音声言語情報処理　22－3
（1998．7．24）

環境適応機能付き音声認識ミドルウエア

小窪 浩明　大淵 康成　天野 明雄　畑岡 信夫

日立製作所　中央研究所
〒185-8601 東京都国分寺市東恋ヶ窪 1-280
Tel.：(042) 323-1111
E-mail：kokubo@hitachi.co.jp

あらまし

我々は，環境騒音と話者の変動に対して頑健なマイコン向け音声認識ミドルウエアを開発した．応用分野にはカーナビゲーションシステムなどがあり，製品開発者は音声認識ミドルウエアを利用することで，容易に音声認識機能を導入することが可能となる．本報告では，音声認識ミドルウエアの概要と本ミドルウエアに搭載された環境適応機能について詳細に述べる．HMM の移動ベクトルに対する補間係数を事前学習して行う話者適応方式については，従来方式である VFS との比較において優位であることを示した．また，HMM 合成方式に基づく雑音適応方式については，実車走行環境下での評価において，語彙数 1000 単語での認識率が 69%から 82%に改善することを確認し，提案方式の有効性を示した．

キーワード　・音声認識ミドルウエア　・マイコン　・話者適応　・雑音適応

Speech Recognition Middleware having Adaptation Functions for Environment

Hiroaki KOKUBO　Yasunari OBUCHI　Akio AMAMO　Nobuo HATAOKA

Hitachi Central Research Labs.
1-280 Higashi-Koigakubo Kokubunji Tokyo, Japan
Tel.：+81 042 323 1111
E-mail：kokubo@crl.hitachi.co.jp

Abstract

We have developed speech recognition middleware on a RISC microprocessor which has robust processing functions against environmental noise and speaker differences. The speech recognition middleware enables developers to use a speech recognition process for any possible speech applications, such as car navigation systems. In this paper we report implementation issues of speech recognition process in middleware of microprocessors and propose robust noise handling functions using noise adaptive models. We also propose a new speaker adaptation algorithm, in which the relationships are provided as a set of pre-trained interpolation coefficients. Experimental evaluation on 1000 word vocabulary speech recognition showed promising results for both robust processing functions of the proposed noise handling method and the proposed speaker adaptation method.

keywords　· speech recognition middleware　· microprocessor　· speaker adaptation · noise adaptation

1 はじめに

近年，音声認識機能付きのカーナビゲーションシステム[1][2]をはじめとし，さまざまな製品に音声認識機能が搭載され始めている．*SuperH*マイクロプロセッサ(以下，SH マイコンと略す)は，カーナビゲーションシステムや携帯情報端末などの CPU として広く採用されており，これらの機器に対する音声入力の要望は強い．従来，音声処理や画像処理には，マイコンの周辺回路として，それぞれ専用の LSI を必要としてきた．しかし，マイコンの高性能化に伴い，ソフトウエアだけで音声処理や画像処理などの機能を実現することが可能となってきた．このような形態をミドルウエアと呼ぶ．ミドルウエアは，専用の LSI を必要としないため，低価格，小型化が可能であり，また，ソフトウエアのみにより機能が実現されているため，開発のフレキシビリティが非常に高い．

我々は，SH マイコンをプラットフォームとした音声認識ミドルウエアを開発した[3][4]．本ミドルウエアは，ほぼリアルタイムの処理で語彙数 1000 語の単語認識を実現する．また，話者適応機能[5][6]と雑音適応機能[7]を採用することにより，話者のバラエティや騒音環境での使用に対する高い頑健性を特長としている．

本報では，はじめに，開発した音声認識ミドルウエアの概要を述べる．次に，ミドルウエアに搭載した話者適応機能と雑音対策に関して，それぞれの方式を詳細に説明する．また，自動車走行中に収録した音声データを用いた評価実験について報告し，騒音環境下での音声認識において，本提案方式の有効性を示す．

2 音声認識ミドルウエア

2.1 基本仕様

音声認識ミドルウエアは，SH マイコンのライブラリセットとして用意されている．表 2.1に本ミドルウエアの仕様を示す．この仕様は，カーナビゲーションシステムなどの製品ターゲットを念頭におき，さらに動作周波数 60MHz の SH-3 で実時間処理が可能であることを前提に決定された．

表2.1 音声認識ミドルウエアの基本仕様

サンプリング	11.025kHz / 12kHz 16bit
音響モデル	半連続 HMM 387 音素片 2 状態 3 混合 ｺｰﾄﾞﾌﾞｯｸｻｲｽﾞ 256
分析パラメータ	14 次 LPC ｹﾌﾟｽﾄﾗﾑ +14 次Δｹﾌﾟｽﾄﾗﾑ
分析フレーム長/ フレーム周期	20ms/10ms
語彙サイズ	1000~2000 words
処理速度	60MHz の SH-3 で実時間
話者対応	事前知識利用話者適応方式 (補間係数事前学習)
騒音対策	高速 HMM 合成方式 (選択的雑音適応)
メモリサイズ	256 kbyte (ROM) 500 kbyte (work)

音響モデルにはコードブックサイズ 256 の半連続 HMM を採用し，さらに，ビーム幅を制限したビタビサーチを行うことで，処理量の削減をはかっている．

2.2 環境適応機能

本ミドルウエアの特長は，話者適応機能[5][6]と雑音適応機能[7]を採用していることである．

(1)話者適応機能

SH マイコンは，ゲーム機から携帯情報端末まで幅広い応用が想定され，音声認識ミドルウエアとしては，あらゆる年齢層のバラエティや男性，女性の声質の違いに対応することが必須となっている．今回，補間係数事前学習に基づく話者適応方式を搭載した．本適応方式では，10 単語程度の単語を発声することで，発声者に適合した高精度な適応モデルが作成される．

(2)雑音適応機能

音声認識ミドルウエアは，カーナビゲーションシステムや携帯端末での応用など，騒音環境下での使用が想定され，雑音対策は必須となっている．音声認識の雑音対策には，スペクトル上で推定雑音成分を除去するスペクトルサブトラクション方式[10]と，音声 HMM に対して推定した雑音 HMM を重畳する HMM 合成方式[11]が広く用いられている．両方式について，並行して本ミドルウエアへの導入を検討しているが，今回の SH-3 向けには，音声認識時の処理量の制約から HMM 合成方式を採用した．HMM 合成方式は，事前に HMM を雑音適応しておくことで，音声認識処理中の負荷は増加しない．

以下，話者適応機能については第 3 章で，騒音適応機能については第 4 章で，それぞれ詳細に述べる．

3 話者適応機能

3.1 事前知識を利用した話者適応方式

話者適応は，利用者が発声した少数の音声データから不特定 HMM を修正し，適応化モデルを作成する．このとき，少量の発声データからすべての音韻モデルを適応するために必要な修正パラメータ(移動ベクトル)を抽出することはできない．このため，これまでに移動ベクトル場のスムージングを仮定し補間する手法が提案されている[8][9]．

本ミドルウエアには，独自に開発した補間係数事前学習に基づく話者適応方式を採用した[5]．本方式では，学習話者の移動ベクトル相互間の関係を線形予測係数の形で事前知識として抽出することで，精度の高い適応化モデルの作成を可能とする．

図 3.1に本方式のブロック図を示す．話者適応は，連結学習/補間/再推定の 3 ステップに分けられる．まず，不特定モデルを用いて発声単語と単語 HMM

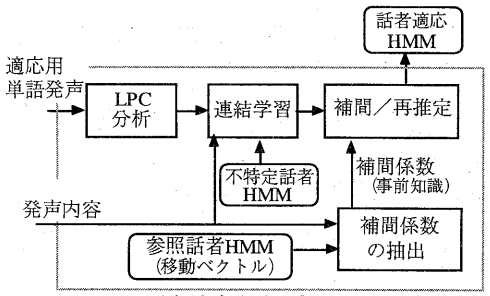


図3.1 話者適応方式のブロック図

とのビタビ照合を行なう．適応単語に含まれる分布については，対応する区間の特徴量の平均ベクトルを適応後の平均ベクトルとするとともに，適応前の平均ベクトルの差(移動ベクトル)を計算する．次に適応単語に含まれない分布(未学習モデル)については，以下の補間式により移動ベクトルを推定する．

$$\mathbf{V}_p = \sum_{q \in N^{(l)}} \mathbf{C}_{pq}^{(l)} \cdot \mathbf{V}_q \qquad (1)$$

ここで，$\mathbf{V}_p$ は未学習分布の移動ベクトル，$\mathbf{V}_q$ は学習された分布の移動ベクトルである．また，$\mathbf{C}_{pq}^{(l)}$ は次元毎の線形結合係数を要素に持つ対角行列，$N^{(l)}(p)$ は分布 p の近傍に存在する既学習分布の集合である．ここで，線形結合係数と近傍集合については，36 人分の特定話者モデルから予め計算しておき，事前知識として利用する．次に，すべての分布の移動ベクトルを次式に基づき再推定する．

$$\mathbf{V}_p' = \Big(\sum_{q \in N^{(R)}(p)} \mathbf{C}_{pq}^{(R)} \mathbf{V}_q + \mathbf{V}_p \Big) / 2 \qquad (2)$$

ここでは，p はすべての分布を含み，近傍集合 $N^{(R)}(p)$ もすべての分布を対象とする．また，ここでも線形結合係数と近傍集合とを事前知識として持っておく．

3.2 シミュレーション評価

JR 駅名 1000 単語を認識タスクとし，ワークステーション上でのシミュレーション実験をおこなった．評価条件を表 3.1に示す．実験では，評価話者 6 名が発声した JR 駅名 300 単語に対して，50 語を適応用の単語とし，残り 250 語を評価用データとして用いた．話者適応は，適応単語数を 1 単語から 50 単語まで順に変化させ，その都度認識率を求

表3.1 話者適応評価実験条件

学習データ (不特定話者モデル)	216 音韻ﾊﾞﾗﾝｽ単語 (男性３０名)
評価話者	男性 6 名
適応データ	JR 駅名 50 単語
評価データ	JR 駅名 250 単語
認識対象語彙	JR1000 駅名

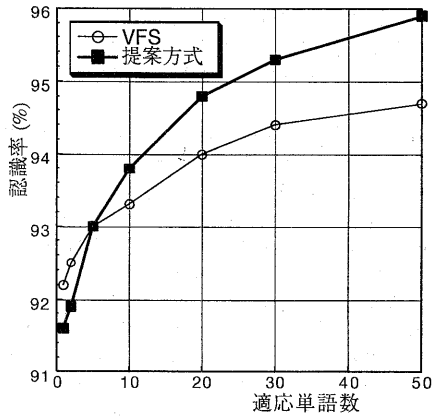


図3.2 適応単語数と認識率の関係

めた．また，比較のため，移動ベクトル平滑化法(VFS)[9]よる適応についても同じ条件で評価した．

実験結果を図 3.2に示す．それぞれの認識率は評価話者 6 名の平均値である．5 単語以下の適応単語数では，VFS のほうが高い認識性能を示しているが，適応語彙数を増やしていくにしたがって，提案方式が高い認識率を示すようになる．これは，提案方式の場合，補間時に特定少数の既学習分布に強く依存する傾向があるため，少ない適応発声でいくつかの分布を誤って適応させてしまった場合，その影響が他の分布まで広がりやすいのではないかと考えられる[6]．しかし，適応データ量が増えるにしたがって，提案方式の詳細な補間や再推定の効果が現れ，VFS の性能を上回る認識率を示すようになる．

また，評価話者中最も認識率の悪かった話者に対しては，話者適応なしで 77%であった認識率が，提案方式により，10 単語の適応で 84.5%，50 単語では 90.5%まで認識率が向上がした(VFS の場合はそれぞれ 82%，87%)．このように，話者適応機能は認識率の良くない話者に対しては特に有効である．

4 雑音対策機能

4.1 分布間距離尺度に基づく選択的モデル適応

HMM 合成方式は，音声認識処理中の負荷は増加しないものの，適応に必要な処理量は非常に大きく，騒音環境の変動に応じて逐次的に適応することは困難である．このため，半連続HMM に対する HMM 合成の処理量削減を目的とし，分布間距離尺度に基づく選択的モデル適応方式[7]を開発した．半連続 HMM における HMM 合成は，コードブックに格納されている全代表分布に対して雑音 HMM を合成することよっておこなわれる．HMM 合成の処理量は適応する分布の数にほぼ比例するため，コードブック内のすべての代表分布に対して HMM 合成を適用するのではなく，適用する代表分布を制限すれば，処理の削減は可能である．本方式は，コードブ

ックに存在する代表分布のうち，雑音によって変形を受けやすい分布を予め選択しておき，その分布のみを HMM 合成することにより処理の高速化をはかった．

図 4.1に提案する選択的雑音適応方式の概念図を示す．選択的雑音適応では，コードブックに格納されている代表分布を雑音による影響を受けやすい分布(カテゴリーA)と影響を受けにくい分布(カテゴリーB)とに予め分類しておき，雑音の影響を受けやすいカテゴリ A のみを HMM 合成することにより，少ない処理量で雑音適応コードブックを作成する．

雑音による影響の受けやすさの評価値としては，各代表分布と，その代表分布に対して雑音モデルを合成した後の分布との間で計算した分布間距離とした．この分布間距離尺度には Kullback Divergence[12]を用いた．コードブック中のコードワード m に対応する代表分布を $N_m[x]$ とし，この分布を HMM 合成によって雑音適応した分布を $N_{m,adapt}[x]$ とすると両者の距離は，次式で計算される．(ただし，$N_m[x], N_{m,adapt}[x]$ はともに無相関ガウス分布である．)

$$D_m = \int (N_m[x] - N_{m,adapt}[x]) \cdot (\log(N_m[x]) - \log(N_{m,adapt}[x]))dx$$

$$= \sum_i \left(\frac{\sigma_m^{\,2}(i) + \Delta_m(i)^2}{\sigma_{m,adapt}^{\,2}(i)} + \frac{\sigma_{m,adapt}^{\,2}(i) + \Delta_m(i)^2}{\sigma_m^{\,2}(i)} - 2 \right) \quad (3)$$

$$\Delta_m(i) = \mu_{m,adapt}(i) - \mu_m(i) \quad (4)$$

ここで，$\mu_m(i), \mu_{m,adapt}(i)$ はそれぞれ $N_m[x]$, $N_{m,adapt}[x]$ の平均ベクトルの第 i 成分，$\sigma_m^{\,2}(i)$, $\sigma_{m,adapt}^{\,2}(i)$ は共分散行列の第 i 対角成分である．このようにして，各代表分布毎に計算された分布間距離を値の大きい順に一定個数を選択し，対応する代表分布をカテゴリーA に登録する．この時，代表分布の出現頻度や，コードブック作成時のクラスタリング情報などにより重みを付けることも考えられるが，今回は考慮していない．

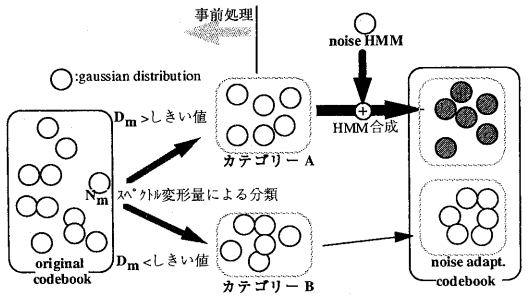
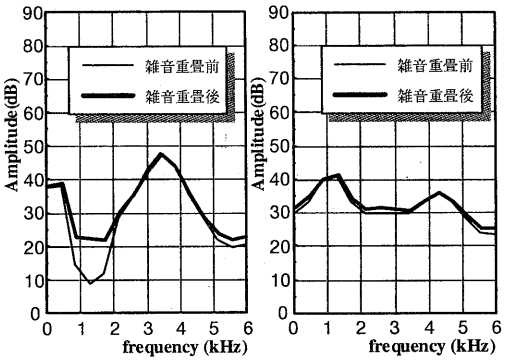


図4.1　選択的雑音適応方式の概念図



(a) 変形大 (D_m=9.34)　　(b) 変形小 (D_m=0.92)

図4.2　雑音重畳によるスペクトルの変形

図 4.2に，コードブック内の代表分布に対して HMM 合成を行った前後のスペクトルの変化を示す．雑音モデルには，自動車走行騒音から計算した 1 状態単一ガウス分布を用いた．図中，(a)は分布間距離 D_m が大きく(D_m =9.34)カテゴリ A に分類された分布，(b)は分布間距離 D_m が小さく(D_m =0.92)カテゴリ B に分類された分布である．また，細線が HMM 合成を行う前のスペクトルであり，実線が HMM 合成後のスペクトルである．(a)，(b)のスペクトルを比較すると，(a)ではスペクトルに大きな谷が存在し，HMM 合成によって雑音成分がその谷を埋めてしまうことでスペクトルが大きく変形する．一方，(b)のスペクトルは，比較的起伏が少ないために，雑音成分が重畳されてもその形状はあまり変形を受けない．これ以外にも，パワの大きな代表分布が HMM 合成による変形を受けにくく，パワの小さな代表分布ほど HMM による変形を受けやすい傾向があることを確認した．このように，コードブックに存在する代表分布において，重畳雑音による変形の受けやすさは，主に代表分布のスペクトル形状やパワの大きさによって決定され，雑音成分が特定の周波数帯域に集中するような特殊な状況を除いて，雑音の特性には依存しないと予想される．この仮定に基き，重畳雑音による変形の受けやすさの指標 D_m を求める際に使用する雑音モデルは，使用環境に応じて複数用意するのではなく，自動車走行騒音から計算した 1 状態単一ガウス分布のみとした．

4.2 評価実験

提案手法の有効性を検証するため，雑音重畳音声を用いた評価実験をおこなった．評価タスクは，不特定話者による JR 駅名 1000 単語の認識である．評価音声は，雑音のない環境で収録した 6 名の音声 1000 単語に対して，自動車の走行騒音を所定の SNR になるよう計算機上で重畳し作成した．

実験では，このコードブックに格納されている代表分布のうち 1/2 の分布のみを HMM 合成により雑音適応したコードブックを用いて認識を行った．こ

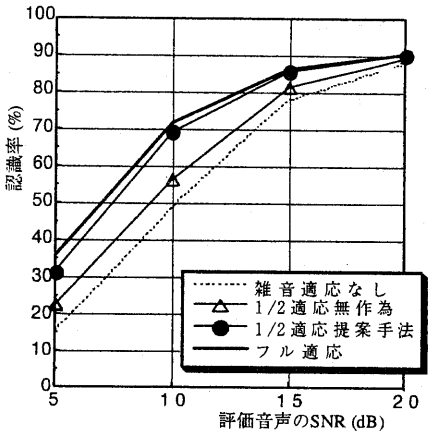


図4.3　雑音モデル適応の性能評価

のとき，雑音適応の処理量は処理する分布数にほぼ比例して1/2となる．また，このHMM合成を適用する代表分布の選び方として，以下の2つの選択基準を設けた．

(1)提案手法 (分布間距離D_mの大きい順に選択)
(2)無作為選択 (コードワード1,3,5,7,..を選択)

実験結果を図4.3に示す．図中の認識率は話者6人の平均値である．雑音適応の方式を比較すると，コードブック中にあるすべての分布に対してHMM合成をおこなった場合(フル適応)が，雑音環境下での認識性能の改善効果が一番大きい．例えば，SNR 10dBでの認識結果では，雑音適応なしで49%であった認識率が雑音適応で71.9%に向上している．次に，HMM合成をおこなう分布の選択基準による認識率の違いを比較する．SNR 10dBでは，任意に選択した場合の認識率が56.6%とフル適応の認識率に比べ15.3%の差があるのに対し，提案手法の認識率は69.4%とフル適応からの劣化は僅かである．以上の結果，本提案手法の有効性が示された．

5　実車環境での性能評価

第4章では，計算機上で雑音を重畳した評価音声を用いたシミュレーション実験により，HMM合成に基づく雑音適応の効果を確認した．シミュレーション評価では，大量の評価データを比較的容易に準備できるという利点がある反面，発声変形[13]の問題など，実際の使用状況と完全に条件を一致させることはできない．そこで，本ミドルウエアを搭載した評価用ボードを作成し，走行中の車内に持ち込んで評価を進めている．本章では，作成した評価ボードの構成と実車環境での性能評価について述べる．ただし，実機の評価では，騒音対策あり/なしで騒音条件を一致させた比較評価が困難なため，今回は，実機評価時に収録した音声を用いた計算機上でのシミュレーション実験についてのみ報告する．

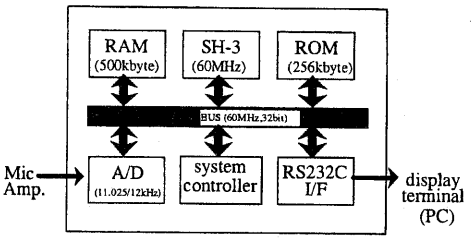


図5.1　試作ボードの構成

5.1　試作ボードの構成

音声認識ミドルウエアの機能および性能を評価するために，試作ボードを作成した．試作ボードの構成を図5.1に示す．SHマイコンは60MHzのクロックで動作するSH-3である．このSHマイコンの周辺に，プログラムやHMM，単語辞書などを格納するROM，ワークメモリとしてRAM，音声を取り込むためのA/Dコンバータ，ディスプレイ端末(PC)と接続するためのRS232Cインタフェースが配置され，各々が60MHz,32bitのバスで接続されている．

試作ボードを起動すると，ROMからプログラムとHMMがRAMに展開され，音声の取り込みが開始される．音声認識は，push to talk方式で，トリガーボタンを押すことで認識プログラムが実行される．認識した結果は，RS232Cインターフェースを介してPCのターミナル画面に表示される．

5.2　実車収録音声を用いた評価実験

評価ボードを用いた実機評価では，周囲騒音など評価条件に再現性が無いため，適応機能あり/なしの評価を同一条件で比較することは困難である．そこで，実機評価実験の際に発声した音声をDATで収録し，その収録音声を用いて計算機上でのシミュレーション評価を行った．

評価音声の収録条件を表5.1に示す．本評価ではできるだけ騒音の激しい環境を想定し，エンジン音の大きなRVタイプの車両を使用した．さらに変速装置のオーバードライブ機構をオフにし，一段低いギアで高速道路を走行した．使用した車種はdata1とdata2とで異なる．高速走行時に両車種を比較すると排気量の大きい2400ccの車両の方が騒音は大きい．発声者は2つのデータセットとも同一男性であり，それぞれJR駅名100単語を発声している．認識実験では，語彙数を1000語に設定した場合と，

表5.1　評価用音声データ

	data1	data2
走行条件	高速走行 100km/h,ｵｰﾊﾞｰﾄﾞﾗｲﾌﾞｵﾌ	
マイク位置	サンバイザ	
発声車	助手席男性	
発声内容	JR駅名100語	
使用車種	1BOXワゴン (2400cc)	1BOXワゴン (2000cc)

140語に絞った場合の2つの条件で行った．語彙数140語は，東京都内のJR駅名に相当する．

5.3 評価結果

評価結果を図5.2に示す．横軸は，雑音適応モデルのSNRであり，右端のデータが雑音適応を行わない場合の認識率を示す．HMM合成に基づく雑音適応による認識率の改善効果は，認識対象語彙が1000単語の場合，data1で69%から82%，data2で80%から84%，とそれぞれ認識率の向上が確認された．今回の実験では，RVタイプの車両を一段低いギアを選択して高速道路を走行させるなど，かなり厳しい条件での評価であった．このため，語彙数1000語での認識率は80%程度であるが，セダンタイプの乗用車であれば，90%近い認識率が達成されると予想される．また，語彙を140単語にした場合には，どちらのデータとも90%以上の認識率が得られている．

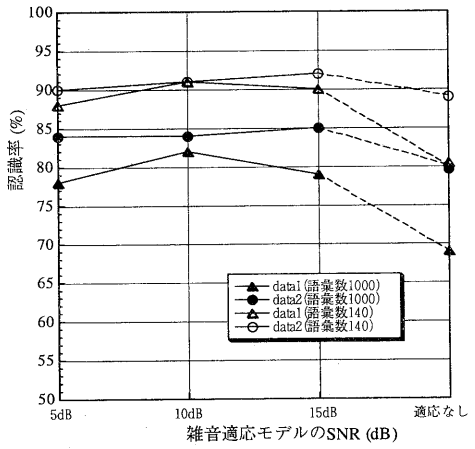


図5.2 走行車内で収録した音声を用いた認識実験結果

6 むすび

本稿では，SHマイコン向け音声認識ミドルウエアについて報告した．本ミドルウェアは，話者適応機能と雑音適応機能を搭載し，話者の変動や騒音環境下での認識に対して頑健であるという特長を持っている．シミュレーションによる評価では，本ミドルウエアに搭載した補間係数事前学習法に基づく話者適応方式が，従来手法(VFS)に対して優位な方式であることを示した．

走行中の車内で収録した音声を用いた評価実験では，RVタイプの車両を一段低いギアで高速走行させるなど，かなり不利な条件にもかかわらず，語彙数が140語の場合で90%以上の認識率が得られた．しかし，語彙数が1000語の場合には80%程度の認識率であり，さらに性能改善が望まれる．この結果は，騒音対策がまだ不充分であることもさることながら，処理量，メモリ量の制約のために，高精度な音響モデルが実現されていないことも原因の一つとしてあげられる．実際，半連続HMMのコードブックサイズを現状の256から1024に変更することで認識率が5～10%程度改善することを確認している．

今後は，スペクトルサブトラクション方式との性能比較や，両方式の併用などを検討すると共に，ミドルウエアの制約の中で，さらに高精度な音響モデルを実現し，騒音環境下での頑健性の向上を目指す．また，試作ボードを用いた実環境での評価を充実させたうえで，さまざまな製品への展開を図っていく予定である．

謝辞

本研究に関して協力，ご助言をいただいた日立製作所半導体事業部の大場信弥氏，鳴島正親氏，近藤和夫氏，塔下哲司氏，脇坂新路氏に感謝いたします．

文献

[1] 石井，他："カーナビゲーション用音声認識ユニット"，音響学会講演論文集,pp.189-190，1996.9

[2] 赤羽："カーナビゲーションと音声技術"，音響学会誌54巻3号，pp.223-228，1998.3

[3] 鳴島，他："システムインテグレーションを支える*SuperH*用音声合成・認識ミドルウエア",日立評論vol.79,No.11,pp.45-50,1997.11

[4] Hataoka, et al.："Development of robust speech recognition middleware on microprocessor" ,proc. of ICASSP98, pp.837-840, 1998.5

[5] Ohbuchi, et al.："A Novel speaker adaptation algorithm and its implementation on a RISC microprocessor", IEEE workshop on ASRU, 1997.12

[6] 大淵，他："移動ベクトルの相関に関する事前知識を利用した話者適応"，音講論pp.23-24,1997.9

[7] 小窪，他："分布間距離尺度に基づく選択的モデル適応に基づくHMM合成の高速化"，音講論 pp.105-106,1998.3

[8] Tonomura, et al.："Speaker Adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation", proc. of ICASSP-95, pp.688-691,1995

[9] 大倉，他："混合連続分布移動ベクトル場平滑化話者適応方式"，信学論(D-II),J76-D-II,12,pp.2469-2476,1993.12

[10] Boll："Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. on Acoustics, Speech and Signal processing, Vol. ASSP-27, No.2,pp.113-120 (1979.4)

[11] Martin, et al.："Recognition of noisy speech by composition of Hidden Markov Models", 信学技報SP92-96, pp.9-10,1992.12

[12] 渡辺,他："木構造確率分布を用いた音声認識",音講論,pp.13-14,1993.10

[13] Rajasekaran, et al.："Recognition of speech under stress and in noise", proc. of ICASSP-86, pp.733-736,1986.