SHマイクロプロセッサ向け音声認識ミドルウエアの開発

6N－5

小窪 浩明，大淵 康成，天野 明雄，畑岡 信夫
日立中央研究所

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[1]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．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の開発を行っている[2]．

本ミドルウエアは，ほぼリアルタイム処理で語彙数1000語の単語認識を実現する．また，話者適応機能と雑音適応機能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徴としている．

本報では，開発した音声認識ミドルウエアを実装した試作システムの概要を述べるとともに，搭載した話者適応機能と雑音対策について報告する．

2．システムの概要

2.1 音声認識ミドルウエア

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1にミドルウエアの仕様を示す．処理量の制約から音響モデルには半連続HMMを採用している．認識語彙1000語での処理速度は，動作クロック60MHzのSH-3を用いた場合で14 ms/frame (1 frame = 10ms)と，単語認識としてはほぼ実時間で認識処理を完了する．また，駅名1000単語を語彙とする評価実験での認識率は約93%であった．

表1 音声認識ミドルウエアの仕様

A/D	12kHz,16bit
音響モデル	387 Phoneme unit 2 state 3mixture Semi-continuous HMM
分析パラメータ	14次LPC cepstrum +14次Δcepstrum
分析フレーム長/ フレーム周期	20ms/10ms
語彙サイズ	1000 word
処理速度	14ms/frame
メモリサイズ	256 kbyte (ROM) 500 kbyte (work)

Development of speech recognition system on SH microprocessor
Hiroaki Kokubo, Yasunari Ohbuchi, Akio Amano,
Nobuo Hataoka (Hitachi CRL)
1-280 Higashi-koigakubo, Kokubunji, Tokyo 185,Japan

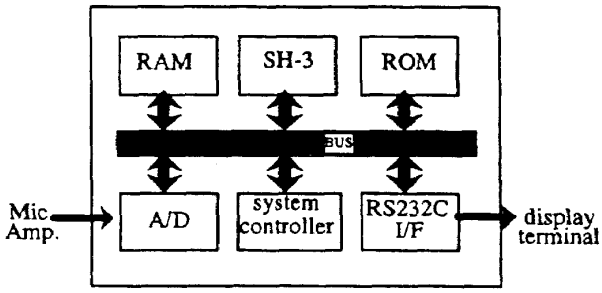

図1 試作ボードの構成

2.2 装置構成

音声認識ミドルウエアを搭載した試作ボードの構成を図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，音声認識結果をディスプレイ端末(PC)に出力するための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．

3．話者適応

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る．あらゆる年齢層や性差の違いに対応するため，話者適応機能を搭載した．

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[3]．

本ミドルウエアには，独自に開発した補間係数事前学習法に基づく話者適応方式を採用した[4]．図2に本方式のブロック図を示す．話者適応は，連結学

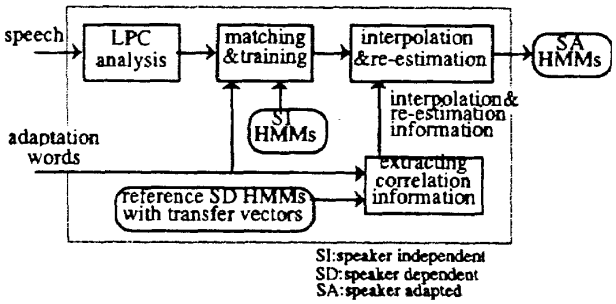

図2 話者認識方式のブロック図

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 HHH 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．

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

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，$\mathbf{C}_{pq}^{(I)}$ は次元毎の線形結合係数を要素に持つ対角行列，$N^{(I)}(p)$ 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．

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

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．

ワークステーションでのシミュレーションでは，特定話者モデルから抽出した事前知識を用いた提案手法は，移動ベクトルの平滑化を仮定した従来手法に比べて優位な結果が得られた[4].

4．雑音対策

カーナビや携帯端末などの応用では，さまざまな騒音環境での使用が想定されるため，雑音対策は必須である．

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と，音声 HMM に対して推定した雑音 HMM を重畳することで雑音適応する HMM 合成方式[5]が広く用いられている．本ミドルウエアには，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 を雑音適応しておけば良いため，音声認識処理中の負荷は増加しない．

HMM 合成方式の適用で問題となるのは，雑音環境適応をおこなった時点と実際に音声認識を開始する時点での騒音環境が変化する場合である．特に，カーナビゲーションへの応用を想定した場合には，騒音環境は常に変動している．そこで予備検討として，ワークステーション上での事前評価を行った．実験では，計算機上で雑音(走行騒音)を重畳した音声(駅名 1000 単語)を評価データとして用い，その評価音声に対して，1)クリーンな音声 HMM，2)評価音声と同一環境の騒音を HMM 合成した場合，3)評価音声と異なる環境(エアコン騒音)で HMM 合成した場合，の3条件について認識率を比較した．実験結果を図3に示す．SNR15dB の条件では，雑音対策なしの認識率が 78%であるのに対し，対策を行うことにより 86.9%と約 9%の認識率の改善が見られた．また，雑音が異なる条件で雑音適応を行った場合も，騒音環境が一致している場合とほぼ同等な認識率が得られた．この結果から，騒音環境が変化するような環境においても充分な性能改善効果が期待できる．

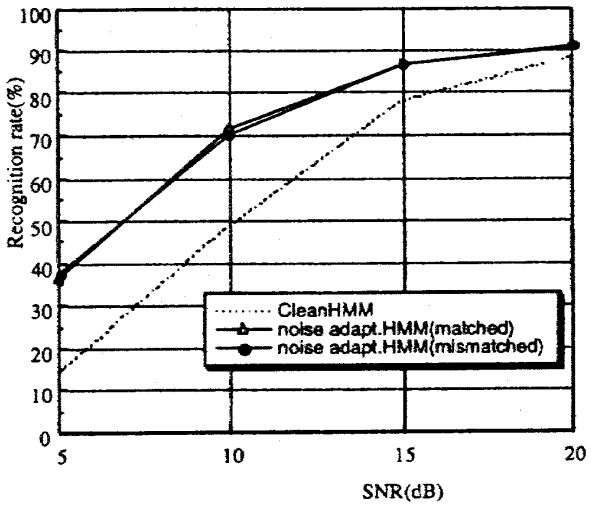

図3 雑音重畳音声(走行騒音)の認識率

5．むすび

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，話者や環境の変動に対して頑健であるという特徴を持つ．今後は実環境での評価をふまえ，さまざまな製品への展開を図っていきたい．

謝辞

本研究に関して助言をいただいた日立半導体事業部の大場氏，近藤氏に感謝する．また，SH マイコンへのインプリメントは日立半導体事業部の塔下氏，脇坂氏の協力による．

参考文献

[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]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

[3]Tonomura, et al., "Speaker Adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation", proc. of ICASSP95,pp.688-691,1995

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

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