\psi_{\underline{y}}<\psi_y<\psi_{\overline{y}}$, $\psi_{\underline{z}}<\psi_z<\psi_{\overline{z}}$ とする．



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

Figure 2: Difference of FTDLI between 1D and 3D

FDLI では，式 (6) にトリリニア補間を以下のように導入することで三次元伝達関数の補間に拡張する．

$$\begin{aligned}
\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xyz})&=\begin{bmatrix}1-D_y & D_y\end{bmatrix}\begin{bmatrix}A_m(\omega,\psi_{\underline{x}yz}) & A_m(\omega,\psi_{\overline{x}yz})\\ A_m(\omega,\psi_{\underline{x}\overline{y}\underline{z}}) & A_m(\omega,\psi_{\overline{x}\overline{y}\underline{z}})\end{bmatrix}\begin{bmatrix}1-D_x\\ D_x\end{bmatrix}\\
\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xy\overline{z}})&=\begin{bmatrix}1-D_y & D_y\end{bmatrix}\begin{bmatrix}A_m(\omega,\psi_{\underline{x}y\overline{z}}) & A_m(\omega,\psi_{\overline{x}y\overline{z}})\\ A_m(\omega,\psi_{\underline{x}\overline{y}\overline{z}}) & A_m(\omega,\psi_{\overline{x}\overline{y}\overline{z}})\end{bmatrix}\begin{bmatrix}1-D_x\\ D_x\end{bmatrix}\\
\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xyz})&=(1-D_z)\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xy\underline{z}})+D_z\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xy\overline{z}})
\end{aligned} \tag{11}$$

ここで，D_x, D_y, D_z はそれぞれ，補間係数である ($0\le D_x,D_y,D_z\le 1$)．同様に TDLI では，式 (7) にトリリニア補間を以下のように導入する．

$$\begin{aligned}
\hat{A}_{m[\mathrm{T}]}(\omega,\psi_{xyz})=&A_m^{(1-D_x)(1-D_y)(1-D_z)}(\omega,\psi_{\underline{x}\underline{y}\underline{z}})A_m^{(1-D_x)(1-D_y)D_z}(\omega,\psi_{\underline{x}\underline{y}\overline{z}})\\
&A_m^{(1-D_x)D_y(1-D_z)}(\omega,\psi_{\underline{x}\overline{y}\underline{z}})A_m^{(1-D_x)D_yD_z}(\omega,\psi_{\underline{x}\overline{y}\overline{z}})\\
&A_m^{D_x(1-D_y)(1-D_z)}(\omega,\psi_{\overline{x}\underline{y}\underline{z}})A_m^{D_x(1-D_y)D_z}(\omega,\psi_{\overline{x}\underline{y}\overline{z}})\\
&A_m^{D_xD_y(1-D_z)}(\omega,\psi_{\overline{x}\overline{y}\underline{z}})A_m^{D_xD_yD_z}(\omega,\psi_{\overline{x}\overline{y}\overline{z}})
\end{aligned} \tag{12}$$

三次元伝達関数に対する FTDLI は，式 (11) で得られる $\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xyz})$ と，式 (12) で得られる $\hat{A}_{m[\mathrm{T}]}(\omega,\psi_{xyz})$ とを，B3) と B4) に従って統合し，補間伝達関数 $\hat{A}_m(\omega,\psi_{xyz})$ を得る．最後に，$\hat{A}_m(\omega,\psi_{xyz})$ を式 (3) の $\boldsymbol{A}(\omega,\psi_{xyz})$ として使用することで，事前計測伝達関数よりも細かな空間解像度の音源定位（超解像音源定位）が実現できる．

4 OH-SSL による音源探索コストの削減

4.1 OH-SSL のアルゴリズム

GSVD-MUSIC では，式 (3) の空間スペクトルの算出と音源探索にかかるコストが，$\boldsymbol{A}(\omega,\psi_x)$ の空間解像度に比例した探索数に依存する．OH-SSL では，探索数を削減するために音源探索を階層化する．以下簡略化のため，ψ_x 軸上の一次元音源探索の階層化を扱うが，ψ_y 軸と ψ_z 軸についても同様に階層化が可能である．$\psi_x^{[l]}$, K, $d_{x[k]}$ をそれぞれ，$\psi_{xyz}^{[l]}$ の ψ_x，階層数，k 階層目の ψ_x の空間解像度とする．OH-SSL は，音源探索を以下のように階層化する．

C1) 伝達関数を $d_{x[k]}$ 間隔となるように選ぶ．必要なら FTDLI により伝達関数を補間する．

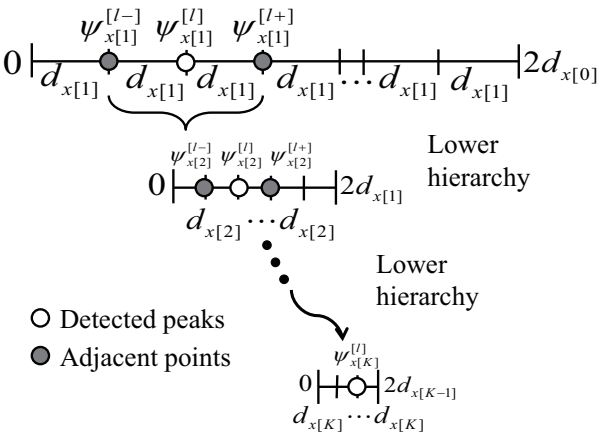


Figure 3: Hierarchical Structure of OH-SSL

C2) $\bar{P}(\psi_{xyz}, f)$ により定位する．ここで，$\psi_{x[k]}^{[l]}$ を k 階層目の $\psi_x^{[l]}$ とする．

C3) $\psi_{x[k]}^{[l]}$ から ψ_x 軸に沿って，その点を囲む二点を選ぶ．その二点を $\psi_{x[k]}^{[l-]}$ と $\psi_{x[k]}^{[l+]}$ とする ($\psi_{x[k]}^{[l-]} < \psi_{x[k]}^{[l]} < \psi_{x[k]}^{[l+]}$)．

C4) $\hat{A}_m(\omega, \psi_{x[k]}^{[l-]})$ と $\hat{A}_m(\omega, \psi_{x[k]}^{[l]})$ を使って，$\psi_{x[k]}^{[l-]}$ と $\psi_{x[k]}^{[l]}$ の間の伝達関数 $\hat{A}_m(\omega, \psi_x)$ を $d_{x[k+1]}$ の空間解像度で式 (10) を使って補間する．

C5) $\hat{A}_m(\omega, \psi_{x[k]}^{[l]})$ と $\hat{A}_m(\omega, \psi_{x[k]}^{[l+]})$ を使って，$\psi_{x[k]}^{[l]}$ と $\psi_{x[k]}^{[l+]}$ の間の伝達関数 $\hat{A}_m(\omega, \psi_x)$ を $d_{x[k+1]}$ の空間解像度で式 (10) を使って補間する．

C6) C4) と C5) で生成された伝達関数を用いて，式 (3) により定位する．

C7) C3)-C6) を k が K になるまで繰り返す．

階層化によって探索の粒度を制御できるため，定位性能を維持しつつ，計算コストを削減できる．

4.2 OH-SSL による探索コストの最小化

Figure 3 に OH-SSL の階層構造を示す．本章では図中の探索点数を最小化する最適な K と $d_{x[k]}$ について述べる．図より，k 階層目の探索点数 $g(k)$ は $g(k) = \frac{2d_{x[k-1]}}{d_{x[k]}}$ と求まる．従って，階層数を K とした時の全探索点数は以下となる．

$$G(K) = \sum_{k=1}^{K} g(k) = \sum_{k=1}^{K} \frac{2d_{x[k-1]}}{d_{x[k]}} \tag{13}$$

$G(K)$ を最小化する粒度を求める．$K=2$ の場合，$d_{x[1]}$ のみが変数となるため，$G(2)$ は $\frac{\partial G(2)}{\partial d_{x[1]}} = 0 \Rightarrow d_{x[1]} = \sqrt{d_{x[0]} d_{x[2]}}$ の時に最小となる．この時，$g(1) = g(2) = \sqrt{d_{x[0]}/d_{x[2]}}$ となるため，各階層の粒度が等しい時に $G(K)$ が最小となる．$K > 2$ も同様に，各階層の粒度が等しい時に $G(K)$ が最小となる．$g(1) = g(2) = \cdots = g(K)$ の条件下で $G(K)$ は以下のように求まる．

$$\tilde{G}(K) = K \left(\frac{d_{x[0]}}{d_{x[K]}} \right)^{\frac{1}{K}} \tag{14}$$

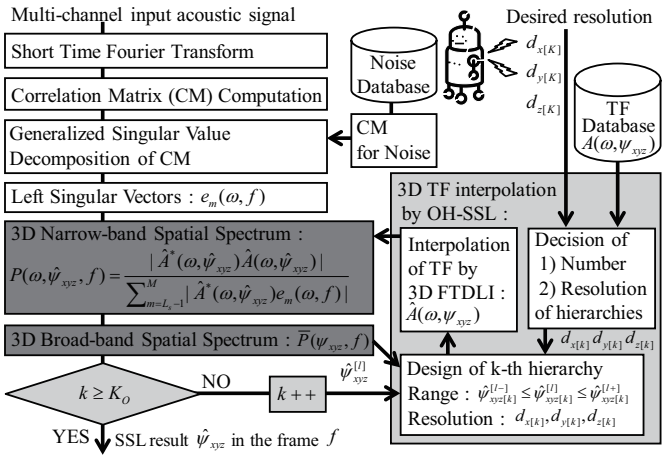


Figure 4: Block Diagram of Noise-robust Real-time 3D Super-resolution SSL

この時，$G(K)$ は最小となる．また，$\tilde{G}(K)$ を最小化する K は，$\frac{\partial \tilde{G}(K)}{\partial K} = \left(\frac{d_{x[0]}}{d_{x[K]}} \right)^{\frac{1}{K}} \left(1 - \frac{1}{K} \log \left(\frac{d_{x[0]}}{d_{x[K]}} \right) \right) = 0$ で以下のように求まる．

$$K = \log \left(\frac{d_{x[0]}}{d_{x[K]}} \right) \tag{15}$$

最後に，各階層の粒度 $d_{x[k]}$ は以下となる．

$$d_{x[k]} = d_{x[0]}^{\frac{K-k}{K}} d_{x[K]}^{\frac{k}{K}} \tag{16}$$

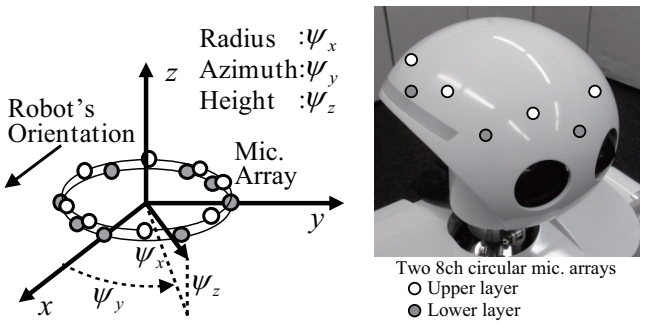
OH-SSL では式 (15)-(16) の K と $d_{x[k]}$ を音源探索の階層化に使用する．以降，式 (15) の K を K_O と表記する．

5 システム構成

Figure 4 に GSVD-MUSIC, FTDLI, OH-SSL を適用した三次元音源定位システムの処理フローを示す．評価ではこのシステムをオープンソースのロボット聴覚ソフトウェア HARK[19] 上に実装し，2.0GHz の Intel Core i7 の CPU を持つ計算機で実時間動作することを確認した．本稿では，マイクロホンアレイを搭載したロボットを残響時間が 0.2 秒 (RT_{20}) の 7m×4m の部屋の中央に，x 軸方向を向くように設置した．マイクロホンアレイは Figure 5(b) のように 8 チャネルの円状アレイが二つ上下に重なった形状の 16 チャネルのものを用いた．$\boldsymbol{A}(\omega, \psi_{xyz})$ は円筒座標上に事前に計測した（距離 1m，方位角 5° 毎，高さ -0.4m〜0.4m を 0.2m 毎に計測）．Figure 5(a) のように ψ_x, ψ_y, ψ_z の各軸はそれぞれ，円筒座標系の半径，方位角，高さとした．入力音響信号は 16kHz, 16 ビットでサンプリングした．音響信号処理のフレーム長とシフト長はそれぞれ，512, 160 サンプルとした．

6 評価実験

本章では，以下の実験を行い，各機能の評価を行う．



(a) Coordinates for SSL (b) Microphone Array

Figure 5: Conditions for Experiment

D1) GSVD-MUSIC の雑音ロバスト性評価

D2) FTDLI の伝達関数の補間性能評価

D3) OH-SSL の音源探索コスト評価

D4) ロボットへの応用と音源定位性能評価

D1) では，SEVD-MUSIC[2], GEVD-MUSIC[14], GSVD-MUSIC の 3 手法を比較し，提案手法の雑音ロバスト性を評価する．D2) では，FDLI, TDLI, FTDLI の 3 手法の補間誤差を比較し，統合の有効性を確認する．D3) では，OH-SSL といくつかの既存手法の音源探索回数を比較し，提案手法の有効性を検証する．最後に，D4) において，Figure 4 の音源定位システムを実際のロボットに適用する．評価の簡略化のため，D2) と D4) の評価では二次元空間（方位角と仰角）を，D1) と D3) の評価では一次元空間（方位角）を対象とした．

6.1 GSVD-MUSIC の雑音ロバスト性評価

実験には，60° に目的音（白色雑音）を，180° に雑音（定常状態でのロボットファン雑音）を配置した．ロボットと各音源との距離は 1[m] とした．また，事前情報のステアリングベクトル $\boldsymbol{A}(\omega, \psi_{xyz})$ の方位角は 5° ごとに計測し，使用した．式 (1) の相関行列の平均化のためのフレーム数はそれぞれ，$T_R = 25$ とした．

評価指標には，信号対雑音比 (Signal-to-Noise Ratio, SNR) 及び，定位正解率 (SSL Correct Rate) を用いた．SNR は以下のように定義した．

E1) M 個のマイクロホンの平均入力音響信号のスペクトル $X_{s_a}(\omega)$ の，パワースペクトル密度 (PSD) $P_{s_a}(\omega)$ を求める（k_{wl} は窓長）．$P_{s_a}(\omega) = \frac{1}{k_{wl}} X_{s_a}(\omega) X^*_{s_a}(\omega)$.

E2) 雑音についても同様に PSD を求める $P_{n_a}(\omega)$.

E3) 音源定位で使用する周波数帯に相当する周波数ビンを用いて平均 SNR を求める．

$$\mathrm{SNR} = 10\log_{10}\left(\frac{1}{j_h - j_l + 1}\sum_{j=j_l}^{j_h}\frac{P_{s_a}(\omega_{[j]})}{P_{n_a}(\omega_{[j]})}\right) \qquad (17)$$

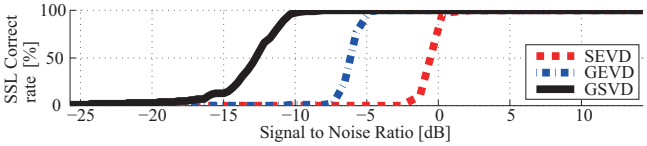


Figure 6: SSL correct rate of SEVD-, GEVD-, and GSVD-MUSIC

定位で用いた周波数帯は 500[Hz]〜2800[Hz] とした．

また定位正解率は，各フレームにおいて，空間スペクトル（式 (4)）が最大となる方向を音源数個取得し，その方向が目的音の方向から 10° 以内に入っているかを判定し，100 フレーム中で正解と判定されたフレーム数の割合として定義した．

Figure 6 に定位性能の比較結果を示す．図の横軸は SNR を，縦軸は定位正解率を表す．図中の実線は GSVD-MUSIC を，鎖線は GEVD-MUSIC を，点線は SEVD-MUSIC の結果を表す．図より，GSVD-MUSIC が GEVD-MUSIC や SEVD-MUSIC に比べ，低い SNR で高い定位精度を維持（SEVD-MUSIC に比べて約 10dB，GEVD-MUSIC に比べて約 5dB の向上）していることから，GSVD-MUSIC の高い雑音ロバスト性が確認できた．

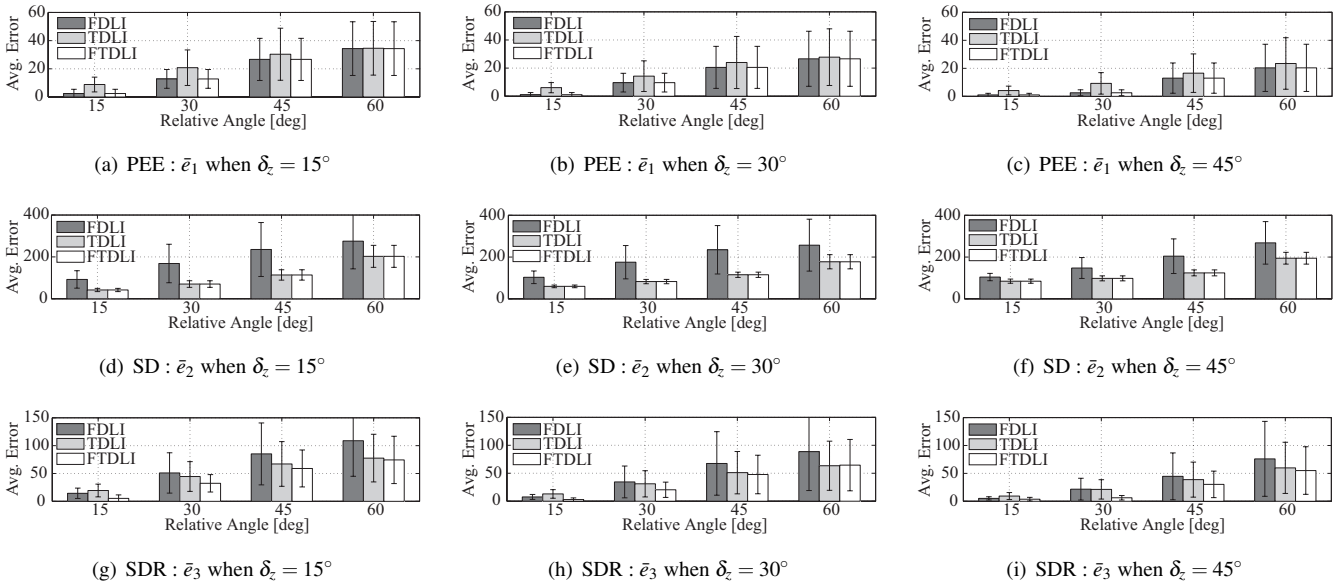
6.2 FTDLI の伝達関数の補間性能評価

FDLI, TDLI, FTDLI による三次元伝達関数の補間誤差を評価し，統合の有効性を確認した．本稿では，FTDLI の汎用性を確認するため，板倉らの伝達関数データベース [18]でまず評価し，次にロボットの事前計測伝達関数を用いて評価を行う．

板倉らの伝達関数データベースは両耳音響信号処理のための頭部伝達関数として知られており，球状の計測器によって録音されている．補間性能の評価のため，三次元座標を球座標（ψ_x, ψ_y, ψ_z をそれぞれ，半径，方位角，仰角とする．）で定義する．実験では，ψ_x=1.2m に固定し，ψ_y と ψ_z を変化させたときの $\hat{\boldsymbol{A}}(\omega, \psi_{xyz})$ と $\boldsymbol{A}(\omega, \psi_{xyz})$ の誤差を評価した．補間に使用する 4 つの事前計測伝達関数の位置 ($\psi_{\underline{xyz}}$, $\psi_{\underline{xy}\bar{z}}$, $\psi_{\underline{x}\bar{y}\underline{z}}$, $\psi_{\underline{x}\bar{y}\bar{z}}$) は，それらがなす中点が $[\psi_x, \psi_y, \psi_z] = [1.2\mathrm{m}, 0°, 0°]$ となるように取った．すなわち，4 点は，$[\psi_x, \psi_y, \psi_z] = [1.2\mathrm{m}, \pm\delta_y, \pm\delta_z]$ と表せる．ただし，δ_y と δ_z はそれぞれ，事前計測伝達関数の方位角と仰角である．評価では，$\delta_y = \{15°, 30°, 45°, 60°\}$ と，$\delta_z = \{15°, 30°, 45°\}$ を用いた．$\hat{\boldsymbol{A}}(\omega, \psi_{xyz})$ を ψ_y と ψ_z について 5° ごとに推定し，計測して得た 5° 毎の伝達関数との誤差を次式で表される加算平均により求めた．

$$\bar{e} = \frac{1}{i_\psi}\sum_{i=1}^{i_\psi}\frac{1}{j_h - j_l + 1}\sum_{j=j_l}^{j_h} f(\omega_{[j]}, \psi_{xyz[i]}) \qquad (18)$$

ここで，$f(\omega_{[j]}, \psi_{xyz[i]})$ は，補間点 $\psi_{xyz[i]}$ での $\omega_{[j]}$ に相当する周波数ビンの補間誤差である．i_ψ は評価に用いた補間

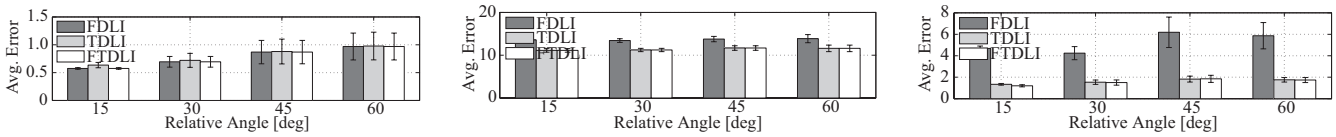


(a) PEE : $\bar{e}_1$ when $\delta_z = 15°$ (b) PEE : $\bar{e}_1$ when $\delta_z = 30°$ (c) PEE : $\bar{e}_1$ when $\delta_z = 45°$

(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 $\delta_z = 45°$

Figure 7: Interpolation error of PEE, SD, SDR using Itakura's TF Database[18]



(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

Figure 8: Interpolation error of PEE, SD, SDR using TFs of a robot-embedded microphone array

点個数であり，$-\delta_y < \psi_y < \delta_y$ と $-\delta_z < \psi_z < \delta_z$ の範囲内の事前計測伝達関数 $\boldsymbol{A}(\omega, \psi_{xyz})$ の個数で決まる．

式 (18) の $f(\omega_{[j]}, \psi_{xyz[i]})$ の誤差指標として，位相推定誤差 (PEE)，スペクトル歪み (SD)，信号対歪み比 (SDR) を用いた．PEE は以下で表される位相誤差指標である．

$$f_1(\omega, \psi_{xyz}) = \sum_{m=1}^{M} \left| \frac{A_m(\omega, \psi_{xyz}) \cdot \hat{A}_m(\omega, \psi_{xyz})}{|A_m(\omega, \psi_{xyz})||\hat{A}_m(\omega, \psi_{xyz})|} - 1 \right| \quad (19)$$

SD は以下で表され，振幅誤差を示す．

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

SDR は以下で表され，伝達関数自体の誤差を示す．

$$f_3(\omega, \psi_{xyz}) = \sum_{m=1}^{M} \frac{|A_m(\omega, \psi_{xyz}) - \hat{A}_m(\omega, \psi_{xyz})|^2}{|A_m(\omega, \psi_{xyz})|^2} \quad (21)$$

以下，$\bar{e}_1, \bar{e}_2, \bar{e}_3$ をそれぞれ，PEE, SD, SDR を用いた時の $\bar{e}$ とする．

Figure 7 に δ_z を変化させた時の $\bar{e}_1, \bar{e}_2, \bar{e}_3$ を示す．図の横軸は δ_y を表す．図より，FDLI が振幅誤差にあたる SD において，TDLI が位相誤差にあたる PEE において，それぞれ補間性能が劣化していることがわかる．FTDLI は FDLI と TDLI を統合することにより，$\bar{e}_1, \bar{e}_2, \bar{e}_3$ の全てにおいて最小の誤差となった．

Table 1: Comparison of computational cost

Condition		$G(K)$				
$d_{x[0]}$	$d_{x[K]}$	H1	H2	H3	SG	OS
360	10.0	36	11	9	11	8
360	1.0	360	36	21	17	12
360	0.1	3600	120	45	25	16
360	0.01	36000	1200	213	37	26

同様にロボットの伝達関数についても評価を行った．板倉らのデータベースと異なり，事前計測伝達関数を円筒座標系で計測したため，ψ_z を音源高さとして定義した．Figure 8 に，δ_z=0.2m に固定した時の，$\bar{e}_1, \bar{e}_2, \bar{e}_3$ の比較を示す．図より，FTDLI が全ての誤差指標において最小の誤差となり，統合の有効性を確認できた．

6.3 OH-SSL の音源探索コスト評価

計算コスト評価のため，OH-SSL と既存手法による総探索数 $G(K)$ を比較した．2 章より，$d_{x[0]}$ と $d_{x[K]}$ を変化させた時の $G(K)$ を計算した．評価対象として，以下の 5 つの手法を比較した．

H1) 階層化を行わない手法

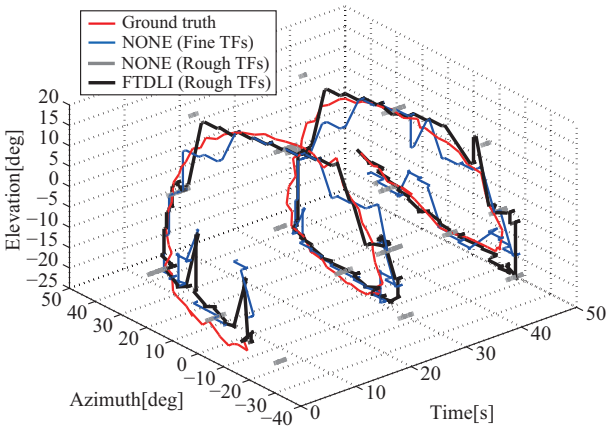


Figure 9: Trajectory of the 3D Localization

H2) 2 階層（各階層の粒度は式 (16) で決定）

H3) 3 階層（各階層の粒度は式 (16) で決定）

SG) Spherical grid 法[10]

OS) OH-SSL（K と $d_{x[k]}$ を式 (15)-(16) で決定）

Table 1 に求まった $G(K)$ を示す．表より，全ての所望の解像度に対して OS が $G(K)$ を最小化していることがわかり，有効性を確認できた．

実際の音源定位計算コスト評価のため，1000 フレーム分の処理を行い，音源定位の計算にかかる平均処理時間を算出した．所望の空間解像度を 1° に設定し，OH-SSL の有無 (OS と H1) による平均処理時間を比較したところ，それぞれ，5.2[ms] と 23.4[ms] となり，78%の計算コストを削減できた．また，フレーム周期である 10ms 以下で処理が実現できていることから，フレーム毎実時間処理を達成できた．

6.4 ロボットへの音源定位の応用と定位性能評価

FTDLI の有無による動的音源に対する音源定位の方向推定誤差を評価する．本稿では，1 つの白色雑音源を動かして動的音源とした．評価対象として，以下の 3 条件を比較した．

F1) 5° ごとの方位角，0.2m ごとの高さの事前計測伝達関数を用いた音源定位（高い空間解像度）

F2) 30° ごとの方位角，0.4m ごとの高さの事前計測伝達関数を用いた音源定位（低い空間解像度）

F3) F2) を用いて，5° ごとの方位角，0.01m ごとの高さの伝達関数を FTDLI を用いて推定した時の音源定位

また，リファレンスデータを得るため，超音波タグ位置測位システムを別途用いた（計測精度：20〜80mm）．Figure 9 に比較結果を示す．図の x 軸，y 軸，z 軸はそれぞれ，時間，推定された音源方位角 ψ_y，推定された音源仰角 ψ_z を示す．図中の赤線，青線，灰色線，黒線はそれぞれ，リファレンスデータ，F1 の結果，F2 の結果，F3 の結果を示す．F2(灰色線) の結果より，事前計測伝達関数の空間解像度が粗い場合，音源軌道が切断されているのが確認できる．また，F2 はリファレンスデータから離れた位置の音源として推定されており，推定結果とリファレンスデータとの平均方向誤差 $\bar{e}_\psi$ は $\bar{e}_\psi = 10.5°$ となった．F1(青線) は連続した軌道が確認できるものの，軌道にばらつきがあり，$\bar{e}_\psi = 7.2°$ となった．これは事前計測伝達関数の ψ_z の解像度が十分でなかったためと考えられる．F3(黒線) は，FTDLI により ψ_z 軸上に高い解像度の伝達関数を生成できるため，$\bar{e}_\psi = 6.5°$ となり，F1 に比べてばらつきの少ない軌道を実現できた．このことから，FTDLI の三次元音源定位での有効性を実環境で確認することができた．

また，OH-SSL の有効性を検証するため，OH-SSL の有無による音源定位の平均処理時間を比較した．所望の解像度を ψ_y 軸上で 5°，ψ_z 軸上で 0.01m に設定し，1000 フレーム分の定位を行い，平均処理時間を算出した．結果，OH-SSL の有無による平均処理時間はそれぞれ，0.028 秒，1.073 秒となった．これは，およそ 97%の処理時間の削減に相当し，OH-SSL の有効性を確認できた．

7 結論

本稿では，ロボットによる三次元音源定位について以下の問題点を扱った．

1) 実環境下のロボットが持つ自己雑音等の大きなパワーを持つ雑音に対するロバスト性

2) 音源定位の三次元空間での十分な空間解像度

3) 三次元空間内の音源探索の実時間性

1) に対して GSVD-MUSIC による音源定位における雑音の白色化を，2) に対してトリリニア補間を用いた FTDLI による三次元伝達関数の空間解像度の向上を，3) に対して OH-SSL による音源探索コストの軽減を提案した．評価では，1) GSVD-MUSIC が，SEVD-MUSIC に比べて約 10dB，GEVD-MUSIC に比べて約 5dB の SNR を向上すること，2) FTDLI が既存の補間手法に比べて誤差の小さな推定を実現すること，3) OH-SSL が音源探索数を約 78%軽減することを確認できた．最後に，これらの手法を実環境下のロボットに適用し，FTDLI による超解像音源定位と OH-SSL による実時間音源定位を確認でき，提案法の有効性を確認した．

今後の課題として，音源定位から得られる三次元位置情報を，画像から得られる定位情報と統合することによる定位のロバスト性の向上などが考えられる．

参考文献

[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 *Proc. of IEEE/RSJ IROS*, pp. 2009–2014, 2007.

[6] K. Nakamura *et al.*, “Real-time Super-resolution Sound Source Localization for Robots,” in *Proc. of 2012 IEEE/RSJ IROS*, accepted.

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

[8] H. Nakashima *et al.*, “A Localization Method for Multiple Sound Sources by Using Coherence Function”, in *Proc. of 18th EUSIPCO*, pp. 130–134, 2010.

[9] B. Rudzyn *et al.*, “Real time robot audition system incorporating both 3D sound source localization and voice characterization”, in *Proc. of IEEE ICRA*, pp. 4733–4738, 2007.

[10] J.-M. Valin *et al.*, “Localization of simultaneous moving sound sources for mobile robot using a frequency- domain steered beamformer approach”, in *Proc. of IEEE ICRA*, vol. 1, pp. 1033–1038, 2004.

[11] J.-S. Hu *et al.*, “Simultaneous localization of mobile robot and multiple sound sources using microphone array”, in *Proc. of IEEE ICRA*, pp. 29–34, 2009.

[12] C. T. Ishi *et al.*, “Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments”, in *Proc. of IEEE/RSJ IROS*, pp. 2027–2032, 2009.

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

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

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

[16] 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.

[17] W. G. Gardner and K. D. Martin, “HRTF measurements of a KEMAR”, *J. Acoust. Soc. Am.*, vol. 97, pp. 3907–3908, 1995.

[18] T. Nishino *et al.*, ”Interpolating head related transfer functions,” in *Proc. of 7th WESTPRAC VII*, 1A-1-3, pp.293-296, 2000.

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