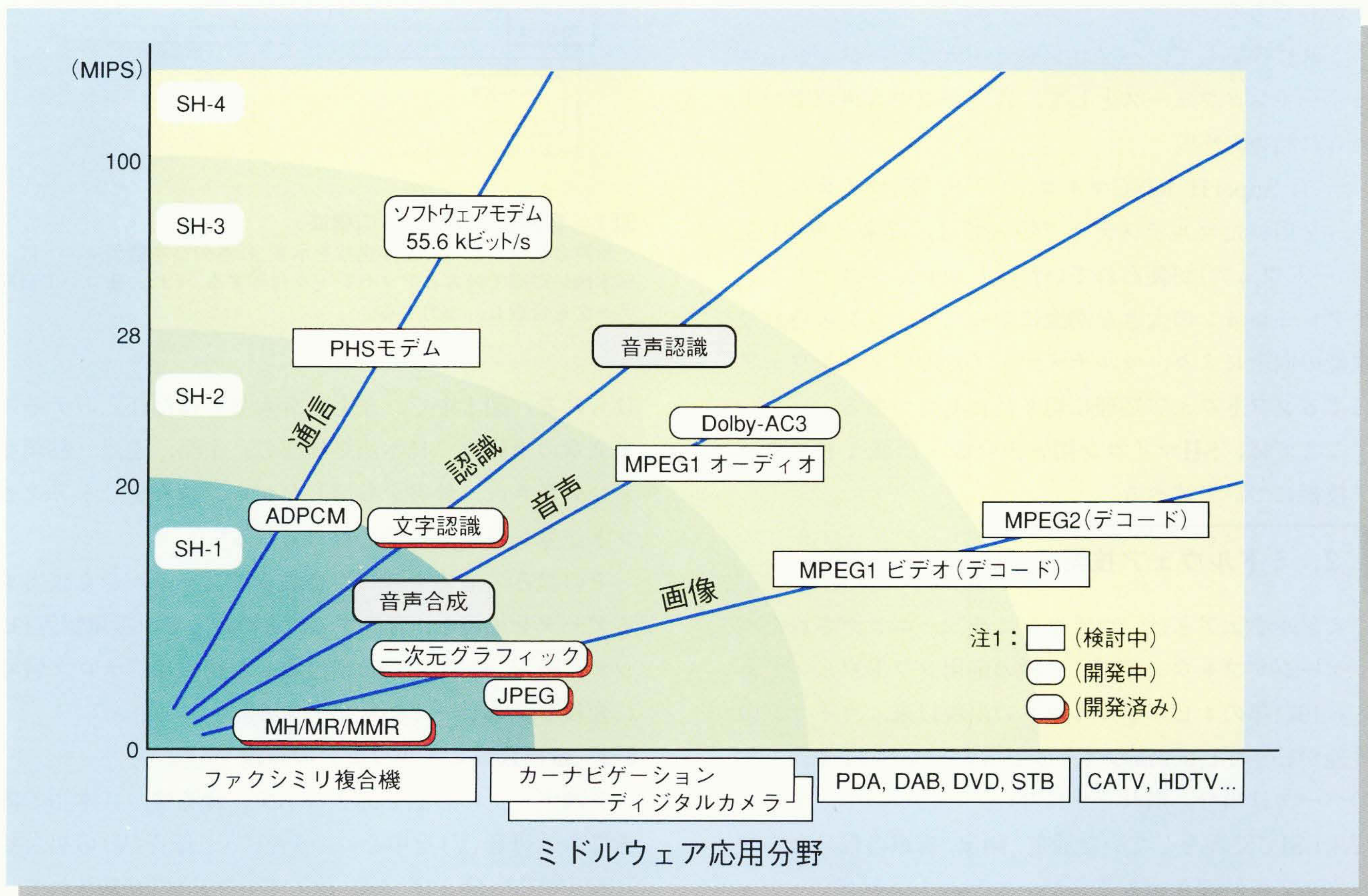


システムインテグレーションを支える SuperH用音声合成・認識ミドルウェア

Speech Synthesizing and Speech Recognition Middleware for
SuperH Microcomputers Supporting System Integration

鳴島正親 Masachika Narushima 畑岡信夫 Nobuo Hataoka
矢島俊一 Shun'ichi Yajima 近藤和夫 Kazuo Kondô



注2: 略語説明 PHS (Personal Handyphone System), ADPCM (Adaptive Differential Pulse Code Modulation), MH (Modified Hoffman) MR (Modified Read), MMR (Modified MR), JPEG (Joint Photographic Experts Group) MPEG (Moving Picture Experts Group), PDA (Personal Digital Assistant), DAB (Digital Audio Broadcasting) DVD (Digital Video Disc), STB (Set-Top Box), CATV (Cable Television), HDTV (High-Definition Television)

SuperHマイコン用ミドルウェアの展開

SuperHマイコン(マイクロコンピュータ)用ミドルウェアの開発マップを示す。従来、専用LSIで処理していた通信、認識、音声、画像の信号をミドルウェアで処理させることがシステムインテグレーションの潮流になりつつある。

マイコン(マイクロコンピュータ)の急激な性能向上に伴い、これまで専用LSIで処理していた音声信号を、ミドルウェアで処理する傾向が増大している。

現在、高音質を実現するために、生の波形を記録する方式による音声合成ミドルウェアを開発中であり、製品化予定である。これにより、任意の日本語テキストを読み上げることができるようになる。

また、高い認識率を得るために、音素片による認識手法を採用した音声認識ミドルウェアもあわせて製品化を予定している。このミドルウェアでは、雑音を音響モデ

ルに重畳する雑音抑圧法、認識語彙(い)高速絞り込み方式、話者適応などの機能を取り込んだ。その結果、SuperH RISC(Reduced Instruction Set Computer)マイコンでは、1,000語認識時の性能として、認識率95%、認識時間0.5秒、マイコン負荷50 MIPS(Million Instructions per Second)の見通しを得られた。

さらに、100語以下の小語彙領域では計算に用いる次元数を低減し、50語認識は15 MIPS程度で処理することができるようになる。

1. はじめに

20年以上前から続いている音声の研究がようやく大きな花を咲かせようとしている^{1),2)}。それは、(1) 音声処理のアルゴリズムの進化、(2) 処理を行うマイコンの急激な性能向上と低価格化、(3) ハンズフリーの自動車電話などに見られる社会からのニーズの高まりなどが整合しつつあるからである。

これに呼応して、パソコンやカーナビゲーションのユーザーインターフェースとして、音声合成や音声認識が用いられ始めた^{3),4)}。

一方、SuperH RISCマイコン(以下、SHマイコンと言う。)を用いたマルチメディア処理では、従来、専用LSI(ハードウェア)が使われていた。しかし、システムインテグレーションの大きな潮流に沿ったマイコンの急激な性能の向上により、マルチメディア処理がミドルウェアによるソフトウェア処理に切り替わりつつある。

ここでは、SHマイコン用音声合成・認識ミドルウェア技術について述べる。

2. ミドルウェア技術

ミドルウェアとは、マイコンにチューニングされたパッケージソフトウェアで、一種の応用ソフトウェアである。1971年の4ビットマイコンの出現以来、マイコンの性能が15年で1,000倍のペースで増大しており、今後もこのペースは当分、崩れそうにない。したがって、従来、専用LSIで処理をしていた通信、画像、音声の信号をミドルウェアで処理させることがシステムインテグレーションの潮流になりつつある。

ミドルウェアによって部品点数を削減できることから、(1) システム全体のコストダウンと低消費電力化、(2) 市場からの多様化へのニーズに容易に対応、(3) 新製品開発期間の短縮などが図れる。

したがって、ミドルウェア開発にあたっては、上にあげた条件を満たすように設計しなければならない。例えば、ミドルウェアのメモリ(プログラム、データ、ワークエリア)の最小化、マイコンと外部メモリとの高速化接続、マイコンへの負荷の低減などが必要である。

3. 音声合成

3.1 システムの概要

音声合成システムの構成を図1に示す。

システム全体は、日本語処理部と規則合成処理部に分

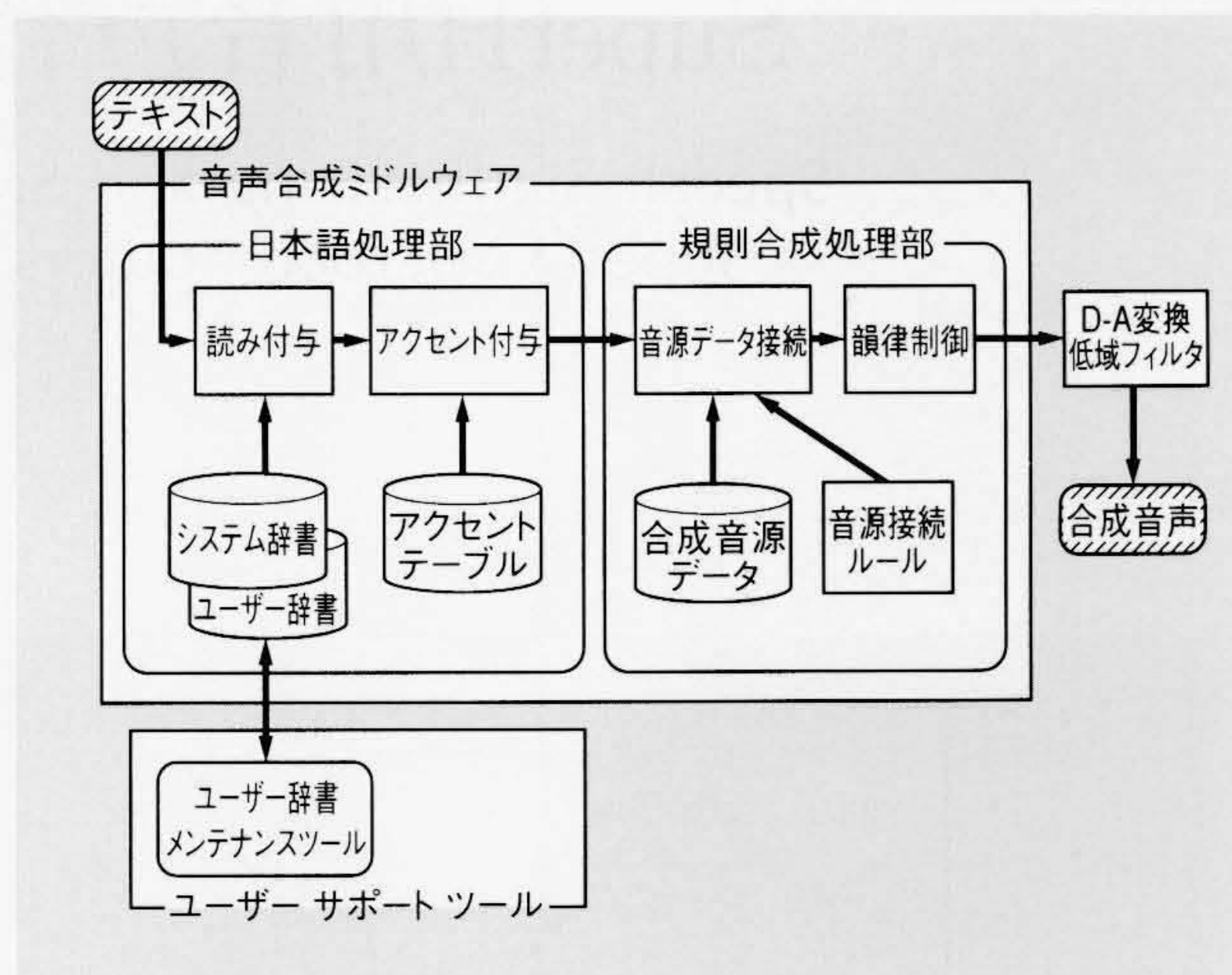


図1 音声合成システムの構成

音声合成システムの全体構成を示す。任意の日本語テキストに、日本語処理部で読みとアクセントを付与する。それに基づいて音源データを合成し、出力する。

けられる。図1中で、左上から入力された任意のテキスト文章の単語は、日本語処理部で、主語、述語、動詞などに分類され、辞書を参照しながら、読み記号とアクセント記号が各単語単位に付けられる⁵⁾。

その読みの情報を基に、合成音源ファイルから該当するデータを抽出し、合成する。そのあとで、韻律制御(ピッチ、速さ、抑揚)を行う。生成した情報はアナログ信号に変換し、スピーカから出力する。

3.2 音声合成ミドルウェアの特徴

音源データの形態を図2に示す。音源は、日本語の特徴である母音(V)を中心に、子音(C)と母音(V)の組合せ(CV, VCV, 例えば, ka, asa, ibuなど)で記録されている。このシステムでは、高音質を得るために、音の情報は生波形の形で処理する。

音声合成ミドルウェアの概要を表1に示す。基本的な合成方式は、波形重畳方式で、任意の日本語入力(漢字かな交じり文)に対応することができる。プログラムサイズは、日本語処理部が約100 kバイト、規則合成処理部は約40 kバイトである。このシステムの特徴は、以下のとおりである。

- (1) 生波形記録音源〔明瞭(りょう)性、肉声感〕
- (2) 高精度音声制御機能(ピッチ、速さ、抑揚)
- (3) 任意文章の音声合成〔テキスト、VICS(Vehicle Information and Communication System)対応〕
- (4) 形態素解析処理(漢字かな交じり文対応)

3.3 音質向上

音声合成で最も重要な項目は、音質である。このミド

表2 音声認識ミドルウェアの概要

プログラムやデータのサイズは合計約300 kバイトとコンパクトであり、音素片音声認識方式の開発によって高認識率が得られる。

項 目	内 容
認 識 方 式	音素片音声認識方式
認識対象言語	日本語の単語
認識対象語数	不特定話者 最大1,000語
認識速度	0.5 s
プログラム容量	約40 kバイト
辞書データ、音響モデル	約250 kバイト

半音節は、子音(C)と母音(V)の組合せを単位として認識する。これらに対して、この論文で述べるミドルウェアでは、音素片を単位として認識している。この音素片とは、音韻と半音節を組み合わせたもので、C, V, CV, VCを単位として認識処理を行う。この方式では、ある波形のサンプリング周波数が2倍になった効果が期待される。しかし一方で、マイコンへの負荷軽減もミドルウェアには重要な項目なので、実際の計算では、状態数を他の方式に比べて $\frac{1}{2}$ 以下に設定している。後述の図5に示すように、結果的にマイコンの負荷を増大することなく、音素片認識方式では他の方式に比べて高い認識率が得られている。

4.3 音声認識ミドルウェアの性能

音声認識のミドルウェアをSuperHマイコン“SH-3”に適用した場合の認識率と認識単位との関係を図5に示す。図5の縦軸は認識率を示し、横軸は図4に示した認識単位である。また、図5中の黒丸は約1,000語の駅名を、白丸は800語の人名を認識させた場合のそれぞれの認識率である。これによると、図4の認識単位で示した音韻、

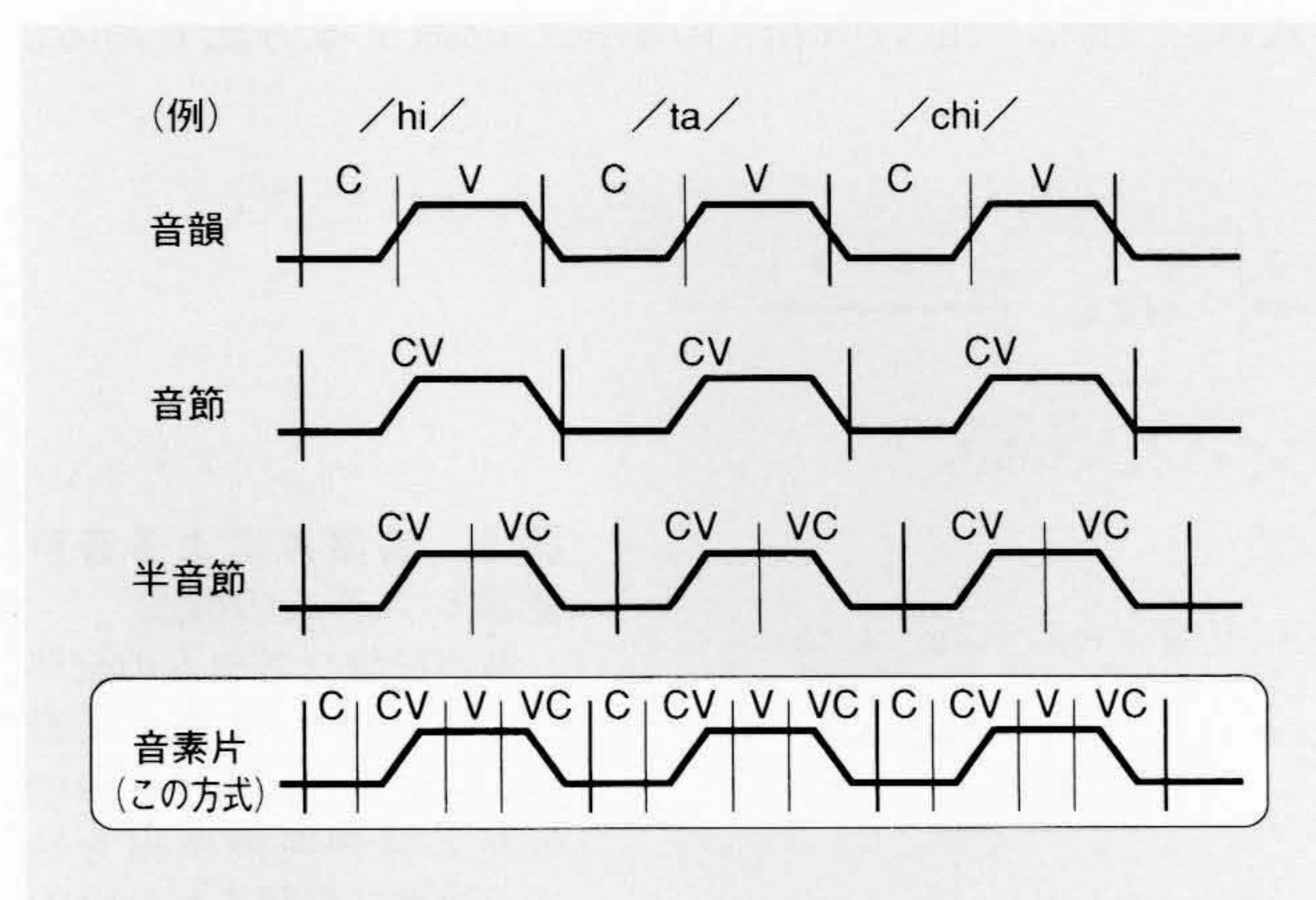


図4 音素片認識方式

“hi-ta-chi”を例にした場合の認識単位を示す。この論文のミドルウェアでは、音素片認識方式を採用している。

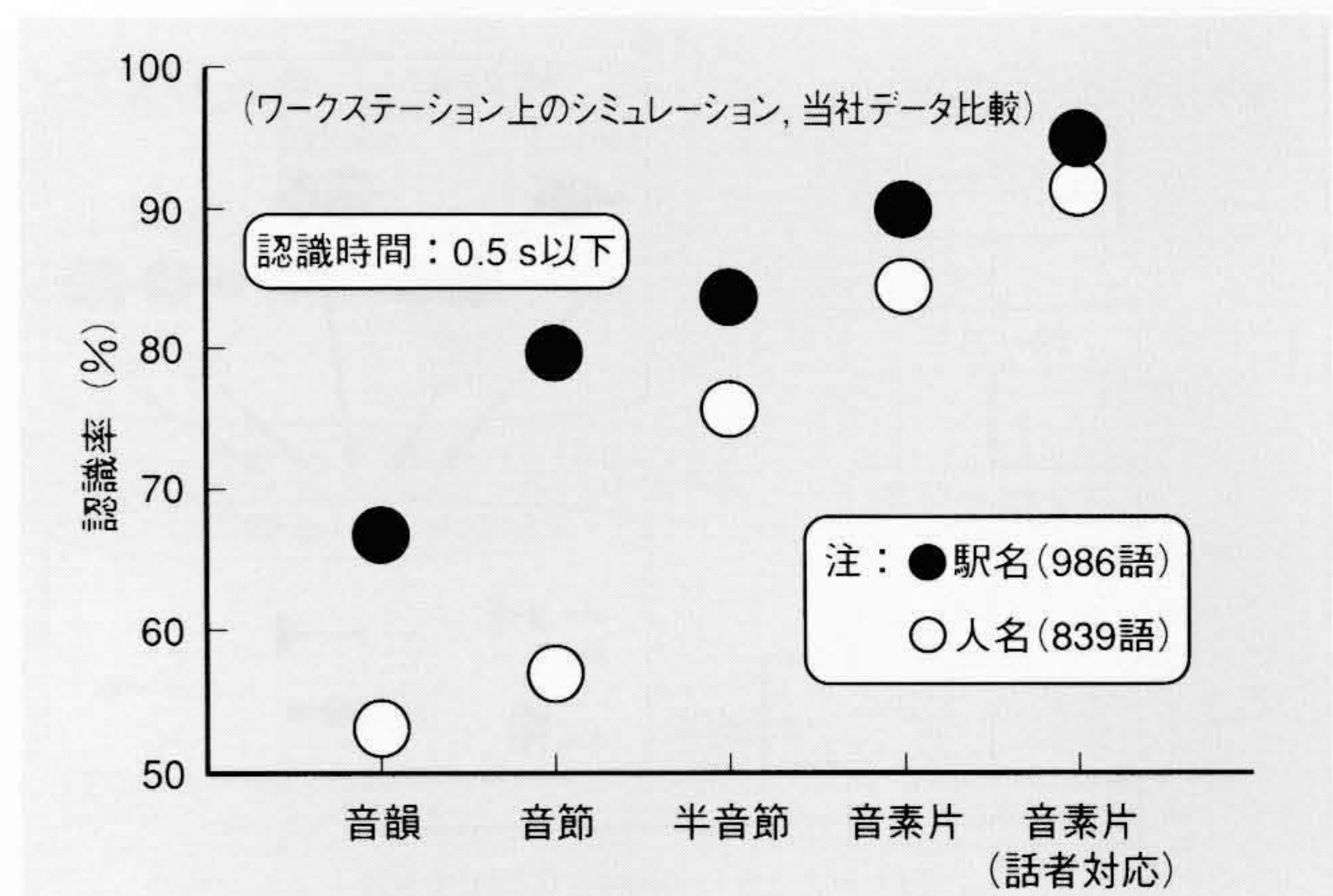


図5 各種認識方式の確認性能

認識単位と認識率の関係を示す。音韻、音節、半音節、音素片の順で認識率の向上が見られる。

音節、半音節、音素片の順で認識率が向上していることがわかる。すなわち、音素片認識方式により、マイコンの負荷を増大することなく、高い認識率が得られる。

認識語数とマイコンに必要な性能〔MIPS(Million Instructions per Second)〕との関係を図6に示す。

同図中のデータから、マイコンにかかる負荷には、認識語数に依存する項と依存しない項があることがわかる。認識語数に依存しない項としては、音声入力・分析処理と雑音対策処理がある。一方、語数に依存する項は、音声認識用データ検索部分の照合処理である。結果として、2,000語で80 MIPS, 1,000語認識で50 MIPS程度のマイコン性能で音声認識をそれぞれ処理できる。

4.4 音声認識処理の雑音対策と高速化

マイコンの負荷を下げ、処理速度を上げるには、(1) マイコン負荷の軽い雑音対策^{3), 4)}, (2) 認識計算回数の削減, (3) 外部メモリへの高速メモリアクセスが特に有効である。以下、これらの各項目について述べる。

(1) 雑音対策

音声認識では、使用環境の雑音が認識率を大きく下げる要因となっている。このミドルウェアでは、雑音モデルをHMM音響モデルに重畳する方式を採用している^{3), 4)}。これは、得られる性能が高い割にはマイコンにかける負荷が小さいからである。

実験によると、対策無しの場合、信号対雑音比が10 dB程度の環境では60%の認識率であるが、これに雑音重畳モデル方式を適用すると83%にまで認識率が改善される。

このほかの雑音抑圧方法には、スペクトラム領域で差分をとる方式などが提案されているが、マイコンへの負

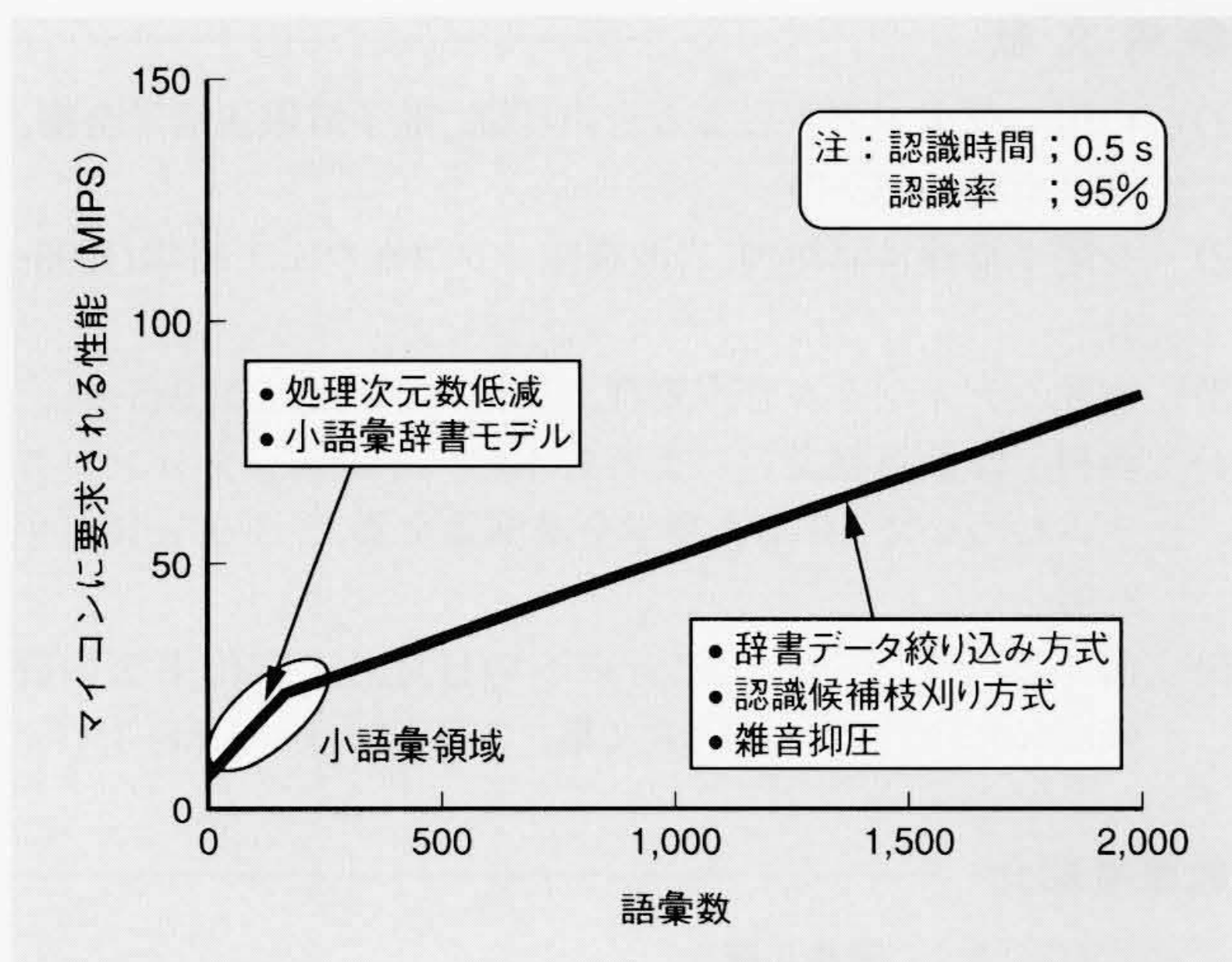


図6 マイコンに要求される性能と語彙数の関係

音声認識ミドルウェアを動作させたときの、認識語数とSuperHマイコンに必要な負荷の関係を示す。例えば、2,000語で80 MIPS、50語で15 MIPSがそれぞれ必要となる。

荷が大きい割りには雑音抑圧効果は少ない。

カーナビゲーションではもちろんのこと、自動車走行雑音への対策が不可欠である。例えば、走行雑音は100 Hz以下の成分が多いことから、アナログフィルタを挿入することも雑音を抑えて認識率を向上させるには有効である。

さらに、単にマイクロホンをついて差をとるだけでなく、二つのマイクロホンの雑音の差が最小になるように逐一計算させるANC(Adaptive Noise Canceller)なども効果的である。

(2) 認識計算回数の削減

認識計算回数の削減とは、認識候補語を計算途中で絞り込んでいく方式である。例えば、1,000語の音声認識を行うということは、別の表現をすれば、999語の単語を認識しないことを計算していることになる。すなわち、確率計算過程で確率の相当低い単語は認識対象語から外し、計算しないようにすることが、計算時間の短縮に有効である^{1), 3)}。

(3) 高速メモリアクセス

多数の単語の確率データと外部メモリに格納されている辞書とを参照する際、マイコンと外部メモリとの情報のやり取りが処理時間にたいへん大きく影響してくる。

4.5 小語彙領域特性

このミドルウェアの別の特徴は、100語以下の小語彙領域での処理のスリム化である。100語以上の大語彙領域では、認識率と認識時間の関係で、三十次元を用いた確率

計算をしているが、小語彙領域ではそれを十六次元に減らして計算を行っている。認識計算では、特に、回帰係数が重要であることがわかっており、認識率を低下させることなく次元数を減らせ、わずかなマイコン負荷で音声認識を処理することができる。このシステムでは、例えば50語認識の場合、10 MIPSで0.5秒、95%の特性が得られている。

これにより、例えば、常時は少ない語数のコマンドに合致するかどうかの小語彙モードで認識処理を作動させ、あるコマンドと合致したら、2,000語の音声認識の大語彙モードに切り替えるシステムも考えられる。こうすれば、認識前にボタンを押すなどの人手による動作をしなくても認識処理を自動的に行わせることができる。

4.6 話者適応

実験をしてみると、認識率はかなり個人に大きく依存することがある。この極端に認識率の悪い人のためには話者適応システムが効果的である。

これはあらかじめ決めてある単語の分析結果を見て、それをHMMモデルに反映し、あるパラメータを少しずつシフトするものである。これにより、例えば、認識率が70%の人でも、10語の事前認識で83%に、20語で87%程度に改善できる。

フラッシュメモリを併用すれば、電源を切ってもそのパラメータは保持されるので、次に使用する場合は認識率の高い状態から使用できる。

5. 今後の展開

(1) 音声合成では、合成音を人間の声にいかに近づけるかが開発のポイントである。このために、200 Mバイトのメモリを用いる方式が提案されているが、マイコンを用いるシステムではまだ実用的ではない。今後、音質向上のためには、現在の単調なピッチだけではなく、マルチピッチ手法やホルマント遷移モデル確率が重要であると考える。

(2) 音声認識の今後の傾向は、2極に分化すると思われる。一つはパーソナル端末を指向した、認識処理の軽量化である。特に、PHS(Personal Handyphone System)やPDA(Personal Digital Assistant)の携帯型端末向けには、雑踏の中でも使用できる雑音特性が要求されている。しかも、低消費電力があわせて重要であり、結果的にもっと軽いソフトウェアで処理する必要がある。

別の傾向はディクテーションから翻訳へ向かう高機能化の路線である。現在では、十分な学習時間を施しても

満足するような認識率がまだ得られているわけではないが、いずれ、声によるメールの入力システムが望まれている。このためには単語認識とは異なる、ネットワークサーチ方式に基づく文章認識のアルゴリズムが必要になってくる。

さらに、将来、マイコンの性能が400 MIPS以上になると、人類の永年の夢である簡易自動翻訳も可能になる。

これらの傾向に対応するためには、従来のマイコンの機能だけでなく、DSP(Digital Signal Processor)機能を強化した回路を内蔵することが重要である。

6. おわりに

ここでは、SuperH RISCマイコンに特化した音声合成と音声認識のミドルウェアについて、その概要と特徴を述べた。

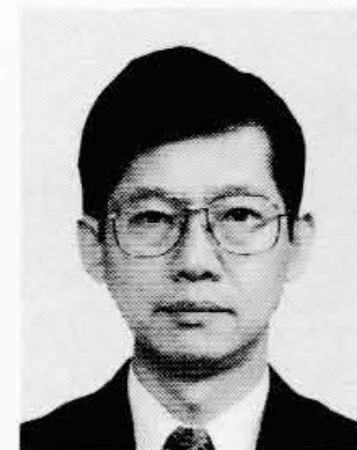
音声合成では、最も重要な「高音質」を実現するために、音源として子音と母音を組み合わせた要素を生波形のまま記録する方式を採用し、良好な音質を得る見通しがついた。音声認識では、音素片単位で認識するアルゴリズムにより、低いマイコン負荷で高い認識率を得ることができ、この高速化、小語彙領域対応についても触れた。

今後も、より高度なミドルウェアの開発に注力していく考えである。

参考文献

- 1) 中川：確率モデルによる音声認識，電子情報通信学会編，コロナ社(昭63-7)
- 2) 長尾：自然言語処理，岩波講座 ソフトウェア科学(1996-4)
- 3) 古井：ディジタル音声処理，東海大学出版会(1985-9)
- 4) 西村：離散単語発声による日本語ディクテーションシステムについて，日本音響学会講演論文集，3-3-9, p.103(平8-9)
- 5) 南：雑音と歪を含んだ音声へのHMM適応化手法の評価，日本音響学会講演論文集，2-5-14, p.85(平8-3)

執筆者紹介



鳴島正親

1971年日立製作所入社，半導体事業部 応用技術本部 ミドルウェア開発センタ 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



矢島俊一

1974年日立製作所入社，半導体事業部 応用技術本部 ミドルウェア開発センタ 所属
現在，音声合成システムの開発に従事
電子情報通信学会会員，日本音響学会会員
E-mail：syajima@crl.hitachi.co.jp



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 半導体技術開発センタ マルチメディアLSI開発部 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