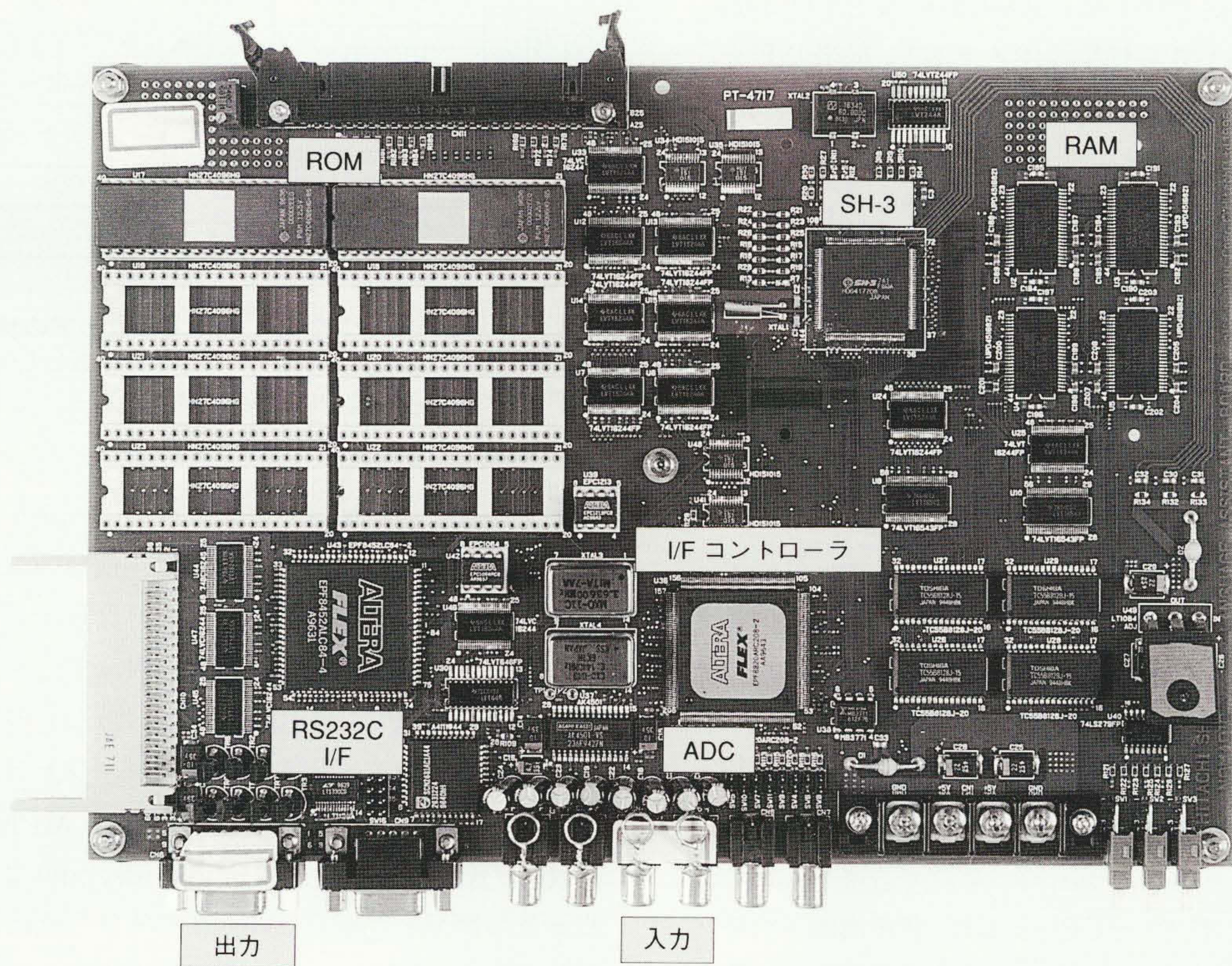


SuperHマイコン用音声ミドルウェア

Speech Middleware on “SuperH” Microprocessor

畑岡信夫 Nobuo Hataoka 鳴島正親 Masachika Narushima
近藤和夫 Kazuo Kondô 石川泰代 Yasuhiro Ishikawa



注：略語説明 ROM(Read-Only Memory), RAM(Random Access Memory), I/F(Interface), ADC(Analog-to-Digital Controller)

SuperHシリーズマイコン(マイクロコンピュータ)“SH-3”をCPU(中央処理装置)とする音声ミドルウェアを搭載した音声認識・合成回路基板

ROMに格納されたミドルウェア(プログラムとデータ)を処理時にRAMに展開し、SH-3をCPUとした高速な音声処理を実現している。

マイコン(マイクロコンピュータ)の性能向上に伴い、これまでは専用LSIやDSP(Digital Signal Processing)を用いて実現していた音声や画像などのメディア処理が、マイコンのソフトウェアで実現が可能となってきた。アプリケーションとハードウェアとを仲介し、かつマイコンの命令体系、メモリアクセスなどに依存して最適化されたソフトウェアをミドルウェアと呼んでいる。

このたび、日立製作所の「SuperHマイコン」をCPU(Central Processing Unit)とした音声ミドルウェアを開発し、その応用製品をトータルソリューションとして提案している。この音声ミドルウェアは、音声認識と音声合成の二つの機能を持っており、環境雑音と話者変動に対して認識率の劣化が少ないロバスト(頑強)な認識技

術と、高品質な合成音の特徴とする合成技術の成果である。音声認識や音声合成は、電話応用だけでなく、パソコンや情報家電などのマルチメディア機器でのユーザーインタフェースとして重要な機能となっている。今回開発した音声ミドルウェアの応用展開分野としては、カーナビゲーションや携帯端末、ゲーム機などがあり、いずれもユーザーインタフェースの向上を図った入出力機能として用いられる。

「SuperHマイコンシリーズ」のSH-3またはSH-4をプラットフォームとし、日立製作所半導体事業部と株式会社日立超LSIシステムズを販売窓口として、顧客への対応を図っている。

1 はじめに

従来、音声認識や音声合成は、処理量、メモリ規模、およびアナログ入出力機能の具備という三つ課題のために、専用装置と高性能なワークステーションでしか実現できなかった。しかし現在は、マイコン(マイクロコンピュータ)の処理規模が100 MIPS(Million Instructions per Second)を超え、さらに半導体メモリも、従来のディスクメモリと遜色がない規模になっている。この結果、ユーザーインタフェースとして不可欠な技術である音声認識・合成技術が、情報家電やカーナビゲーションなどのマイコン搭載製品へ利用できる環境が整ってきた。

日立製作所は、マイコンをCPUとした、アプリケーションに依存しない汎用の音声認識・合成機能である音声ミドルウェアの開発を推進し、顧客向けの具体的な応用に関するトータルソリューションとしての製品を提供している。応用の対象は、SuperHマイコン(以下、SHマイコンと言う。)がCPUとして搭載されているカーナビゲーション、PDA(Personal Digital Assistant)などの携帯端末、ゲーム機などへの音声入出力機能である。今後のミドルウェアの市場規模はマイコン市場としてとらえられ、現在の情報家電分野のマイコン市場は月額100億円、これが平成12年度には月額300億円の規模に達すると見られている。

ここでは、SHマイコンをCPUとした、音声認識・合成エンジンを備えた音声ミドルウェアの技術の特徴と応用展開について述べる。

2 音声ミドルウェアの概要

2.1 マイコン向けミドルウェア

ユーザーのアプリケーションとCPUであるマイコンの間に介在し、マイコンの処理機能に最適化したソフトウェアを「ミドルウェア」と呼んでいる。ソリューションとしてのミドルウェアの特徴を図1に示す。従来は専用のマイコンをCPUとして、さらに周辺回路として専用の音声処理LSIや画像処理LSIで行っていた音声処理や画像処理が、ミドルウェアだけで実現することができるようになっている。

ミドルウェアの特徴は、多様化対応、低価格、小型・低消費電力、さらには開発の短期化である。ROM(Read-Only Memory)のプログラムを換えることにより、音声処理や画像処理などの多様化に対応でき、結果として低価格化、開発の短期化が可能となる。

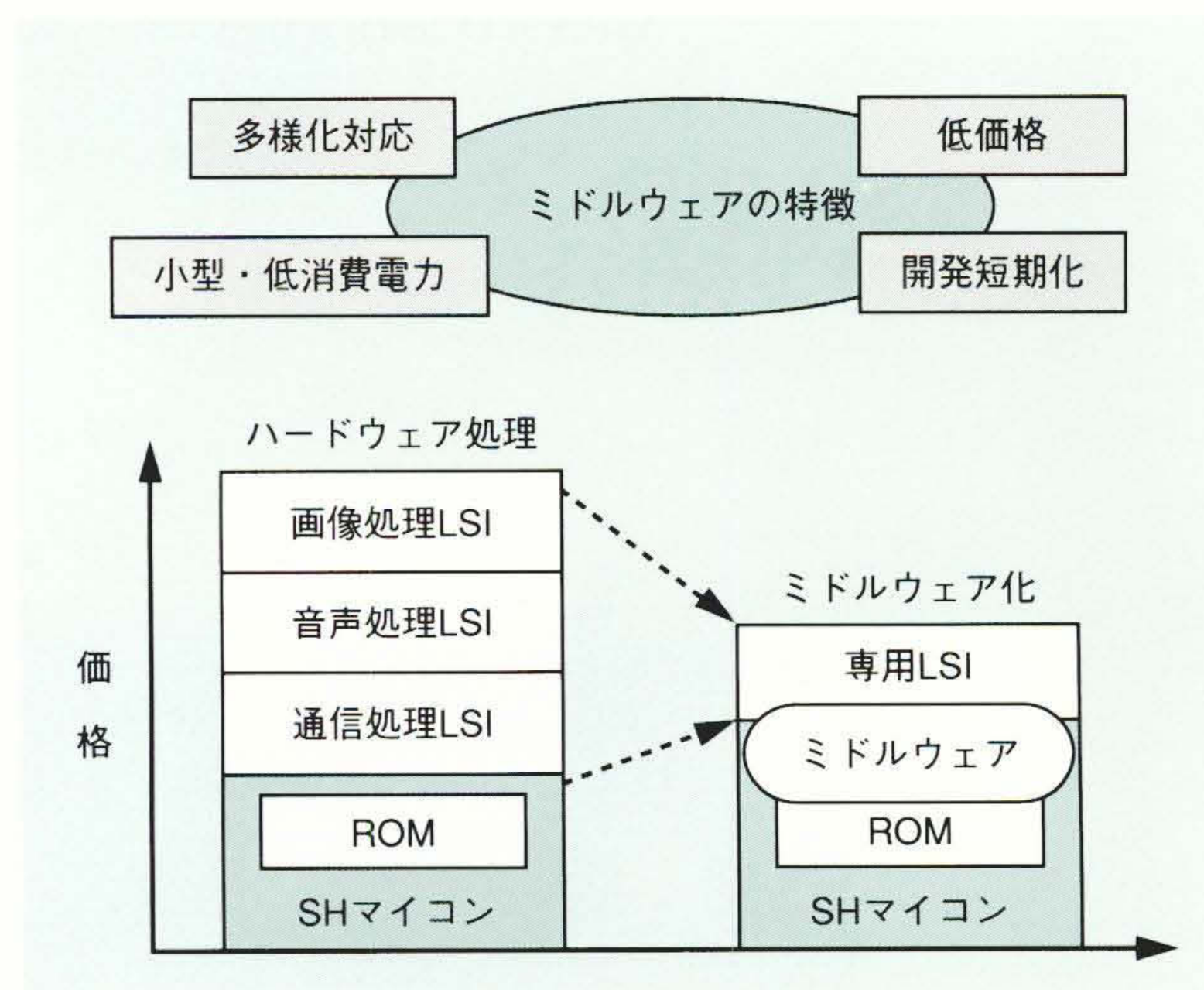


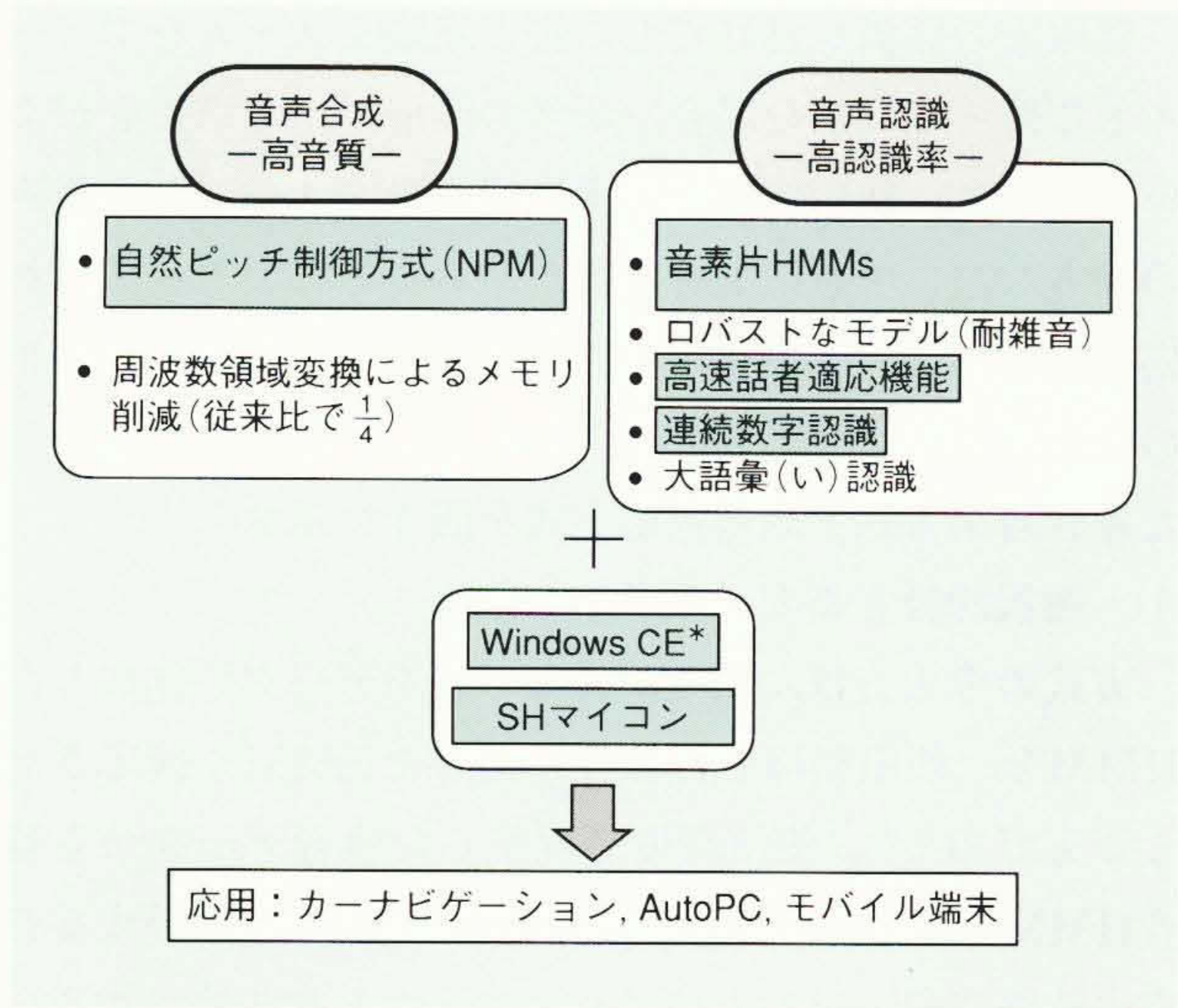
図1 ソリューションとしてのミドルウェアの特徴
多様化対応、低価格など顧客のニーズに合致している。

マイコンを使用するユーザーへのトータルソリューションを提供するために、音声、画像、通信の分野を対象として、ミドルウェアの品ぞろえを目標に開発を行っている。すでに、JPEG(Joint Photographic Experts Group)、文字認識、ADPCM(適応差分符号化)などのミドルウェアを実現しており、現在は音声ミドルウェア、通信用符号化G.723、G.729、H.263(codec)、MPEG(Moving Picture Experts Group)1 Video(decode)などのミドルウェアの開発と拡販を展開している。今後は、MPEG4、CG(Computer Graphics)、立体音響などへの展開を計画している。

2.2 SH音声ミドルウェア開発の基本戦略

SH音声ミドルウェア開発の基本戦略を図2に示す。音声ミドルウェアの決め手は、音声合成ミドルウェアでは「高音質」、音声認識ミドルウェアでは「高認識率」である。このため、開発のポイントを高音質合成と高認識率の実現に絞った。この結果、音声合成ミドルウェアでは、肉声からピッチパターンの類型化を行い、自然な合成音を作成できる自然ピッチ制御方式〔NPM(Natural Prosody Mapping)〕を実現した。

音声認識ミドルウェアでは、独自の認識単位としてHMM(Hidden Markov Model:隠れマルコフモデル)による音素片モデルを完成し、さらに使用環境雑音と使用話者変動に対応するロバスト(頑強)な認識手法を開発した。これらの詳細に関しては第3章で述べる。また、連続数字認識や大語彙(い)認識などの音声ミドルウェアの機能拡張とともに、汎用OS(Operating System)とし



注 1：略語説明ほか
NPM (Natural Prosody Mapping)
HMMs (Hidden Markov Models)
* Windows CEは、米国およびその他の国における米国Microsoft Corp. の登録商標である。
注 2：□ (日立製作所優位技術)

図 2 SH音声ミドルウェア開発の基本戦略
音声合成では高音質が、音声認識では高認識率がそれぞれキーポイントである。

てのWindows CEの利用とSHマイコンの性能向上とを背景に、カーナビゲーション、AutoPC、モバイル通信などの情報家電の分野での応用展開を積極的に推進している。

2.3 SH音声ミドルウェアの基本仕様

開発したSHマイコンをCPUとしたSH音声ミドルウェアの仕様を表 1 に示す。

認識方式は音素片HMM方式であり、環境変化と使用者の話者変動に強い機能として、雑音対策と話者適応を備えた認識仕様としている。音声合成の基本方式は波形音源を保持し、韻律制御としては、肉声からパターン化した韻律を用いることにより、自然な音声合成が可能となっている。

CPUは日立製作所のSH-3(60 MHz)マイコンである。仕様の決定にあたっては、SH音声ミドルウェア製品の典型的な応用であるカーナビゲーションを念頭に置いた。例えば、認識語彙数は地名单語認識をタスクとし、60 MIPS程度の処理能力で認識可能で、かつ現実的にカーナビゲーションのサービスが可能となる値として2,000単語を設定した。このために、認識処理全般の高速化を図り、特にHMM照合部では高速照合を可能とする方式を開発した。この結果、2,000語程度の語彙数としては、応答性の良いシステムの提供を実現している。

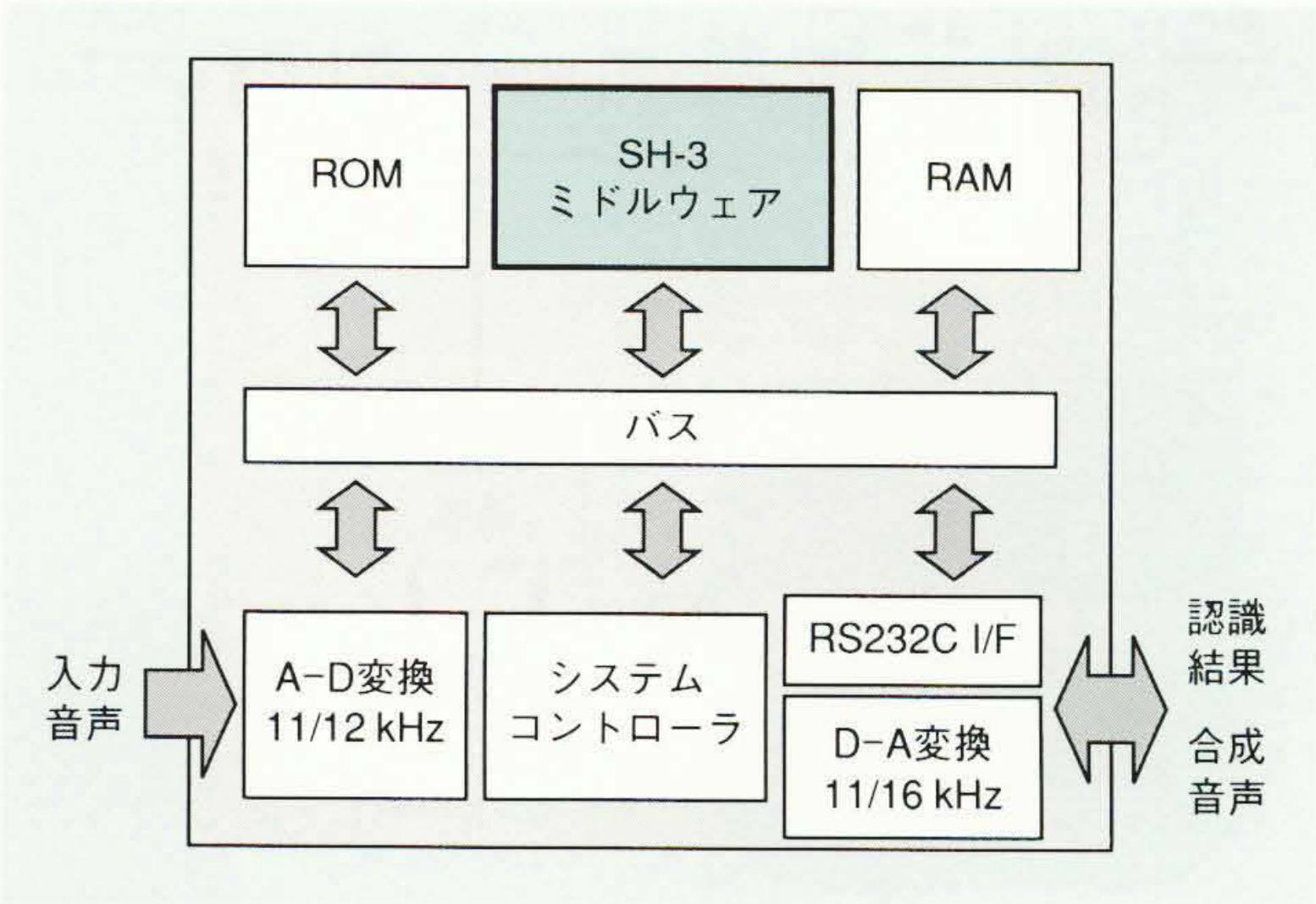
表 1 SH音声ミドルウェアの仕様
60 MHz版のSH-3を使用した場合の仕様を示す。

項 目		内 容
処理サイクル		60 MHz
外部バス		60 MHz, 32ビット
サンプリング周波数		11 kHz, 12 kHz, 16 kHz
音声認識	音響モデル	音素片・半連続HMM
	フレーム周期	10 ms
	フレーム長	20 ms
	処理時間	フレーム当たり12 ms
	応答時間	～0.6 s
	語彙数	2,000語
	メモリサイズ	200 kバイト (音響モデル, 辞書) 500 kバイト (ワーク)
音声合成	音源	母音・子音・母音波形
	韻律付与	肉声韻律
	合成単位	定型文・任意文章
	メモリサイズ	700 kバイト (音源, 辞書) 150 kバイト (ワーク)

2.4 SH音声認識・合成ボードの構成

SH音声ミドルウェアを搭載した評価用音声認識・合成ボードの構成を図 3 に示す。

入力音声は11 kHzまたは12 kHzでサンプリングして、CPUであるSH-3で音響モデルとプログラムをROMからRAM(Random Access Memory)に展開し、認識処理を実行する。認識結果は、RS232Cインタフェースを経由して、表示機能付き端末で表示する。音声合成の場合は、音源をROMから読み込んで、CPUで規則合成処理を実行し、D-A(Digital-to-Analog)変換した合成音声を出力する。



注：略語説明
A-D (Analog-to-Digital), D-A (Digital-to-Analog), I/F (Interface)
図 3 SH音声認識・合成ボードの構成
実際、ミドルウェアはROMに格納されており、処理実行時にRAMに展開される。

SH音声ミドルウェアを搭載した評価用SHボードを31ページの図に示す。

3 音声認識ミドルウェア

3.1 基本認識方式

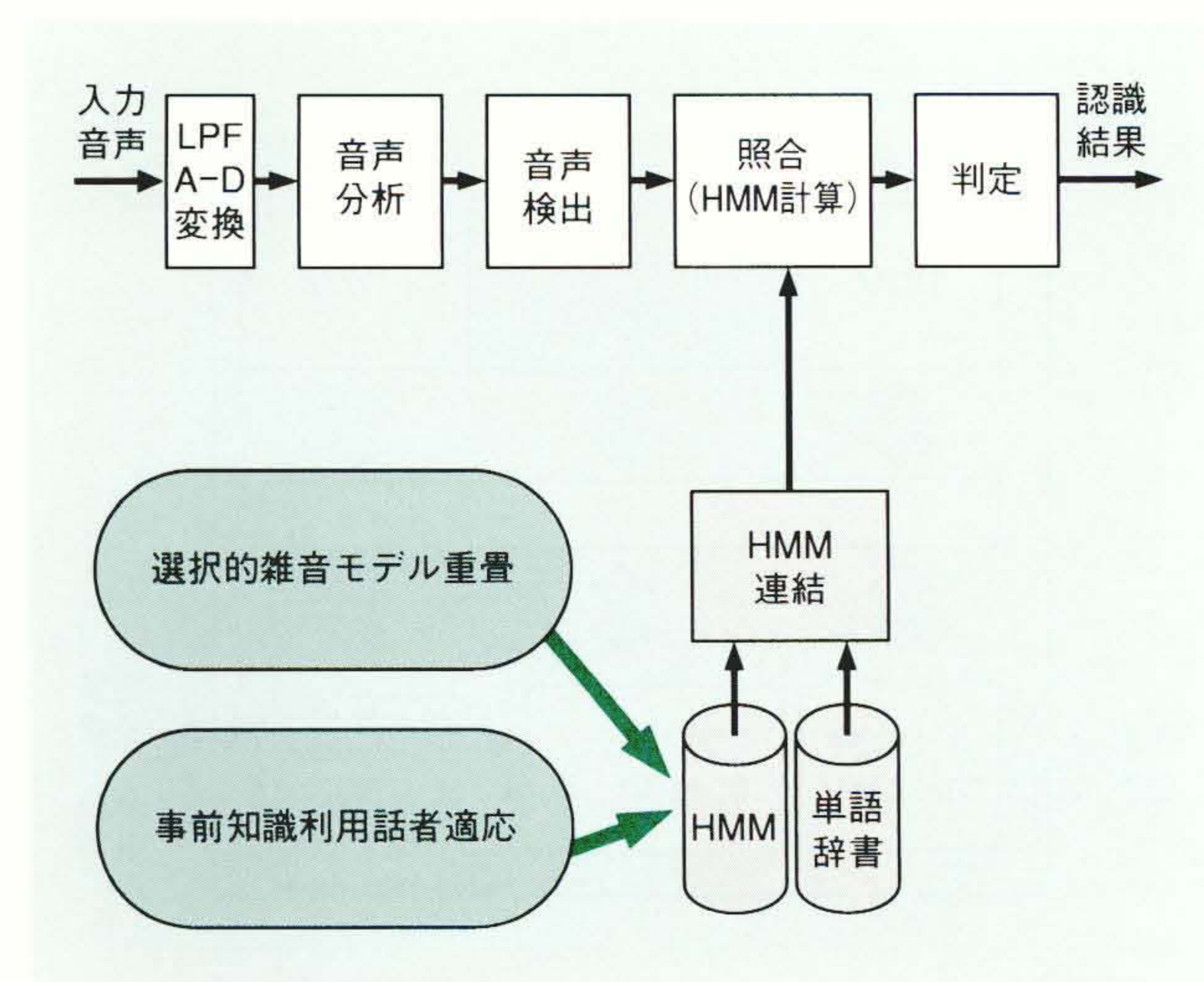
離散型HMMでは、学習の過程で、認識基本単位を音響パターンとしてあらかじめコード化しておき、HMM計算時にはテーブルを参照して処理を行う。したがって、処理速度は速いが、コードを記憶しておくデータメモリ容量が膨大になるという問題があった。

一方、拡張方式として連続型HMMがあり、音響パターンを正規分布の平均と分散の各パラメータとして持つことで、パターンをデータとして格納する必要がなく、記憶メモリの格段な削減が可能となる¹⁾。

SHミドルウェアでは、処理の高速化とデータメモリの削減が課題となっていた。そこで、まずデータメモリの大幅な削減が可能となる連続型HMMを用い、処理量の課題は、照合でのサーチ時間の削減を図ることによって解決することにした。さらに、認識単位の予備評価を行い、認識基本単位を独自の音素片モデルとした。ここでの音素片とは、子音、母音の各遷移部と定常部とを別々に表現した音響単位である。

3.2 環境対応技術

環境対応技術は、実際に使用されるフィールド環境での認識率劣化を防ぐために重要である。環境対応には、大別して、雑音対策と使用者の話者変動対策がある。



注：略語説明 LPF (Low-Pass Filter)

図4 雑音対策と話者適応

既存のHMMモデルを、環境雑音と使用話者へ適応する二つの手法を開発した。

実験室の評価が良好でも、音声認識システムをフィールドに持っていった場合、所望の性能が出ないことが多い。これは、使用現場での環境雑音や、ユーザーの発声の仕方などに予想以上の変動があることによる。環境対応技術は、音声認識を現実的に実用化するためには必須の技術である。SH音声ミドルウェア開発の中で開発した雑音対策方式と話者適応方式を図4に示す²⁾。

(1) 選択的雑音モデル重畳

方式の考え方は、環境が異なる音声データで作成したHMMを、使用環境での条件に適したHMMに適応させることである³⁾。使用環境で収集した雑音データから雑音HMMを作成し、既存のHMMに重畳する。重畳する音声の特徴空間の選び方により、加法性と乗法性の雑音のどちらにも対応が可能となっている。ここでは、車載での雑音対策を目的とし、主に加法性の雑音を対象とした。この際、雑音環境下で特徴が大きく変動する音響モデルだけに適応する選択的方式を開発した。評価結果を図5に示す。

車載雑音としては、走行雑音と空調雑音の2種類を想定し、評価タスクは駅名1,000語認識で、28名発声の評価音声データを使った。SN (Signal-to-Noise) 比は、5 dB、

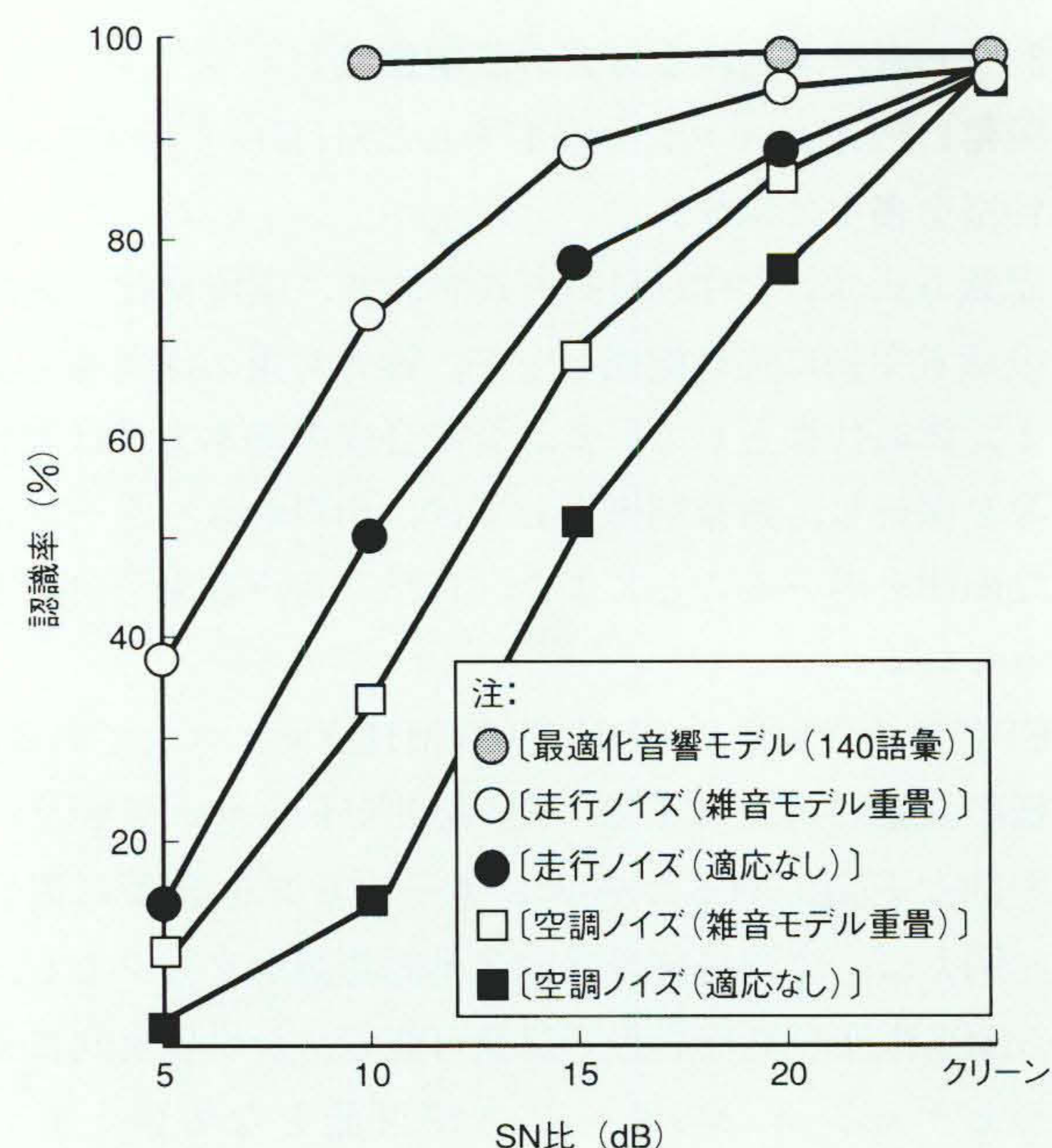


図5 選択的雑音モデル重畳の認識評価

走行ノイズでは雑音モデル重畳により、誤認識の約半分が改善され、さらに、走行ノイズに対して劣化のない最適化音響モデルを実現した。

10 dB, 15 dB, 20 dBの4種となるように雑音データをクリーンな音声に波形レベルで加えた。

空調雑音のほうが、走行雑音の場合よりも認識率が悪く、雑音モデル重畳方式の効果も少ないことがわかった。これは、走行雑音は車のエンジンノイズが主であり、エンジンの回転では数百ヘルツ以下のノイズ成分が多く、音声の帯域との重なりが少ない結果、ノイズ重畳モデルが良好に働いている結果だと類推される。一方、空調雑音は音声の帯域にまんべんなく加わっている結果、特に子音特徴の変動を起こし、認識率劣化の原因となっていると思われる。

走行雑音では、劣化の約半分のエラーが雑音モデル重畳によって改善されている。さらに、10 dB前後の高速道路上の走行時でも、140語彙で認識率の劣化がない最適化音響モデルを開発した。

今後の展開として、雑音成分をスペクトラム上で取り除くスペクトラム減算(Spectrum Subtraction)方式との併用を検討している。

(2) 事前知識利用話者適応

環境雑音に対するロバスト性のほかに、話者の変動に対するロバスト性が重要になっている。日立製作所は、指定した単語音声を発声することにより、認識システムを使用話者の音声の特徴に適応させる方式、いわゆる「教師あり話者適応」を開発した。話者適応は、使用話者の音声データから、不特定話者用音響モデルを使用話者に合った音響モデルに適応する機能である。

技術課題は、音響モデルの一部しか含まない少数の音声データから、使用話者の音響モデルの全体をいかに推定するかである。今までは、音響モデル間の補間や平滑化による手法、または、複数の音響モデルを用意し、それらの中から使用話者の特徴に近い音響モデルを変形させる手法を用いることが多かった。

今回開発した新方式は、多数話者データから音響モデル間の相関関係を事前に解析評価し、事前知識としてこの相関関係データを蓄え、話者適応時には少数の音声データから全体の音響モデルの変化を推定する方式である。事前データとして蓄えるメモリの量が若干多くなるが、少数の適応データから、全体の音響モデルを精度よく推定することが可能になる。これを、1,000駅名、話者数28名の評価タスクで評価した。認識評価結果を図6に示す。

この結果、適応単語数の増加とともに新方式の効果が大きくなることと、適応単語を50語近辺とすることによ

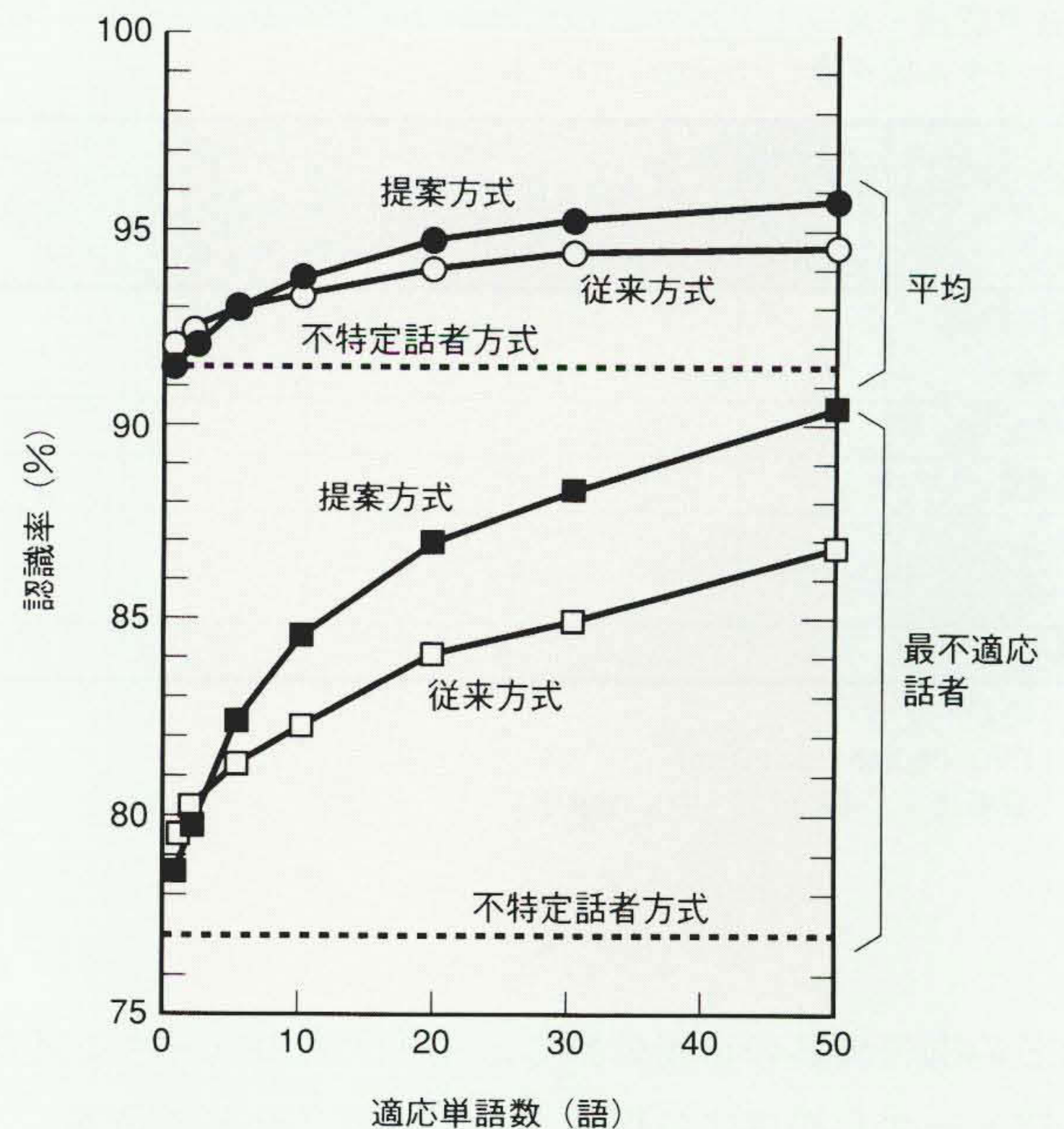


図6 話者適応の認識評価結果

話者適応により、最も認識率が低い話者(最不適応話者)の改善が著しい。

り、不特定話者型では80%以下であった話者の認識率が約90%に改善されることがわかった。

4 音声ミドルウェアの応用展開

ミドルウェアは、画像、音声、通信の融合を目標としたシステムの多様化への対応と、低価格化、小型化、低消費電力化などの廉価対応、および製品サイクルの速さに応じた開発短縮を可能とするシステム対応である。ユーザーがいかに新しいシステムを構築できるかが焦点であり、ハードウェアとソフトウェアの連携によるトータルソリューション対応が必須となっている。

このように、システムソリューションに対する要求への対応の観点から、ミドルウェアは、マイコンの拡販展開に大きな影響を与えるキーテクノロジーとなりつつある。市場規模では、今後のマイコンの売り上げの50%はミドルウェアの寄与によるものとなることが予想される。

市場でのミドルウェアの分野別ニーズを表2に示す。この中で、音声認識・合成を核技術とする音声ミドルウェアは、カーナビゲーションなどの車載機器から、セットトップ端末やゲーム機などの据置型機器、さらに携帯端末(HPC(Handheld Personal Computer))、携帯電話

表2 市場での分野別ミドルウェアのニーズ

音声認識・合成ミドルウェアは、カーナビゲーションをトップにデジタル家電品に必須のものである。

機能	カーナビゲーション	セットトップボックス	ゲーム機	DVD	ハンドヘルドパソコン	デジタルカメラ
音声認識・合成	◎	△	○	○	○	△
JPEG	△	○	○	—	△	○
MPEG	△	○	○	○	—	△
グラフィックス	○	○	○	—	△	—
モデム	△	△	△	—	○	—

注：略語説明ほか

DVD(Digital Video Disc)

重要度：◎(大), ○(中), △(小)

などの携帯機器の市場分野で、不可欠なヒューマンインタフェースを提供する機能として重要になっている。

カーナビゲーションの分野では、地図情報の検索や目的地のルート選定で音声認識機能は必要であり、ルートガイダンスやVICS(Vehicle Information and Communication System)情報の出力では、音声合成が必須の技術となっている⁴⁾。さらに、視覚と両手は運転操作に占有されており、安全性を考慮すれば、正に音声認識・合成技術が必須の分野となっている。カーナビゲーションの分野では、汎用のマイコンがCPUとして利用されており、音声ミドルウェアによるトータルソリューションの提供が強く望まれている。

5 おわりに

ここでは、SHマイコンをCPUとした音声ミドルウェアの特徴と応用展開について述べた。

音声合成ミドルウェアでは、肉声の韻律情報を類型化して利用する自然ピッチ制御方式により、自然な合成音声を実現した。音声認識ミドルウェアでは、音素片HMM方式をベースに、使用環境に対するロバスト性向上を目的とした雑音対策と話者適応、および高速な照合手法を実現した。この結果、60 MIPSのSH-3のマイコンをCPUとして、2,000語彙でほぼ実時間応答の音声認識ミ

ドルウェアを完成した。

今後は、文音声合成の自然性向上と大語彙認識を実現し、顧客のニーズにこたえるトータルソリューションへの対応を目指す考えである。

参考文献

- 1) 中川：確率モデルによる音声認識，電子情報通信学会(1988-7)
- 2) N. Hataoka, et al.: Development of Robust Speech Recognition on a RISC Microprocessor, Proc. of IEEE International Conf. on Acoustics, Speech and Signal Processing(1998-5)
- 3) F. Martin, et al.: Recognition of Noisy Speech by Using Composition of Hidden Markov Models, Proc. of Japan Acoustical Society, 1-7-10, pp.65~66(1992-10)
- 4) 赤羽：カーナビゲーションと音声技術，音響学会誌，Vol. 54, No.3, pp.223~228(1998)

執筆者紹介



畑岡信夫

1978年日立製作所入社，中央研究所 マルチメディアシステム研究部 所属
現在，音声・音響処理，ヒューマンインタフェース技術の研究開発に従事
工学博士
電子情報通信学会会員，音響学会会員，IEEE会員
E-mail: hataoka@crl.hitachi.co.jp



近藤和夫

1974年日立製作所入社，半導体事業部 システムLSI本部 マルチメディアLSI開発センタ 所属
現在，SuperHマイコン用音声認識ミドルウェアの開発に従事



鳴島正親

1971年日立製作所入社，半導体事業部 システムLSI本部 第1システムLSI設計センタ 所属
現在，SuperHマイコン用音声ミドルウェアの開発に従事
E-mail: narushim@cm.musashi.hitachi.co.jp



石川泰代

1970年日立製作所入社，株式会社日立超LSIシステムズ 情報システムセンタ 第1情報システム部 所属
現在，音声ミドルウェア，ボードシステムの開発に従事