社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE

大規模マイクロホンアレイによる発話方向推定の検討

菊池 慶子[1]　醍醐 徹[1]　中島 弘史[2]　中臺 一博[2]　長谷川 雄二[2]　金田 豊[1]

1 東京電機大学工学部 〒101-8457 東京都千代田区神田錦町 2-2
2（株）ホンダ・リサーチ・インスティチュート・ジャパン 〒351-0188 埼玉県和光市本町 8-1
E-mail: 1{08gmc09@ms, 07gmc09@ms, kaneda@c}.dendai.ac.jp
2{nakajima, nakadai, yuji.hasegawa}@jp.honda-ri.com

あらまし　本稿では，大規模マイクロホンアレイを用いた発話方向推定について報告する．筆者らは，先に音源の指向特性に基づいたビームフォーミング法を提案し，スピーカや人の発話方向推定について報告した．しかし，この手法において，ビームフォーマーの設計に用いる伝達関数は，対象となる音源の伝達関数を用いないと，性能が劣化するという問題点があった．また，発話区間の検出（VAD）も手動で行われているという問題点があった．前者の問題は異なる音源の伝達関数の位相差が主な原因と考えて，振幅伝達特性のみを利用したヒストグラムによる発話方向推定手法を提案した．後者については，非発話区間など信頼できない時間周波数特徴量のみを自動的にマスクするために，内積値に基づく音声周波数成分検出と自己相関を利用した発話区間検出を導入した．評価実験を通じて，スピーカの伝達関数を用いた場合でも，提案手法により人間の発話方向推定性能が大きく向上することを示した．

キーワード　発話方向検出，発話区間検出，マイクロホンアレイ

Estimation of sound source orientation using a 96 channel microphone array

Keiko KIKUCHI[1]　Tohru DAIGO[1]　Hirofumi NAKAJIMA[2]
Kazuhiro NAKADAI[2]　Yuji HASEGAWA[2]　Yutaka KANEDA[1]

1 Faculty of Engineering, Tokyo Denki University　2-2 Nishikicho, Chiyoda-ku, Tokyo, 101-8457 Japan
2 Honda Research Institute Japan Co., Ltd.　8-1 Honcho, Wakoh-shi, Saitama, 351-0188 Japan
E-mail: 1{08gmc09@ms, 07gmc09@ms, kaneda@c}.dendai.ac.jp
2{nakajima, nakadai, yuji.hasegawa}@jp.honda-ri.com

Abstract　This paper addresses sound source orientation estimation using a 96ch microphone array. We proposed a beam-forming method with estimation of sound source directivity, and reported orientation estimation of a speech source such as a loudspeaker or an actual human. However, in this method, a transfer function to design a beam-former should be the same as that of target sound source. Otherwise the performance deteriorated due to a mismatch between these two transfer functions. In addition, voice activity detection (VAD) was manually performed. To solve the former, we proposed amplitude-based orientation estimation using a histogram to relax the effect of the mismatch problems mainly caused by phase errors and outliers. For the latter, speech frequency component detection based on inner product and automatic VAD based on auto-correlation are introduced to form a frequency-temporal masking pattern.　Preliminary experiments showed that sound source orientation estimation with automatic VAD for actual human voices drastically improved even when using a loudspeaker-based transfer function.

Keyword　Sound orientation estimation,　Voice activity detection (VAD),　Microphone array

1. はじめに

マンマシンインタフェースや人 - ロボットインタラクションへ適用することを目的とした音響信号処理では，これまで音源（人）の定位，音声の分離抽出といった技術に注目が集まってきた．しかし，発話方向推定もこうした用途に利用可能な重要な技術である．

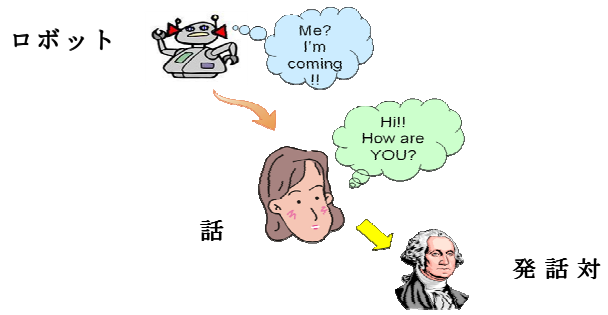


図 1 発話方向の重要性を示す例

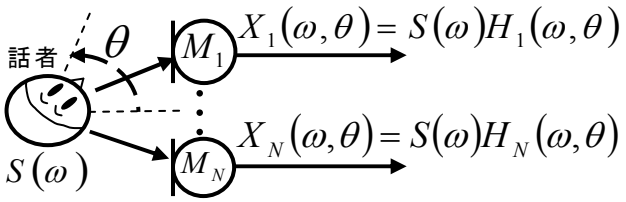


図 2 発話方向推定モデル

実際に発話方向が推定できれば様々な応用が考えられる．例えば，図 1 ではロボットが自分に話しかけられているかどうかを判断する必要があり，そのために発話方向（話者の顔の向き）の推定が有効である．同様に音声認識を利用した自動券売機などの装置や，テレビ会議で誰と誰が話しているかを検出する際に利用できる．

発話方向推定は画像処理を利用して行うこともできる [1,2]．しかし，画像情報は必ずしも利用できるわけではない．そこで，本稿ではマイクロホンアレイから得られる多チャンネル音響信号から発話方向を推定する手法を提案し，その有効性を検証する．

従来，多チャンネル音響信号を利用した音源位置推定の研究例は存在するが[3]，発話方向推定の報告例は少ない[4,5,6]．そこで我々はビームフォーミング（以下 BF）を発話方向にも拡張した手法を提案して,こうした問題への解決を目指した研究を行っている[5]．この手法は，それぞれ音源の位置と向きが異なる伝達関数群を元に BF を設計し，その出力が最大となる BF の焦点方向を発話方向として推定する．

これは，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，BF の設計および出力の計算は，それぞれ伝達関数データベースの作成および作成データベースと入力信号との照合と見なすこともできる．筆者らは先に，室内壁面に設置した大規模マイクロホンアレイを用い，BF で用いる伝達関数として推定対象の人間の実発話に基づく伝達関数を用いることで精度が向上することを示した[6]．しかし，この手法（従来法とする）では発話者全員の伝達関数を測定する必要があり実用的とはいえない．また発話区間検出に必要なパラメータを発話毎に手動で調整する必要があった．

そこで本報告では，スピーカを用いて測定した伝達関数に基づく発話方向推定処理の改良とパラメータの調整を必要としない発話区間検出の導入を行う．

2. 発話方向の検出方法（従来法）

2.1. 発話方向を考慮した伝播モデル

図 2 にマイクロホンアレイと受音信号の関係を示す．図 2 において話者の位置 (x,y) は固定，$S(\omega)$ は発話音声の周波数特性，$M_1 \sim M_N$ は N 個のマイクロホン，$H_1(\omega,\theta) \sim H_N(\omega,\theta)$ は話者が θ 方向を向いている時の話者 - マイクロホン間の伝達関数である．また，$X_1(\omega,\theta) \sim X_N(\omega,\theta)$ は各マイクロホンでの受音信号を表しており，$X_i(\omega,\theta) = S(\omega)H_i(\omega,\theta)$ と表すことができる．従って，各変数をベクトル化すると，下記のように表現できる．

$$\mathbf{h}(\omega,\theta) = [H_1(\omega,\theta),\cdots,H_N(\omega,\theta)]^T \quad (1)$$

$$\begin{aligned}\mathbf{X}(\omega,\theta) &= [X_1(\omega,\theta),\cdots,X_N(\omega,\theta)]^T \\ &= [S(\omega)H_1(\omega,\theta),\cdots,S(\omega)H_N(\omega,\theta)]^T \\ &= S(\omega)\mathbf{h}(\omega,\theta)\end{aligned} \quad (2)$$

但し，T は転置である．

2.2. 伝達関数データベース

スピーカを音源として測定した伝達関数には，スピーカの周波数特性が含まれる．スピーカの周波数特性を含まない伝達関数を得るため，次式により伝達関数 $\mathbf{h}(\omega,\theta)$ を各周波数および各方向で正規化した．

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

（H ：エルミート転置）

この $\mathbf{h}_0(\omega,\theta)$ を全ての周波数と方向について予め計算しデータベースとして記録した．

2.3. データベースを元にした発話方向の推定

話者が発話方向 $\hat{\theta}$（未知）に向いて発話した時の話者 - マイクロホン間の伝達関数をベクトル化したものを $\mathbf{h}(\omega,\hat{\theta}) = [H_1(\omega,\hat{\theta}),\cdots,H_N(\omega,\hat{\theta})]^T$ とおく．また，受音信号ベクトルを正規化したものを $\mathbf{X}_0(\omega,\hat{\theta})$ とおくと，$\mathbf{X}_0(\omega,\hat{\theta})$ は以下のように記述できる．

$$\mathbf{X}_0(\omega,\hat{\theta}) = \frac{S(\omega)\mathbf{h}(\omega,\hat{\theta})}{|S(\omega)|\|\mathbf{h}(\omega,\hat{\theta})\|} = \frac{S(\omega)}{|S(\omega)|}\mathbf{h}_0(\omega,\hat{\theta}) \quad (4)$$

（$\hat{\theta}$ は発話方向）

式（3）と式（4）の内積の絶対値

$$\begin{aligned}\left|\mathbf{h}_0(\omega,\theta)^H\mathbf{X}_0(\omega,\hat{\theta})\right| &= \left|\frac{S(\omega)}{|S(\omega)|}\mathbf{h}_0(\omega,\theta)^H\mathbf{h}_0(\omega,\hat{\theta})\right| \\ &= \left|\mathbf{h}_0(\omega,\theta)^H\mathbf{h}_0(\omega,\hat{\theta})\right|\end{aligned} \quad (5)$$

は，$\mathbf{h}_0(\omega,\hat{\theta})$ と $\mathbf{h}_0(\omega,\theta)$ の方向類似度を表すことになる．本稿では，これを周波数で平均したものを $\mathbf{C}(\theta)$ と表す．

$$\mathbf{C}(\theta)=\sum_{\omega} w(\omega,t)\left|\mathbf{h}_0(\omega,\theta)^H \mathbf{X}_0(\omega,\hat{\theta})\right| \qquad (6)$$

但し，$w(\omega,t)$は時間・周波数に対する重み関数であるが，従来法では$w(\omega,t)=1$であった．また，これらの処理は短時間フレーム単位で行われるため，入力信号$\mathbf{X}_0(\omega,\theta)$，内積値$\mathbf{C}(\theta)$は時間の関数でもあるが，簡略化のため変数$t$は省略した．従来法では，この$\mathbf{C}(\theta)$の最大値を取る$\theta$を発話方向$\hat{\theta}$の推定値とする．

以上説明した従来法の処理ブロック図を図 3 に示す．

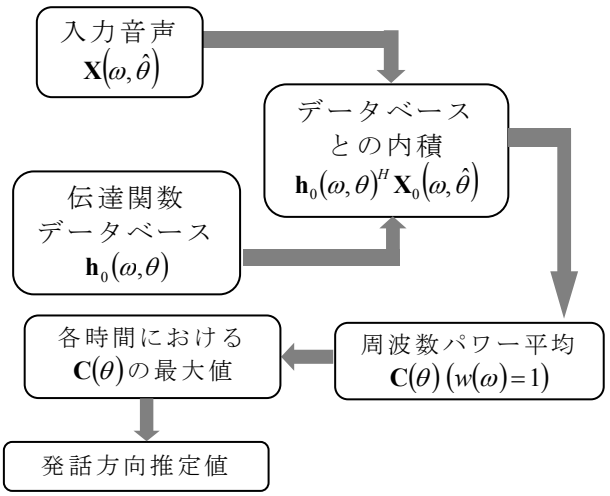


図 3 従来法における処理ブロック図

3. 方向推定処理の改良

本章では，スピーカを用いた伝達関数に基づく発話方向推定の精度向上を目的として，従来法に対し，新たに 4 つの処理を加えた手法を提案する．

3.1. 振幅特性の内積

従来の推定方法では，式（6）に示すように，伝達関数の複素成分（振幅情報と位相情報）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情報は系の変動に敏感で変化しやすい．例えば，話者の身長や位置ズレによって位相は変化し，この位相変化は推定誤差の原因となる．これを防ぎ，位置ズレにも頑健に対応するために，提案法では伝達関数および受音信号の振幅成分のみの内積計算を行った．

具体的には，伝達関数の振幅成分をベクトル化したものを

$$\mathbf{h}_a(\omega,\theta)=\left[|H_1(\omega,\theta)|,\cdots,|H_N(\omega,\theta)|\right]^T \qquad (7)$$

とし，これを正規化したものを

$$\mathbf{h}_{a_0}(\omega,\theta)=\frac{\mathbf{h}_{a_0}(\omega,\theta)}{\sqrt{\mathbf{h}_{a_0}(\omega,\theta)^H\mathbf{h}_{a_0}(\omega,\theta)}} \qquad (8)$$

と表す．また，各受音信号の振幅成分をベクトル化したものを

$$\mathbf{X}_a(\omega,\hat{\theta})=\left[|X_1(\omega,\hat{\theta})|,\cdots,|X_N(\omega,\hat{\theta})|\right]^T \qquad (9)$$

とおき，これを正規化したものを

$$\mathbf{X}_{a_0}(\omega,\hat{\theta})=\frac{\mathbf{X}_a(\omega,\hat{\theta})}{\sqrt{\mathbf{X}_a(\omega,\hat{\theta})^H\mathbf{X}_a(\omega,\hat{\theta})}} \qquad (10)$$

とする．式（6）と同様に周波数平均$\mathbf{C}_a(\theta)$は次式

$$\mathbf{C}_a(\theta)=\sum_{\omega} w(\omega)\mathbf{h}_{a_0}(\omega,\theta)^H\mathbf{X}_{a_0}(\omega,\hat{\theta}) \qquad (11)$$

で計算し，この$\mathbf{C}_a(\theta)$の最大値を取るθを発話方向$\hat{\theta}$の推定値とする．

3.2. 発話区間検出

従来法では目視による発話区間検出を行ってきたが，今回は，入力信号の自己相関関数を利用した発話区間検出方法を導入した．この方法は音声の母音の周期性を根拠とするもので，周期性雑音が無い環境では有効に動作が期待できる．

具体的には，まず，基準とするチャンネルで受音した信号を分析窓長（1024 点）で切り出し，その自己相関関数$\phi(\tau)$を計算する．信号が周期性を有する場合，$\phi(\tau)$はその周期に対応するτにピークを有する．$\phi(\tau)$が最大となる$\tau=0$のピークを除き，$\tau>0$の範囲における$\phi(\tau)$のピークの最大値が閾値（αと設定）を超えた場合に，周期性を持つ信号であるとし，発話区間と判定した．なお，今回は$\phi(\tau)$がβ以下の値をとった時間以降の最大値を「$\tau>0$の範囲における$\phi(\tau)$のピークの最大値」と判定した．

$$w_t(\omega,t)=\begin{cases}1 & (\text{発話区間})\\ 0 & (\text{非発話区間})\end{cases} \qquad (12)$$

パラメータαおよびβは，実験的に定め，今回はα=0.5，β=-0.2 とし，すべての条件や発話でこの値を利用した．

ただしこのままの方法では，音声の残響部分も含めて発話区間と推定してしまう．残響は音源からの直接音とは異なり，発話方向情報を含まないので，推定の妨害成分となる．よって，残響部分を発話区間から除外することが重要である．最適な除外方法の検討は今後の課題とし，今回は検出した発話区間（式(12)）の最後尾 2 フレーム（時間データ 1024 点に相当）を除外したものを最終的な発話区間とした．

3.3. 周波数マスク

音声があまり含まれていない周波数成分を利用すると推定誤差の原因となる．そこで，そのような周波数成分を計算から除外した．具体的には，高い内積値を持つ周波数成分は音声成分であると仮定して式(13)のようなバイナリ値の重み関数$w_\omega(\omega,t)$（周波数マスク）を設定した．

$$w_\omega(\omega,t)=\begin{cases}1 & (p_{in}(\omega,t)/p_{mean}(\omega)\geq p_{threshold})\\ 0 & (p_{in}(\omega,t)/p_{mean}(\omega)\leq p_{threshold})\end{cases} \qquad (13)$$

ただし，$p_{in}(\omega,t)$はある時刻ある周波数の内積値，$p_{mean}(\omega)$はその周波数の時間軸方向に平均した内積値である．また，$p_{threshold}$は閾値で，今回は$p_{threshold}=1.5$と定めた．このマスクと，推定結果のデータを乗算すると，閾値以上のデータは残り，閾値以下のデータは0となるので，有効な周波数成分のみを抽出できる．

3.4. ヒストグラムの導入

従来法では，各周波数ωおよび各方向θで算出した方向類似度を周波数方向に平均した$\mathbf{C}(\theta)$から発話方向を推定していた（式(6)）．しかし，方向類似度（内積値）の絶対的大きさは周波数に依存するため，平均の結果も絶対的な内積の大きさで加重される．これに対して，周波数マスクで音声成分と判定された周波数成分については，均等加重で評価を行うという方法が考えられる．

具体的には，平均した方向類似度$\mathbf{C}(\theta)$ではなく，各周波数について方向類似度が最大となる方向を算出し，その数をカウントしたヒストグラム$\mathbf{C}_{Hist}(\theta)$

$$\mathbf{C}_{Hist}(\theta)=\sum_{\omega} w_{\omega}(\omega,t)w_t(\omega,t)U(\omega,\theta) \qquad (14)$$

$$U(\omega,\theta)=\begin{cases}1 & \theta_m=\arg\max_{\theta}[a(\omega,\theta)]\\ 0 & \text{otherwise}\end{cases}$$

$$a(\omega,\theta)=\left|\mathbf{h}_0(\omega,\theta)^H\mathbf{X}_0(\omega,\hat{\theta})\right|$$

をが最大値を取るθを発話方向$\hat{\theta}$の推定値とした．ここで$\arg\max_{\theta}[\ ]$は，括弧内の関数が最大となるθを示す．

以上3.1～3.4節で説明した4つの処理を含んだ提案法の処理ブロック図を図4に示す．

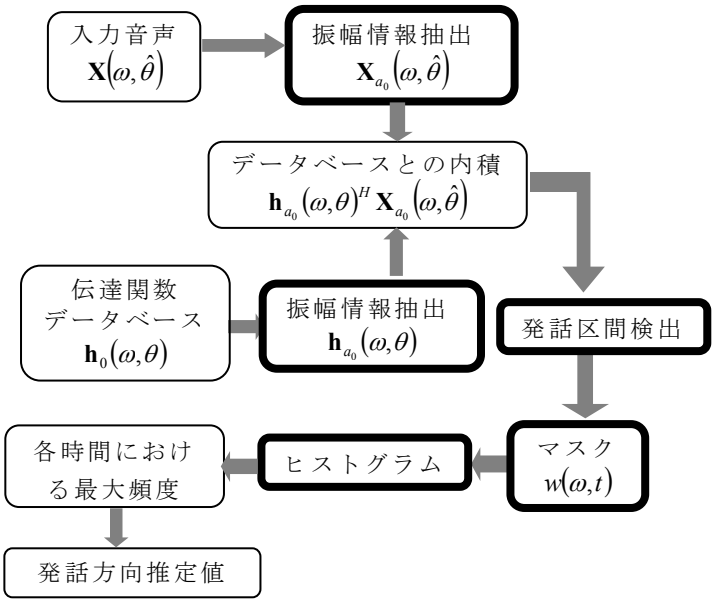


図4 提案法の処理ブロック図

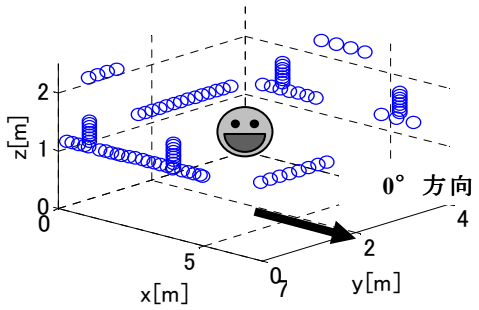


図5 マイクロホンの配置

4. 評価実験

4.1. 伝達関数データベースの作成

実験は広さ7m×4m，高さ3.5m，残響時間が約230msで，壁三面が吸音素材，一面がガラス面の実験室で行った．マイクロホン数は96で，図5に○印で示すように室内の壁面に配置した．マイクロホン配置は，本研究目的に特化させたものではなく，水平および垂直方向に概ね等間隔に配置したアレイを周囲四壁に分散した配置となっている．

伝達関数データベース作成に用いたスピーカはGENELEC 1029Aを用いた．スピーカは部屋の中央(x=3m，y=2m)に配置し，図5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）．測定は，サンプリング周波数16kHz，信号長2^{14}のTSP信号を利用してインパルス応答を測定し，そのインパルス応答の初期部分（1024点）を切り出してフーリエ変換を施し，伝達関数を得た．

4.2. 評価用音声の収録

伝達関数データベースの作成時に利用した実験室で行った．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．

4.3. 方向推定の処理結果

実験は，3章で説明した4つの処理（表1）に関して，それぞれの有効性を検証することを目的とした．

図6には，受音した信号より求めたスペクトログラムを示す．図より音声の成分は8kHzまで含まれており，また，500Hz以下の低周波域には雑音が存在することがわかる．

図7は検出された発話区間を示す．今回は残響の影響を受けていそうな部分を多めに除外したので，図6で見られる，音声の（残響を含めた）区間より小さな区間となっている．

図8は，発話方向推定の元となる各時間および各周波数における方向類似度（内積）の最大値を表した図である．図8を図6と比較すると，2kHz以上の高周

波数域では，音声の存在する区間で内積値は大きくなっていることがわかる．しかし 1kHz 以下の低周波数帯域では，音声区間と非音声区間によらず内積の値に差が見られず，この帯域を利用することは誤差の原因となる可能性が考えられる．

図 9 は，3.3 節で述べた周波数マスクの結果であり，白い箇所が $w_\omega(\omega,t)=1$，黒い箇所が $w_\omega(\omega,t)=0$ である．図より周波数マスクが，単に音声が強い部分を抽出しているのではなく，方向類似度の差が高く，方向推定に有効な高周波成分を多く抽出できることがわかる．

図 10 は，音声区間検出および周波数マスク処理を導入したうえで，パワー平均により求めた方向類似度ー時間特性である．実際の音源方向は 270° であり，概ね正しい方向で方向類似度が高い値をとっていることがわかる．

図 11 は，表 1 に示す 4 つの処理の有無を組み合わせた計 16 種類の場合の，発話方向推定結果の誤差と標準偏差を示している．図 11（a）は，手動調整により得られた発話区間を用いた結果，図 11（b）は 3.2 節で述べた発話区間検出法を用いた結果である．各図において濃色の棒は「振幅と位相を用いた内積処理」の結果，淡色の棒は「振幅のみを用いた内積処理」の結果を表している．さらに各図の左から，「付加処理なし」「周波数マスク処理を行った場合」「ヒストグラム処理を利用した場合」「周波数マスクとヒストグラムの両方の処理を行った場合」を表している，

図 11(a)と(b)を比較して，これら 2 つはほぼ同程度の誤差の大きさとなっている．このことより，提案する自動発話区間検出法は，手動調整とほぼ同程度の性能を達成できることが確認された．

次に，各図における濃色棒と淡色棒を比較すると，すべての条件で淡色棒の方が誤差が小さくなっていることがわかる．これより，「振幅と位相の両方」を用いる処理（濃色棒）より「振幅のみ」を用いた処理（淡色棒）が優位であることが示された．

次に図 11（a）（b）における 4 条件を比較してみると，いずれの図でも「周波数マスクとヒストグラムの両方の処理を行った場合」が最も誤差が小さくなっていることがわかる．このことより，今回改良のために提案した周波数マスクとヒストグラムの 2 つの処理は，共に推定誤差を低減する効果があることが確認できた．

従来法である，図 11（a）の「付加処理なし」の「位相＋振幅」の条件では，平均誤差は約 30° であるのに対して，改良法である図 11（b）の「周波数マスクとヒストグラムの両方の処理」を用いると，平均誤差が約 4°，標準偏差が約 7° と大幅な性能改善を達成することができた．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. まとめ

本報告では，室内に設置された大規模マイクロホンアレイを用いた発話方向の推定方法に関してその性能向上の検討を行った．従来法は，受音信号とあらかじめ求めた伝達関数データベースとの周波数領域の内積の大きさに基づいた方法であるが，今回は，1) 内積を振幅成分のみの内積に変更，2) 残響の影響部分を排除した発話区間検出の導入，3) 音声が多く含まれる周波数成分の抽出処理（周波数マスク），4) 最終決定に周波数成分ごとの方向推定ヒストグラクの導入，の 4 つの処理を提案した．実際の室内における実験の結果，これら 4 つの処理を加えた改良型処理では，従来法の誤差 30° を大きく下回る約 4° の平均推定誤差で，発話方向推定を行うことができた．

今後は，発話方向検出処理の改良，周波数マスクの改良やマイクロホン数の削減，画像情報との統合などにより，さらに実用的で高精度な推定を検討したい．

文　献

[1]松本 吉央, 怡土 順一, 竹村 憲太郎, 小笠原 司, ”リアルタイム顔・視線計測システムの開発と知的インタフェースへの応用”, 情報処理学会論文誌コンピュータビジョンとイメージメディア, vol.47, No.SIG 15 (CVIM16), pp.10-21, 2006.

[2]大塚 和弘, ”非言語行動の観測に基づく複数人物の会話シーン分析人工知能学会”, SIG-SLUD-A602-01, pp. 1-6, 2006.

[3]H. F. Silverman and W. R. Patterson, ”The Huge Microphone Array”, Technical report, LEMS, BROWN University, 1996.

[4]P. C. Meuse and H. F. Silverman, “Characterization of talker radiation pattern using a microphone array” , ICASSP94, IEEE, vol. 2, pp. 257-260, 1994.

[5]K. Nakadai, H. Nakajima, K. Yamada, Y. Hasegawa, T. Nakamura and H. Tsujino, ”Sound Source Tracking with Directivity Pattern Estimation Using a 64ch Microphone Array”, IEEE/RSJ Intl. Conference on Intelligent Robots and Systems (IROS 2005), pp. 196-202, 2005.

[6] 醍醐 徹, 菊池 慶子, 中島 弘史, 中臺 一博, 長谷川 雄二, 金田 豊, ”室内残響を考慮した大規模マイクロホンアレイによる発話方向の推定”, 音講論集, pp.627-630. 2007.

表 1 提案する 4 つの処理

処理	内容の説明
振幅特性の内積	3 章 1 節
発話区間検出	3 章 2 節
周波数マスク	3 章 3 節
ヒストグラム*	3 章 4 節

*ヒストグラムなしの場合は周波数パワー平均を用いた.

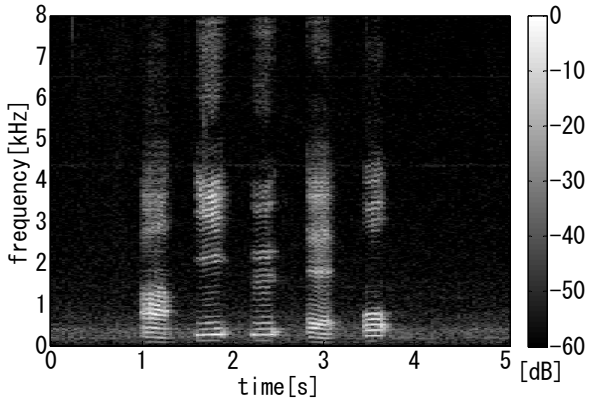


図 6 受音した音声のスペクトログラム

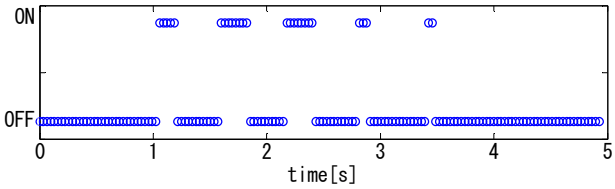


図 7 発話区間検出結果

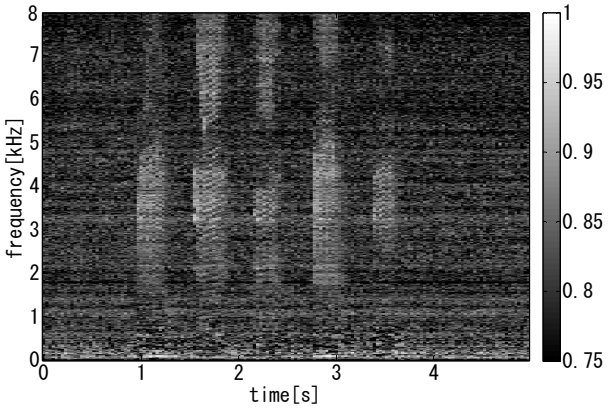


図 8 時間 - 周波数における方向類似度の最大値

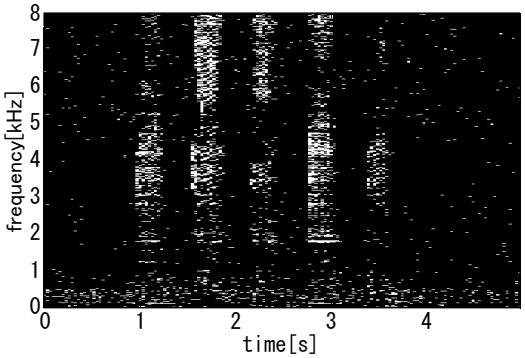


図 9 周波数時間マスク $w_{\omega}(\omega,\theta)$

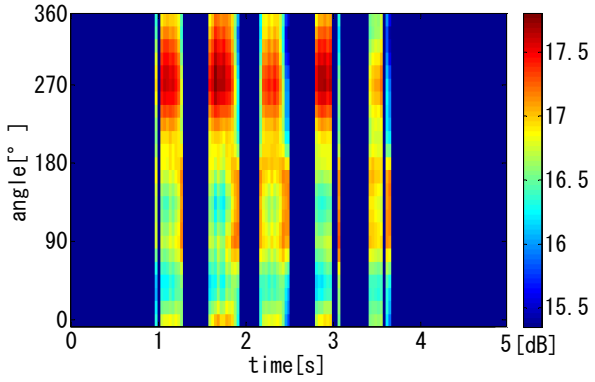
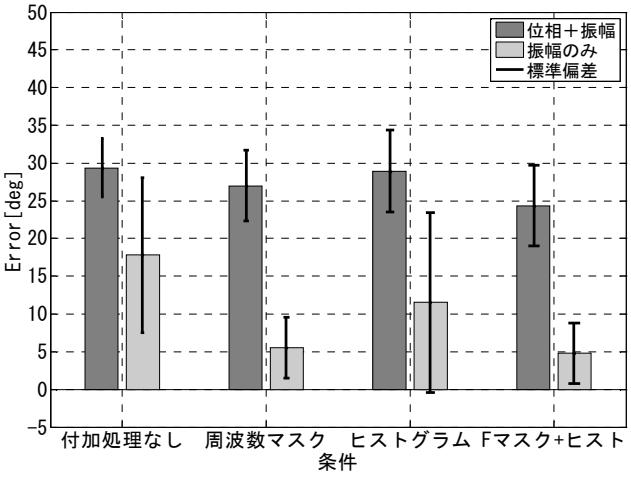
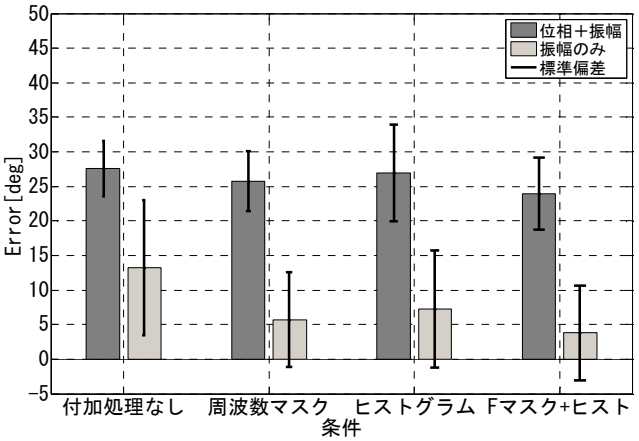


図 10 方向類似度－時間特性



（a）手動発話区間検出を用いた場合



（b）自動発話区間検出を用いた場合

図 11 発話方向推定結果の誤差と標準偏差