音声言語情報処理 34－6
(2000. 12. 21)

マイクロホンアレーを用いたHMMに基づく音源識別の評価

西浦 敬信[†‡] 中村 哲[†] 鹿野 清宏[‡]

[†] ATR 音声言語通信研究所

〒619-0288 京都府相楽郡精華町 2-2-2

[‡] 奈良先端科学技術大学院大学 情報科学研究科

〒630-0101 奈良県生駒市高山町 8916-5

E-Mail:{tnishi,nakamura}@slt.atr.co.jp, shikano@is.aist-nara.ac.jp

あらまし テレビ会議システムや音声による機器の制御において，発話者から離れた位置にあるマイクロホンで発話者の音声を高音質に受音することは極めて重要である．そこで発話者から離れた位置にあるマイクロホンでも発話者の音声を高音質に受音する方法としてマイクロホンアレーが注目されている．マイクロホンアレーを用いて高音質に発話者の音声を受音するためには，発話者の方向を推定することが必要となる．しかし，これまでの方向推定に関する研究では，複数の音源方向を推定することは多数試みられているが，その中から話者方向を推定することは困難であった．そこで本稿では，HMMに基づく音源識別を用いた話者位置推定法を提案する．まず，マイクロホンアレーを用いて音源方向を推定したのちに，ビームフォーミングを行い，その音を高音質に受音する．その後，HMMを用いた音声および環境音モデルにより音声・非音声の識別を行うことにより話者方向を推定する．また，本システムの音声認識性能も合わせて評価する．評価実験の結果，本手法により残響下でも良好に音声・非音声が識別でき，話者の方向を推定できることがわかった．

キーワード マイクロホンアレー，音源識別，HMM，話者位置推定，音声認識，RWCP音声音響データベース

Evaluation of Sound Source Discrimination Based on HMMs Using a Microphone Array

Takanobu NISHIURA[†‡] Satoshi NAKAMURA[†] Kiyohiro SHIKANO[‡]

[†] ATR Spoken Language Translation Research Laboratories

2-2-2 Seika, Soraku, Kyoto, 619-0288 JAPAN

[‡] Graduate School of Information Science, Nara Institute of Science and Technology

8916-5 Takayama, Ikoma, Nara, 630-0101 JAPAN

E-Mail:{tnishi,nakamura}@slt.atr.co.jp, shikano@is.aist-nara.ac.jp

Abstract It is very important for a hands-free speech interface to capture distant talking speech with high quality. A microphone array is an ideal candidate as an effective method for capturing distant talking speech. However, it is necessary to localize the target talker before capturing distant talking speech using a microphone array. In the conventional method of talker localization, it is difficult to estimate the target talker position accurately among localized sound sources, while the sound sources can be easily localized in a multiple sound source environment. To cope with this problem, we propose a talker localization algorithm by discriminating the sound sources using statistical speech and noise models based on HMMs (Hidden Marcov Models). First, the directions of signal arrival are estimated using a microphone array. Then, the desired sound signals are enhanced by steering the directivities to the estimated directions of signal arrival. The talker can be localized after discriminating between "speech" or "noise" from the desired sound signals using statistical speech and noise HMMs. In this paper, we evaluate the discrimination performance for the source position-known condition and position-unknown condition. The system recognizes the input from a sound source which is discriminated as being "speech" using statistical speech and noise HMMs. As a result, we confirm that the talker position is localized accurately because speech and noise can be discriminated efficiently in reverberant environments.

Key words Microphone array, Sound source discrimination, HMM, Talker localization, Speech recognition, RWCP-DB.

1 はじめに

テレビ会議システムや音声による機器の制御において，話者から離れた位置にあるマイクロホンで話者の音声を受音することを考えると，残響や背景雑音の影響により受音した話者の音声が歪みを受け，音質が低下するという問題がある．そこで，話者から離れた位置にあるマイクロホンでも，話者の音声を高音質に受音する方法として，マイクロホンアレーが注目されており，盛んに研究が行われている．これらの研究では，マイクロホンアレーを用いて話者の方向に指向性を形成することにより，高音質な音声の受音を実現している．マイクロホンアレーを用いた指向性は大きくわけると，話者の方向に対して超指向性を形成する遅延和アレー [1, 2] と雑音や残響の方向に対して死角を形成する適応形アレー [3, 4, 5] に分類できる．両手法とも指向性雑音を効果的に抑圧できることから，近年では音声認識などに盛んに利用が検討されている．マイクロホンアレーを用いて高音質に発話者の音声を受音するためには，発話者の方向を推定することが必要となる．しかし，これまでの方向推定に関する研究では，音源方向を推定 (例えば文献 [6]) することは多数試みられているが，検出された音源方向の中から発話している話者方向を推定することは困難であった．

そこで本稿では数ある音源の中から音声・非音声を識別することにより話者の方向を推定する．著者らはこれまでに，HMM を用いた環境音の識別 [7] を試みてきた．その結果，HMM を用いることにより良好に音源を識別できることがわかった．しかし，これまでの研究において，ドライソースにおける音源識別性能は明らかとなったが，残響下や雑音環境下における性能は未知数であった．そこで本稿では，複数の音源が存在する環境下において，マイクロホンアレーを用いて音源方向が既知および未知である場合の音源識別性能を実験的に検討する．まず HMM を用いた音声および環境音モデルにより音声・非音声の識別を行い，音源が話者であるかどうか識別を行う．さらに本システムの音声認識性能も合わせて検討する．

2 ビームフォーミングと音源方向推定

2.1 ビームフォーミング

マイクロホンアレーを用いて高音質に音声を受音するためには，ビームフォーミングが必要となる．本稿では目的信号に対して鋭い指向性を形成できる遅延和アレーを用いた．図 1 に示すように，目的の信号が θ 方向から到来し，マイクロホン数 M，マイクロホン間隔 d の等間隔直線配列マイクロホンアレーで受音される状況を考える．ここでは簡単のために平面波音場を仮定する．遅延和アレーでは，θ 方向から到来する信号を同相化して加算することにより，

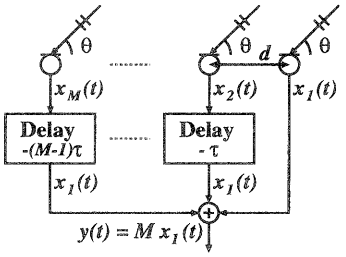


図 1: 遅延和アレー

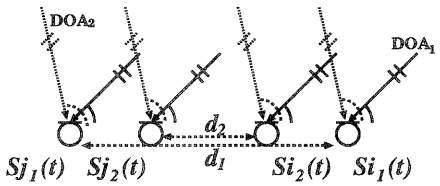


図 2: CSP 法による到来時間差の推定

θ 方向から到来する信号は M 倍になって出力される．一方，θ 方向以外の方向から到来する信号は同相化されず M 倍にはならない．よって，θ 方向に感度が高く，それ以外の方向に感度が低い指向性が形成される．

2.2 音源方向推定

音源方向を推定する方法として信号の相関を利用する CSP(Cross-power Spectrum Phase analysis) 係数加算法 [9] が提案されている．この手法はマイクロホンペアによる CSP 係数を加算することにより，異なる音源同士の相関や残響などに頑健な方向推定が行えることが特徴である．図 2 に示すようにマイクロホン i_n, j_n で信号 $s_{i_n}(t)$, $s_{j_n}(t)$ を受音すると，CSP 係数は式 (1) により推定できる．

$$\begin{aligned} &\mathrm{CSP}_{i_n,j_n}(k_n) \\ &= \mathrm{IDFT}\left[\frac{\mathrm{DFT}\left[s_{i_n}(t)\right]\mathrm{DFT}\left[s_{j_n}(t)\right]^*}{\left|\mathrm{DFT}\left[s_{i_n}(t)\right]\right|\left|\mathrm{DFT}\left[s_{j_n}(t)\right]\right|}\right] \end{aligned} \quad (1)$$

式 (1) では，2ch の受音信号をフーリエ変換して，振幅で正規化を行う．そして位相差を求めて逆フーリエ変換を行い CSP 係数を求めている．さらに式 (2) により CSP 係数を加算する．

$$\begin{aligned} \mathrm{CSP}_{i,j}(\theta) &= \sum_{n=1}^{N} \mathrm{CSP}_{i_n,j_n}(\theta), \\ subject\quad to\quad & \theta = \cos^{-1}\left(\frac{c\cdot k_n/F_s}{d_n}\right) \end{aligned} \quad (2)$$

ここで，N は加算の回数，d_n はマイクロホンの間隔，c は音速，F_s はサンプリング周波数を示す．m 個の音源方向 DOA_m は式 (2) により加算した CSP 係数の m 個の最大値

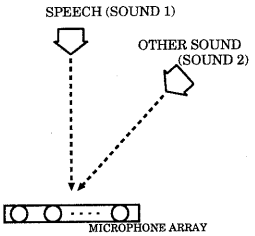


図 3: 2 音源が存在する環境

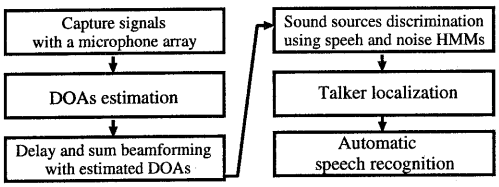


図 4: 音源識別アルゴリズム

から式 (3) により推定することができる.

$$\mathrm{DOA}_m = \underset{\theta}{\mathrm{argmax}}(\mathrm{CSP}_{ij}(\theta)) \qquad (3)$$

CSP 係数加算法は異なる CSP 係数を加算するため，残響下においても頑健な音源方向推定が行える．よって本稿でも，CSP 係数加算法を用いて音源方向を推定する．

3 音源識別アルゴリズム

図 3 のようにマイクロホンアレーに対して正面方向から音声，右方向から非音声が到来する場合を考える．ここで高音質に音声を受音するためには，音源方向を推定し，その後，推定方向に対してビームフォーミングを行う必要がある．しかし，この状況において音源方向を推定することができても，どの方向に話者が存在するのか推定できないという問題があった．そこで，話者の位置を推定し音声のみを高音質に受音するために音源を識別することを検討する．図 4 にその流れ図を示す．最初に音をマイクロホンアレーで受音し，CSP 係数加算法により音源の到来方向を推定する．次に推定方向に対してビームフォーミングを行い，高音質に音を抽出する．ここで，様々な環境音から作成した環境音モデルと音声モデルを用いて尤度を計算し，音声および非音声の識別を行うことにより，音源が話者であるかどうか識別する．さらに音源が話者であればその音声に対して認識を行う．

4 環境音データベース

HMM を用いた音声および環境音モデルにより音源識別を行うためには，モデル作成のためにあらかじめ様々な環境音が必要となる．そこで本研究では，新情報処理開発機構 (RWCP:Real World Computing Partnership) に設置されている知的資源ワーキンググループの実環境音響データベースサブワーキンググループによって作成された RWCP 実環境音声・音響データベースを [10](以下 RWCP-DB) に含まれる環境音のドライソースを用いて環境音のモデルを作成した．RWCP-DB の方針は無響室で収録した種々の音源データ (ドライソース) と種々の部屋のインパルス応答 (音響伝達特性) を収録し，それらを畳み込むことで，多種類の音環境データを得るというものである．

表 1: RWCP データベース一例.

	音源の系統	音源の例
衝突系	木質	木板を木棒で叩くなど
	金属	金板を金棒で叩くなど
	プラスチック	プラケースを木棒で叩くなど
	セラミック	ガラスを叩くなど
動作系	粒子落下系	豆を箱に注ぐなど
	ガス噴射系	スプレーの噴射など
	摩擦系	ノコギリを引くなど
	破裂破壊系	割箸を折るなど
	弾性音系	拍手など
特徴的	金属小物系	鈴を鳴らすなど
	紙系	紙を破るなど
	楽器系	ラッパの音など
	電子音系	電話の呼出音など
	機械系	ゼンマイの音など

RWCP-DB 中の環境音データは非音声の認識の研究のために一種類の音源について 100 サンプルを基準として収録されており，音源の設置方法や発生方法を変化させることによって，ある程度のバラエティーがもたされている．表 1 に環境音データベースの内容を簡単に示す．表 1 中の「衝突系」は硬い物体の単発的な衝突に起因するタイプの音源を表しており，「動作系」は音だけから明確な音源種類を特定することは難しいが特徴的な音色を持つタイプの音源を表している．また，「特徴的」は音色が音源種類そのものを特徴的に表すタイプの音源である．表 1 では 15 種類の系統を示しているが，全体のデータとしては約 90 種類，10,000 サンプルがあり，実験に使用した環境音は 90 種類の音源データ各 50 データの約 4,500 サンプルである．

また，RWCP-DB には，様々な環境においてマイクロホンアレーを用いて測定した音響伝達特性 [11] も存在する．そこで RWCP-DB 中の音響伝達特性と環境音・音声を用いて仮想実環境実験を行った．著者らはこれまでにも RWCP-DB を用いた音源識別に関する研究 [7] を行っている．しかし，この研究は音響伝達特性を考慮しておらず，ドライソースを用いた音源識別にとどまっていた．そこで本研究では RWCP-DB 内の環境音のドライソースと音響伝達特性を用いて雑音・残響下での音源識別について検討を行う．

表 2: データ収録条件

マイクロホンアレー	素子数 14, 素子間隔 2.83 cm
音源方向推定	CSP 係数加算法 [9]
Beamformer	遅延和アレー [2]
音響伝達特性	RWCP-DB
残響時間 $T_{[60]}$	0.0, 0.3, 1.3 sec.
SNR	-5dB, ～, 30dB, clean

5 評価実験

本稿では，音源方向が既知である場合および CSP 係数加算法により音源方向を推定した場合の，音源識別性能を実験的に評価した．図 5 に実験環境を示す．音源はマイクロホンアレーに対して正面方向に目的音源，右方向に雑音源が存在する．音源とマイクロホンアレーとの距離は 2m である．この環境において，部屋の残響時間および目的音源と雑音源の SNR を変化させたときの音源識別率および音声認識率を評価した．

5.1 実験条件

表 2 のデータ収録条件と表 3 に示す音源識別実験条件により評価実験を行った．この実験条件下においてシングルマイクロホンおよびマイクロホンアレーを用いて，SNR が (-5dB, ～, 30dB, clean)，および残響時間 $T_{[60]} = 0.0, 0.3, 1.3$ sec. における音源識別性能を評価した．さらに識別結果が音声である場合には，表 2 の収録条件と，表 4 の音声認識条件により，その音声に対する音声認識性能を評価した．

5.2 性能評価

本実験では，音声 216 単語 ×2 と環境音 92 種類 ×2 の合わせて 616 音に対して，音声・非音声に対する音源識別率により性能を評価する．さらに識別結果が音声の場合には音声認識（単語認識）を行い，音声認識性能の評価を行う．

5.3 音源方向が既知である場合の実験結果

まず，音源方向が既知である場合の実験結果について説明する．図 6(a) にシングルマイクロホン，(b) にマイクロホンアレーを用いた場合の実験結果を示す．図中の棒グラフは音源識別率，折れ線グラフは音声認識率を示す．

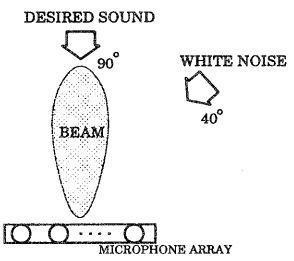


図 5: 実験環境

表 3: 音源識別における実験条件

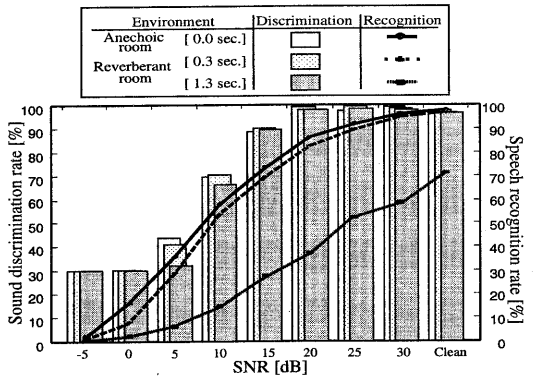
フレーム長	32 msec.（ハミング窓）
フレーム周期	8 msec.
特徴ベクトル	MFCC, Δ MFCC, Δパワー
HMM	Gaussian mixture 型 HMM
音響モデル数	音声 1 モデル
	非音声 1 モデル
音声 DB	ATR 音声データベース SetA
音声モデル学習	男女 16 名　約 2000 語
非音声 DB	RWCP-DB
非音声モデル学習	環境音 92 種類 × 20
テスト (オープン)	音声: 音素バランス 216 語 × 2
	非音声: 環境音 92 種類 × 2

表 4: 音声認識における実験条件

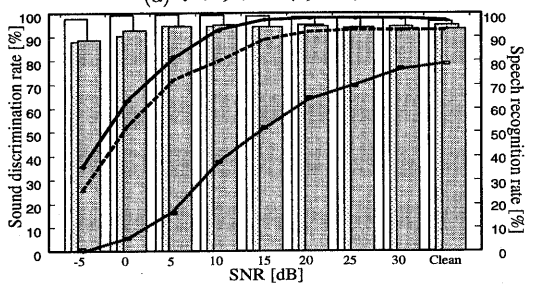
フレーム長	25 msec.（ハミング窓）
フレーム周期	10 msec.
特徴ベクトル	MFCC, Δ MFCC, Δパワー
音響モデル	IPA 1998 年度 音響モデル
テスト (オープン)	音声: 音素バランス 216 語

最初に，図 6(a)(b) 中の音源識別率を示す棒グラフに注目する．図 6(a)(b) を比較するとシングルマイクロホンよりもマイクロホンアレーを用いて既知音源方向にビームフォーミングを行ったほうが，特に低 SNR において音源識別率が飛躍的に向上しており，マイクロホンアレーを用いることにより低 SNR の環境であっても，高い音源識別率が得られる．つぎに音源識別性能の耐残響性について述べる．図 6(a)(b) の棒グラフより，マイクロホンアレーを用いたときの低 SNR 環境においては残響時間が長くなると僅かながら音源識別率が低下する傾向があるのもの，各 SNR においてほとんど差がないことがわかる．これより，マイクロホンアレーを用いた音源識別手法は残響について頑健であることがわかる．これより音源の方向が予めわかっているならば，その音源が話者であるかどうか，高残響下でも十分に識別できることがわかる．

次に，図 6(a)(b) 中の音声認識率を示す折れ線グラフに注目する．図 6(a)(b) を比較するとシングルマイクロホンよりもマイクロホンアレーを用いて既知音源方向にビームフォーミングしたほうが，特に低 SNR の時に音声認識率が向上していることがわかる．顕著な例として SNR が 10dB を例に説明する．残響時間が $T_{[60]} = 0.0sec.$ の環境において SNR が 10dB のときシングルマイクロホンによる音声認識率は図 6(a) より 58.3%であったのに対し，マイクロホンアレーを用いると図 6(b) より 92.6% となり認識率が約 35%改善されていることがわかる．また残響時間が



(a) シングルマイクロホン



(b) マイクロホンアレー (既知音源方向)

図 6: 音源識別率と音声認識率

$T_{[60]} = 1.3sec.$ の環境において SNR が 20dB のときシングルマイクロホンによる音声認識率は図 6(a) より 37.1%であったのに対し，マイクロホンアレーを用いると図 6(b) より 64.3%となり認識率を約 27%改善できることから，残響下においても性能改善が期待できることがわかる．

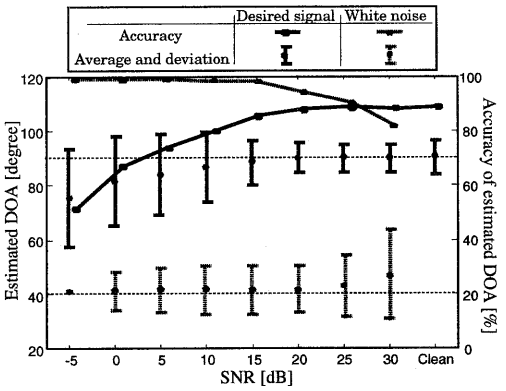
5.4 音源方向が未知である場合の実験結果

続いて音源方向が未知である場合の実験結果について説明する．まず CSP 係数加算法を用いて音源方向を推定する．本実験では，方向を推定するために 2 つのマイクロホンペアの CSP 係数を加算することにより方向を推定した．図 7(a)(b)(c) にその実験結果を示す．推定した方向に対して平均値と偏差および式 (4),(5) により推定音源方向の正解率を計算する．式 (4),(5) 中の N は全総数，Q_{tru} は真の音源方向，Q_{est} は推定した音源方向，Er は許容方向推定誤差を示し，本実験では $Er = 5°$ とした．

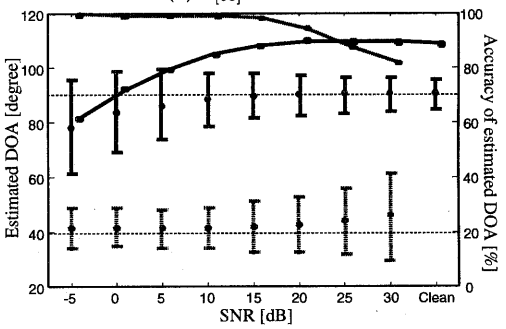
$$\text{Accuracy} = \frac{\sum_{n=0}^{N} I_{cor}[n]}{N} \tag{4}$$

$$I_{cor}[n] = \begin{cases} 1 & \|Q_{est} - Q_{tru}\| \leq Er \\ 0 & \|Q_{est} - Q_{tru}\| > Er \end{cases} \tag{5}$$

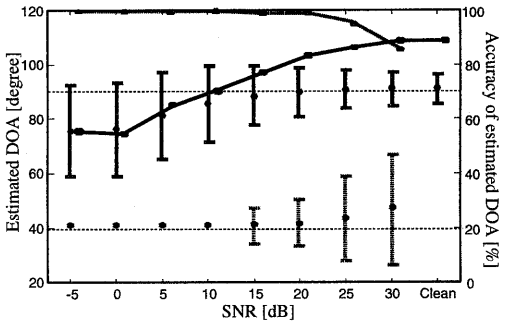
図 7(a)(b)(c) より平均値と偏差から，SNR が高くなるにつれて目的信号に対する推定方向が真の音源方向に近づくことがわかる．また反対に SNR が低くなるにつれて雑音信号に対する推定方向が真の音源方向に近づくこともわかる．これは，2 つの方向を推定する場合，推定方向の信頼性は CSP 係数加算法による CSP 係数が信号のエネルギーに依存することが影響している．



(a) $T_{[60]} = 0.0sec.$



(b) $T_{[60]} = 0.3sec.$



(c) $T_{[60]} = 1.3sec.$

図 7: 音源方向推定結果

また，図 7(a)(b)(c) から目的方向に対する方向推定結果の正解率に比べて，雑音方向に対する方向推定結果の正解率は高 SNR 環境においても非常に高いことがわかる．この結果は音声と白色雑音の統計的性質に大きく依存していると考えられる．CSP 係数加算法は信号の相関を基にして

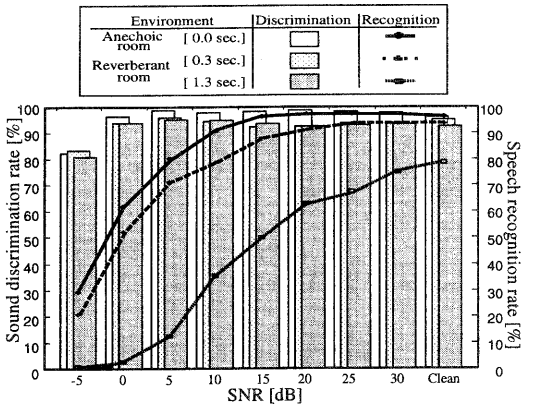


図 8: 推定音源方向を用いた音源識別率と音声認識率

音源方向を推定する．そのため音声のように自己相関の高い信号よりも，白色雑音のように自己相関が無い信号のほうが精度良く音源方向を推定できる．

次に，残響に対する頑健性について述べる．図 7(a)(b)(c) より目的信号および雑音信号の方向推定精度は $T_{[60]} = 0.0sec.$ も $T_{[60]} = 1.3sec.$ もほとんど変わらない．この結果から CSP 係数加算法が残響に対して頑健であることがわかる．

さらに，推定した音源方向を用いて音源識別実験および音声認識実験を行った．本実験では推定した 2 方向に対してより正面方向に近い方向を目的音の方向としてビームフォーミングを行い音源識別および音声認識を行う．図 8 に推定音源方向を用いた実験結果を示す．図中の棒グラフは音源識別率，折れ線グラフは音声認識率を示す．

図 6 中の音源識別率を示す棒グラフに注目する．図 6(a) と図 8 を比較するとシングルマイクロホンよりもマイクロホンアレーを用いて推定音源方向にビームフォーミングを行ったほうが，高音源識別率であることがわかる．また図 6(b) と図 8 を比較すると，推定音源方向を用いた場合には，低 SNR では方向推定精度の劣化により音源識別率も低下が見られる．しかし，高 SNR では方向推定精度の向上に伴い，音源識別率は既知音源方向を用いた場合と同程度の性能になることがわかる．この傾向は残響時間にかかわらず同じである．これより音源の方向が未知である場合でも，その音源が話者であるかどうかマイクロホンアレーを用いることにより高残響下でも十分に識別できることがわかった．また，音声認識率においても音源識別実験と同様の傾向があることが図 8 中の折れ線グラフにより確認できる．

以上の評価実験から話者の位置を推定するにあたり，音源の方向が未知であっても CSP 係数加算法を用いて音源の方向を推定し，その方向にビームフォーミングを行い音源の識別を行うことによって，その音源が話者であるかどうか良好に判別できることが確認できた．

6 まとめ

本稿では，話者の位置を推定するために，音声および環境音の HMM モデルにより，音声・非音声を識別することを検討した．その結果，高残響下でも良好に音源を識別でき，話者方向の判別ができることを RWCP-DB を用いて確認した．また音源の方向を CSP 係数加算法により推定し同様の実験を行った．その結果，本提案手法は音源方向が未知である場合でも良好に話者の方向を推定できることがわかった．

謝辞: 本研究の機会を与えてくださった，ATR 音声言語通信研究所 山本誠一社長に感謝致します．

付録: RWCP 実環境音声・音響データベース

RWCP 実環境音声・音響データベースは，現在平成 10 年度版を配布中である．正式版は平成 13 年春頃リリース予定となっている．現在，以下のサイトにて無償配布の受付が行われている．

http://tosa.mri.co.jp/sounddb/

参考文献

[1] J. L. Flanagan, J. D. Johnston, R. Zahn and G. W. Elko, "Computer-Steered Microphone Arrays for Sound Transduction in Large Rooms," J. Acoust. Soc. Am., Vol. 78, No. 5, pp. 1508–1518, Nov. 1985.

[2] S. U. Pillai, "Array Signal Processing," Springer-Verlag, New York, 1989.

[3] L. J. Griffiths and C. W. Jim, "An Alternative Approach to Linearly Constrained Adaptive Beam-forming," IEEE Trans. AP, Vol. AP-30, No. 1, pp. 27–34, 1982.

[4] Y. Kaneda and J. Ohga, "Adaptive Microphone-array System for Noise Reduction," IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-34, No.6, pp. 1391–1400, Dec. 1986.

[5] H. Saruwatari, S. Kajita, K. Takeda and F. Itakura, "Speech Enhancement Using Noise Adaptive Complementary Beamforming," IEICE Technical Report, SP99-77, pp. 1–8, 1999.

[6] 安部正人, " 多数センサによる音源推定," 音学誌, Vol. 51, No. 5, pp. 384–389, 1995.

[7] 三木 一浩, 西浦 敬信, 中村 哲, 鹿野 清宏, "HMM を用いた環境音識別の検討," 信学技報, SP99-106, Dec. 1999.

[8] C. H. Knapp and G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay," IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-24, No. 4, pp. 320–327, 1976.

[9] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano, "Localization of Multiple Sound Sources Based on a CSP Analysis with a Microphone Array," Proc. ICASSP2000, pp. 1053–1056, Jun. 2000.

[10] S. Nakamura, K. Hiyane, F. Asano, T. Yamada, and T. Endo, "Data Collection in Real Acoustical Environments for Sound Scene Understanding and Hands-Free Speech Recognition," Proc. Eurospeech99, pp. 2255–2258, Sep. 1999.

[11] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada,"Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition," Proc. LREC2000, pp. 965–968, May. 2000.