人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202-11

複数のマイクロホンアレイおよび空間情報と反射音を利用した音源定位の検討

Investigation on sound localization using multiple microphone arrays, reflection and spatial information

○石井カルロス寿憲（ＡＴＲ知能ロボティクス研究所）
Jani Even（ＡＴＲ知能ロボティクス研究所）
萩田紀博（ＡＴＲ知能ロボティクス研究所）

* Carlos Toshinori ISHI, Jani EVEN, Norihiro HAGITA (Intelligent Robotics and Communication Labs., ATR)

carlos@atr.jp, even@atr.jp, hagita@atr.jp

Abstract -複数のマイクロホンアレイにおいて音源方向推定を行い、空間の情報と反射音の方向の情報を利用し音源定位（3次元空間の位置推定）に利用する枠組みを提案し、人の声とスピーカから再生した音声を音源とした評価実験をおこなった。マイクロホンアレイの位置、壁の位置、音源の種類、音源の位置と向きに応じて、観測される直接音や反射音が変化し、反射音が重要となる条件を分析した。

1　はじめに

家庭, オフィス, 商店街など, 異なった環境では, 場所や時間によって多様な雑音特性を持つため，音声などの特定の音を対象としたアプリケーションでは，使用される環境の雑音の種類や度合いにより，期待した性能が得られないという問題がある．

本研究では，音環境の事前知識の習得およびその利用を総称して「音環境知能」と呼ぶ．また，のことを「音環境地図」と呼ぶ．実環境では，異なった場所で発生する複数の音が混合して観測されるため，音環境地図の生成において，騒音計で空間をスキャンするような従来の単純な方法は不十分である．音環境の事前知識として役立つと考えられる音源の位置や種類を特徴付けた音環境地図の生成には，空間的情報（通常の地図）に加え，音源の定位，分離及び分類が必要となる．そこで，本研究では，複数の音源を定位するため，複数のマイクロホンアレイを連携させ，空間内の特定の音源に対する音環境地図を生成し，音環境を構造化することを目的としている．本論文では、この目的を達成するための第一ステップとして、複数のマイクロホンアレイによる音源位置推定の問題に焦点を当てる。

マイクロホンアレイ処理における一つの問題として、アレイの周りに壁や天井やガラス窓やディスプレイなどの音を反射する表面が存在する場合、音源の直接音と同時に音源の反射音も観測されることがある。我々はマイクロホンアレイを天井に取り付けて集音を試みているが、特に音源との距離が大きい場合、強い反射音も頻繁に観測している。

これまでの音源定位や音源分離に関するほとんどの研究[1 ~ 10]では、反射音は悪影響を与えるものとして扱われてきたが、本研究では、反射音を利用して、音源位置推定に役立てる枠組みを提案し、その効果を評価した。

本論文は以下のように構成される。次ぐ2章では、提案手法を説明する。3章では、データ収集と提案手法による音源位置推定における分析結果を述べる。4章でまとめと今後の課題を記す。

2　提案手法

提案手法では、複数のマイクロホンアレイを用いて複数の音源方向を推定し、空間の情報を用いて反射音の方向を推定し、これらの情報を統合して音源定位（3次元空間の位置推定）を行う。音源方向推定においては、先行研究で提案した手法を採用し、2.1節で述べる。空間情報と反射音を利用した音源定位の提案手法は2.2節で述べる。

2.1 MUSIC スペクトル

MUSIC（Multiple Signal Classification）とは、音源定位において分解能が高い特徴を持つ手法の一種である。Fig. 1 にMUSICスペクトルの推定法のブロック図を示す。まずフーリエ変換(FFT)により多チャンネルのスペクトル X(k,t) をフレーム毎に求め、スペクトル領域でチャンネル間の空間的相関行列 R_k をブロック毎に求め、相関行列の固有値分解により指向性の成分と無指向性の成分のサブ空間を分解し、無指向性のサブ空間に対応する固有ベクトル

E_k^n と、対象の検索空間に応じて予め用意した方向ベクトル a_k を用いて（狭帯域の）MUSICスペクトルP(k)を周波数ビンごとに求め、特定の周波数帯域内の周波数ビン毎のMUSICスペクトルを統合して広帯域MUSICスペクトルが求まる。アルゴリズムの詳細は付録に記載している。

ここでは、広帯域MUSICスペクトルを単に「MUSICスペクトル」と呼び、MUSICスペクトルの時系列を「MUSICスペクトログラム」を呼ぶ。

音源定位においては、MUSICスペクトルのピークを探索することにより、音源の方向が求まる。

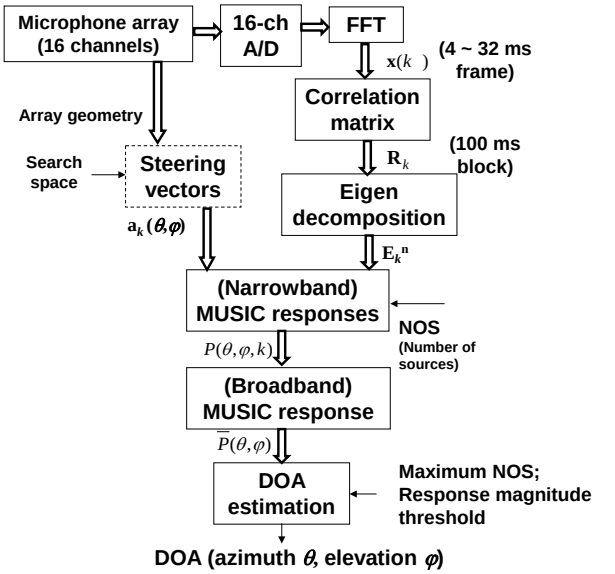

Fig. 1. The MUSIC-based sound localization algorithm, and related parameters.

ただし、MUSIC法を用いた音源定位の実用化においては、主に2つの問題が挙げられる。一つ目は、チャンネルの数および探索空間が大きくなるにつれて、処理時間が重くなり、通常のパソコンでは、実時間処理が追いつかないことである。もう一つは、MUSICスペクトルを求めるには、音源数を予め与える必要があることである。

著者らの先行研究[10]で、実時間処理を可能にするため、MUSICスペクトルの推定においていくつかのパラメータを分析した。その結果、FFTのフレーム長を64～128点（4～8msに対応）、ブロック長を100msに設定することにより、2GHzのCore2DuoのCPUを用いて、音源方向推定の精度を保ちつつ、実時間処理が可能であることを示した。

狭帯域MUSICスペクトルの推定において、その時刻に発している指向性を持つ音源数（NOS）を与える必要があるが、音源数の推定は難しいため、先行研究[10]で提案した通り、固定数を与え、MUSICスペクトル上で、特定の閾値を超えたピークのみを指向性のある音源とみなす方法を用いる。

2.2　空間情報と反射音を利用した複数アレイによる音源定位

本節では、複数のアレイにおいて、2.1節で説明したMUSICスペクトルによる音源方向推定を行い、空間情報とアレイの位置情報を用いて反射音の方向も推定し、これらの情報を統合して複数の音源位置の推定を行う手法を説明する。概要図をFig. 2に示す。

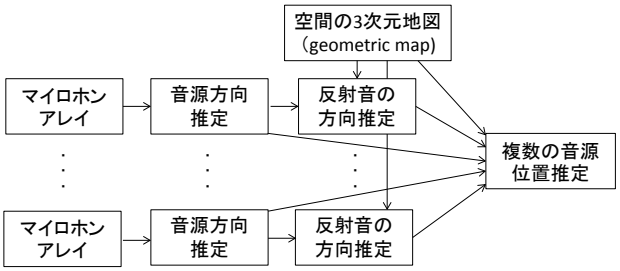

Fig. 2. The proposed sound localization system using multiple arrays and sound reflections.

複数のマイクロホンアレイを用いて音源方向を推定し、空間内のアレイの位置と向きが既知である場合、それぞれのアレイで推定された音源方向が重なった位置に音源が存在する確率が高いというのが本手法の基本的な概念である。

また、空間内のアレイの位置、音が反射しやすい天井や壁やディスプレイなどとの位置関係によって、アレイで反射音が測定される場合があり、一つのアレイでも反射音と直接音の方向が検出された場合、反射音を壁や天井で反転させた方向と直接音が重なった位置に音源が存在する確率が高いと予想される。従来のマイクロホンアレイ処理では、反射音は音源定位や音源分離に悪影響を与えるが、本手法では、逆に反射音の情報を利用することとなる。

定位された音源が反射音であるか否かは予め分からないため、まず推定されたすべての音源方向を壁や天井で反転させる。反射は空間内で複数生じ得るが、本研究では、2度目以降の反射は強度も指向性も衰える可能性があるため、反転は1度のみ行うこととする。

また3次元空間を考慮し、方位角および仰角で音源方向を表現する。

推定された方向には、角度の誤差（Angle uncertainty: AU）があり、アレイからの距離に応じて推定位置の誤差（Position uncertainty: PU）が大きくなる。幾何学的に、推定位置誤差を以下の式で求めることができる。

$$PU(d) = \pm AU / 360 * 2\pi * d \quad (1)$$

d はアレイの中心からの距離で、AUは推定角度の誤差を度単位で表したものである。例えば、球面上で5度の分解能で音源方向が検知された場合（AU = 5）、アレイから1メートル離れた位置に音源がある場合

（d = 1m）、その方向に直線を1メートル伸ばした際の推定位置誤差は ±8.7 cm となる。2メートルの場合、誤差はその倍の ±17.4 cm となる。

検知された2つの方向が上述の誤差を考慮して空間上で重なっているか否かを判定する方法として、それぞれの方向に直線を引き、2直線の最短距離を幾何学の公式を用いて推定する。この最短距離がそれぞれの直線における誤差（PU）を足した値よりも小さい場合、これらの直線は重なっていると判定する。また、検出された方向の重なりが生じた位置に音源が存在する可能性が高いとみなす。

検出されたすべての直接音と反射音の方向に引いた直線の距離をペア毎に求め、方向の重なりを複数探索する。重なりがあった場合は、平均位置を音源の推定位置とする。重なりがない場合は方向情報を保留とし、重なりが生じた時点で、位置を割り当てる。

3　データ収集および分析結果

3.1 マイクロホンアレイと音源方向推定の設定

本実験に用いた 16 素子のマイクロホンアレイの形状を Fig. 3 に示す。3 次元空間における方位角および仰角を求めるため、マイクは直径 30cm の半球面上に Fig. 3 に示すように配置した。

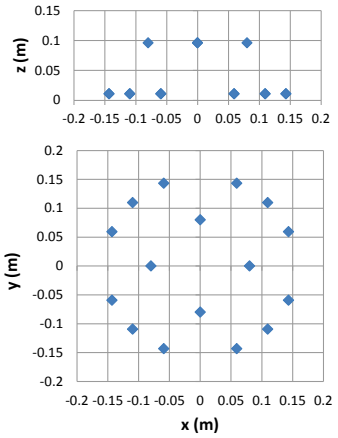

Fig. 3 The geometry of the 16-element microphone array.

多チャンネルオーディオキャプチャデバイスとして、東京エレクトロンデバイス社の16-channel A/D 変換機 TD-BD-16ADUSB を使用した。マイクは、Sony の ECM-C10 を用い、16 kHz/16 bitsでサンプリングを行った。

MUSICスペクトルによる音源方向推定のパラメータとして、音源の固定数を3、MUSICパワーの閾値を2.5dB、同時に発する音源の最大数を6 に設定した。

また、MUSICスペクトルを求める際に用いる周波数帯域に関しては、空間的歪み（spatial aliasing）と低周波数帯域における低い分解能を避けるため、1000 ~ 5000 Hzの帯域を用いた。

音源方向推定の探索空間は、3 次元空間で球面上 5 度間隔の分解能に設定し、アレイを天井に取り付けるため、方位角は 0 ～ 360 度、仰角は-5 度～-80 度に制限した。-85 ~ -90 度（アレイの真上の方向）には、アレイの形状より音源が存在しない場合にも MUSIC スペクトルにピークが生じるため、その領域を探索空間から除外している。これは使用したキャプチャの特性により、すべてのチャンネルで同位相の雑音が生じるためである。

3.2 評価データの収集

本実験では、Fig. 4 に示すように、2つのアレイを天井に取り付けた。アレイと天井の間には吸音素材を入れ込み、天井での反射は扱わないこととした。また床は反射しにくいタイルカーペットであり、反射が生じたとしても天井に設置したアレイへの距離が大きいため、床での反射も扱わないこととした。従って本研究では、推定された音源方向を壁で一回のみ反転させることとした。

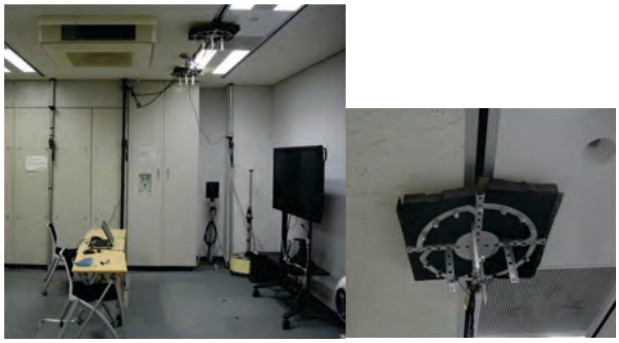

Fig. 4 The microphone arrays attached in the ceiling.

音源の向きによって、その音源の指向性が変化し、同じ位置でもアレイに対する向きによってアレイで観測される指向性の強度が変化することが考えられる。また、音源の種類によっても、指向性が異なることが予想され、本研究では、人が発した音声と、音声をスピーカから流した場合の2つの種類を対象音源とした（これらの音源をこれ以降それぞれ “Human” および “Loudspeaker” と呼ぶ）。また、環境に固定されたエアコン（Fig. 4 の左上）もアレイに対して指向性を持つ雑音源となる。

対象音源の位置として、Fig. 4の机の周り6か所を固定し、各位置において、前後左右の4つの向きでデータを収録した。エアコンはスイッチオンの状態にした。正確に音源の位置を固定することは難しいが、向きを変えた際に、口の位置ができるだけ変わらないようにした。

スピーカとして、ONKYOのGX-77Mを用いた。スピーカの高さは、話者が椅子に座った時の口の高さと一致するようにした。話者には各位置および各向きで同じ文を同じような発話スタイルで発声する

よう指示した。スピーカからは同じ話者の声を再生した。スピーカの音量は人の声の強度に近づけるよう調整した。表 1 に、設定した音源の位置とマイクロホンアレイの位置を記す。Fig. 5 に音源の位置および向きとアレイの位置を部屋の上面図に重ねて示す。x = 0 および y = 0 の平面には壁が存在する。x = 7400 mm および y = -5600 mm にも壁が存在するが、アレイから離れているため、本実験の反射音推定には用いなかった。

表 1．対象音源の位置およびマイクロホンアレイの位置情報

	1	2	3	4	5	6
x (mm)	1000	2000	3000	3000	2000	1000
y (mm)	-2700	-2700	-2700	-1200	-1200	-1200
z (mm)	1160	1160	1160	1160	1160	1160

	array1	array2
x (mm)	1410	3560
y (mm)	-1430	-1430
z (mm)	2630	2630

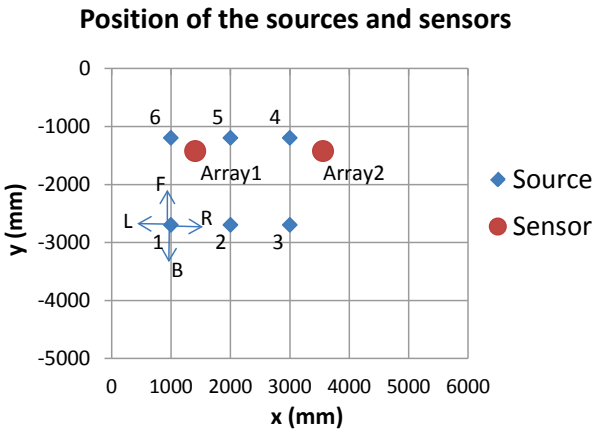

Fig. 5 Position (1 ~ 6) and orientation (F: front, L: left, B: back, R: right) of the target sources and the microphone array sensors (Array1, Array2) in the room.

3.3 音源の種類およびアレイに対する音源の向きの影響

本実験は、それぞれのアレイで測定された音源方向と反射音が実際発した音源位置をどの程度精度よく推定可能であるかを評価することを目的とする。そのため、評価尺度として、各アレイで検出された各音源方向に対する直線と、表 1 に示した対象音源の座標位置との距離を測定した。ここでは、音源方向推定誤差による位置推定誤差の他に、スピーカの直径が 9 cm で、対象音源の位置が正確ではないことも考慮し、位置推定誤差が40 cm以内であれば、その方向は対象音源が発しているものとみなすこととした。各アレイで観測された各方向に対し、上述の条件を満たしたブロックの数を発話区間のブロック数で割ったものを検出率とする。

Fig. 6 に “human” および “loudspeaker”の2種類の音源に対する結果を音源の位置（“1” ~ “6”）と向き（“F”, “L”, “B”, “R”）の条件ごとに表示している。それぞれのアレイ（“Array1”, “Array2”）で検出された音源方向は、直接音（“d”）、平面 y=0 での反射音（“ry”）および平面 x=0 での反射音（“rx”）に分けて結果を表示している。

Fig. 6 の結果より、まず音源の位置と方向によってそれぞれのアレイで直接音（d）および反射音（ry, rx）が観測される率が変化することが分かる。これは、音源の位置と向きによって、アレイが「見えている」のか、壁が「見えている」のかに依存する。例えば “human” 音源におけるFig. 6の上図の “1L” の条件では、Array1の直接音 d と反射音 rx がおよそ 0.8 の率で検出されている。また、Array2では、反射音 rx がおよそ 0.6 の率で検出され、直接音はほとんど検出されていない。

“human”と “loudspeaker”の結果を比較すると、全体的に人が発声した場合の方向の検出率が高い結果が得られた。これは人よりもスピーカの方が、指向性が強いことが原因である。

音源位置推定においては、同じ音源に対し、複数（少なくとも2つ）の方向が検出されれば、その重なった位置に音源が存在すると判定することができる。例えば、“human”音源の “6R”の条件で、0.9 以上の率で両アレイの直接音が重なって観測されている。 “loudspeaker”音源の場合でも、0.8前後の検出率が得られている。

“human”音源で、直接音が高い率で上位を占めている条件は、{2F, 3F, 4F, 5F, 6F, 3L, 4L, 4B, 5B, 1R, 2R 5R, 6R} で、全条件のおよそ半分を占めいている。平面x=0での反射音（rx）が上位に入っている条件は、Array1の場合{1L, 2L, 5L, 6L, 6B}となっている。これらの条件は、平面x=0 の壁に近く、その方向を向いている条件である。また、平面y=0での反射音（ry）が上位に入っている条件は、 Array1の場合は{1F, 6F}で、Array2の場合は{3F, 4F, 4R}となっている。

その一方、 “loudspeaker”音源では、直接音が高い率で上位を占めている条件は、{6R} のみの条件となっている。反射音（rx もしくは ry）が最も高い率で検出されている条件は、{4F, 5F, 6F, 1L, 2L, 5L, 6L} となっている。そのうち、{4F, 1L, 6L}の条件では、直接音がほとんど観測されず、反射音のみが上位を占めている。これらは、両アレイに背いているが、壁が近いため反射音が直接音よりも強く観測される条件である。従って、音源の指向特性に応じて、反射音の特定は音源定位に大きな役割を果たすことが示されている。

{1B, 2B, 3B} の条件では、音源が両アレイに背いている状態であるため、両アレイで直接音も反射音

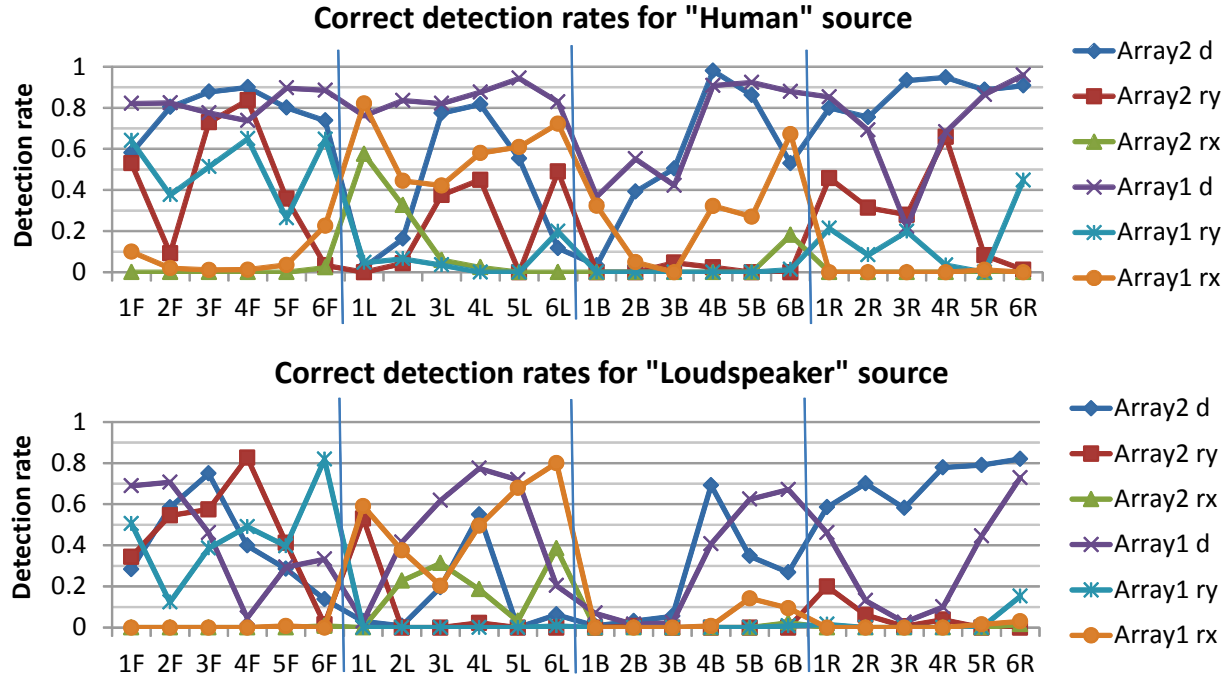

Fig. 6 Correct detection rates for direct path (d) reflection at plane y=0 (ry) and reflection at plane x=0 (rx) by each array (Array1, Array2), for each position (1 ~ 6) and orientation (F: front, L: left, B: back, R: right) of the target sources ("human" and "loudspeaker").

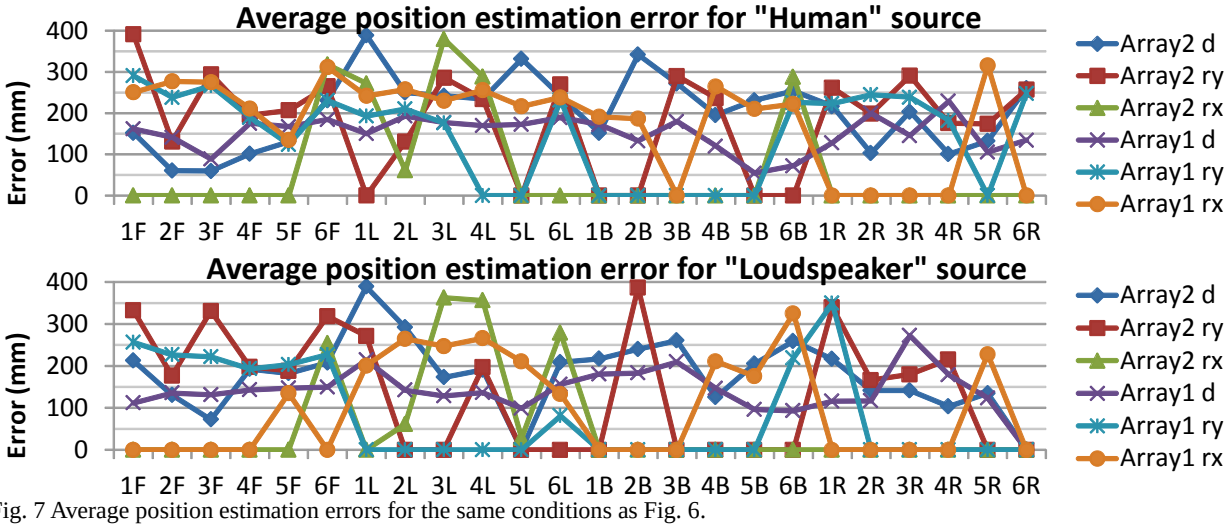

Fig. 7 Average position estimation errors for the same conditions as Fig. 6.

も検出率が低い（0.5 前後）。指向性の高いスピーカでは、検出率が 0 となっている。これらの条件の対処として、部屋に設置するアレイの数を増やす必要があると考えられる。

Fig. 7 に、Fig. 6 の条件に対応する平均位置推定誤差を示している。ただし、これらのグラフで誤差が0の点は、その条件で方向が検出されなかった場合を示している。

Fig. 7 の結果より、直接音でも反射音でも平均誤差は100 ~ 300 mmの範囲で検出されていることが読み取れる。いくつかの条件（例えば "Array2 ry 1F", "Array2 d 1L", "Array2 rx 3L 4L"）では、誤差が300 mm 以上となっているが、これらの条件は、音源とアレイとの距離が長く（直接音は 1 番目の位置からおよそ3 m、反射音の場合折り返しの距離も含めて 1 番目の位置からおよそ 5 m、3 番目と4 番目の位置からおよそ 7 m となり）、方向推定の誤差が大きくなることが原因と考えられる。この誤差を小さくする対処法として、局所的に分解能を上げることが考えられるが、今後の課題とする。

4　おわりに

本研究では、複数のマイクロホンアレイにおいて音源方向推定を行い、空間の情報と反射音の方向の

情報を利用し音源定位（3 次元空間の位置推定）に利用する枠組みを提案した。

人の声とスピーカから再生した音声を音源とした評価実験を行った結果、マイクロホンアレイの位置、壁の位置、音源の種類、音源の位置と向きに応じて、観測される直接音や反射音が変化し、反射音が重要となる条件を明らかにした。スピーカは人よりも指向性が強いことが分かり、アレイとの距離よりも壁とアレイに対する向きに応じて、反射音のみが観測される場合があり、人の声とは異なる観測パターンが得られた。また、このような結果は、通常スピーカを用いて「実験室実験」を行う研究が多いが、人が実際に発声した際の指向性特定が異なることを考慮すべきであることを示している。

今後の課題として、異なった音源や向きのより詳細な分析を行い、本研究の音源定位法を LRF などによる人位置検出の結果と統合させ、誰がいつどこで発話したのかを記述する音環境知能技術に発展させる予定である。

付録：MUSIC 法

M個のマイク入力のフーリエ変換$X_m(k,t)$は、式(1)のようにモデル化される。

$$\mathbf{x}(k,t) = [X_1(k,t),...,X_M(k,t)]^{\mathrm{T}} = \mathbf{A}_k\,\mathrm{s}(k,t) + \mathbf{n}(k,t) \quad (1)$$

ベクトルs(k,t)はN個の音源のスペクトル$S_n(k,t)$から成る：s$(k,t) = [S_1(k,t), \ldots, S_N(k,t)]^{\mathrm{T}}$ 。kとt はそれぞれ周波数と時間フレームのインデックスを示す。ベクトル$\mathbf{n}(k,t)$は背景雑音を示す。行列 $\mathbf{A}_k$ は変換関数行列であり、(m,n)要素はn番目の音源からm番目のマイクロホンへの直接パスの変換関数である。$\mathbf{A}_k$ のn列目のベクトルをn番目の音源の位置ベクトル（steering vector）と呼ぶ。

まず、式(2)で定義される空間相関行列$\mathbf{R}_k$を求め、式(3)に示す$\mathbf{R}_k$の固有値分解により、固有値の対角行列$\mathbf{\Lambda}_k$および固有ベクトルから成る$\mathbf{E}_k$が求められる。

$$\mathbf{R}_k = E[\mathbf{x}(k,t)\mathbf{x}^H(k,t)] \quad (2)$$

$$\mathbf{R}_k = \mathbf{E}_k\mathbf{\Lambda}_k\mathbf{E}_k^{-1} \quad (3)$$

固有ベクトルは $\mathbf{E}_k=[\mathbf{E}_k^{\mathrm{s}} \mid \mathbf{E}_k^{\mathrm{n}}]$ のように分割出来、$\mathbf{E}_k^{\mathrm{s}}$と$\mathbf{E}_k^{\mathrm{n}}$はそれぞれ支配的な$N$個の固有値に対応する固有ベクトルと、それ以外の固有ベクトルである。

MUSIC空間スペクトルは式(4)と(5)で求める。rは距離、θとφはそれぞれ方位角と仰角を示す。式(5)は、スキャンされる点 (r,θ,φ) における正規化した位置ベクトルである。

$$P(r,\theta,\varphi,k) = \frac{1}{|\,\tilde{\mathbf{a}}_k^H(r,\theta,\varphi)\mathbf{E}_k^n\,|^2} \quad (4)$$

$$\tilde{\mathbf{a}}_k(r,\theta,\varphi) = \frac{\mathbf{a}_k(r,\theta,\varphi)}{\left\|\mathbf{a}_k(r,\theta,\varphi)\right\|} \quad (5)$$

空間スペクトル（本稿ではMUSIC応答と呼ぶ）は、MUSIC空間スペクトルを式(6)のように平均化したものである。

$$\overline{P}(r,\theta,\varphi) = \frac{1}{K}\sum_{k=k_L}^{k_H} P(r,\theta,\varphi,k) \quad (6)$$

k_Lとk_Hは、周波数帯域の下位と上位の境界のインデックスであり、$K = k_H - k_L + 1$。音源の方位は、MUSIC応答のN個のピークから求められる。

謝辞

本研究は総務省の戦略的情報通信研究開発推進制度（SCOPE）の研究委託により実施したものである。

参考文献

1) F. Asano, M. Goto, K. Itou, and H. Asoh, “Real-time sound source localization and separation system and its application on automatic speech recognition,” in *Eurospeech 2001*, Aalborg, Denmark, 2001, pp. 1013–1016.
2) K. Nakadai, H. Nakajima, M. Murase, H.G. Okuno, Y. Hasegawa and H. Tsujino, "Real-time tracking of multiple sound sources by integration of in-room and robot-embedded microphone arrays," in *Proc. of IROS 2006*, Beijing, China, 2006, pp. 852–859.
3) S. Argentieri and P. Danès, "Broadband variations of the MUSIC high-resolution method for sound source localization in Robotics," in *Proc. of IROS 2007*, San Diego, CA, USA, 2007, pp. 2009–2014.
4) M. Heckmann, T. Rodermann, F. Joublin, C. Goerick, B. Schölling, "Auditory inspired binaural robust sound source localization in echoic and noisy environments," in *Proc. of IROS 2006*, Beijing, China, 2006, pp.368–373.
5) T. Rodemann, M. Heckmann, F. Joublin, C. Goerick, B. Schölling, "Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping," in *Proc. of IROS 2006*, Beijing, China, 2006, pp.860–865.
6) J. C. Murray, S. Wermter, H. R. Erwin, "Bioinspired auditory sound localization for improving the signal to noise ratio of socially interactive robots," in *Proc. of IROS 2006*, Beijing, China, 2006, pp. 1206–1211.
7) Y. Sasaki, S. Kagami, H. Mizoguchi, "Multiple sound source mapping for a mobile robot by self-motion triangulation," in *Proc. of IROS 2006*, Beijing, China, 2006, pp. 380–385.
8) J.-M. Valin, F. Michaud, and J. Rouat, "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," *IEEE ICASSP 2006*, Toulouse, France, pp. IV 841–844.
9) B. Rudzyn, W. Kadous, C. Sammut, "Real time robot audition system incorporating both 3D sound source localization and voice characterization," Procs.of *ICRA 2007*, Roma, Italy, 2007, pp. 4733–4738.
10) C. T. Ishi, O. Chatot, H. Ishiguro, N. Hagita, “Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments,” in *Proc. of the 2009 IEEE/RSJ Intl. Conf. on Intelligent Robots and System*, St. Louis, USA, 2009, pp. 2027–2032.