大規模マイクロホンアレイを用いた発話方向実時間推定*

☆春原 卓也[1], 中島 弘史[2], 中臺 一博[2], 金田 豊[1]
1 東京電機大・工 2 (株) ホンダ・リサーチ・インスティチュート・ジャパン

1 はじめに

近年, 身近な存在となっているロボットが, 人間社会に溶け込むためには, 人と同様に音声によるスムーズなコミュニケーションが必要である. そのため, 人の聴覚と同様の機能を有する聴覚システムの構築を目指した研究 (ロボット聴覚) が行われている. この研究では, 音源定位, 音源分離, 音声認識機能が主な対象とされている. しかし, 人・ロボット間のスムーズなコミュニケーションを実現するためには, これらの機能だけで十分とは言えず, 音源 (話者) の向きを推定する機能 (発話方向推定) も重要である. 図 1 に, 発話方向推定が必要となる例を示す. 図中の女性話者はロボットではなく, 別の男性に話しかけている. しかし, 発話方向推定を有さないロボットは, 女性の発話に誤って反応してしまい意思通りのコミュニケーションを取ることはできない. つまり, ロボットはこの女性の発話対象が自分ではないことを判断する必要がある.

筆者らは, 室壁に設置したマイクロホンアレイを用いた発話方向推定手法について高性能化を進めてきた. 本稿では, 2009 年に構築した実時間発話方向推定システムを改良し, より高精度なシステムを構築したので報告する.

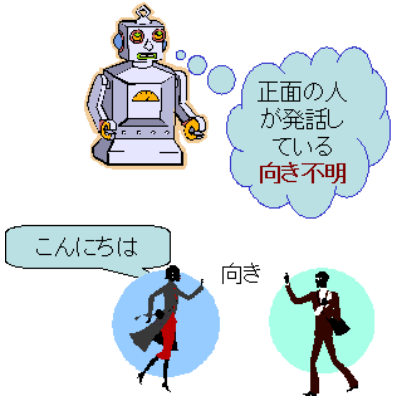


Fig. 1 発話方向推定の重要性

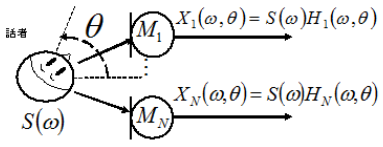


Fig. 2 発話方向推定モデル

2 発話方向推定システム

2.1 発話方向推定に関する従来研究

音響処理を利用した発話方向推定法の従来研究としては, 中島らによる拡張ビームフォーミング (BF) に基づく手法 [1] や丹羽らによる空間相関行列の固有値に着目した手法 [2] がある. 筆者らは, 拡張 BF に基づく発話方向推定法をもとにより高精度でロバストな推定法を構築する研究を進めてきた. 図 3 に提案してきた手法を示す. 始めに, 室壁に設置した 96ch 大規模マイクロホンアレイを用いて拡張 B F 法により発話方向の推定が可能であることを実証した (2007 年法)[3]. なお, 本稿ではこの手法をベース手法とする. しかし, この手法は 推定対象者全員の伝達関数を予めを測定する必要があるという問題点があった. この問題を解決するために, 発話方向推定処理に, 振幅成分の抽出, 発話区間検出処理, 周波数マスク付きヒストグラムの導入の 3 つの処理を加えることにより, スピーカを用いて測定した伝達関数を使用した場合でも高精度推定を行えることを示した (2008 年法)[4]. しかし, この手法は計算量が膨大であり, 実時間性が低いという問題点があった. そこで, 計算量がシンプルである 2007 年法をもとに実時間で動作する推定システムを開発した (2009 年法)[5]. しかし, この推定システムは使用する周波数帯域が限定されていることに起因すると思われる精度低下が見受けられること, およびこの精度低下が本当に周波数帯域不足によるものなのか, またどの程度の精度低下が起きているのかに関する詳細な評価が行われていないという問題があった.

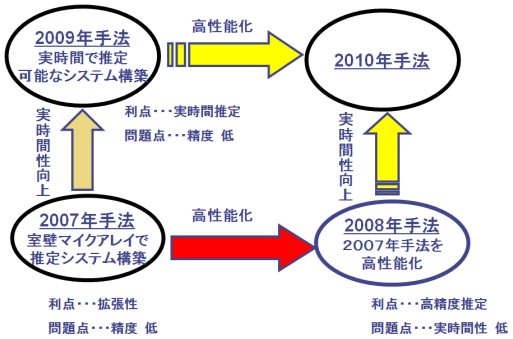


Fig. 3 提案した推定手法

2.2 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 (BF) 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)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. 中島らは位置だけでなく, 音源の向きに対しても焦点を形成する BF を設計できることを示した (拡張 BF)[1]. これにより, 音源の位置だけでなく, 音源の向きも推定できることを示した (拡張走査 BF).

遅延和による拡張走査 BF は, 伝達関数と発話者の音声信号との内積値を算出し, その最大値によって方向を推定することと等価である. すなわち, 拡張 BF の設計, および走査は, 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音声信号の照合に相当する.

2.3 発話方向を考慮した伝播モデル

図 2 に N 素子のマイクロホンアレイを用いた発話方向推定モデルを示す. $S(\omega)$ は発話音声の周波数特性, M_k は k 番目のマイクロホン, $H_k(\omega,\theta)$ は話者が方向 θ 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. 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 $X_k(\omega)$ は,

$$\mathbf{X}_k(\omega,\theta) = S(\omega)\mathbf{H}_k(\omega,\theta) \quad (1)$$

なお, 実環境において受音信号には雑音が含まれるが今回は簡略化のために省略した. 各変数をベクトルで表現

*Real-time speaker orientation estimation using a 96 channel microphone array. by HARUBARA, Takuya[1], NAKAJIMA Hirofumi[2], NAKADAI Kazuhiro[2], KANEDA Yutaka[1](1 Tokyo Denki University, 2 Honda Research Institute Japan)

すると

$$\begin{aligned}\mathbf{h}(\omega,\theta) &= [H_1(\omega,\theta),...,H_N(\omega,\theta)]^T \quad (2)\\ \mathbf{X}(\omega,\theta) &= [X_1(\omega,\theta),...,X_N(\omega,\theta)]^T \\ &= [S(\omega)_1(\omega,\theta),...,S(\omega)H_N(\omega,\theta)]^T \\ &= S(\omega)\mathbf{h}(\omega,\theta) \quad (3)\end{aligned}$$

と表すことができる. 但し, T は転置である.

2.4 伝達関数データベース

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.

$$\mathbf{h}_0(\omega,\theta) = \frac{\mathbf{h}(\omega,\theta)}{\sqrt{\mathbf{h}(\omega,\theta)^H\mathbf{h}(\omega,\theta)}} = \frac{\mathbf{h}(\omega,\theta)}{|\mathbf{h}(\omega,\theta)|} \quad (4)$$

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を含まない伝達関数を得ることができる. この $h_0(\omega,\theta)$ を全ての周波数と方向に予め算出しデータベースとした.

2.5 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 θ_s(未知) に向いて発話した時の受音信号ベクトル $\mathbf{X}_0(\omega,\theta_s)$ を, 正規化したものを $\mathbf{X}_0(\omega,\theta_s)$ とおけば,

$$\mathbf{X}_0(\omega,\theta_s) = \frac{S(\omega)\mathbf{h}(\omega,\theta_s)}{|S(\omega,\theta_s)\mathbf{h}(\omega,\theta_s)|} = \frac{S(\omega)}{|S(\omega)|}\mathbf{h}_0(\omega,\theta_s) \quad (5)$$

となる. 式 (4) と式 (5) の内積の絶対値 $C_\omega(\omega,\theta)$ は

$$\begin{aligned}C_\omega(\omega,\theta) &= |\mathbf{h}_0(\omega,\theta)^H\mathbf{X}_0(\omega,\theta_s)| \\ &= \left|\frac{S(\omega)}{|S(\omega)|}\mathbf{h}_0(\omega,\theta)^H\mathbf{h}_0(\omega,\theta_s)\right| \\ &= |\mathbf{h}_0(\omega,\theta)^H\mathbf{h}_0(\omega,\theta_s)| \quad (6)\end{aligned}$$

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

$$C(\theta) = \sum_\omega C_w(\omega,\theta) \quad (7)$$

として計算し, この $C(\theta)$ が最大となる方向 θ を発話方向の推定値とする.

3 発話方向推定の高性能化

ここでは, 2008 年に菊池らによって, 改良された発話方向推定の処理手法について説明する. 具体的には, 2007 年法に対し, 新たに　振幅特性の抽出,　自己相関関数に基づく発話区間検出処理,　周波数マスク付きヒストグラムの導入の 3 つの処理を加えた方法となっている.

3.1 振幅特性の抽出

2007 年法は, 式 (7) に示すように, 伝達関数の複素成分 (振幅成分と位相成分) での内積計算を行っていた. しかし, 高周波帯域において位相成分は話者の口の高さや位置の変動などに敏感に変化しやすいという性質がある. そのため, 内積値の抽出時に, 位置の変動に対してロバストな振幅成分を抽出する処理を加える. これにより, 方向推定精度が向上すると考えられる. よって, 振幅成分による内積を利用した発話方向推定法を用いる. 伝達関数の振幅成分をベクトル化したものを

$$\mathbf{h}_a(\omega,\theta) = [|H_1(\omega,\theta)|,...,|H_N(\omega,\theta)|]^T \quad (8)$$

とし, これを正規化したものを

$$\mathbf{h}_{a0}(\omega,\theta) = \frac{\mathbf{h}_a(\omega,\theta)}{|\mathbf{h}_a(\omega,\theta)|} \quad (9)$$

とする. 同様にして, 受音信号の振幅成分をベクトル化したものを

$$\mathbf{X}_a(\omega,\theta) = [|X_1(\omega,\theta_s)|,...,|X_N(\omega,\theta_s)|]^T \quad (10)$$

とおき, これを正規化したものを

$$\mathbf{X}_{a0}(\omega,\theta) = \frac{\mathbf{X}_a(\omega,\theta_s)}{|\mathbf{X}_a(\omega,\theta_s)|} \quad (11)$$

とする. 式 (7) と同様に内積値の絶対値 $C_{aw}(\omega,\theta)$ は

$$C_{aw}(\omega,\theta) = |\mathbf{h}_{a0}(\omega,\theta)^H\mathbf{X}_{a0}(\omega,\theta_s)| \quad (12)$$

として計算され, これをもとに平均類似度 $C_a(\theta)$ を

$$C_a(\theta) = \sum_\omega C_{aw}(\omega,\theta) \quad (13)$$

として計算し, この $C_a(\theta)$ が最大となる方向 θ を発話方向の推定値とする.

3.2 発話区間検出

2007 年法では, 入力信号のレベルに対し, 手動で定めた閾値を用いて発話区間検出を行っていた. しかし, 話者や発話方向の変化で入力信号レベルも変化してしまうため, 複数の話者や発話方向に対して, 高精度に発話区間を検出するためには, その都度, 手動で閾値を最適に調整する必要があり, 実用的な処理ではなかった. 改良手法では, 入力信号のレベルではなく, 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ている. この手法の利点を以下に示す.

1. 音声のレベル変化に影響を受けない
2. 非周期性雑音に頑健
3. パラメータの手動調整が不要

処理内容を以下に示す.

1. 受音信号を分析窓長 L で切り出し自己相関関数 $\phi(\tau)$ を算出. τ は遅延時間.
2. $\phi(\tau) < \beta$ となる最小の遅延時間 τ_{min} を探す, β はパラメータ.
3. $\tau > \tau_{min}$ の範囲内で相関値が最大となる $\phi(\tau)_{max}$ を取得.
4. $\phi(\tau)_{max} \quad > \alpha$ の場合, 発話区間, それ以外では非発話区間とした. α はパラメータ.

上記の処理で判定された音声区間から 2 値の時間マスク $w_t(t)$ を生成した.

$$w_t(t) = \begin{cases} 1 & if\ 発話区間 \\ 0 & if\ 非発話区間 \end{cases} \quad (14)$$

パラメータ α および β は実験的に定め, 本稿では $\alpha = 0.5$, $\beta = -0.3$ とし, すべての条件や発話でこの値を利用している. また, 改良手法では上記の処理に加えて音声の残響部分を発話区間から除外する処理を導入し, 高精度化をはかっている.

3.3 周波数マスク付きヒストグラム

2007 年法では, 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.

$$w(\omega) = \begin{cases} 1 & (p(\omega)/p_{mean}(\omega) \geq \gamma) \\ 0 & (p(\omega)/p_{mean}(\omega) < \gamma) \end{cases} \quad (15)$$

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.

4　発話方向の実時間推定システムの構築

ここでは, 2009 年に中島らによって構築された実時間システムについて説明する. 図 4 に開発したシステムのブロック図を示す. ベース手法では MATLAB を用いたオフラインシステムであったが, 実時間システムは, C 言語で実装を行った. また, 音取得部, 音源位置・向き推定部, 推定結果 3 D表示部から構成されており, 各モジュールは MMI 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. しかし, この手法は使用周波数帯域の制限などによる精度低下が見られた. また, 詳細な評価が行われておらず精度低下の原因が不明であるという問題があった.

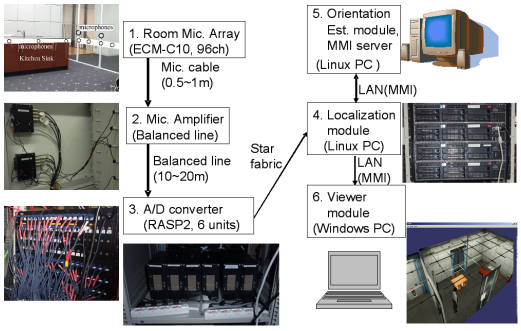
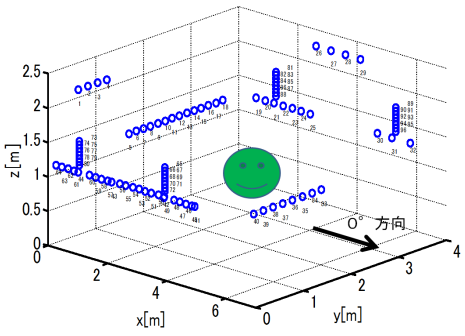
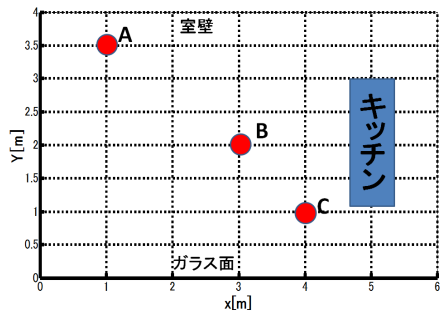


Fig. 4　実時間システムのブロック図



(a) マイクロホン配置図

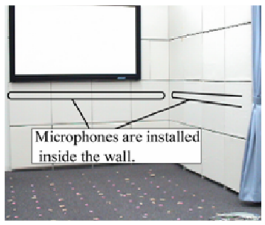


(b) 発話地点

Fig. 5　実験室の模式図

5　評価実験

4 章で記した問題点を解決するために実時間システムの評価実験 (実験 1) を行った. また, 3 章で説明したオフラインシステム改良手法の評価実験 (実験 2) を行い, 実験 1 の結果と比較した. 実験は図 5 に示すように, キッチンが設置された部屋で行った.



(a) 実験室 (壁付近)

(b) 実験室 (ガラス面付近)

Fig. 6　実験室

5.1　伝達関数データベースの構築

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高速フーリエ変換 (FFT) することにより作成した. スピーカは, GENELEC 社の 1029A を用いた. スピーカは, 部屋の 221 点に配置し, 図 5 の 0 °方向から実時間システムでは 45° 刻みで 315° まで, 改良手法では 15° 刻みで 345° まで反時計周りに回転させた (以降, DB 刻み角度と記す). 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, 音源の高さは 1.5m とした.

5.2　評価用音声の収録

実際の人の発話を録音し, 評価用音声とした. 話者は男性 1 名で, 口の高さを伝達関数測定時のスピーカと同じ高さ (1.5m) と伝達関数測定時のスピーカよりも 20cm 高い場合 (1.7m) とし, 発話位置を図 5 に示すように部屋の壁付近 (x=3.5m, y=1m), 部屋の中央 (x=3m, y=2), ガラス面付近 (x=4, y=1m) の 3 地点 (本稿ではそれぞれを地点 A, B, C, とする) で 0°, 90° , 180°, 270° の 4 方向に向き「あ, い, う, え, お」と発話した.

5.3　実験結果

5.3.1　実時間システムの詳細評価 (実験 1)

発話者の口の高さ, 発話位置の変動による推定精度について調査するために口の高さを 2 種類, 発話位置 3 地点で評価した. 図 7 は高さと位置を変化させたときの発話方向推定結果の誤差と標準偏差を示す. なお, 横軸は各測定地点, 縦軸が推定誤差である. 図 7 は口の高さが伝達関数測定時のスピーカと同じ場合 (以降, 「伝達関数と同じ」と記す), 口の高さが伝達関数測定時のスピーカと違う場合 (以降, 「伝達関数と違う」と記す) の推定誤差を示す. まず話者の口の高さの変動による影響, 次に発話位置の変動による影響について考える.

1. 話者の口の高さの変動による影響
   図 7 より話者の口の高さが伝達関数と違う場合, 推定誤差は A 地点では 90° 以上, B, C 地点でも 60° 以上である. しかし, 話者の口の高さが伝達関数と同じ場合 A 地点では約 30°, C 地点でも約 10° 程度誤差が小さくなっている. 高さの違いによる影響を低減することは今後の課題である.
2. 発話位置の変動による影響
   図 7 より口の高さが伝達関数と違う場合 A, C 地点の推定誤差は B 地点の推定誤差よりも大きくなっている. これは, A, C 地点はそれぞれ部屋の壁, ガラス面付近であり, 壁に近いために部屋の中心である B 地点よりも誤差が大きくなっていると考えられる.

5.3.2　オフラインシステム改良手法との比較 (実験 2)

オフラインシステムとの差異を明らかにするために, 同じ入力データを用いて, 周波数帯域と DB 刻み角度を変化させた時の誤差を評価した. 図 8 は下記の 5 つの条件での誤差と標準偏差を示す. ここで以下の要因に着目し推定精度の比較を行う.

1. 改良手法の有無
   改良手法の有無による比較を行うために, 使用周波数帯域と DB 刻み角度が同じである条件 (1) と (3) の結果を比較する. 3 つの発話地点において, 実時間システムよりも改良手法オフラインシステムの方が A 地点では約 50°, B 地点では約 20°, C 地点では約 30° 誤差が小さい. これより, 改良手法が誤差低減に大きく寄与 (20° – 50°) していることが分かる.
2. 使用周波数帯域
   使用周波数帯域による推定精度の比較を行うために, 条件 (3) と (5) の結果を比較する. どちらも改良手法オフラインシステムであり, 使用周波数帯域が違うだけである. 3 つの発話地点にすべてにおいて使用周波数帯域が 0-8kHz の場合が誤差 20° 以下であり, 0-1.5kHz の場合よりも 20° 以上誤差が小さい. これより, 使用周波数帯域の制限 (0-1.5kHz) が 20° – 25° 程度の誤差増加に寄与していると分かる.
3. DB 刻み角度
   DB 刻み角度による推定精度の比較を行うために, 条件の (2) と (3) の結果を比較する. DB 刻み角度が 45° の場合, 誤差が 10° 以下であり, すべての条件中, 最も高精度となっている. これは, 誤差が 15° である場合においても, DB 刻み角度が 45° の場合は角度分解能が低いために正解方向として処理されるためである. よって, 今回の結果だけでは DB 刻み角度は精度低下の原因として認められない.

上記の事から, 実時間システムの精度が低い主な原因は, 処理方法が改良手法でないことである. また, 使用周波数が 0-1.5kHz であることも原因となっている.

6 実時間システムの改良

実時間システムを高精度化するために, 従来のシステムに改良手法を導入し, その評価を行った.

6.1 導入した改良手法

今回, 実時間システムには以下の 2 つの改良手法を導入した.

1. 振幅特性の抽出
2. 周波数マスク付きヒストグラムの導入

なお, 発話区間検出処理はシステム構成上の理由から音源定位時に SN 比をもとに音声区間を判定する手法を導入した. 性能は改良手法のものとほぼ同等である. また, 改良手法の導入だけでなく, 処理の効率化を行い, 計算量を削減した. これにより処理は従来の実時間システムよりも増加しているのにも関わらず, 使用周波数帯域を 0-1.5kHz から 0-4kHz に広げた.

6.2 改良実時間システムの評価実験

従来の実時間システムの評価実験時と同じの入力データを用いて, 方向推定精度の評価実験を行った. 実験条件は 5 章と同じである. 図 9 に従来の実時間システムの推定精度との比較を示す. 図は推定誤差と標準偏差を示している. 図 9 より, 各発話地点において改良実時間システムは従来の実時間システムよりも推定誤差が 20° 以上低減している. また, 実際に実時間性を確認した結果, 高速化により従来の実時間システムよりも推定のサンプリング周期が短くなることでスムーズな動作が実現できることを確認した.

7 まとめ

本稿では, 発話方向実時間推定システムの評価実験を行い, その結果から精度の低い原因を調査した. また, 調査した原因をもとに改良実時間システムを提案し, その評価実験をしたところ, 推定精度, 実時間性ともに従来の実時間システムよりも優れていることが分かった. しかしながら, 推定誤差は未だに十分に小さいとは言えない. 今後は, 実時間性を低下させずに, さらに高精度な推定が可能な手法を構築する.

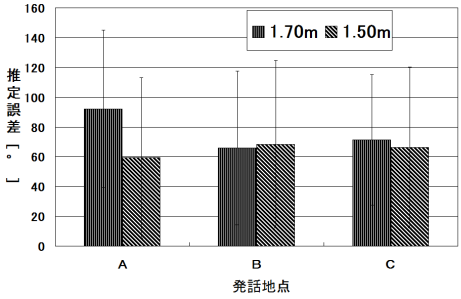


Fig. 7 実時間システムによる推定誤差

Table 1 実験 2 の条件

	使用周波数帯域 [kHz]	DB 刻み角度 [°]	手法
(1)	0-1.5	45	ベース
(2)	0-1.5	15	改良
(3)	0-1.5	45	改良
(4)	0-8	15	改良
(5)	0-8	45	改良

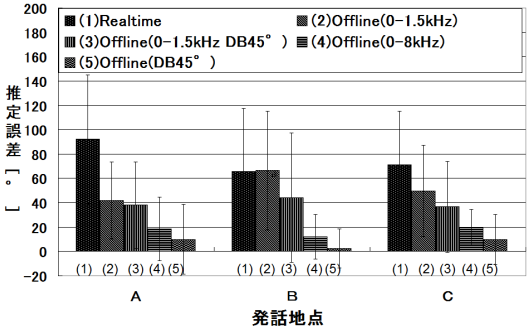


Fig. 8 推定誤差の比較

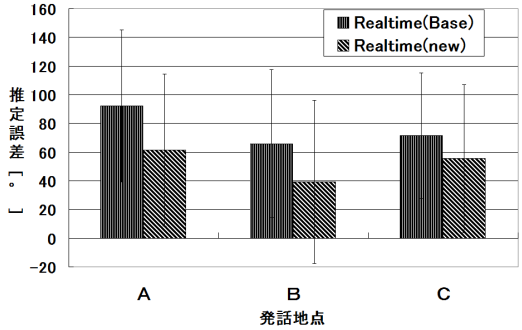


Fig. 9 改良実時間システムの推定誤差

参考文献

[1] 中島 弘史, " 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, 日本音響学会, 2005, pp. 619-620.

[2] 丹羽 健太ら, "空間相関行列の固有値の比率に着目した発話者向きの推定," 日本音響学会講演論文集 (春), 2009, pp.775-776.

[3] 醍醐 徹ら, " 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," 日本音響学会秋期研究発表会, 日本音響学会, 2007, pp.627-630.

[4] 菊池 慶子ら, " 大規模マイクロホンアレイによるた発話方向推定の検討," 信学技報 EA2008-37, 電子情報通信学会, 2007, pp. 13-18.

[5] Hirofumi Nakajima, et al. Real-time sound source orientation estimation using a 96 channel microphone array. IROS 2009: 676-683.