大規模マイクロホンアレーを用いた発話方向の実時間推定

○菊池慶子(東京電機大学院)　中島弘史((株)HRI-JP)　中臺一博((株)HRI-JP)
長谷川雄二((株)HRI-JP)　金田豊(東京電機大学院)

Real-time sound source orientation estimation using a 96 channel microphone array

Keiko KIKUCHI[1], Hirofumi NAKAJIMA[2], Kazuhiro NAKADAI[2]
Yuji HASEGAWA[2], Yutaka KANEDA[1]
1 Tokyo Denki University, 2 Honda Research Institute Japan Co., Ltd.

Abstract — This paper addresses sound source orientation estimation using a 96ch microphone array. We reported an weighted delay-and-sum beam-forming method for orientation estimation of a speech source such as a loudspeaker or an actual human. However, the performance of this method was degraded due to noise and reverberation. In addition, it has a difficulty in real-time processing. To solve, these problems, we propose two methods, that is, frequency component selection and time interval selection respectively. We implemented a real-time sound source orientation estimation system with these proposed methods using C language. Experimental results showed that the proposed methods drastically improved the performance under noisy and reverberant environments in real time.

Key Words: Source orientation, Microphone array, Beam forming, Intelligent environment

1. はじめに

ロボット聴覚の研究分野では，人‐ロボットインタラクションを可能にするための聴覚システムの構築を目指した研究が行われている[1][2]．主に音源定位（位置推定），音源分離，音声認識といった聴覚機能が研究されている．しかし，これらの機能以外にも，例えば，音源(話者)の向きを推定する機能(以後, 発話方向推定と記す)も重要である. 図1 は，発話方向推定が必要となる例である．図中の女声話者は，ロボットでは無く，別の男性に話しかけている．しかし，発話方向推定を有しないロボットは，女性の発話に反応してしまう．つまり，ロボットはこの女声の発話対象が自分では無い事を判断する必要がある．

このような問題に対して，画像処理で解決する方法もある．しかし，画像処理だけでは，オクルージョンなどにより，誰が誰に発話しているのか判別することが困難な状況が出てくる．

本稿では，ロボット本体に搭載したマイクロホンからの入力情報に限らず，周囲環境に設置したマイクロホンからの入力情報も積極的に利用するロボット聴覚システムを考える．我々は，これまでに部屋の壁に設置したマイクロホンアレーを利用した発話方向の推定法を提案した[3][4]．しかし，このシステムは，(1)全周波数帯域利用による雑音帯域の利用，(2)発話区間検出時における残響の未考慮，(3)低い実時間処理性という課題があった．本稿では，これらの課題に対して，それぞれ周波数選択による雑音帯域の削除，時間区間選択による残響部分の削除，C言語を用いた実時間音源方向推定システムを導入した．本稿では，これらの処理を導入した実時間音源方向推定システムの構築とその評価結果について報告する．

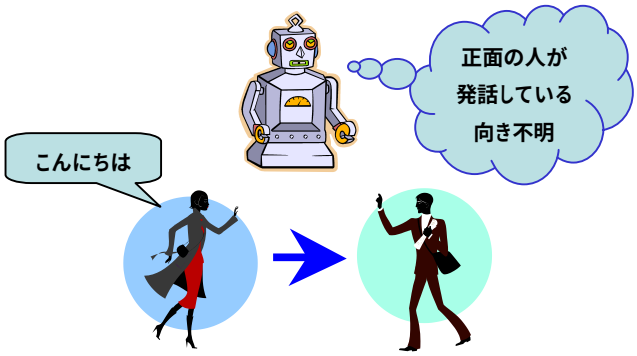

図1　発話方向推定が必要な例

2. 従来の発話方向推定方法とその課題

本章では，筆者らが発話方向の推定法として用いている拡張ビームフォーミングによる音源方向の推定方法について説明する．

2.1 音源方向への拡張ビームフォーミング

ビームフォーミング(以後，BFと記す)は，空間的な指向性を形成する技術である．BFは，その指向性の焦点を走査することで，音源パワーの空間分布を推定することができ，その最大値から音源位置を推定できる(走査BF)[4]．BFは，ある特定の位置に対して焦点を形成するよう設計するのが一般的である．筆者らは伝達関数を音源の向きによって変化する関数に拡張することにより，位置だけでなく，音源の向きに対しても焦点を形成するBFが設計できることを示した(拡張BF)[5]．この拡張BFを用いると，一般的な走査BFと同様の処理により，音源の位置だけでなく，音源の向きも推定することができる(拡張走査BF)．具体的には，それぞれ音源の位置と向きが異なる伝達関数を元にBF を設計し, その出力が最大となるBFの焦点位置と方向を，それぞれ発話者の位置と

向きとして推定する．

拡張走査BF は，BFを遅延和で設計した場合，伝達関数と入力信号の内積の最大値によって方向を推定するのと等価である．従って，拡張BFの設計，および拡張BFの走査は，それぞれ位置と向きが異なる複数の伝達関数を集めたデータベースの作成(以後，伝達関数データベースと記す)，および伝達関数データベースと入力信号の照合と見なすことができる．本章では簡略化のため，音源の向きのみに対する拡張走査BF について，データベースの作成と照合の観点から説明する．

2.2 音源の向きに拡張した伝達関数モデル

図2にN素子のマイクロホンアレーを用いた音源方向推定のモデルを示す．$S(\omega)$は周波数ωでの音源(話者)の周波数特性，M_kはk番目のマイクロホン$(k=1,2,\cdots,N)$，$H_k(\omega,\theta)$は話者がθ方向を向いている時の話者‐k番目マイクロホン間の伝達関数である．ここでは，話者の位置は既知であるとした．マイクロホンM_kでの受音信号$X_k(\omega,\theta)$は，

$$X_k(\omega,\theta)=S(\omega)H_k(\omega,\theta) \tag{1}$$

各変数をベクトルで表現すると

$$\mathbf{h}(\omega,\theta)=\left[H_1(\omega,\theta),\cdots,H_N(\omega,\theta)\right]^T \tag{2}$$

$$\begin{aligned}\mathbf{X}(\omega,\theta)&=\left[X_1(\omega,\theta),\cdots,X_N(\omega,\theta)\right]^T\\&=S(\omega)\left[H_1(\omega,\theta),\cdots,H_N(\omega,\theta)\right]^T\\&=S(\omega)\mathbf{h}(\omega,\theta)\end{aligned}$$

と表すことができる．

2.3 伝達関数データベースによる発話方向推定法

伝達関数データベースは，各向きの音源から各マイクロホンまでの伝達関数を集めたものである．方向推定においては，伝達関数ベクトル$\mathbf{h}(\omega,\theta)$のベクトルの向きのみが必要となるため，次式により各周波数および各方向で大きさを 1 に正規化した$\mathbf{h}_0(\omega,\theta)$を用いた．

$$\mathbf{h}_0(\omega,\theta)=\frac{\mathbf{h}(\omega,\theta)}{\sqrt{\mathbf{h}(\omega,\theta)^H\mathbf{h}(\omega,\theta)}}=\frac{\mathbf{h}(\omega,\theta)}{|\mathbf{h}(\omega,\theta)|} \tag{3}$$

ここでHは複素共役転置を示す．正規化により，伝達関数に含まれる出力機器の特性(スピーカの周波数特性など)を含まない伝達関数を得ることができる．同様に，話者が発話方向$\hat{\theta}$(未知)に向いて発話した時の受音信号ベクトル$\mathbf{X}(\omega,\hat{\theta})$は，

$$\mathbf{X}(\omega,\theta)=\left[X_1(\omega,\hat{\theta}),\cdots,X_N(\omega,\hat{\theta})\right]^T \tag{4}$$

と表され，これを正規化したものを$\mathbf{X}_0(\omega,\hat{\theta})$とする．

$$\mathbf{X}_0(\omega,\hat{\theta})=\frac{\mathbf{X}(\omega,\hat{\theta})}{|\mathbf{X}(\omega,\hat{\theta})|} \tag{5}$$

式(3)と式(5)の内積値$C(\omega,\theta)$は

$$C(\omega,\theta)=\mathbf{h}_0(\omega,\theta)^H\mathbf{X}_0(\omega,\hat{\theta}) \tag{6}$$

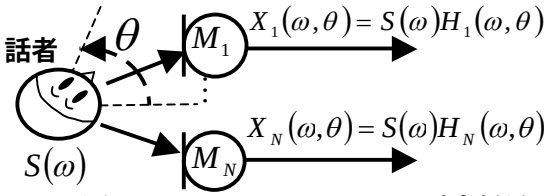

図2　マイクロホンアレーと受音信号

$$\begin{aligned}&=\frac{S(\omega)}{|S(\omega)|}\mathbf{h}_0(\omega,\theta)^H\mathbf{h}_0(\omega,\hat{\theta})\\&=\mathbf{h}_0(\omega,\theta)^H\mathbf{h}_0(\omega,\hat{\theta})\end{aligned}$$

となり，方向はθと$\hat{\theta}$の伝達関数の類似度を示す．この$C(\omega,\theta)$を周波数で平均した平均類似度$C_a(\theta)$を

$$C_a(\theta)=\sum_{\omega}C(\omega,\theta) \tag{7}$$

として計算する．この$C(\omega,\theta)$が最大となるθを音源方向として推定した．

2.4 伝達関数ベクトルにおける振幅抽出

2.1節から2.3節で述べた推定手法は，伝達関数の複素成分(振幅成分と位相成分)の内積から発話方向推定を行っていた．しかし，高周波帯域において位相成分が系の変動に敏感で変化しやすいという課題がある．系の変動は，例えば，話者の口の高さや位置の変化によって起こる．位置の変化量が少ない場合，伝達関数の変動は，特に高周波において位相が大きく変化する．そのため，内積値の抽出の際に，位置の変化に対して比較的ロバストな振幅成分を抽出する処理を加えることで，方向推定精度が向上する．

具体的には，伝達関数ベクトル$\mathbf{h}_0(\omega,\theta)$，受音信号ベクトル$\mathbf{X}_0(\omega,\hat{\theta})$の各要素の振幅値を求めたベクトル$\mathbf{h}_{a0}(\omega,\theta)$，$\mathbf{X}_{a0}(\omega,\hat{\theta})$を使用する．

2.5 従来システムの課題

従来システムは，2.3 節，2.4 節の処理を行い，音源向きの推定を行う[5]. しかし, 従来システムでは，手法について(1)雑音の多い帯域も含んだ全周波数帯域利用，(2)発話区間検出における残響部分の含有，実装について(3)実時間性の低下といった問題が挙げられた．これらの課題に対して，それぞれ，周波数選択による雑音帯域の削除，時間区間選択による残響部分の削除，C 言語を用いた実時間音源方向推定システムを提案する．

3. 提案システムについて

3.1 周波数選択(周波数マスク付ヒストグラム)

従来手法では，音声の周波数特性を考慮せず，全ての周波数帯域で一律に平均した平均類似度の最大値から発話方向を推定していた．この方法では，音声がほとんど含まれず，正しい方向が得られない周波数帯域の成分も含まれるため，推定精度が低下す

る．そこで，短時間区間ごとに DFT を行い，各周波数での推定結果のヒストグラムをとり，最大頻度の方向をその時刻の発話方向と推定した．

ただし，音声がほとんど含まれない帯域では内積値が低下することに着目して，内積値の低い周波数成分をヒストグラムの計算から除去する周波数マスクを導入した。周波数マスクは次式の周波数重み $w(\omega)$として定義される．

$$w(\omega)=\begin{cases}1 & (p(\omega)/p_{mean}(\omega)\geq\gamma)\\ 0 & (p(\omega)/p_{mean}(\omega)<\gamma)\end{cases} \quad (8)$$

ここで，$p(\omega)$は内積値を，$p_{mean}(\omega)$はその時間平均値を表す．またγは閾値を表し，今回はγ=1.5 とした．

3.2 時間区間選択

従来の手法では，発話方向推定を行うべき時刻を特定するために，音声が存在する時間区間の検出を行った．今回は，母音の周期性の有無に基づいた音声区間検出法[7]を利用した．しかし，この手法で検出した音声区間では，発話方向誤検出が発生した．

誤検出の発生部分を調べた典型的な結果を図 3 に示す。図において横軸は時間で縦軸は方向，色は内積値を表し，赤色の強い方向が推定方向となる．図に示すように，誤推定は音声区間の後半に多いことが判明した．これより，音声区間として検出された時間区間の後部は，発話区間ではなく，音声が残響として残っている区間であり，発話方向検出に利用するのは不適切な区間であると判断した．この結果に基づき，今回は音声区間の後部の 2 フレームを削除した区間を発話区間と定義することとした．

4. 発話方向の実時間推定システム

提案手法をもとに実時間で動作する発話位置・向き推定システムを開発した．図 4 に開発したシステムのブロック図を示す．

マイクロホンアレーは広さ 7m×4m，高さ 3.5m，残響時間が約 230ms の実験室に設置した．実験室の暗騒音レベルは約 40dB であった。マイクロホン数は 96 で，図 4 に○印で示すように室内の壁面に配置されている。

従来システムは，MATLAB を用いたオフラインシステムであった．提案システムは，C 言語で実装を行った．音取得部，音源位置・向き推定部，推定結果３Ｄ表示部から構成されており，各モジュールは MMI[8]を用いてネットワークで接続されている．このため，2 台の PC を用いて負荷分散を行い，実時間処理を可能としている．

5. 評価結果

提案手法の有効性を示すための評価実験を行った．実験では，提案手法と従来手法による推定精度の比較のため，実験条件は両手法で処理可能な条件に

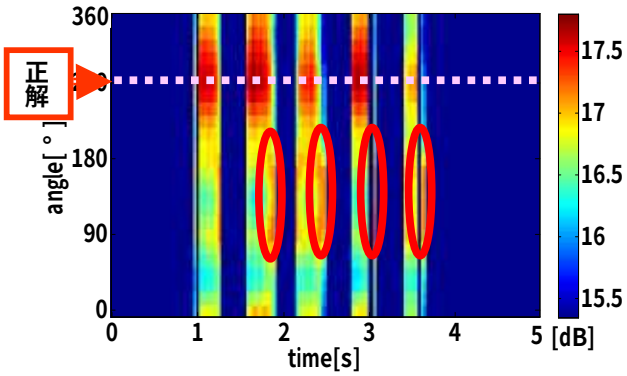

図 3　推定結果の誤推定部分

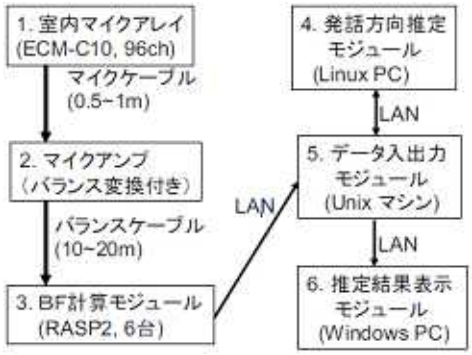

図 4　発話方向の実時間推定システムのブロック図

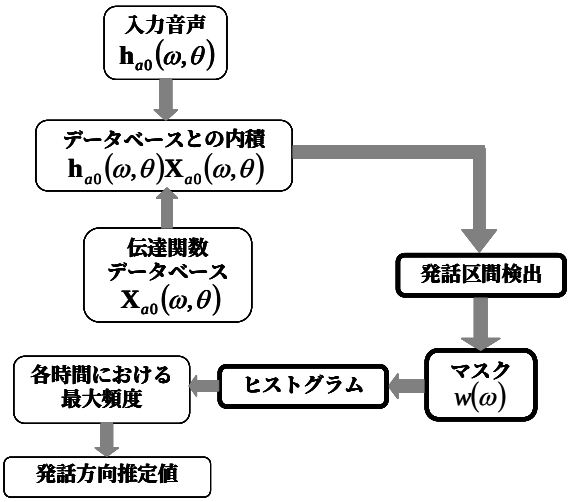

図 5　提案手法の処理フロー

設定した．そのため今回は，位置推定モジュールの誤差や周波数帯域制限に関する方向推定誤差の評価は行わなかった．また，全体の処理の流れは図5の通りである．

5.1 伝達関数データベースの作成

伝達関数データベースは，スピーカを音源として測定したインパルス応答をFFTにより周波数解析作成した．スピーカは，GENELEC社の1029A を用いた．スピーカは部屋の中央(x=3m，y=2m)に配置し，図6の0°方向から反時計回りに15°刻みで345°まで回転させた(計24 方向)．インパルス応答収録時のサンプリング周波数は16kHz，音源信号は信号長2^{14}のTSP信号とした．伝達関数は，このインパルス応答の初

期部分(1024 点)を切り出し，FFTを行うことにより計算した．

5.2 評価用音声の収録

話者は男性1 名で，部屋の中央(x=3m, y=2m)で0 °，90 °，180 °，270 °の4方向に向き，「あ，い，う，え，お」と発話した音声を収録した．

5.3 推定精度の評価

提案法の有効性を示すため，従来法と提案法の発話方向推定結果の誤差を評価した．また提案法で導入した2つの処理の寄与度について検討するため，それぞれの処理を導入した場合と導入しない場合の処理について評価した．全ての処理の導入しない処理が従来手法であり，全ての処理を導入した処理が提案手法である．図7は，各処理で推定した発話方向推定結果の誤差と標準偏差を示している．各図の棒グラフは，左から，処理なし，周波数マスク付ヒストグラムのみ導入，周波数マスク付ヒストグラムと時間区間選択を導入した場合となっている．処理なしの場合では平均誤差35 °となっているのに対し，今回の提案手法を付加した場合は平均誤差7 °となった．これにより，周波数選択，時間区間選択共に推定精度向上に有効であることがわかる．

また，この誤差の絶対的な大きさも，今回のデータベースが角度15 °おきのものであり，推定結果も15 °間隔で求めたことから，推定の量子化角度以内の誤差であったと評価できる．

5.4 発話方向の実時間推定システム

本システムの実際の利用状況は図8の通りである．従来法では図7のように棒グラフのみで推定精度を示していたものが，図8のようにグラフィックかつリアルタイム表示が可能となった．

6. まとめ

本稿では，発話方向を高精度に推定できる手法を提案した．提案手法は，拡張BF に基づく従来の方向推定法に対し(1)周波数選択による雑音帯域の除去，(2)発話区間検出時の残響部分の削除という2つの処理の追加と改良を行うことにより，高精度な方向推定を実現した．また提案手法を元に(3)実時間で動作する発話方向推定システムを開発した．推定誤差を評価した結果，従来手法で35 °あった方向推定誤差が，提案手法では7 °程度に低減できることがわかった．

マイクロホン数の削減や主要周波数の選択などによる計算量の削減，画像情報の取得によるロバスト化などが今後の課題である．

参 考 文 献

[1] K. Nakadai et al., “Active audition for humanoid,” *AAAI 2000*. AAAI, 2000, pp. 832–839.

[2] H. Nakajima et al., “High performance sound source separation adaptable to environmental changes for robot audition,” IROS 2008. IEEE/RSJ, 2008, pp. 2165–2171.

[3] K. Nakadai et al., “Sound Source Tracking with Directivity Pattern Estimation Using a 64ch Microphone Array,” *IROS 2005*. IEEE/RSJ, 2005, pp. 192–202.

[4] 菊間, ”アレーアンテナによる適応信号処理,” 科学技術出版, 1999.

[5] 醍醐他, ”室内残響を考慮した大規模マイクロホンアレーによる発話方向の推定,” **日本音響学会秋期研究発表会**, *ASJ*, 2007, pp.627–630.

[6] 中島, ”音源の方向を推定可能な拡張ビームフォーミング,” **日本音響学会秋期研究発表会**, *ASJ*, 2005, pp. 619–620.

[7] 伊藤, 水島, ” 音声／非音声識別機能を有する環境騒音抑圧法の検討,” 信学技法, EA95-59, pp.17-25, 1995.

[8] 鳥井他, ”人・ロボットインタラクションシステムの為のミドルウェアの開発, ” SI-2006, SICE, 2006

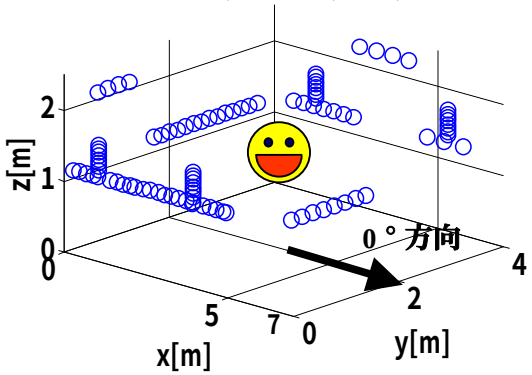

図 6　マイクロホンの配置図

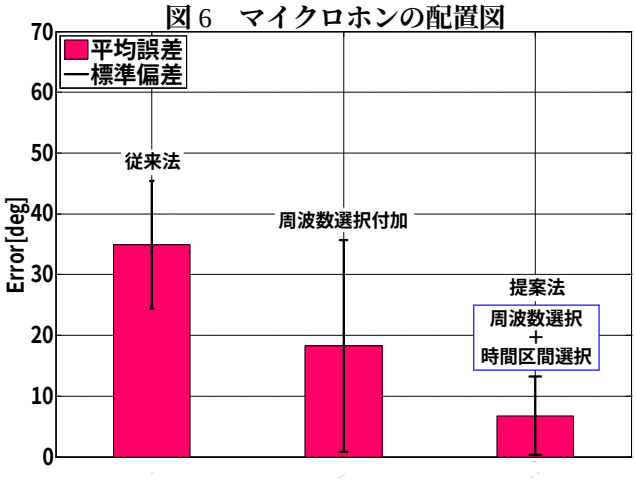

図 7　推定精度の評価結果

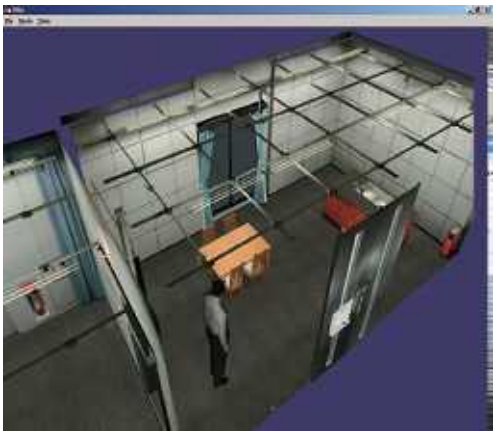

図 8　実時間発話方向推定システムの例