複数移動ロボットを用いた音源分離における音源配置に応じたロボットの最適配置探索

○関口 航平　坂東 宜昭　糸山 克寿　吉井 和佳
京都大学 大学院情報学研究科

1. はじめに

近年の通信技術の発達により様々な方法で遠隔地とのコミュニケーションが可能となった．その一つがテレプレゼンスロボットである．テレプレゼンスロボットとは，移動機構にディスプレイやマイクロホンを搭載して，遠隔地にいる操縦者がまるで現地にいるかのように感じさせることができるロボットである．例えば，在宅勤務者が自宅から社内の人とコミュニケーションをとるなどの目的でテレプレゼンスロボットが使用され始めている [1, 2]．

遠隔地とのコミュニケーションにおいて，目的音以外の雑音への対策が不可欠となる．雑音が存在する場合，ロボットで録音される音は複数の音を含む混合音となり，操縦者は目的音の認識が困難になる．このような状況に対処するため，混合音をマイクロホンアレイ処理を用いて各音源に分離する研究が行われている．水本らは音源分離を用いて，操縦者が指定した方向の音だけを聞くテレプレゼンスロボットを開発した [3]．

音源分離にはマイクロホンと音源の位置関係が重要となり，音源の位置によっては分離が困難となる問題がある．例えば，ロボットと複数の音源が一直線に並んだ場合などである．注目音源が一つならば音源に近づくことで聞きやすくなるが，注目音源が複数存在する場合には，この方法では他の音源が聞きにくくなってしまう可能性がある．

本研究では，複数の子機ロボットによる音源分離支援システムの開発を行う (図 1)．操縦者に聞きたい音源を指定させ，その音源の配置に応じてマイクロホンアレイを搭載した子機ロボットを適切な位置に移動させることで，分離精度を向上させる．このとき，複数のロボットに搭載されたマイクロホンアレイ全体を一つの大きなマイクロホンアレイとみなし，すべてのマイクロホンでの観測音を用いて音源分離を行う．子機ロボットの適切な位置は自明ではないため，ロボットの配置をどのように決定するのかが問題となる．本研究では各ロボット配置での音源分離精度を予測することによって最適配置を決定する．

2. 音源分離に最適なロボット配置の探索

本稿では音源が複数存在する環境において，マイクロホンアレイを搭載した複数のロボットで録音した混合音から，注目音源を精度良く分離することを目的とする．ここで問題となるのは，音源分離に最適な複数ロボットの配置は自明でないことである．また，ある配置で実際に音を録音して音源分離を行っても，元信号がないため分離音から分離精度を計算できず，最適配置を探索することができない．したがって，複数ロボットの最適配置探索には，実際に音源分離を行わず各ロボット配置での音源分離精度を予測することが必要となる．本研究では，音源分離に幾何制約付きブラインド音源分離手法の一つである GICA[4] を使用する．この手法は分離性能や環境適応性が高く計算量も少ないため，実時間での動作が望まれるロボット聴覚に適した手法である．一方，この手法では音源分離精度の予測が困難であるという問題がある．そこで，GICA と分離精度について相関のある遅延和ビームフォーミング（DSBF）の利得を用いて音源分離精度の推定を行う．利得とは分離音中に含まれる目的音と雑音の比率であり，音の混合過程と分離過程を推定することによって求めることができる．利得を用いた評価関数により遺伝的アルゴリズムで最適配置を決定する．

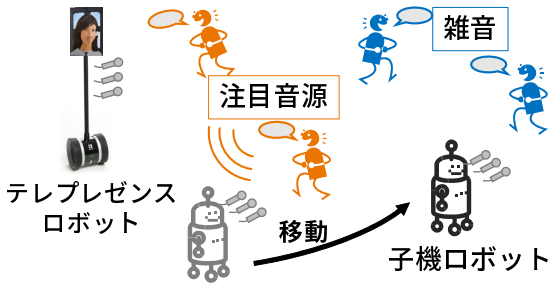


図 1　複数ロボットの配置最適化の一例

本稿で扱う配置最適化問題を以下のように定める．

入力 $\boldsymbol{X}_t = [\boldsymbol{x}_{t1}, \cdots, \boldsymbol{x}_{tM}]^T \in \mathbb{C}^{M\times F}$
N 個の音源が混合した M チャンネル観測音

出力 (1) $\boldsymbol{Y}_t = [\boldsymbol{y}_{t1}, \cdots, \boldsymbol{y}_{tN'}]^T \in \mathbb{C}^{N'\times F}$
注目している N' 個の音源の分離音
(2) $\boldsymbol{B}^* = [\boldsymbol{b}_1^*, \cdots, \boldsymbol{b}_R^*] \in \mathbb{R}^{R\times 2}$
R 台のロボットの最適配置の座標

仮定 (1) 各マイクロホンはすべて同期済み
(2) N 個の音源座標 $\boldsymbol{C} = [\boldsymbol{c}_1, \cdots, \boldsymbol{c}_N] \in \mathbb{R}^{N\times 2}$
は音源定位と三角測量により既知 [5]

ここで，音源の総数を N とし，そのうち注目する音源の数を N' と定める．$\boldsymbol{X}_t, \boldsymbol{Y}_t$ はそれぞれ，録音した音響信号，分離音の t フレーム目を短時間フーリエ変換して得る．F は周波数ビンの数を表し，$\boldsymbol{x}_{tm} = [x_{tm1}, \cdots, x_{tmF}]$，$\boldsymbol{y}_{tm} = [y_{tm1}, \cdots, y_{tmF}]$ である．

マイクロホンアレイの配置最適化の関連研究には佐々木らの手法 [6] と Martinson らの手法 [7] がある．前者はマイクロホンアレイを搭載した 1 台のロボットを用いる．DSBF の利得を用いて，すべての方向に対して高い分離精度をもつ，音源配置によらないマイクロホ

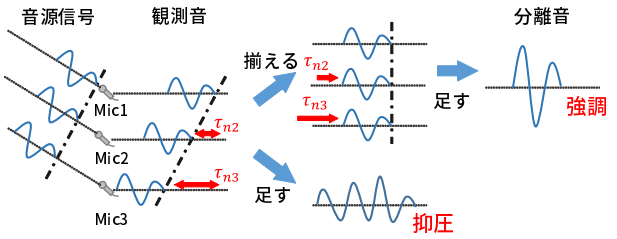


図 2　DSBF の概要

ンアレイの最適配置を探索する．後者は 1 チャンネルマイクロホンを搭載した複数のロボットを用いる．幾何制約によって音源定位に最適な配置を探索する．

2.1　音の混合過程

音源の信号 $\boldsymbol{S}_t = [\boldsymbol{s}_{t1}, \cdots, \boldsymbol{s}_{tN}] \in \mathbb{C}^{N\times F}$ とマイクロホンによる観測音 $\boldsymbol{X}_t$ の関係について説明する．ここで，$\boldsymbol{s}_{ti}$ は音源 i の t フレーム目の音源信号の短時間フーリエ変換を表す．音の伝搬を線形時不変システムと仮定すると，音源信号と観測音の関係は以下のように表される．

$$\boldsymbol{x}_{t\cdot f} = \boldsymbol{H}_f \boldsymbol{s}_{t\cdot f} \tag{1}$$

ここで，$\boldsymbol{x}_{t\cdot f} = [x_{t1f}, \cdots, x_{tMf}]^T \in \mathbb{C}^M$，$\boldsymbol{s}_{t\cdot f} = [s_{t1f}, \cdots, s_{tNf}]^T \in \mathbb{C}^N$ であり，$\boldsymbol{H}_f \in \mathbb{C}^{M\times N}$ は混合行列である．雑音と残響を考慮せず，音の距離減衰と到達時間差のみを考慮した場合，x_{tmf} と s_{tnf} の関係は次のように表される．

$$x_{tmf} = \sum_{n=1}^{N} \frac{1}{d_{nm}} s_{tnf} \exp(-j2\pi f \tau_{nm}) \tag{2}$$

ここで，d_{nm} は音源 n とマイクロホン m の間の距離を表し，τ_{nm} は音源 n のマイクロホン m への到達時間を表し，$\tau_{nm} = d_{nm}/c$（c は音速）で計算される．音の振幅は距離に反比例するため，$1/d_{nm}$ の項は距離減衰を表す．式 (1) と式 (2) を比較すると，混合行列 $\boldsymbol{H}_f$ の (m,n) 成分 h_{mnf} は以下のように表される．

$$h_{mnf} = \frac{1}{d_{nm}} \exp(-j2\pi f \tau_{nm}) \tag{3}$$

2.2　音源分離

マイクロホンでの観測音 $\boldsymbol{x}(t)$ と分離音 $\boldsymbol{y}(t)$ の関係について説明する．音の混合過程と同様に，音源分離が線形時不変システムで表されると仮定すると，観測音と分離音の関係は以下の式で表される．

$$\boldsymbol{y}_{t\cdot f} = \boldsymbol{W}_f \boldsymbol{x}_{t\cdot f} \tag{4}$$

ここで，$\boldsymbol{y}_{t\cdot f} = [y_{t1f}, \cdots, y_{tNf}]^T \in \mathbb{C}^N$ であり，，$\boldsymbol{W}_f \in \mathbb{C}^{N\times M}$ は分離行列を表す．式 (1) と式 (4) から，$\boldsymbol{W}_f = \boldsymbol{H}_f^{-1}$ のとき，$\boldsymbol{y}_{t\cdot f} = \boldsymbol{W}_f \boldsymbol{x}_{t\cdot f} = \boldsymbol{W}_f \boldsymbol{H}_f \boldsymbol{S}_{t\cdot f} = \boldsymbol{S}_{t\cdot f}$ となり，分離音は音源信号と等しくなる．

GICA は ICA を基にした手法であり，音源信号の独立性を過程して，分離音が独立となるような分離行列 $\boldsymbol{W}$ を推定する．分離行列 $\boldsymbol{W}$ を推定するために，以下の二つのコスト関数を用いる．

$$\boldsymbol{J}_{\mathrm{ICA}}(\boldsymbol{W}) = \int p(\boldsymbol{y}) \log \frac{p(\boldsymbol{y})}{\prod_k p(y_k)} \tag{5}$$

$$\boldsymbol{J}_{\mathrm{GC}}(\boldsymbol{W}) = \|\boldsymbol{W}\boldsymbol{H} - \boldsymbol{I}\|^2 \tag{6}$$

ただし，$p(\boldsymbol{y}) = p(y_1, \cdots, y_N)$ である．$\boldsymbol{J}_{\mathrm{ICA}}(\boldsymbol{W})$ は $p(\boldsymbol{y})$ と $\prod_k p(y_k)$ の KL-divergence であり，独立性の尺度となっている．$\boldsymbol{J}_{\mathrm{GC}}$ は幾何制約を表す．実際の環境での混合行列 $\boldsymbol{H}$ は未知であるため，ここで与える $\boldsymbol{H}$ はあらかじめ録音したインパルス応答や幾何的に計算したインパルス応答から作成 する．本研究ではリアルタイムで音源分離を行うために以下の更新式を用いて，逐次的に分離行列 $\boldsymbol{W}$ を推定する．

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \alpha \boldsymbol{J}'_{\mathrm{ICA}} - \beta \boldsymbol{J}'_{\mathrm{GC}} \tag{7}$$

ただし，α, β はステップサイズパラメータであり，$\boldsymbol{J}'_{\mathrm{ICA}} = \nabla_{\boldsymbol{W}^*} \boldsymbol{J}_{\mathrm{ICA}}$，$\boldsymbol{J}'_{\mathrm{GC}} = \nabla_{\boldsymbol{W}^*} \boldsymbol{J}_{\mathrm{GC}}$ である．$\{\}^*$ は複素共役を，∇ は微分作用素を表す．GICA では観測音によって分離行列 $\boldsymbol{W}$ が異なるため，事前に分離行列を推定することが困難である．したがって，利得を計算することができない．

本研究では遅延和ビームフォーミング (DSBF) という手法に注目する．DSBF と GICA の分離精度には高い正の相関がある．本研究では，この相関を利用して，GICA での分離精度の予測に DSBF の利得を用いる．音源分離手法として DSBF を用いた場合，分離行列 $\boldsymbol{W}_f$ の要素はマイクロホンと音源の位置関係から決定される．DSBF とは注目音源の座標から各マイクへの到達時間差を推定し，観測信号を到達時間差だけ時間シフトして足し合わせることにより注目音を強調する音源分離手法である (図 2)．本研究では，各マイクロホンと音源の距離を考慮し，音源に近いマイクロホンの観測音の比率を高く，音源と遠いマイクロホンの観測音の比率を低くして足し合わせる．したがって，分離音と観測音の関係は周波数領域では次のように表される．

$$y_{tnf} = \sum_m \frac{1}{d_{nm}} x_{tmf} \exp(j2\pi f \tau_{nm}) \tag{8}$$

式 (4) と式 (8) から，分離行列 $\boldsymbol{W}_f$ の (n,m) 成分 w_{nmf} は以下の式で表される．

$$w_{nmf} = \frac{1}{d_{nm}} \exp(j2\pi f \tau_{nm}) \tag{9}$$

2.3　目的関数

複数ロボットの配置最適化における目的関数を DSBF の利得の調和平均として定める．ロボット配置 $\boldsymbol{B} = [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_R]$ における目的関数の値を $f(\boldsymbol{B})$ とすると，

$$f(\boldsymbol{B}) = \frac{N'}{\sum_{n\in D} \frac{1}{g_n(\boldsymbol{B})}} \tag{10}$$

ここで，D は注目音源の集合を表す．$g_n(\boldsymbol{B})$ は音源 n の利得を表し，音源 n の分離音中の音源 n と雑音の比率として定める．利得の調和平均を目的関数としたの

は，本研究ではすべての注目音源の高精度な分離を目的とするためである．もし一つでも分離精度が悪い音源が存在する場合，目的関数の値は大きく低下する．

利得は分離音と音源信号の関係式から計算することが可能である．式 (1) と式 (4) から，分離音と音源信号の関係は周波数領域で以下のように表される．

$$\boldsymbol{y}_{t\cdot f} = \boldsymbol{A}_f \boldsymbol{s}_{t\cdot f} \tag{11}$$

ここで，$\boldsymbol{A}_f \in \mathbb{C}^{N\times N}$ は利得行列であり，$\boldsymbol{A}_f = \boldsymbol{W}_f \boldsymbol{H}_f$ として定める．利得行列 $\boldsymbol{A}_f$ の対角成分は分離音中に含まれる目的音源の比率を，非対角成分は雑音の比率を表している．したがって，音源 n の利得 $g_n(\boldsymbol{B})$ は以下のようになる．

$$g_n(\boldsymbol{B}) = \frac{\sum_f a_{nnf}}{\sum_{n\neq k}\sum_f a_{nkf}} \tag{12}$$

ここで，a_{nkf} は利得行列 $\boldsymbol{A}_f$ の (n,k) 成分であり，音源 n の分離音に含まれる音源 k の割合を示す．

DSBF の利得を用いた場合には，式 (8) と式 (2) から a_{nkf} は以下のようになる．

$$a_{nkf} = \left| \sum_{m=1}^{M} \frac{1}{d_{nm} d_{km}} \exp(j2\pi f(\tau_{nm} - \tau_{km})) \right| \tag{13}$$

2.4 最適配置探索

本研究では遺伝的アルゴリズムを用いて最適配置探索を行う．これは，グリッドサーチによる全探索を用いた場合，ロボットの台数に応じて指数的に計算量が増加してしまうためである．ロボットの座標と向きの組を個体とみなし，個体の組み替えは現在位置の近傍へ移動することで行い，突然変異によりランダムに移動することで局所最適解に陥ることを防ぐ．ロボットの向きはランダムに与える．ただし，ロボット間の距離が離れすぎた場合，１つの時間フレーム内に含まれる音源信号の区間がロボット間で大きく異なってしまい，分離精度が低下してしまう問題がある．そのため，ロボット間の距離が一定距離以内に収まるように制約を設ける．選択はエリート選択とルーレット選択を併用する．このとき，複数台のロボットに搭載した全てのマイクを１つのマイクロホンアレイとみなして目的関数の計算を行う．世代交代を一定回数行った後，評価関数の値が最大の個体を最適配置とする．

3. 評価実験

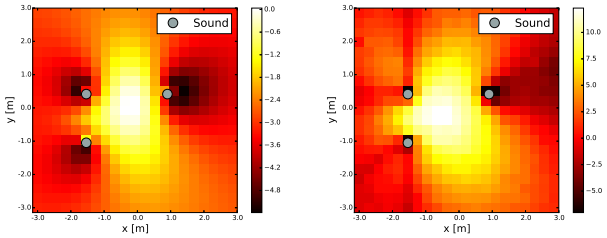
DSBF と GICA の分離精度の比較実験と，提案法による最適配置での分離精度の評価実験を行った．

3.1 DSBF と GICA の分離精度の評価実験

GICA の分離精度予測に DSBF の利得を用いることの妥当性を評価するために，実際に各手法で音源分離を行う実験を行った．

3.1.1 実験条件

１辺６ m の正方形の部屋に音源３つ，ロボット１台がある場合を想定する．３つの音源の配置を６種類用意し，各音源配置について８チャンネルマイクロホンアレイを搭載したロボットを，0.2m 間隔の格子の各点に配置する．各ロボット配置での観測音をシミュレーション混合を用いて作成し，DSBF と GICA を用いて音源分離を行う．音源信号は JNAS 音素バランス文を用いた [8]．音源分離精度の指標は signal-to-distortion ratio(SDR)[9, 10] を用いた．SDR は値が大きいほど分離精度が高いことを示す．２つの手法による分離音の分離精度に相関があるかを調べるために，相関係数を計算した．



(a) DSBF　　(b) GICA

図 3　音源３つの場合の各点での分離精度の一例

3.1.2 実験結果

図 3 は各点での分離精度を示している．明るい部分は分離精度が良く，暗い部分は分離精度が悪いことを示す．DSBF と GICA では分離精度は大きく異なるが，分離精度が高くなる位置は近いことがわかる．この場合の分離精度の相関係数は 0.85 であり，６種類の音源配置について同様に相関係数を計算したところ平均で 0.83 となった．この結果から DSBF と GICA の分離精度には高い正の相関があることが確かめられた．

3.2 提案法の評価実験

提案法による最適配置での分離精度を評価するために，シミュレーション混合を用いた評価実験を行った．

3.2.1 実験条件

１辺５ m の正方形の部屋に音源６つ，ロボット２台がある場合を想定する．６つの音源の配置を６種類用意し，各音源配置について６つの音源のうち３つを注目したい音源，残りの３つの音源を雑音とみなす．各ロボットは８チャンネルマイクロホンアレイを搭載し，１台のロボットを座標 $[0,-1]$ に固定して，もう１台の配置を提案法によって最適化する ($M=16$, $N=6$, $N'=3$, $R=2$)．最適配置での観測音を幾何学的に計算したインパルス応答を使ったシミュレーション混合を用いて作成し，GICA で音源分離を行って分離精度を計算する．音源信号は JNAS 音素バランス文を用いた．

この分離精度を，１台のロボットを提案法と同じ配置に固定し，もう１台のロボットをランダムに配置した場合，正解配置に配置した場合での分離精度と比較した．正解配置はグリッドサーチで各点について実際に分離精度を計算して，分離精度が最大となった配置とした．音源分離精度の指標は SDR の注目音源についての調和平均を用いた．各音源配置について提案法による最適化とランダムに配置する操作を３０回行い，SDR の調和平均の平均を計算した．

3.2.2 実験結果

図 4 に，各音源配置にてランダムにロボットを配置した場合の分離精度の平均値，提案法による最適配置

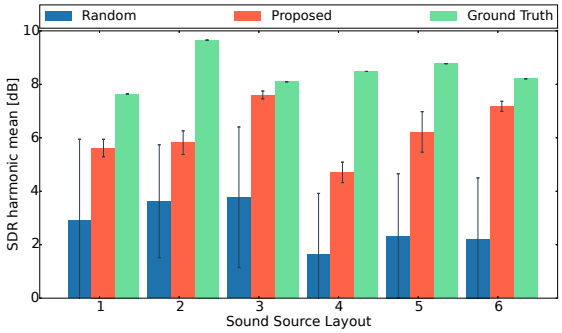


図 4　各音源配置での注目音源の分離精度

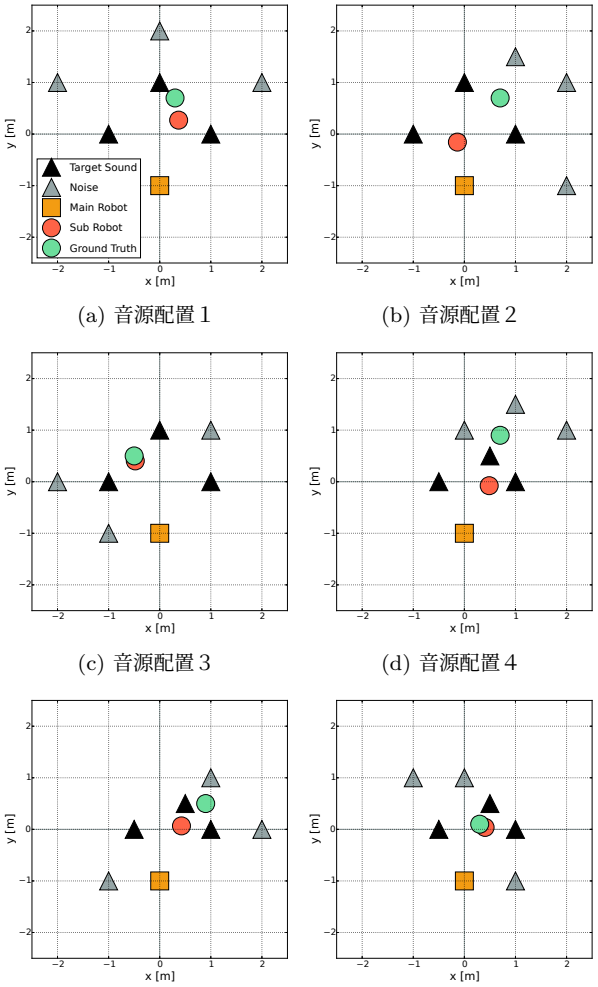
での分離精度の平均値，正解配置での分離精度を示す．全ての場合において提案法による最適配置での分離精度はランダムに配置した場合の分離精度を上回っており，平均で 3.4 dB 分離精度が向上した．図 5 は各音源配置での提案法による最適配置と正解配置の一例を示している．座標 $[0, -1]$ の黄色の四角は固定したロボットを示し，赤丸が子機ロボットの提案法による最適配置を表す．また，緑の丸が子機ロボットの正解配置，黒の三角が注目音源，灰色の三角が雑音を示す．

図 5(c) と図 5(f) は音源配置 3 と 6 の場合の例である．これらの例では正解配置と提案法による最適配置が近く，分離精度も正解配置での分離精度に近い値を得ることができている．一方，音源配置 1，2，4，5 では，ランダムに配置した場合より分離精度は向上しているが，正解配置と比較すると配置は異なり，分離精度も正解配置での分離精度を下回っている．これは DSBF と GICA の性質の違いによるものだと考えられる．音源配置 1，2，4，5 の例では，正解配置は提案法による最適配置と比較して雑音側に寄った配置となっている．これは，GICA では雑音を多く含んだ観測音を利用することで，分離音から雑音を取り除くことができ，雑音に近い位置でも高い分離精度を得ることができるためである．一方，提案法による最適配置は正解配置と比較して，雑音から離れた配置となっている．これは，DSBF では注目音源の位相に合わせて信号を時間シフトするだけで，雑音を多く含んだ観測があっても分離精度の向上に役立たないためである．GICA の特性を考慮した分離精度の推定が今後の課題である．

4.　まとめ

本稿では GICA を用いた音源分離に最適な複数ロボットの最適配置を探索する手法を開発した．GICA の利得は計算することが困難であるため，GICA と分離精度について相関のある DSBF の利得を用いて GICA の分離精度を予測した．シミュレーション混合音を用いた評価実験で，ランダムにロボットを配置した場合と比較して提案法により音源分離精度が平均 3.4 dB 向上することを確認した．今後は GICA の特性を考慮した分離精度推定手法の開発と，ロボットや音源数が変化した場合や，実環境での録音を用いた評価実験を行う予定である．

謝辞 本研究の一部は，科研費 24220006，および ImPACT「タフ・ロボティクス・チャレンジ」の支援を受けた．



(a) 音源配置 1　(b) 音源配置 2

(c) 音源配置 3　(d) 音源配置 4

(e) 音源配置 5　(f) 音源配置 6

図 5　各音源配置での提案法による最適配置と正解配置

参　考　文　献

[1] R. Berri et al. Telepresence robot with image-based face tracking and 3d perception with human gesture interface using kinect sensor. In *JCRIS*, pages 205–210, 2014.

[2] R. Yan et al. An attention-directed robot for social telepresence. In *HAI*, pages III–1–2, 2013.

[3] T. Mizumoto et al. Design and implementation of selectable sound source separation on the texai telepresence system using HARK. In *ICRA*, pages 2130–2137, 2011.

[4] H. Nakajima et al. Blind source separation with parameter-free adaptive step-size method for robot audition. *IEEE Trans. Audio, Speech and Language Processing*, 18(6):1476–1485, 2010.

[5] Y. Sasaki et al. Multiple sound source mapping for a mobile robot by self-motion triangulation. In *IROS*, pages 380–385, 2006.

[6] Y. Sasaki et al. 32-channel omni-directional microphone array design and implementation. *J. Robotics and Mechatronics*, 23(3):378–385, 2011.

[7] E. Martinson et al. Optimizing a reconfigurable robotic microphone array. In *IROS*, pages 125–130, 2011.

[8] Y. Sagisaka and N. Uratani. ATR spoken language database. *J. The Acoustic Society of Japan*, 48(12):878–882, 1992.

[9] E. Vincent et al. Performance measurement in blind audio source separation. *IEEE Trans. Audio, Speech and Language Processing*, 14(4):1462–1469, 2006.

[10] C. Raffel et al. mir_eval: A transparent implementation of common MIR metrics. In *ISMIR*, pages 367–372, 2014.