音声言語情報処理 29－14
(1999. 12. 20)

HMM を用いた環境音識別の検討

三木 一浩　西浦 敬信　中村 哲　鹿野 清宏

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　TEL:0743-72-5287
E-Mail:{kazuhi-m,takano-n,nakamura,shikano}@is.aist-nara.ac.jp

あらまし　実環境において音声によるコミュニケーションを行なう場合、人間は音場の中から対象となる音源つまり、発話者を同定し、その方向から来る音に聞き耳を立て、相手の音声の認識を行なう。このとき、人間はそれらの環境音の中であっても音声を識別し発話を認識することができる。したがって、自律移動ロボットなどのシステムにおける音声認識を考える場合についても、環境音と音声を識別することは大変重要になる。本稿では種々の環境音が存在する状況において音声認識を行なうための第一歩として、HMM を用いた環境音の認識実験および環境音と音声の識別を試みる。3 状態 HMM による 90 種類の環境音に対する認識実験では 95.4%の認識率が得られた。また、本稿では環境音のクラスモデルを用いた HMM 合成法を提案する。提案法を用いて環境音の重畳した音声に対する認識実験を行なった結果、音声区間の前の雑音を使用して雑音モデルを学習する従来の HMM 合成法に対して高い認識性能が得られた。

キーワード　環境音、音声認識、HMM

Environmental Sound Discrimination Based on Hidden Markov Model

Kazuhiro MIKI　Takanobu NISHIURA　Satoshi NAKAMURA
Kiyohiro SHIKANO

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0101 JAPAN　TEL: 0743-72-5287
E-Mail:{kazuhi-m,takano-n,nakamura,shikano}@is.aist-nara.ac.jp

Abstract

In real acoustic environments, human communicates each other through speech by focusing on the target speech sound among environmental sounds. We can easily discriminate the target sound from other environmental sounds. For hands-free speech recognition, the discrimination of the target speech from environmental sounds is inevitable. This mechanism might also be important even for the self moving robot to sense acoustic environments and communicate with human. This paper proposes HMM-based environmental sound recognition. Environmental sounds are modeled by three state HMMs and evaluated using 90 kinds of environmental sounds. The recognition accuracy was 95.4%. This paper also proposes recognition of the environment sound-added isolated words by HMM composition. The experiments show the HMM composition of speech HMMs and a HMM of class environmental sounds outperforms the conventional HMM composition of speech HMMs and a noise HMM trained using noise periods prior to the target speech in the environments.

key words　envaironmental sound, speech recognition, HMM

1 はじめに

実環境において音声によるコミュニケーションを行なう場合、人間は音場の中から対象となる音源つまり、発話者を同定し、その方向から来る音に聞き耳を立て、相手の音声の認識を行なう。このとき、人間は日常的に音声以外の様々な物音を聞いているはずであるが、それらの環境音の中であっても音声を識別し発話を認識することができる。このような環境において自律移動ロボットや計算機で音声認識を行なう場合についても、やはり人間と同様に音声と環境音の識別ということが重要になる。

様々な環境音の中から目的の音を抽出する方法、つまり計算機に「聞き耳を立てさせる」方法の一つとしてマイクロホンアレー [1] を用いる方法がある。マイクロホンアレーはマイクロホン素子を空間的に並べたものであり、マイクロホンアレーの指向性を目的音に向けることにより高品質の音を収録するものである。しかし、マイクロホンアレーを音声認識の入力システムとして用いる場合には音声の到来方向を推定する必要がある [1, 2]。マイクロホンアレーを用いて比較的簡単に目的音の方向を推定する方法として、目的信号のパワーが最大となる方向を目的音の到来方向とする方法がある [1]。しかし、実環境においては音声以外の環境音も多々存在し、音声のもつパワーが常にそれらの環境音の中で最大になるとは限らず、その結果、音声以外の音にマイクロホンアレーの指向性を向けてしまう可能性もある。このような場合にもパワー以外の情報を用いて音声と環境音を識別することで音声の方向を推定することが可能になる。

雑音環境下で音声認識を行なうための様々な提案がなされている [3, 4, 5]。簡単かつ強力な方法としてスペクトルサブトラクション [3] がよく用いられる。この方法は雑音区間から、雑音のスペクトルを推定し雑音が重畳した音声信号のスペクトルから雑音のスペクトルを減算することで雑音の除去を行なう方法である。しかし、この方法は定常な雑音が存在する場合には有効であるが、単独の雑音が連続することで発生する非定常雑音が存在する場合には十分な性能が得られない。

また、他の方法としては認識時の HMM を雑音環境に適応させる HMM 合成法 [4, 5] がある。HMM 合成法はクリーン音声を用いて学習したクリーン HMM と雑音を用いて学習した雑音 HMM を合成し雑音重畳音声に対する HMM を作成する方法である。この方法を用いる場合には音声区間の前の雑音から雑音モデルを作成する方法がよく用いられるが、雑音が非定常な場合や数種類の雑音が存在する場合については、実際の雑音モデルと雑音のミスマッチが起こり認識性能の低下を招く。このような場合にも環境音の認識を行なうことにより合成する環境音を選ぶことで認識性能の低下を抑えることができる。

ここでは、上記の研究を進める第一歩として独立に生じる環境音および、環境音が連続して起こる場合について HMM を用いた環境音間の認識実験、環境音と音声の識別実験をおこない環境音がモデル化できるかを知る。また、連続した環境音が重畳している音声に対し HMM 合成法を用いた認識実験を行なう。本稿では、まずはじめに今回の実験に用いた実環境音声・音響データベースについて説明し、その後、各実験について説明する。

表 1: 環境音データ

	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

2 環境音データベース

本稿で実験に用いた環境音は、技術研究組合 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ものである (以下 RWCP-DB と記す)[6, 7]。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ことである。本稿ではその中から環境音のドライソースを用いて実験を行なった。

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持たされている。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。

3 環境音の認識、音声と環境音の識別

前述した RWCP-DB の環境音を用いて、単独の環境音について環境音間の認識実験を行なった。また、実際の環

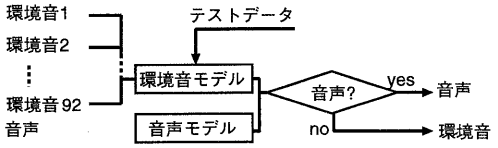


図 1: 全体環境音モデルを使用して識別

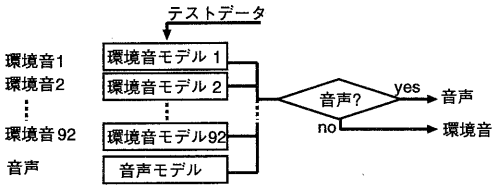


図 2: 個別の環境音モデルを使用して識別

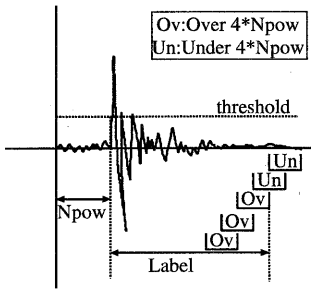


図 3: ラベルの作成

境のように同様の環境音が連続で起こる状況を考え、単独の環境音をつなげることで連続環境音を作成し、認識実験を行なった。さらに、単独環境音、連続環境音に音声を組み合わせたテストデータによって音声か環境音かの識別実験を行なった。音声と環境音の識別の方法として、

1. 図 1のように全種類の環境音を使用して作成した一つの環境音モデルを使用する方法
2. 図 2のように各々の環境音モデルを用いる方法

を用いた。

3.1 環境音 HMM の作成

RWCP-DB の環境音には HMM で学習を行なうためのラベルが与えられていないため、まずモデル学習のために環境音のラベルつけを行なった。ここでは、環境音区間の切り出しに環境音のパワーを用いた。実際には、図 3のように各環境音について目視によって閾値を設定し、その閾値を越えた点を環境音の開始点とした。また、データの始点から環境音の開始点までの暗騒音平均パワーを

表 2: 個別、全体環境音 HMM

状態数	3 状態
特徴量	MFCC ΔMFCC ΔPower
出力確率分布	4 4 2 混合連続分布型
サンプリング周波数	12kHz
データベース	RWCP 実環境音響データベース
学習データ (個別)	45 サンプル x 92 環境音
学習データ (全体)	20 サンプル x 92 環境音

表 3: 音素 HMM

状態数	3 状態
特徴量	MFCC ΔMFCC ΔPower
出力確率分布	256 256 128 混合 TiedMixture 分布型
サンプリング周波数	12kHz
データベース	ATR 日本語音声データベース
学習データ	特定話者 2620 単語

表 4: テストデータ

NSET	単独環境音
CNSET	連続環境音
SNSET	216 単語 + 単独環境音
SCNSET	216 単語 + 連続環境音

Npow とし、データの終点から前方に検索を行ない、環境音のパワーが 3 フレーム連続して暗騒音平均パワーの 4 倍を越えた点を環境音の終点とした。このとき、環境音開始点から環境音終点までに環境音のラベルがうたれ、その他についてはポーズがうたれる。このようにして作成したラベルを用いて HMM の学習を行なった。HMM は 3 状態の left-to-right モデルであり、特徴量として 16 次のメルケプストラム、16 次の Δ メルケプストラム、1 次の Δ パワーを用いた。学習データにはサンプリング周期 48kHz のデータをダウンサンプリングして得られた 12kHz のデータを用いた。個別の環境音 HMM の学習には、92 の環境音に対して各 45 サンプルを用い、合計 4,140 サンプルを使用した。また、全種類の環境音から作成される一つの全体環境音 HMM の学習には各環境音に対して 20 サンプルの合計 1,840 サンプルを用いた。これらの条件を表 2にまとめる。同様に、音声と環境音の識別実験に用いた音素 HMM の条件を表 3に示す。

3.2 テストデータ

環境音間の認識実験用のテストデータとして環境音データベースの中でモデルの学習に使用しなかった環境音の中から各環境音をひとつずつ持つ単独環境音のテストデータを 5 セット作成した (NSET1 ～ 5)。また、同じ単独環境音の繰り返しで作られる 92 個の連続環境音とランダムに数個の環境音をつなげた 3 個のデータからなるテスト

表 5: 環境音間の認識結果

(a) 単発環境音		(b) 連続環境音	
テストデータ	認識率 [%]	テストデータ	認識率 [%]
NSET1	96.7	CNSET1	69.5
NSET2	94.6	CNSET2	67.4
NSET3	96.7	CNSET3	72.6
NSET4	94.6	CNSET4	60.0
NSET5	94.6	CNSET5	60.0
平均	95.4	平均	65.9

表 6: 連続環境音の誤認識結果の一部

正解	認識結果
ガラス瓶を叩く 1 の繰り返し	ガラス瓶を叩く 1 , ガラスコップを叩く 1 の繰り返し
ガラス瓶を叩く 2 の繰り返し	ガラス瓶を叩く 2 , 陶器を叩くの繰り返し
複数のコインをまく 1 の繰り返し	複数のコインをまく 1, サイコロをふる 1 の繰り返し
ガラスコップを叩く 1 の繰り返し	ガラスコップを叩く 1, ガラスコップを叩く 2 の繰り返し

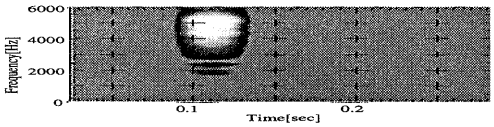
データを 5 つ作成した (CNSET1 ～ 5)。つぎに、音声と環境音の識別実験用のテストデータとして環境音の単独、連続テストデータに ATR の音韻バランス 216 単語を加えたもの ([SNSET1 ～ 5],[SCNSET1 ～ 5]) を作成した。テストデータを表 4に示す。

3.3 実験結果

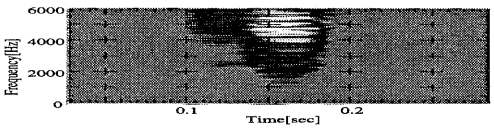
3.3.1 環境音間の認識実験結果

テストデータ NSET を用いた単独環境音の認識結果およびテストデータ CNSET を用いた連続環境音の認識結果を表 5に示す。この結果、単独の環境音の認識については各テストデータにおいて高い認識率を得ることができた。このとき各テストデータにおいてはエアキャップをつぶす音と紙やすりを使う音、クリップの音と割箸を折る音の誤認識が多くあった。図 4にクリップと割箸を折る音のスペクトログラムを示す。人間の感覚では異なると思われるこれらの音であるが HMM が時間方向の揺らぎを吸収することを考えると認識誤りを起こしやすい構造になっていることがわかる。

連続環境音の認識率については単独環境音の認識結果に比べ大きく劣化している。大きな原因として、認識系に認識時のモデルの繰り返しを許したため、一つの環境音を正解の環境音に近い別の 2 つの環境音と誤認識してしまうということがあげられる。例えば、ガラス瓶を横から木棒、スプーンで叩く音の繰り返しをガラス瓶を横から木棒、スプーンで叩く音とガラスコップを木棒、スプーンで叩く音の繰り返しと誤認識するなどである。図 5にガラス瓶を叩く音とガラスコップを叩く音のスペクトログラムを示す。連続環境音では単環境音間ほど高い性能は得られていないが、表 6のように比較的近いカテゴリ間で認識誤りを起こす。したがって、環境音のクラスタリングを考えた場合にはそれらの誤りは同一のクラスに吸収される。この結果は、連続環境音のクラスタリングの可能性を示している。この方法を用いてクラスタリングを行なうことにより、後に述べるクラスモデルを用いた HMM 合成が行なわれる。

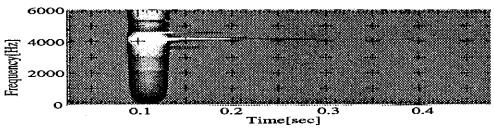


(a) クリップをはさむ音

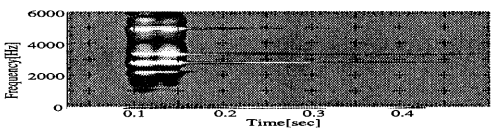


(b) 割箸を折る音

図 4: 誤認識単独環境音のスペクトログラム



(a) ガラス瓶を叩く音



(b) ガラスコップを叩く音

図 5: 誤認識連続環境音のスペクトログラム

3.3.2 環境音と音声の識別結果

テストデータ SNSET について図 1のように環境音全体から学習した全体環境音モデルを用いた単独環境音と音声の識別実験の結果を表 7に示す。また図 2のように個々の環境音モデルを使った単独環境音と音声の識別実験の結果を表 8に示す。表 7、8が示すように、単独の環境音については音声、環境音の識別は高い精度で行なえることがわかった。また、図 1のように多くの環境音を用いて一つのモデルを作ったものに比べ、図 2のように細かな種類で精度の高い環境音モデルを作る方が音声と環境音の識別率は高くなった。

同様にテストデータ SCNSET での識別結果について

表 7: 全体環境音モデルでの単独環境音、音声識別結果

テストデータ	識別率 [%]	音声を環境音 [個]	環境音を音声 [個]
SNSET1	99.7	0	1
SNSET2	100.0	0	0
SNSET3	99.7	0	1
SNSET4	100.0	0	0
SNSET5	99.7	0	1
平均	99.9	0	0.6

表 8: 個別環境音モデルでの単独環境音、音声識別結果

テストデータ	識別率 [%]	音声を環境音 [個]	環境音を音声 [個]
SNSET1	100.0	0	0
SNSET2	100.0	0	0
SNSET3	100.0	0	0
SNSET4	100.0	0	0
SNSET5	100.0	0	0
平均	100.0	0	0

表 9: 全体環境音モデルでの連続環境音、音声識別結果

テストデータ	識別率 [%]	音声を環境音 [個]	環境音を音声 [個]
SCNSET1	99.7	0	1
SCNSET2	99.7	0	1
SCNSET3	99.7	0	1
SCNSET4	99.7	0	1
SCNSET5	100.0	0	0
平均	99.8	0	0.8

表 10: 個別環境音モデルでの連続環境音、音声識別結果

テストデータ	識別率 [%]	音声を環境音 [個]	環境音を音声 [個]
SCNSET1	100.0	0	0
SCNSET2	100.0	0	0
SCNSET3	100.0	0	0
SCNSET4	100.0	0	0
SCNSET5	100.0	0	0
平均	100.0	0	0

も、全体で学習した環境音モデルを使ったものと個別の環境音モデルを使ったものをそれぞれ表 9、10に示しておく。この結果から連続環境音が存在する環境であっても環境音と音声の識別は精度良く行なえることがわかる。この場合についても環境音全体で作成したモデルに比べ個々に学習されたモデルを用いた方が音声と環境音の識別率は高くなった。この結果を用いることにより、マイクロホンアレーでの方向推定や、HMM 合成などでの雑音区間の切り出しなどが可能となる。

4 HMM 合成を用いた環境音重畳音声の認識

次に、音声データにいくつかの環境音データを重畳させることによって作成したテストデータに対してクリーンモデルに様々な環境音モデルを HMM 合成することによる認識率の改善を評価した。

従来法においては雑音 HMM は音声区間の前の雑音を使用して学習される。この方法は雑音が定常な場合には有効であるが、実環境に存在するような非定常な雑音については、効果が弱くなると考えられる。したがって、本稿では音声区間の前の時点において環境音を認識しクリーンモデルに前もって作成済みの環境音モデルを合成することで認識率の向上を試みる。

表 11: 環境音 HMM

状態数	2 状態エルゴディック
特徴量	MFCC
出力確率分布	連続分布型
サンプリング周波数	12kHz
データベース	RWCP 実環境音響データベース

表 12: 音素 HMM

状態数	3 状態
特徴量	MFCC
出力確率分布	256 混合 TiedMixture 分布型
サンプリング周波数	12kHz
データベース	ATR 日本語音声データベース
学習データ	特定話者 2620 単語

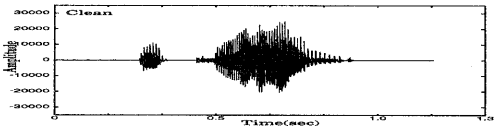


図 6: クリーンデータ /ikioi/

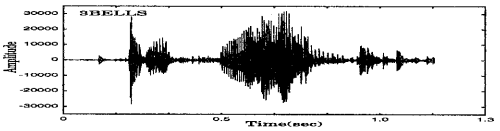


図 7: 複数ベル音声 /ikioi/

4.1 HMM モデル

HMM 合成用の環境音のモデルおよび音声のモデルをそれぞれ表 11、12に示す。

4.2 テストデータ

テストデータには特定話者の ATR 音韻バランス 216 単語を用いる。その単語に対して 1 種類のベルの音 (bells3) の 10 サンプルからランダムに選んだ数個のベルの音を重畳させることで音声を作成した (単独ベル音声)。また、3 種類のベルの音 bells1、bells2、bells3 の 30 サンプルからランダムに選んだ数個のベルの音が重畳した音声も作成した (複数ベル音声)。クリーンな単語と複数のベルの音が重畳しているデータを図 6、7に示す。これらのテストデータ作成用環境音はモデルの学習には使われていない。また、複数ベル音声データにおいて音声データ

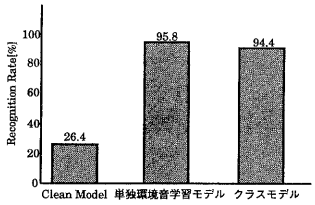


図 8: 単独ベル音声認識結果

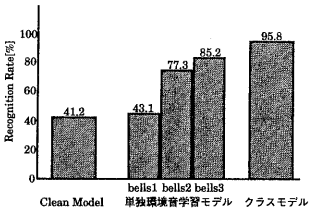


図 9: 複数ベル音声認識結果

の全パワーと環境音の全パワーの SNR は約 5dB 程度である。

4.3 HMM 合成を用いた認識実験

前述した HMM およびテストデータを使って、HMM 合成を用いた認識実験を行なった。比較対象として一つの単独環境音で学習した環境音モデルとクリーンモデルとの合成モデルを用いる。

実環境においては常に定常な環境音が存在する環境は少なく、類似しているものの異なる音が連続して続く状況が存在する (例えば、ドアの閉まる音)。このような状況が複数ベル音声でシミュレートされている。この状況で 1 つのデータのみを用いてモデルの合成を行なう場合には雑音モデルのミスマッチが起こり、認識性能が劣化する。そこで、その環境で起こりうる環境音をクラスタリングしておき、音声区間の前の環境音からいずれかのクラスを決定し、そのクラスのモデルを合成することで、認識性能の向上をはかる。

本稿では、環境音のクラスがベルのクラスであると識別できるとした上で、従来のように一つの環境音で学習した環境音モデルとクリーンモデルとの合成モデル (単一環境音モデル)、識別したクラスの環境音モデルとクリーンモデルとの合成モデル (クラスモデル) について認識性能の比較を行なう。使用したデータは、単独ベル音声と複数ベル音声の 2 種の環境音重畳 216 単語データである。

4.4 実験結果

単独ベル音声の認識結果を図 8に、複数ベル音声の認識結果を図 9に示す。図 8の様に同じ環境音が続く場合においてはクラスモデルに対し単独環境音モデルの方が良い結果となっているが、この場合においてもクラスモデルを用いたことによる大きな性能の劣化はない。また、本稿が目的とするような複数ベル音声の環境については図 9に示されるように一つの環境音で合成を行なった場合にはそれぞれの学習データの間に大きな変動があり、認識性能の劣化が起こっている。これに対し、クラスモデルを用いた場合は、高い認識性能を保っている。

5 おわりに

本稿では RWCP-DB の環境音データを用いて単独環境音、連続環境音の認識実験を行なった。単独環境音の認識については十分高い認識性能が得られたが、連続環境音については、十分な結果が得られなかった。しかし、連続環境音の認識誤りについては同一のクラス間の誤りが多く、入力環境音がどのクラスに属すかを調べる場合には十分な性能が得られると思われる。同様に、単独環境音、連続環境音と音声の識別実験については両環境音とも十分な識別性能が得られ、比較的簡単に両者の識別を行なえることがわかった。

また、複数の環境音が重畳する音声に関して HMM 合成を用いる場合には、雑音モデルの作成法として音声区間の前の一つの環境音を用いる方法は認識性能の劣化をひきおこすことがある。しかし、音声区間の前の環境音に対しクラスタリングを行ない、その環境音のクラスモデルとクリーンモデルを合成することで高い認識性能を得ることができた。

環境音は実環境の音声認識を考える場合には避けられない。今後は、今回示した環境音の認識や識別を用いて、環境音のクラスタリングやマイクロホンアレーの方向推定などを行なっていく予定である。

参考文献

[1] 大賀 寿郎, 山崎 芳男, 金田 豊, “音響システムとディジタル処理”, コロナ社, (1995).

[2] 阿部 正人, “多数センサによる音源推定”, 音響学会誌,51 巻,5 号,pp.384-389,(1995).

[3] 鹿野 清宏, 中村 哲, 伊勢 史郎, “音声・音情報のディジタル信号処理”, 昭晃堂, (1997).

[4] 滝口 哲也, 中村 哲, 鹿野 清宏, “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”, 電子情報通信学会論文誌,Vol.J79-D-II,NO.12,pp2047-2053,Dec.(1996).

[5] F.Martin, K.Shikano, Y.Okabe, “Recognition of Noisy Speech by Composition of Hidden Markov Models”, 電子情報通信学会技術報告, SP92-96,pp.9-16, (1992).

[6] 中村 哲, 比屋根 一雄, 浅野 太, 遠藤 隆, “実環境における音響シーンデータベースの構築”, 日本音響学会講演論文集, 10 月, pp.137-138, (1998).

[7] 新情報処理開発機構 技術研究組合 実環境音響サブワーキンググループ , “実環境音声・音響データベース報告書”, (1998).