人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-061-11

任意の混合音を入力とした マイクロホンアレイ形状のキャリブレーション

Calibration of microphone array shape with arbitrary sound mixtures as input

糸山 克寿 [1,2*]　中臺 一博 [1]
Katsutoshi Itoyama[1,2]　Kazuhiro Nakadai[1]

[1] 東京工業大学
[1] Tokyo Institute of Technology
[2] （株）ホンダ・リサーチ・インスティチュート・ジャパン
[2] Honda Research Institute Japan Co., Ltd.

Abstract: 本稿では，マイクロホンアレイを構成する個々のマイクロホンの位置を観測した混合音からキャリブレーションするための手法について報告する．マイクロホンと音源の位置によって定まる音の伝達特性と音源スペクトルから混合音が観測される過程の確率的生成モデルに基づき，マイクロホン位置の事後確率の最大化によりキャリブレーションを実現する．マイクロホン数，マイクロホンアレイの形状，音源数などを様々に変化させながら提案手法によるキャリブレーションを行うシミュレーション実験により，提案手法が観測混合音に対してキャリブレーションが行えることを確認し，さらに提案手法の性質と限界について議論する．

1　はじめに

近年，マイクロホンアレイがロボットをはじめとする様々な機器に搭載されるようになってきた．また，音源定位や音源分離などの音響信号処理技術も開発・研究されている [1, 2, 3, 4]．特に，ロボット聴覚は，ロボットを含む実世界に配置可能なシステムの聴覚機能の開発を目指す研究者に注目されている [5]．

マイクロホンアレイの普及に伴い，その校正方法は非常に重要である．マイクロホンアレイをロボットに搭載する場合，デバイスの経年劣化，マイクロホン位置の測定誤差，ロボットの動作による環境変化などの理由により，あらかじめ最適に調整されたアレイのパラメータ値と最適値との間にミスマッチが生じる可能性があることが課題である．これは，図 1 に示すように，音源とマイクロホンアレイの間の伝達関数に誤差が生じてしまい，その結果として音源方向が正しく推定されない問題などを引き起こす．

使いやすいキャリブレーションには，主に 2 つの条件が必要である．1）環境音の中には複数の音源が収録されていることが多いので，同時に収録された音源信号でキャリブレーションを行うこと．2）任意の音源信号を用いて校正を行うため，特定の音を用意することなく校正が可能であること．しかし，これまでの研究の多くは，マイクロホンアレイ処理の登場以前に，手間のかかる校正方法を報告している [6, 7, 8, 9]．上記の条件を実現するためには，同時に収録された任意の音源信号で校正できる方法が必要であるが，まだ実現されていない．

そこで，本稿では，伝達関数のミスマッチ問題を回避するために，各マイクロホンの初期位置を中心とした事前分布を仮定した Maximum *A Posteriori*（MAP）推定に基づくマイクロホンアレイのマイクロホン位置の新規キャリブレーション方法を提案する．MAP 推定は，環境音としてのホワイトノイズを含む複数の同時音源を用いた校正や，環境音のような立ち上がりの悪い音源信号を用いた校正が可能である．

2　関連研究

マイクロホンアレイの伝達関数推定やマイクロホン位置推定は，非同期分散マイクロホンアレイやアドホックマイクロホンアレイなどの分野で取り組まれてきた．Thrun は，音のオンセットタイミングを利用したオンライン校正法 [10] を提案し，実際のマイクロホンデバイスを用いてその有効性を実証した．しかし，音源位

*連絡先：東京工業大学 工学院 システム制御系
〒 152-8552 東京都目黒区大岡山 2-12-1
E-mail: itoyama@ra.sc.e.titech.ac.jp

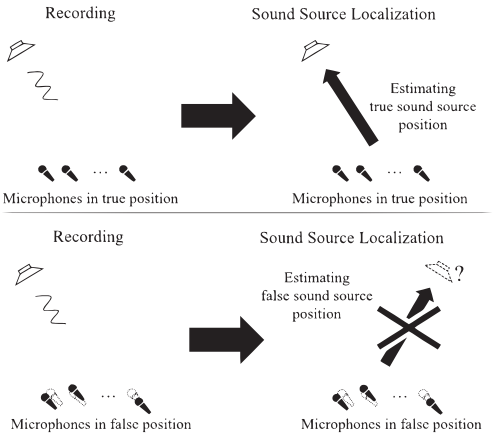

図 1: マイクロホンアレイを構成するマイクロホンの正しい位置が与えられると，それに基づく音源方向の推定結果は正しいものとなる（上段）．一方で，正しくないマイクロホンの位置が与えられると，それに基づく音源方向の推定結果も誤ったものになってしまう．

置があらかじめ決まっていること，マイクロホンが完全に同期していることなどの制約があった．三浦らは，Simultaneous Localization And Mapping (SLAM) に基づく非同期型オンラインマイク位置推定法を提案した [7]．SLAM におけるロボットの位置と地図を，それぞれ音源位置とマイクロホン位置に置き換え，さらに移動中に手拍子をすることで，8 個のマイクアレイのマイク位置を漸進的に推定することに成功した．また，アドホックマイクロアレイでは，到着時間差 (Time Difference of Arrival, TDOA) を用いた距離推定に基づく方法が提案されている [8, 9]．これらの手法では，音源信号が十分にスパースであることや，TSP (time stretched pulse) のような特定の音源信号を前提としていたため，任意の音響信号を用いられるわけではなく，その点で実用性に課題を残していた．

これらの研究では，音声信号が拍手音のようにスパースである，あるいはオンセットが明瞭であるという強い仮定を採用している．本論文ではこれらの仮定を回避するため，音源スペクトルのモデルと音源からマイクロホンまでの音の伝達過程を組み合わせた混合音スペクトルの観測モデルを構築し，そのモデルに基づいた最適化アルゴリズムを提案する．これにより，任意の音源信号の混合音を入力としたキャリブレーション手法を実現する．

3 問題設定

残響や背景雑音のない D 次元空間 (主に $D = 2$ を想定するが，$D = 3$ の状況でもほとんど同様に考えることができる) に M 個の同期されたマイクロホンと N 個の音源が存在するとする．各マイクロホンに $1, \ldots, m$ の番号を割り当て，m 番目のマイクロホンの位置をデカルト座標系で $\mathbf{u}_m \in \mathbb{R}^D$ で表す．これらのマイクロホンは，基準位置 $\bar{\mathbf{u}}_m$ にしたがって設置されているが，実際の位置 $\mathbf{u}_m$ は基準位置からのずれを含んでいる．これは，会議室などで各テーブルに一つずつマイクロホンが置かれており，実際のマイクロホンの位置はテーブル上の任意の位置であるため正確な位置は分からない，というような状況に相当する．これらのマイクロホンの位置を観測された混合音を用いて校正する（推定する）ことが本研究の目的となる．マイクロホンの位置は時不変であるとする．なお，状況を簡単にするため，1 番目のマイクロホンは座標系で原点にあるとする．

マイクロホンと同様に，各音源に $1, \ldots, N$ の番号を割り当てる．各音源はマイクロホン群から十分に遠い距離にある (far-field condition) とするため，マイクロホン群からみた方向のみを考え，音源の方向 $\mathbf{v}_n$ はマイクロホン群の中心点の近傍である座標系原点からみた音源の方向を表す大きさが 1 の D 次元の単位ベクトルで表す．$\mathbf{v}_n = (v_{m1}, \ldots, v_{mD})^\mathsf{T}$, $|\mathbf{v}_n| = 1$ 音源に関して以下の仮定をおく．

1. 音源の数は既知である．
2. 各音源の方向は既知であり，音源は移動しない．

3.1 混合音の観測モデル

音源 n からみたマイクロホン 1 とマイクロホン m の距離の差 d_{nm} は

$$d_{nm} = -\left(u_{m1}v_{n1} + \ldots + u_{mD}v_{nD}\right) \tag{1}$$

で表される．$\mathbf{u}_0 = (0, \ldots, 0)^\mathsf{T}$ なので，$d_{n0} = 0$ であることに留意する．音速を c とすると，音源 n から発せられた音がマイクロホン 1 とマイクロホン m に到達するまでの時間差は d_{nm}/c となる．音源 n から発せられた音が周波数 ω の正弦波であったとすると，2 つのマイクロホン間での位相差は $2\pi d_{nm}\omega/c$ となる．マイクロホンのサンプリング周波数を f_s，STFT フレーム長を F_0，周波数ビン数を $F = F_0/2 + 1$，周波数インデックスを $f = 0, \ldots, F$ とすると，f 番目の周波数 $\omega_f = f \cdot f_s/F_0$ となり，周波数 ω_f における位相差は $\psi_f d_{nm} = 2\pi d_{nm}\omega_f/c$ となり，周波数領域での伝達関数ベクトル $\mathbf{a}_{nf} \in \mathbb{C}^M$ は以下となる．

$$\mathbf{a}_{nf} = (\exp(i\psi_f d_{n1}), \ldots, \exp(i\psi_f d_{nM})) \tag{2}$$

n 番目の音源から発せられた音響信号のスペクトル s_{nft} が各マイクロホンで収録されたものは，

$$\mathbf{a}_{nf}s_{nft} \tag{3}$$

で表される．環境中には N 個の音源が存在するので，各マイクロホンでは全ての音源からの信号が足し合わされたものが観測される．

$$\mathbf{x}_{ft} = \sum_n \mathbf{a}_{nf} s_{nft} \tag{4}$$

ここで行列表記を導入し，

$$\mathbf{X}_f = \left(x_{mft}\right), \mathbf{S}_f = \left(s_{nft}\right), \mathbf{A}_f = \left(a_{nfm}\right) \tag{5}$$

とすると，式 (4) は

$$\mathbf{X}_f = \mathbf{A}_f \mathbf{S}_f \tag{6}$$

と書き表すことができる．

4　キャリブレーション手法

前節で導出した観測スペクトル $\mathbf{X}_f$ に基づいて各マイクロホンの位置を推定する手法について本節では述べる．音源スペクトルが確率的な生成モデルに従うと考えて，その確率モデルを定義し，観測スペクトルが従う分布を導出する．観測スペクトルに対するマイクロホン位置の尤度を導出し，この尤度を最大化することでマイクロホン位置を推定するアルゴリズムについて述べる．

4.1　確率モデル

音源スペクトルのモデルとして，複素ガウス分布を考える．

$$s_{nft} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_s^2) \tag{7}$$

$$p\left(s_{nft}\right) = \frac{1}{\sqrt{\pi}} \exp(-s_{nft}^* s_{nft}) \tag{8}$$

音源スペクトル s_{nft} がマイクロホン m で観測されたときのスペクトルは $a_{nfm} s_{nft}$ となるので，このスペクトルが従う分布は

$$a_{nfm} s_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(0, |a_{nfm}|^2 \sigma_s^2\right) \tag{9}$$

となり，全ての音源スペクトルが足し合わされた観測スペクトル x_{mft} が従う分布は

$$x_{mft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_n |a_{nfm}|^2 \sigma_s^2\right) \tag{10}$$

となる．すべてのマイクロホンの観測スペクトルをベクトル表現すると

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_n \mathbf{a}_{nf}^* \mathbf{a}_{nf} \sigma_s^2\right) \tag{11}$$

となる．

観測スペクトル $\mathbf{x}_{ft}$ は音源スペクトルの線形結合で表現できる．ここで，観測スペクトルに対するステアリングベクトルなどのパラメータの対数尤度を音源スペクトルの推定値に対する対数尤度で近似する．

$$\begin{aligned}\log p\left(\ldots, \mathbf{X}_f, \ldots\right) &\approx \log p\left(\ldots, \hat{\mathbf{S}}_f, \ldots\right) \\ &= \sum_{nft} \log p\left(\hat{s}_{nft}\right)\end{aligned} \tag{12}$$

$$\begin{aligned}&\log p\left(\hat{s}_{nft}\right) \\ &\overset{\mathrm{c}}{=} \Re\left(x_{m_1 ft}^* x_{m_2 ft} \exp\left(i\psi_f(d_{nm_1} - d_{nm_2})\right)\right)\end{aligned} \tag{13}$$

この対数尤度をマイクロホン位置 $\mathbf{u}_m$ に関して最大化することで，与えられた観測スペクトルに対して最も尤もらしいマイクロホンの配置を推定する，すなわち，マイクロホン位置のキャリブレーションを行うことができる．

4.2　最適化アルゴリズム

式 (12) の目的関数を勾配法を用いて最大化することでマイクロホン位置のキャリブレーションを行う．提案する最適化アルゴリズムは以下のように表すことができる．

1. マイクロホン位置の初期値を定める．
2. 前ステップの目的関数の値を $-\infty$ にセットする．
3. 反復ステップ数 t を 0 にセットする．
4. t が事前に定めた最大数 T に達するまで以下を繰り返す．
5. マイクロホン位置の現在値を用いて目的関数の値と勾配を計算する．
6. 目的関数の値の差分を計算し，その絶対値が事前に定めた閾値よりも小さければ反復を停止する．
7. マイクロホン位置を勾配と学習率を用いて更新する．
8. $t \leftarrow t+1$ としてステップ 4 に戻る．
9. マイクロホン位置を推定結果として返す．

5　評価実験

本節では，提案するキャリブレーション手法の定量的な性質およびキャリブレーション性能の限界を明らかにし，その有効性を示すために行った評価実験について述べる．シミュレーションで構築された 2 次元の無残響チェンバーに M 個の同期されたマイクロホンと N 個の音源を配置した．全ての音源から同時にそれぞれ異なる信号を発生させ，それら全てが混合した音を各マイクロホンで録音した．録音された混合音に対して提案手法を適用し，マイクロホンの位置をキャリブレーションした．キャリブレーション性能は，マイクロホンの推定位置と真の位置との平均誤差で評価する．

基本的な実験設定は以下の通りである．

- マイクロホンの基準形状：半径 10 cm の円形 8ch アレイ
- マイクロホンの実際の位置の基準位置からのずれ：平均 0 cm，標準偏差 1 cm の circulary symmetric 正規分布にしたがってサンプルされた乱数
- 音源の数：4
- 音源の配置（方位角）：0° から 360° の一様分布からサンプルされた乱数，各音源の間隔は少なくとも 20°
- 音源信号：CHiME3 challenge 孤立発話音声（7138 発話，平均長 7.64 s）からランダムに選択

以下の 3 つの実験を行った．

1. マイクロホン数または音源数を変化させた場合の性能を評価する．
   (a) 音源数を 4 に固定してマイクロホン数を 2, 4, 6, 8, 12, 16 に変化
   (b) マイクロホン数を 8 に固定して音源数を 2, 4, 6, 8, 10, 12 に変化
   (c) マイクロホン数を 16 に固定して音源数を 2, 4, 8, 12, 16, 17 に変化
2. マイクロホンアレイの基本形状を変化させた場合の性能を評価する．円形 9ch，格子状 9ch，十字形 9ch，直線形 9ch の 4 通り．
3. マイクロホンアレイの配置スケールを変化させる．半径 1 cm, 3 cm, 10 cm, 30 cm, 1 m, 3 m, 10 m の 7 通り．
4. マイクロホンの配置ずれ（真値と所与の値の差）を変化させた場合の性能を評価する．0.1 cm, 0.16 cm, 0.25 cm, 0.4 cm, 0.63 cm, 1 cm, 1.6 cm, 2.5 cm, 4 cm, 6.3 cm, 10 cm の 11 通り．
5. 与えられた音源方向に誤差が含まれる場合の性能を評価する．誤差なしの場合，標準偏差が 1°, 2°, 5°, 10° の正規分布に従う加法的誤差が含まれる場合，刻み幅が 1°, 2°, 5°, and 10° の離散値への丸め誤差が含まれる場合．

5.1　結果と考察

実験 1 の結果を Figures 2, 3, 4 に示す．これらの図は，横軸がマイクロホン数もしくは音源数を，縦軸がマイクロホン位置の平均誤差を意味し，箱ひげ図は各条件で 100 回ずつ実験を試行した結果を表している．図 2 ではマイクロホン数 M が 4 以下のとき，図 3 では音源数 N が 8 以上のとき，図 4 では音源数 N が 16 以上のとき，それぞれ推定誤差が大きく増大していることが分かる．これらをまとめると，マイクロホン数 M と音源数 N の関係が $M \leq N$ である場合に，推定誤差が大きく劣化していることが分かる．すなわち，提案するキャリブレーション手法が有効に機能するためには，音源数 N を超えるマイクロホン数 M が必要であることが分かる．ただし，ここでの音源数 N は，同時に発音している音源の数であり，例えば各音源の発話区間が与えられている場合には，マイクロホン数を超える音源数にも対応できる可能性はある．

さらに，Figure 2 を詳細に観察すると，マイクロホン数 M が 5 の場合は，マイクロホン数 M が 6 以上の場合に比べて推定誤差がわずかに大きい傾向が見られる．Figure 3 を観察すると，音源数 N が 2 または 7 の場合は，音源数 N が 3 から 6 の場合に比べて推定誤差がわずかに大きい傾向がみられる．同様に，Figure 4 を観察すると，音源数 N が 2, 14, 15 の場合は，N が 3 から 13 の場合に比べて推定誤差が大きい傾向がみられる．これらを総合すると，提案手法が有効に機能するのは，音源数 N が 3 以上で，マイクロホン数に対しておよそ 90%未満である場合であると結論づけることができる．

実験 2 の結果を Figure 5 に示す．アレイの半径が 30 cm のときが最も推定誤差が小さく，半径がそれより大きく，もしくは小さくなるほど推定誤差が増大している．特に，半径が 3 cm 以下，もしくは 300 cm 以上の場合は誤差が非常に大きくなっている．この結果から，提案手法はアレイの半径が 6 cm から 200 cm のときに有効に機能するといえる．

半径が 3 cm 以下の場合に推定誤差が増大する原因について考察する．本実験では，マイクロホン位置のずれは，平均が 0 cm，標準偏差が 1 cm の正規分布からランダムにサンプルされている．一方で，半径が 3 cm のとき，隣接するマイクロホン同士の間隔は 1.85 cm であり，これに標準偏差が 1 cm のランダムなずれを足し

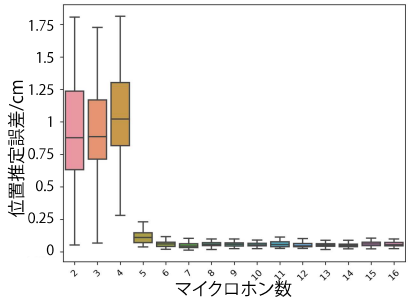

図 2: 実験 1 の結果．音源数 N を 4 に固定しマイクロホン数 M を変化させた場合

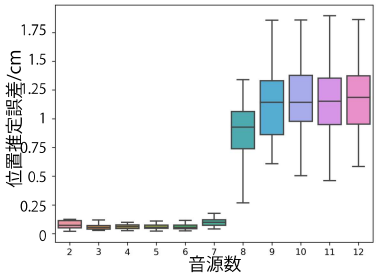

図 3: 実験 1 の結果．マイクロホン数 M を 8 に固定し音源数 N を変化させた場合

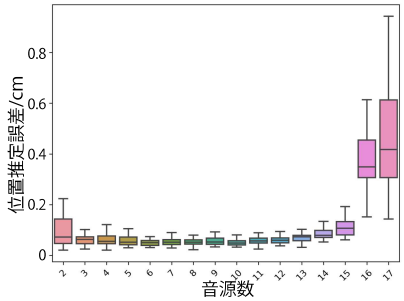

図 4: 実験 1 の結果．マイクロホン数 M を 16 に固定し音源数 N を変化させた場合

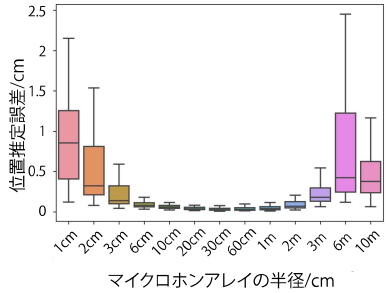

図 5: 実験 2 の結果．マイクロホンアレイの大きさを変化させた場合

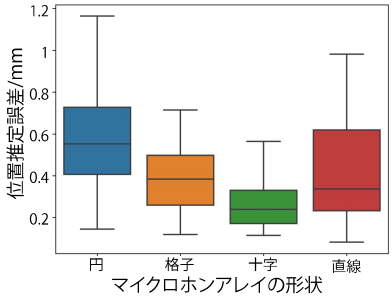

図 6: 実験 3 の結果．マイクロホンアレイの形状を変化させた場合

合わせると，マイクロホンアレイの真の形状は元の形状から大きくゆがんでしまう．校正すべき位置のずれが相対的に大きくなったことで，真の位置を推定することが難しくなり，推定誤差が増大したと考えられる．

半径が 300 cm 以上の場合に推定誤差が増大する原因について考察する．提案手法では短時間フーリエ変換で得られるスペクトログラムを入力として用いる．マイクロホンのサンプリング周波数は 16 kHz，STFT の窓長は 512 点（32 ms に相当）である．したがって，音速が 340 m s^{-1} のとき，32 ms の信号は 10.88 m の距離に渡って存在することになる．一方で，アレイの半径が 300 cm であるため，最も遠いマイクロホン同士の間隔は 6 m となり，このようなマイクロホンは同一時間フレームの信号のうち半分以下の長さしか共有していない．実際に同じ時間区間の音源信号を参照しているのは同一時間フレームの信号のうち半分以下となってしまう．このため，同一フレームであっても大部分は異なる音源信号の区間が観測である，という状況であり，推定誤差が増大したと考えられる．

実験 3 の結果を Figure 6 に示す．円形アレイが最も推定誤差が大きく，十字形アレイが最も推定誤差が小さくなった．円形 9ch アレイの場合が最も推定誤差が大きい原因について考察する．円形アレイと格子状アレイを比べると，X 軸方向・Y 軸方向におけるアレイの全長そのものはどちらも 20cm であるが，格子状アレイではマイクロホン同士の最小間隔が 10cm であるのに対して，円形アレイではマイクロホン同士の最小間隔が 6.84cm であり，すなわち円形アレイはより高密度で小規模であるとみなせる．実験 2 の結果より，半径 30cm 以下のアレイでは，スケールが小さいほど推定誤差が増大しているため，円形アレイの相対的なスケールの小ささが推定誤差の増大に繋がったと考えられる．

格子状アレイと十字形アレイは，いずれもマイクロホン同士の最小間隔は 10cm だが，十字形アレイの方が推定誤差が小さい．十字形アレイの方が端から端までの長さが大きいため，この全長の違いが推定誤差の違いに繋がったと考えられる．

直線形アレイは，80cm の端から端までの長さをもち，これはこれらのアレイの中では最大であるものの，推定誤差は最小ではない．このアレイの場合はマイクロホンが並んでいる X 軸方向における誤差は小さいものの，直交する Y 軸方向の誤差は大きく，Y 軸方向の校正はほとんど行えていない．アレイの非等方的な形状がこの結果を導き出したと考えられる．

6 まとめ

本論文では，マイクロホンアレイで観測された混合音を用いてマイクロホン位置を校正する方法を提案した．シミュレーション実験により，提案手法は任意かつ同時に複数の音源信号が録音された混合音を用いて校正できることが示され，その特性が分析された．一方，提案手法の限界として，マイクロホン数が音源数より少ない場合，アレイサイズが極端に小さい場合，または極端に大きい場合に推定誤差が大きくなることが示された．今後はこれらの限界を克服するために提案手法を改良し，さらに騒音や残響のある実環境での実験を通して，その実用性を評価することを目指す．

謝辞

本研究は科研費 JP19K12017, JP19KK0260 および JP20H00475 の助成を受けた．

参考文献

[1] C. Zhang, D. Florencio, D. E. Ba, and Z. Zhang. Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings. *IEEE Trans. Multimedia*, Vol. 10, No. 3, pp. 538–548, 2008.

[2] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano. Localization of multiple sound sources based on a CSP analysis with a microphone array. In *ICASSP 2000*, Vol. 2, pp. 1053–1056, 2000.

[3] Keisuke Nakamura, Kazuhiro Nakadai, Futoshi Asano, Yuji Hasegawa, and Hiroshi Tsujino. Intelligent sound source localization for dynamic environments. In *IROS 2009*, pp. 664–669, 2009.

[4] Kazuhiro Nakadai, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. Sound source separation of moving speakers for robot audition. In *ICASSP 2009*, pp. 3685–3688, 2009.

[5] Kazuhiro Nakadai, Hiroshi G. Okuno, and Takeshi Mizumoto. Development, deployment and applications of robot audition open source software HARK. *J. Robot. Mechatron.*, Vol. 29, No. 1, pp. 16–25, 2017.

[6] D. Su, T. Vidal-Calleja, and J. V. Miro. Simultaneous asynchronous microphone array calibration and sound source localisation. In *IROS 2015*, pp. 5561–5567, 2015.

[7] H. Miura, T. Yoshida, K. Nakamura, and K. Nakadai. SLAM-based online calibration of asynchronous microphone array for robot audition. In *IROS 2011*, pp. 524–529, 2011.

[8] V. C. Raykar, I. V. Kozintsev, and R. Lienhart. Position calibration of microphones and loudspeakers in distributed computing platforms. *IEEE Trans. Speech and Audio Process.*, Vol. 13, No. 1, pp. 70–83, 2005.

[9] N. Ono, K. Shibata, and H. Kameoka. Self-localization and channel synchronization of smartphone arrays using sound emissions. In *APSIPA ASC 2016*, pp. 1–5, 2016.

[10] Sebastian Thrun. Affine structure from sound. In *NIPS'05*, p. 1353 – 1360, 2005.