音声言語情報処理 29－20
(1999. 12. 20)

〔特別講演〕

実環境におけるハンズフリー音声認識

中村 哲

奈良先端科学技術大学院大学 情報科学研究科

〒630-01 奈良県生駒市高山町8916-5

E-Mail:nakamura@is.aist-nara.ac.jp

あらまし 利用者にマイクロホンの装着の負荷をかけない入力として，マイクロホンから離れてハンズフリーで発話を行うハンズフリー音声認識技術が重要である．しかし，実環境下における音声の認識，特にマイクロホンから離れた発話を入力とするハンズフリーの音声認識では，環境に存在する雑音および部屋の残響が大きな性能劣化をもたらしてしまう．本稿では，このような問題に対し，モデル適応化による手法，マイクロホンアレーによる手法について紹介する．マイクロホンアレーは，超指向性を形成して対象音声を高SNRで受音することでこれらの環境雑音の影響を低減できる．一方，モデル適応化では、部屋の伝達特性を推定してその環境の観測信号にモデルを適応することで性能改善が可能となる．本稿ではさらに，これらの研究を進めるために現在収録を行っている実環境音声・音響データベースについても述べる．

キーワード 音声認識，実環境，ハンズフリー，遠隔発話，マイクロホンアレー，モデル適応化，HMM合成

Hands-Free Speech Recognition in Real Environments

Satoshi NAKAMURA

Graduate School of Information Science, Nara Institute of Science and Technology

8916-5 Takayama, Ikoma, Nara, 630-01 JAPAN

E-Mail:nakamura@is.aist-nara.ac.jp

Abstract Hands-free speech recognition is a very important for a natural human machine interface. The distant-talking speech in real environments is distorted by noise and reverberation of the room. The paper tries to give a prospect of the solution based on previous studies and our research efforts. Especially a microphone array based-method and a model adaptation method are discussed. The microphone array can reduce the influences of the environmental noises by beam-forming. On the other hand, the model adaptation method can estimate the acoustical transfer function and adapt the speech models against the distorted observation signals. Furthermore, this paper also addresses the project collecting research acoustic data for hands-free sound recognition which includes many kinds of dry sounds and impulse responses in real rooms.

keywords speech recognition, real environment, hands-free, distant-talking speech, microphone array, model adaptation, HMM composition

key words

1 はじめに

音声認識技術は，統計的音響モデルや統計的言語モデル，およびそれらを支える大規模な音声データベース，テキストデータベースの整備により，近年，著しい進歩を遂げている．しかし，現状のシステムには接話マイクやハンドマイクの使用が必要で，これらの装備による拘束はインタフェースを不自然なものにしていた．やはり，自然なインタフェースとしては，マイクロホンに拘束されず，任意の位置から動きながら音声で機器に指示するというハンズフリーのインタフェースが必要となる．ところが，実環境下における音声の認識，特にマイクロホンから離れて音声を入力するハンズフリーの音声認識では，環境に存在する雑音および部屋の残響が大きな性能劣化をもたらす．また，デスクトップマイクロホンなどを用いた場合でも，口との距離が少し離れた時や発話者が横を向いた時などに同様の性能劣化が生じる． 本稿では，これらの問題に対処する研究として，モデル適応のアプローチとマイクロホンアレーを用いたアプローチによる研究について紹介する．

2 ハンズフリー音声認識

図1に示すように，ハンズフリーの音声認識の利点は，マイクロホンを装着する必要がなく，自由に移動しながら発話できるということである．これにより，利用者は相手が人の場合と同様に自由に発話をすることができる．しかし，このためには実環境において以下の機能を有する音声認識システムを実現する必要がある．

- 環境に存在する指向性雑音，無指向性の雑音に頑健な音声認識
- 部屋の反射，残響などの音響特性に頑健な音声認識
- 複数の雑音源や発話者を含む音源の中からの選択的な音声の認識
- 発話者の移動，雑音源の移動などの環境，音響特性の変化に頑健な音声認識

これまで，雑音環境下における音声認識の研究は種々行われてきたが，このようなハンズフリー環境における音声認識を実現するためには，雑音だけでなく，部屋の音響条件，複数音源，音源の移動などの問題を扱わなければならないことがわかる．これらは，今までの接話マイクロホンの使用を前提とした音声認識の研究ではほとんど触れられなかった技術である．

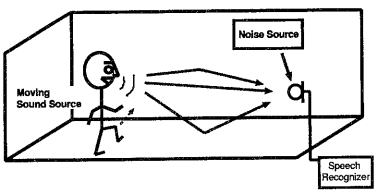

図 1: 実環境下におけるハンズフリー音声認識

3 研究の動向

ハンズフリー音声認識を実現するためのアプローチとしては，単一マイクロホンで受音した信号に対し，前処理的に音声強調 (Speech Enhancement) をする方法と音声認識に用いる HMM などのモデルを観測データに合うように適応化する方法がある．さらに，複数のマイクロホン素子を備えるマイクロホンアレーを用いて，対象音源に指向性を形成して受音する事で加法性雑音や乗法性歪みである残響の影響を除去する方法がある．これまでにも，近接マイクロホン利用時に存在する加法性雑音や電話回線などの乗法性歪みの要因に対処するための研究は多く行なわれてきている．それらの研究は，音声強調による方法とモデル適応化による補償の方法に大別できる．音声強調としては，加法性雑音に対して Spectral Subtraction[1]，電話回線などの歪みに対して Cepstral Mean Subtraction [2] が提案されている．モデル適応化としては，加法性雑音に対して HMM 合成法 [3]，PMC[4]，電話回線などの歪みに対して Stochastic Matching[6] などの手法があげられる．また，両方の要因を取り扱う研究についても行なわれてきている [5, 7, 8]．ハンズフリー環境に対してもこれらの方法を適用することは可能であるが，加法性雑音についてはSNRが大幅に減少すること，乗法性歪みについてはその推定が非常に難しいことや話者が移動することにより時々刻々伝達経路が変わるという問題がある．いずれにしても，受音時に高 SNR で受音する事がこれらの処理にとってきわめて重要となる．

一方，受音時の SNR を改善する方法として複数のマイクロホン素子を用いたマイクロホンアレーを利用する方法がある [9, 17, 10, 11]．マイクロホンアレーを音声認識に適用する場合，まず正確に対象音源を同定しビームフォーミングすることにより忠実に原信号を取り出し，その信号を音声認識部に送る形となる．従って，マイクロホンアレーとその信号処理は，音声認識の前処理と位置付けることができる．これまで，音声認識性能の改善を目的に種々のアレー信号処理部の改善が行なわれてきた．マイクロホンアレー、複数マイクロホンの適用により，対象とする環境にもよるが，概ねＳＮＲで約 10-20dB 程度の改善，それに応じた認識率の改善が得られることが報告されている [12, 13, 14, 15, 16, 18, 19, 26, 27]．

本稿では，ハンズフリー音声認識のために著者らが進め

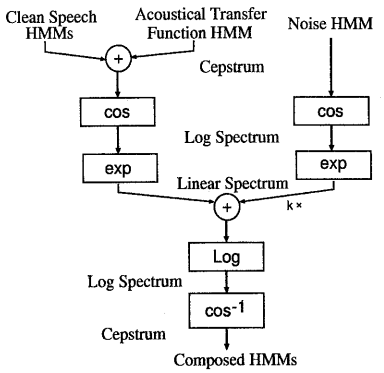

図 2: 出力確率の合成アルゴリズム

ているモデル適応化のアプローチとマイクロホンアレーによるアプローチについて紹介する．

4 モデル適応化手法

モデル適応化手法としては種々の方法が提案されているが，ここでは単一マイクロホン受音の信号に対するHMM合成法を取り上げる．もし，HMMで表現されている情報源がお互いに独立であれば，2つのHMMを合成することが可能となる [3, 4]．音声と加法性雑音は，周波数領域では独立であり加法性が仮定できる．一方，音声と音響伝達特性は，周波数領域では積によって関係づけられているので，ケプストラム領域では独立であり加法性が仮定できる[1]．ゆえに，雑音及び残響のある環境下において観測され得る音声のモデルがHMM合成法により構成できることになる．図2に加法性雑音，乗法性雑音に対するHMM合成の手順を示す．一般に，音声認識用の音素モデルはケプストラムパラメータを用いて学習されているので，ケプストラム領域で乗法性歪みのHMMとHMM合成を行い，その後，フーリエ変換，指数変換を行って周波数領域に変換し，加法性雑音のHMMと合成を行う．

著者らは，ハンズフリー環境でマイクから離れて話者が移動しながら発話する場合においても，エルゴーディックHMMを用意し，それぞれの状態に対象とする部屋の代表的な位置からの音響伝達特性を割り当てておけば，移動話者の音声認識が可能となることを示した [8]．

HMM合成法で問題となるのは，クリーン音声で学習した音声のHMMモデルと合成する加法性雑音のHMMモデルおよび乗法性歪みのHMMモデルをいかに学習するかである．雑音モデルについては，雑音のみを含む部分を用いて雑音HMMを学習することが可能であるが，乗法性歪みのHMMモデルは独立に学習することが難しい．それぞれの位置におけるインパルス応答を測定することも可能であるがあまり現実的でない．そこで，著者らはそれぞれの位置で発話された学習用音声を用いて，音響伝達特性の短区間ケプストラム表現のHMMパラメータの推定を行うHMM分解法を提案している [21]．学習により伝達特性のHMMパラメータを求めることにより，分析フレームより長いインパルス応答の影響もある程度含んだ形で推定することになる．

加法性雑音と乗法性歪みがある場合，HMM分解は2回（周波数領域とケプストラム領域）適用され，また領域変換の際には，特徴パラメータを直接取り扱うのではなく，その統計量が用いられる．さらに，推定された音響伝達特性を基に，いくつかの代表的な音響伝達特性HMMを用意することで，そのエルゴディック HMM により移動音源に対する認識 [8] が可能となる．

観測データを用いて音響伝達特性 HMM を推定する方法には，モデル合成の逆のプロセスであるモデル分解を用いる．音響伝達特性はモデル領域において次式により推定される．

$$\lambda_{H_{cep}} = \mathcal{F}^{-1}[\log\{\ \exp(\mathcal{F}(\lambda_{O_{cep}})) \ominus\ \lambda_{N_{lin}}\}] \ominus \lambda_{S_{cep}}$$

ここで，λ はモデルパラメータの集合を表し，O は観測信号，S はクリーン音声，H は音響伝達特性，N は雑音を表す．添字の cep と lin はそれぞれケプストラム領域と線形スペクトラム領域を表し，$\ominus$ は HMM の分解を表す．$\mathcal{F}$ はコサイン変換，log は対数変換，exp は指数変換を表す．雑音残響環境下では，HMM 分解法は，まず線形スペクトラム領域において適用され，さらにケプストラム領域において適用される．2 段階の分解は，合成モデルの尤度が最大になるようにして行なわれる．

$$1.\ \hat{\lambda}_{SH} = \underset{\lambda_{SH}}{\operatorname{argmax}}\ \Pr(\lambda_O \mid \lambda_{SH}, \lambda_N)$$

$$2.\ \lambda_H = \underset{\lambda_H}{\operatorname{argmax}}\ \Pr(\hat{\lambda}_{S+H} \mid \lambda_S, \lambda_H)$$

図 3 に示すように HMM 分解法では，モデル領域において音響伝達特性のモデルパラメータ推定を行なう．図 3 の例では，音響伝達特性 HMM は 3 状態で定義されているが，実際には，状態数は対象とする部屋の大きさや構造などによって決められる．モデルの各状態をある場所（空間）に対応させ，すべての状態間における遷移を可能にすることにより話者の場所移動にモデル側で対処可能となる．実験では音素 HMM を使用するので，一つの音素 HMM 内では場所移動はないものとし，音素連結の際のみ場所間の遷移を考慮する．

[1] 実際にはインパルスレスポンスの長さが短区間分析の窓長を有意に越える場合には問題が生じる [20]．

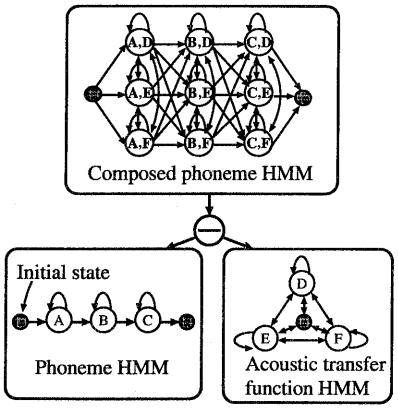

図 3: HMM 分解法による伝達特性 Ergodic-HMM の推定

部屋の伝達特性モデルとして Ergodic-HMM を用い，移動音源の認識性能を評価した．図 4 に実験で使用した部屋を示す．固定音源データとして，発話者が図に示されている場所 g1，g2 及び g3 から発話した音声を用いる．移動音源データとしては，図に示されている “Starting position” から矢印の方向へ動きながら発話した音声を用いる．発話内容は ASJ データベースの 30 文章である．評価は 1 文節を 1 語彙として定義し，文節認識を行なう（306 語彙）．言語モデルは使用していない．クリーン音声 HMM（Tied-mixture HMM）は，ASJ データベースの 7380 文章を用いて学習している．テスト話者の close-talking speech の accuracy は 90.4% である．以下では Ergodic-HMM の有効性を検討するために，Ergodic-HMM を作成せずに各場所の合成 HMM のままで認識した場合（尤度最大による認識）と比較を行なう．

表 1 に移動音源に対する文節認識率を示す．クリーン音声 HMM による認識率は 63.3% である．各合成 HMM (g1, g2, g3) の尤度最大での認識率は 76.7% となり，3 つの合成 HMM の Ergodic-HMM を用いて認識すると，認識率は 82.3% まで改善される．また適応データとして移動音声データ 1 文章を用いて伝達特性 HMM の平均値ベクトルを推定した場合，認識率は 80.9% であり，クリーン音声 HMM による結果よりは改善されているものの，合成 Ergodic-HMM を用いる方法には及ばない．

表 1: 移動音源に対する文節認識率

モデル	Phrase accuracy
クリーン HMM	63.3
尤度最大	76.7
Ergodic-HMM	82.3

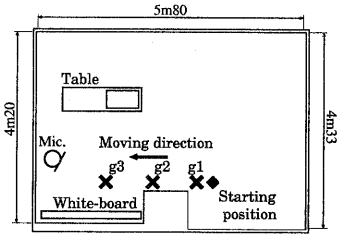

図 4: 収録条件

5 マイクロホンアレーの利用

マイクロホンアレーを備えた音声認識システムの処理の流れを図 5 に示す．図に示すように，マイクロホンアレーを音声認識に適用する場合，まず正確に対象音源を同定しビームフォームすることにより忠実に原信号を取り出し，その信号を音声認識部に送る形となる．しかしながら，このような構成では，アレー処理部が音源同定を誤れば全く認識ができないという致命的な問題が生じる．また，SNR を改善するために指向性を鋭くするほど高い音源位置の同定精度が要求される．これらの問題は，従来のアプローチがマイクロホンアレー処理を単に音声認識の前処理としてしか扱わなかったことに起因している．解決のためには，マイクロホンアレー処理と音声認識の結果をお互いにフィードバックし，統合的に処理することが必要になる．

著者らは，HMM法に基づく音声認識システムに音源方向探索を組み込み，音素系列と音源方向系列を尤度最大基準で同時に求める方法を提案している [22]．この方法では，比較的多くのチャンネルを有し空間分解能の高いマイクロホンアレーを用意し，指向性をいろいろな方向に向けることにより方向毎の信号を取り出す．方向毎に特徴抽出を行ない，方向，時間，HMM状態の 3 次元から

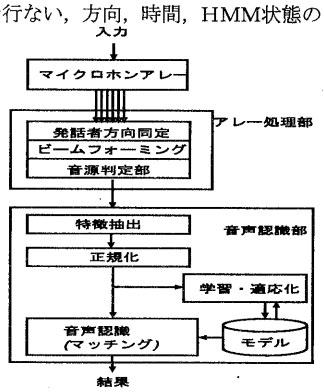

図 5: マイクロホンアレーを備えた音声認識システム

なるトレリス空間を構成した後，尤度最大基準でViterbi探索し，最適音素系列，音源方向系列の組合せを求める．図6に3次元トレリスの例を示す．これにより，音声認

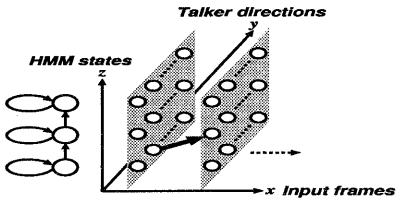

図 6: 3次元トレリス空間のViterbi探索

識用の文法で記述された言語制約の下に尤度が最大となる音源方向系列，音素系列を同時に求めることができる．また，この方法によれば，移動する話者の音声も認識することもできる．図7に移動する音源を認識した場合の例を示す．図は，0〜180度まで移動しながら1単語発話した音声の方向同定結果である．単語区間では良好に音源方向を同定していることがわかる．また，図8に遅延和アレー，適応型アレー（AMNOR）を用いた場合の移動音源に対する216単語の認識結果を示す．AMNORを用いる場合は，10度おき18種類のビームフォーマを予め学習しておき使用した．さらに，複数の音源が移動している状況においてもN-best探索法への拡張を行なえば，複数の音源を一度に認識可能となる．現在，複数の音源が存在する状況で音源軌跡の仮説のクラスタリングを行い音源グループを求め，音源グループのN-best探索を行う手法について検討している[23].

6 実環境音響・音声データベース

実環境におけるハンズフリー音声認識性能を評価するためには，評価のためのデータベースが必要となる．しかしながら，このようなデータベースは現在のところ殆ど存在しない．代表的な実環境を含む共通の多量のデータベースがあれば，システムの評価，モデルの学習など

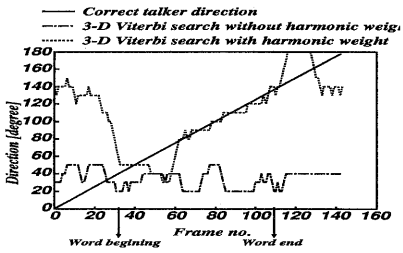

図 7: 3次元Viterbi探索を用いた移動話者の方向追跡の例 (SNR:20dB)

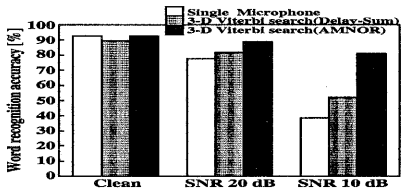

図 8: 3次元ビタビ法の単語認識率 (話者移動時)

に利用できる．著者らは，平成9年度より，新情報処理開発機構（RWCP）の知的資源ワーキンググループの実環境音声音響データベースサブワーキンググループにおいて，実環境における音・音声のマイクロホンアレー収録データベースの構築を進めている[24]．このデータの収録の方針は，無響室での種々の音源データ（Dry Source）と，種々の部屋のインパルス応答（音響伝達特性）を独立に収録し，これらを畳み込むことで種々多様な実環境音源をシミュレートしようとするものである．これらのデータ収集についで実音場の音を収録する．実音場の収録においては，マイクロホンアレーを用い，畳み込みでシミュレートできない移動音源，面音源，屋外の散乱性雑音などを収録する．

1. 音響情景データベース
   - ドライソースデータベース
   - インパルス応答データベース
2. 実音場マイクロホンアレーデータベース

屋外のデータについては，全周カメラ画像，状況のラベルなどを収録する予定である．表2に現在までに収録した環境音データの種類を示し，表3,4にマイクロホンアレーを用いて収録したインパルスレスポンスデータを示す．さらに，平成11年度は，一般の部屋のインパルスレスポンスに加え，日英音素バランス文の移動音源データを可変残響室で収録している．このデータについては，マイクロホンアレー収録データと移動時の再生用スピーカの座標を測定し，音源位置同定の研究，移動話者の音声の認識の研究に供する予定である．また，ドライソースデータベースに関しては，統計的モデルの学習を行うために十分な環境音のサンプルが収録されている．著者らは，環境音の識別，音声と環境音の識別，環境音が音声に重畳した場合の音声認識について，HMMに基づいた方法により検討を行っている[25].

7 まとめ

本稿では，ハンズフリー音声認識を実現するためのモデル適応化およびマイクロホンアレーの応用について述

べた．ハンズフリーの音声認識は，音声というモードを最大に利用する場合にどうしても必要になる技術である．この技術の実現のために，マイクロホンアレー信号処理，音声認識，モデル適応化のそれぞれの性能改善もさることながら，全体を統合する立場からの研究も今後ますます必要になると考えられる．

表 2: ドライソースデータの種類（各 200 サンプル以上）

	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

表 3: インパルスレスポンス収録機材

受音器	61 ch・3 軸直交形アレー（ONSOKU）
	54 ch・球形アレー（ONSOKU）
	16 ch・円形アレー（ONSOKU）
	2 ch・ロボット（WASEDA）
音源	スピーカー（DIATONE DS-7）
	ヘッドトルソー（B&K Type4128）

表 4: 収録音場

部屋	残響時間	dBA	dBC
無響室	0.003 秒	17.3	44.3
残響可変室	0.383 秒	18.5	48.3
	0.313 秒	17.4	46.0
	0.128 秒	18.0	48.0

参考文献

[1] S.F.Boll. “Suppression of Acoustic Noise in Speech Using Spectral Subtraction”, IEEE Trans,ASSP-27,pp.113-120,1979.

[2] Atal.B, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification”, Proc. J.Acoust.Soc.Amer., Vol.55, pp.1304-1312,1974.

[3] F.Martin, K.Shikano and Y.Okabe, “Recognition of Noisy Speech by Composition of Hidden Markov Models”, 信学技報, SP92-96, 1992.

[4] M.J.F.Gales and F.J.Young, “An improved Approach to the Hidden Markov Model Decomposition of Speech and Noise”, Proc.ICASSP, pp.233-236,1992.

[5] M.J.F.Gales and S.J.Young, “PMC for Speech Recognition in Additive and Convolutional Noise”, CUED-F-INFENG-TR154, 12, 1993.

[6] A.Sankar and C-H.Lee, “Robust Speech Recognition Based on Stochastic Matching”, Proc. ICASSP, pp.121-124,1995.

[7] 南，古井，“HMM 合成に基づく尤度最大化適応法”，音声研究会，SP95-24, pp.45-50, 6, 1995.

[8] 滝口，中村，鹿野，“雑音と残響のある環境下での HMM 合成によるハンズフリー音声認識法”，信学論（D-II），vol.J79-D-II, No.12, pp.2047-2053, Dec.1996.

[9] 金田，“騒音下音声認識のためのマイクロホンアレー技術”，音響学会誌 Vol.53,No11,1997

[10] S.Nakamura, K.Shikano, ”Room Acoustics and Reverberation:Impact on Hands-Free Recognition”, Proc. EUROSPEECH97, 1997

[11] 中村，鹿野，“認識技術の進展”，情報処理 Vol.38, No.11 1997

[12] D.V.Compernolle,W.Ma,F.Xie,M.V.Diest, “Speech Recognition in Noisy Environments with the Aid of Microphone Arrays”, Speech Communication, No.9, 1990

[13] H.F.Silverman, S.E.Kirtman, J.E.Adcock, P.C.Meuse, “Experimental Results for Baseline Speech Recognition Performance using Input Acquired a Linear Microphone Array” Proc. DARPA Workshop, 1992

[14] 中村，山田，鹿野，”マイクロホンアレーを用いた音源方向検出による音声認識”，音響講論 1-5-8, 1995,3

[15] D.Giuliani,M.Omologo,P.Svaizer, “Experiments of Speech Recognition in a Noisy and Reverberant Environment using a Microphone Array and HMM Adaptation”, Proc.ICSLP96, 1996

[16] Q.Lin,C.Che,D.Yuk, L.Jin, B.Vries, J.Pearson, J.Flanagan, “Robust Distant-Talking Speech Recognition” Proc.ICASSP96, 1996

[17] 浅野，“話者方向同定と雑音抑制による音声認識性能の改善”，音響学会誌 Vol.53,No11,1997

[18] K.Kiyohara, Y.Kaneda, S.Takahashi, H.Nomura, J.Kojima, “A Microphone Array System for Speech Recognition”, Proc.ICASSP97, 1997

[19] M.Omologo, “On The Future Trends of Hands-Free ASR: Variabilities in The Environmental Conditions and in The Acoustic Transduction”, ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, 1997

[20] 中村，滝口，鹿野，”短区間スペクトル分析における残響補正に関する検討”，音声研究会 SP98-25, 1998

[21] T. Takiguchi, S. Nakamura, Q. Huo and K. Shikano, “Adaptation of Model Parameters by HMM Decomposition in Noisy Reverberant Environments”, ESCA-NATO Workshop, pp.155-158, 1997.

[22] 山田，中村，鹿野，”マイクロホンアレーによる 3 次元トレリス探索に基づく 移動話者の音声認識”，情報処理学会研究会，音声言語情報処理 15-6, 1997

[23] P.Helacleous, S.Nakamura, K.shikano, “An Improvement to 3-D N-best Search Using Path-Distance Based Clustering for Recognizing Multiple Sound Source”, 音声研究会 SP99, 1999,12

[24] S.Nakamura, et al, “Data Collection in Real Acoustical Environments for Sound Scene Understanding and Hands-free Speech Recognition”, Proc. EUROSPEECH99, 1999

[25] 三木，西浦，中村，鹿野，”HMMを用いた環境音識別の検討”，音声研究会 SP99, 1999,12

[26] 猿渡，武田，板倉，“非線型マイクロホンアレーによる音声強調，” 電気音響研究会 EA98-77,1998

[27] 清水，武田，梶田，板倉，”空間音響特性依存HMMによるスペースダイバーシチ型音声認識”，音響講論，1-1-19, 1999, 9