汎用音声処理カードによる大語彙音声認識

黒田明裕† 西村雅史†

種々の音声処理機能を1枚のPC用アドインカードで実現することを目的とした汎用音声処理カードを作成し，HMMに基づく大語彙単語音声認識をこのカード上に実現することを試みた．ラベル出現頻度に基づく予備選択法によって対象単語を絞り込み，かつビームサーチを利用することで，計算量を約0.3%にまで削減した．また，音声処理カード上の記憶量を削減するため，ほとんどの認識パラメータをPCのシステムメモリー上に置き，必要とされるものだけを随時DMA転送によってカードに転送する方法を考案した．これらの手法により，11.5MIPSの汎用DSP，おのおの32Kワードのプログラムおよびデータメモリーで構成されたこの音声処理カードを使って，PC上の応用プログラムの実行に影響を与えることなく，実時間1,000単語特定話者音声認識を実現することが可能となった．本システムによる男性話者2名，女性話者1名に対する薬品名1,000単語の平均認識率は99.7%，平均応答時間は0.4秒であった．

Large-Vocabulary Speech Recognition on a General-Purpose Speech Processing Card

Akihiro Kuroda† and Masafumi Nishimura†

This paper describes a real-time 1,000-word speaker-dependent speech recognizer on a general-purpose speech processing card which consists of an 11.5 MIPS digital signal processor and 64 K words of memory. The amount of computation for HMM-based speech recognition was significantly reduced by using a word pre-selection method and the beam search. The memory requirement of the card was also minimized by using DMA (Direct Memory Access). Experiments with 1,000 drug names were conducted for three speakers. The average recognition rate and response time were 99.7% and 0.4 sec, respectively.

1. はじめに

音声認識を普及させその市場を拡大するためには，音声認識の性能および機能の高度化の追究に加えて，小型かつ低コストで実現するための技術の開発，音声認識を効果的に利用する応用プログラムやユーザーインターフェースの研究開発などが重要である．

初期の音声認識装置は独立したボックス型のものが多かったが，近年小型化が進み，PC (Personal Computer) 内蔵のカード型のものや，ソフトウエアのみによるものが増えてきた[1)~5)]．カード型のものとしては，(A)高機能を実現するため，複数の認識専用LSIやDSP (Digital Signal Processor) を大規模なメモリーと共に搭載したもの[2)] (複数カード構成のものが多い)，(B) DSPまたは認識用LSIを搭載した安価なカード上で，認識のための処理のほとんどを実行しているもの[3)] (現状では，メモリーおよび計算量の制約から，機能を小語彙に限定している)，(C) DSPを使い，認識の前処理として音響処理のみを行うもの[4)] (認識処理の大部分はPC上のソフトウエアで行っている)，などがある．また，(D)ソフトウエアのみによる音声認識システムでは，AD変換器以外の特殊なハードウエアは使用せず，認識のための処理をすべてPC上のソフトウエアで行っている[5)]．

(A)は通常，高機能の実証を主目的としたプロトタイプであり，製品としては他の3者の形態をとることが多い．価格の面からは通常(D)が優れており，機能に関しては，(C)の方式で，大語彙，連続，または不特定話者を対象としたものが現れてきている．一方，(C)および(D)の方式はPCのプロセッサーに対する負荷が大きく，これらに共通する問題点として，応用プログラムや，他の音声処理機能との共存性（コンカレンシー）が低いことがあげられる．ユーザーにとって，音声認識はPC上の応用プログラムのための一入力手段であるから，認識の処理そのものがPCのプロ

† 日本アイ・ビー・エム(株)東京基礎研究所
IBM Research, Tokyo Research Laboratory, IBM Japan Ltd.

セッサーパワーを過度に消費し，応用プログラムの実行速度に悪影響を及ぼすことは，音声認識の応用範囲を狭くすることにつながりかねない．この観点では，(B)が優れており，少なくとも現状では，(C)や(D)と，実時間性の高いプログラムや，計算量の多いプログラムとの共存は困難である．言いかえると，(B)の方式で大語彙等の高機能な音声認識を実現することができれば，音声認識の応用範囲が広がる．

音声認識手法に関しては，従来，DP (Dynamic Programming) マッチングに代表されるようなパターン照合が主流であったが，近年，HMM (Hidden Markov Model) やニューラルネットワークのような統計的な手法を用いれば，発声の揺らぎや経時変化に対して頑強なシステムを構築できることが明らかになってきた．特に，筆者らはフェノニックマルコフモデル[6]と呼ばれる時間構造表現能力に富むHMMを中心に研究を進め，このモデルが大語彙認識時にも必要記憶量が少なく，また，少ない訓練データで高い認識精度が得られることを示してきた[7),8)]．

今回，このような観点から，フェノニックマルコフモデルに基づいた1,000単語実時間音声認識システムを，汎用のDSPを搭載した1枚の音声処理カード上に実現することを試みた．まず，計算量の削減のため，単語予備選択を導入し，かつビームサーチを利用した．次に，カード上の記憶量削減のため，ほとんどの認識パラメータをPCのシステムメモリー上に置き，DMA (Direct Memory Access) 転送によって，必要なものだけを随時カードに転送する方法を考案した．これらにより，大語彙音声認識を，PC上の応用プログラムの実行に影響を与えることなく，低コストの汎用音声処理カード上に実現することが可能になった．

本論文では，今回実現したシステムに関して，そのシステム構成，認識方式，インプリメンテーション上の工夫，および性能の評価結果について報告する．

2. 汎用音声処理カード (SPC)

2.1 ハードウエア構成

音声認識システムのハードウエアは，PC (IBM PS/55) と，それに付加する音声処理カード (SPC : Speech Processing Card) とで構成されている．SPCは，汎用の音声処理を目的として設計されたPC用アドインカードであり，音声認識のほかに，テキスト音声合成，録音／再生，電話の自動発信／応答などの機能も

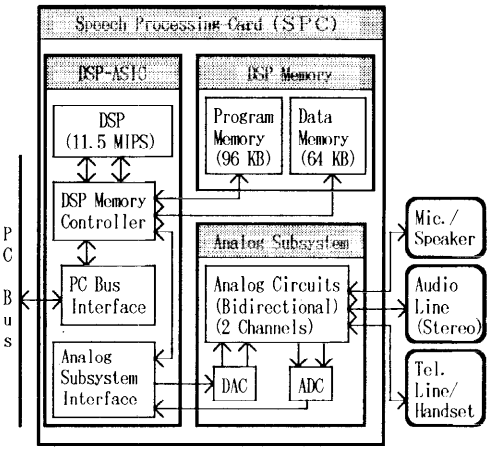


図1 音声処理カード (SPC) のブロック図
Fig. 1 Block diagram of speech processing card (SPC).

実現できる[9)]．**図1**に示すように，SPCは，DSP-ASIC (Application Specific IC)，DSPメモリー，アナログサブシステムの，三つの部分から構成されている．

(1) DSP-ASIC

中核となるDSPとして，Harvardアーキテクチャーに基づいた16ビット固定小数点DSP（ただし，ALU (Arithmetic Logic Unit) まわりは32ビット）[10)]を使用した．命令はすべて実効的に1サイクルで実行され，11.5 MIPS (Million Instructions Per Second) のスピードである．1命令で乗算，ALU演算，メモリーとのデータ転送の3種類の演算を同時に制御することができ，最大34.5 MOPS (Million Operations Per Second) の演算能力がある．PCバスとのインターフェースは，PCのシステムメモリーとDSPメモリーとの効率的なデータ転送を実現するために，バスマスター機能 (DMA転送機能) を持っている．DSPのプログラム (DSPタスク) は，PCメモリーおよびDSPメモリーのアドレス，データ長，転送方向などのひとまとまりの情報を，DSPメモリーの決められた領域にリスト状 (DMAリスト) に配列することにより，容易にDMA転送を要求することができる．DMAハードウエアは，約1ミリ秒ごとにDMAリストをスキャンし，要求されたDMA転送を行い，DSPタスクの次回の起動時までに必要なデータを用意する．

(2) DSPメモリー

アクセスタイム45ナノ秒のSRAM (Static RAM) を用いて，プログラムメモリー96 KB (32 K×24 ビ

ット），データメモリー 64 KB（32 K×16 ビット）を実装した．これは，コストおよびスペースを考慮して決定したが，大語彙音声認識用のパラメーターテーブルなど（本認識システムの場合，1,000 単語で約 850 KB）を保持するには，不十分な大きさである．この問題に対しては，認識アルゴリズムと SPC ハードウエアの協業で解決した．その手法については 3 章で詳述する．

（3）アナログサブシステム

マイク／スピーカー，オーディオラインレベル入出力（ステレオ），電話回線，電話機などを，直接接続することができる．サンプリング周波数は，8, 9.6, 10, 12, 16, 20, 44.1, 48（単位：kHz）のうちから一つを，プログラムで選択できる．

2.2 ソフトウエア構成

ソフトウエアの構成を**図 2** に示す．図において，下の 2 層が SPC（DSP）のプログラムであり，その上に PC のプログラムが数層ある．DSP オペレーティングシステムは，マルチタスク機能を持つリアルタイム OS[11] であり，その上で，いくつかの DSP タスクを同時に走らせることができる．PC タスクは，SPC インターフェースを通して DSP タスクと通信し，共同で音声処理機能を実現する．音声認識の認識時の処理（特徴量抽出，単語境界検出，単語予備選択，詳細マッチングなど）は，そのほとんどを DSP タスクで実行しているので，PC 上の応用プログラムの実行時間への影響はほとんどない．一方，音声認識の訓練時の処理（コードブックの推定，HMM の訓練など）は，そのほとんどを PC タスクで行っている．なお，音声処理 PC タスクは，応用プログラム作成のための API[12]（Application Program Interface）も提供している．

3. 認識システムの構成

本認識システムの構成を**図 3** に示す．

3.1 特徴抽出部

入力音声は，10 kHz, 16 ビットで，標本化，量子化される．このデータについて，一回差分による高域強調を行った後，窓幅 25.6 ミリ秒，フレーム周期 9.6 ミリ秒のハミング窓を用いた FFT を行う．さらに，この出力を 18 チャンネルからなる臨界帯域フィルターにかけ，対数変換したものを，フレーム内対数パワー値とともに 19 次元の特徴量ベクトル $\boldsymbol{S}$ として用

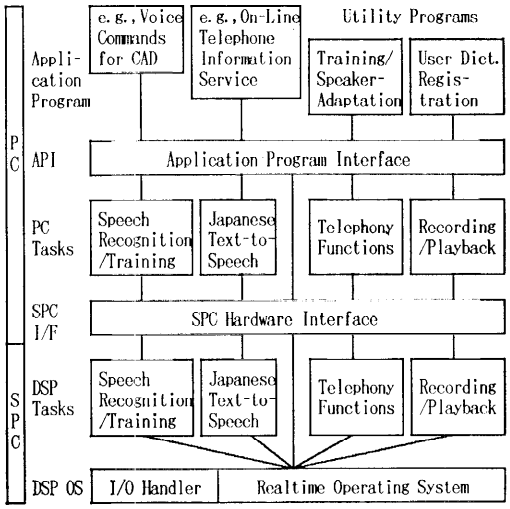


図 2 ソフトウエアの構成
Fig. 2 Software structure.

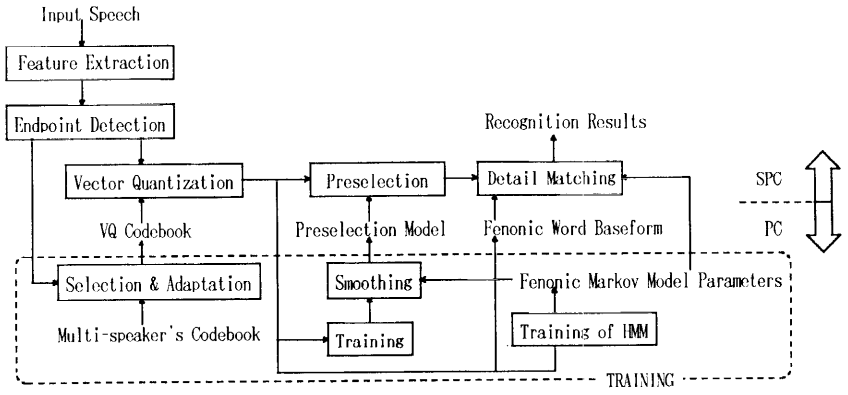


図 3 大語彙音声認識システムの構成
Fig. 3 Block diagram of large vocabulary speech recognition system.

いた．これをスペクトルの静的特徴と呼ぶ．一方，このフレームを中心にした前後3フレームのスペクトル変化量を動的特徴 $\boldsymbol{D}$ として用いた[7]．

フレームごとに得られる静的特徴 $\boldsymbol{S}$ および動的特徴 $\boldsymbol{D}$ は，それぞれベクトル量子化器により，静的特徴のラベル L_s および動的特徴のラベル L_d に変換される．

コードブックとしては，静的特徴 64 種，動的特徴 128 種とした．静的特徴，動的特徴，おのおの 32, 64, 128 種の計9通りの組み合わせで予備実験を行ったところ，最も認識率が高かったのは両者 128 の場合であったが，SPC 上のメモリーの制約から，次に精度の高かった組み合わせを選択した．

3.2 単語境界検出部

単語境界の検出は，基本的には2段閾値法[13]を用いるが，SN 比に応じて適応的に閾値を変更し，また，呼気雑音や舌打ち音，語頭で見られる不明瞭な発声，さらにはマイク位置の変動などに対処するため，複数の始端候補点を検出する処理を行う．

3.3 単語予備選択部

計算量削減のため，時間情報のない，ラベルの出現頻度だけの単語モデルによって，対象単語を数パーセントにまで絞る．具体的には次の処理を行う．認識対象単語 w それぞれに対し，静的特徴のラベル L_s および動的特徴のラベル L_d の出現頻度 $\Pr(L_s|w)$ および $\Pr(L_d|w)$ を推定しておく．入力ラベル対の時系列を $L=(L_s(1), L_d(1)), (L_s(2), L_d(2)), \cdots, (L_s(T), L_d(T))$ とし，入力フレームに同期して，次式で対象単語ごとの対数尤度 $\log \Pr(L|w)$ を求め，尤度の上位 n 個を詳細マッチングの対象とする．

$$\log \Pr(L|w) = \sum_{t=1}^{T} \{\log \Pr(L_s(t)|w) + \log \Pr(L_d(t)|w)\} \quad (1)$$

フレームごとに必要とされる計算は2回の 16 ビット整数の加算にすぎない．また，データ参照についても複雑なアドレス計算を必要としない．具体的には，プログラムの核となる部分は DSP のアセンブリコード7ステップで実現されており，この計算に要するプロセッサーパワーは対象が 1,000 単語の場合でも 0.7 MIPS 程度である．なお，n は対象単語数に応じて変更する．

一方，これらの予備選択モデルのパラメータを SPC 上に展開したのでは，1単語あたり 384 バイト (=2バイト×(64+128)) のメモリーが必要となり，1,000 単語分のデータを保持できない．そこで，パラメータは PC のメモリー上に保持し，入力ラベルが得られるごと（フレームごと）にそのラベルの生起確率だけを 1,000 単語分，PC から SPC に DMA 転送する．転送レートは，400 K バイト/秒程度である．転送が容易になるように，あらかじめ各パラメータは，ラベルごとに対象単語数分を連続して展開しておく．

3.4 詳細マッチング部

フェノニックマルコフモデルの構造を**図4**に示す．本システムでは継続長モデル[14]をその構造に含む(B)のモデルを用いた．そのトレリス上の遷移は，図4に示してあるように，2〜1/2 の傾斜制限を持つ DP パスに相当している．このモデルは従来われわれが使っていたモデル(A)と比べて，特にビタービ計算の際のパスの大幅な回り込みを防ぐことができる．

先にわれわれは，ラベル出力確率のみを独立として取り扱うフェノニックマルコフモデルを，静的特徴のみを扱っていた従来型モデルの一変形として提案し，その有効性を示している[7]．しかし，このモデルは遷移確率に，48 K バイト (=2バイト×3×64×128) の記憶量を必要とし，SPC 上に保持することは現実的でない．そこで，遷移確率まで含めて独立な二つのモデルを組み合わせて表現するモデルを新たに採用した．このモデルを用いた場合，単語 w に対する入力ラベル対 (L_s, L_d) の時系列 L の対数尤度 $\log \Pr(L|w)$ は次式で定義される．

$$\begin{aligned}\log \Pr(L|w) = \sum_{t=1}^{T} \{&\log \Pr(L_s(t)|F_s(i_t)) \\ &+ \log \Pr(L_d(t)|F_d(i_t)) \\ &+ \log \Pr(B_{i_t,i_{t+1}}|F_s(i_t)) \\ &+ \log \Pr(B_{i_t,i_{t+1}}|F_d(i_t))\} \quad (2)\end{aligned}$$

ここで $F_s(i)$ (あるいは $F_d(i)$) はフェノニック単語

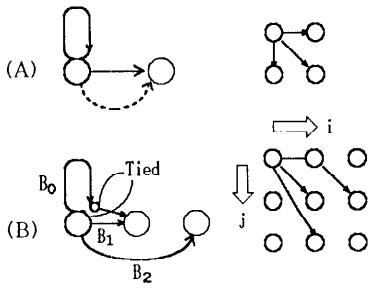


図4 フェノニックマルコフモデルとトレリス上の遷移
Fig. 4 Examples of fenonic Markov models and their transitions on trellis.

ベースフォーム中の状態番号 i におけるフェノンを，$I=i_1, i_2, \cdots, i_T$ はビタービアルゴリズムによって求まる最適パス上の状態番号列を，また，$B_{i,j}$ は状態番号 i から j への遷移を表している．予備実験の結果，モデルの変更に伴う認識率の低下は少なく，遷移確率 ($\Pr(B_{i,j}|F_s)$ および $\Pr(B_{i,j}|F_d)$) のための記憶量も 1,152 バイト (=2 バイト×3×(64＋128)) で済み，現実的な解決策であるといえる．

詳細マッチングは，発声終端候補点において，単語予備選択結果が得られると同時に開始する．なお，音響処理と単語予備選択の処理は発声終端が確定するまで続行する．まず，予備選択結果に基づいて，認識対象単語のフェノニック単語ベースフォームを PC から SPC に DMA 転送する．マッチング処理は入力フレーム同期の形式で行い，ビームサーチを併用する[15]．

この処理はノンリアルタイムタスクとして実行し，音響処理，単語境界検出等のリアルタイムタスクの処理の後に残ったすべてのプロセッサーパワーを使用する．この場合も，単語予備選択と同様に，モデルのパラメータを SPC 上に展開したのでは，出力確率 ($\log \Pr(L_s|F_s)$ および $\log \Pr(L_d|F_d)$) だけでも，40 K バイト (=2 バイト×(64×64＋128×128)) のメモリーが必要になってしまう．そこで，入力ラベルごとにそのラベルの出力確率のみを全フェノン数分 (192 個 (=64＋128)) PC から DMA 転送する．これについても転送が容易になるように，あらかじめ出力確率を，出力ラベルごとにフェノンの数だけ連続して PC のメモリー上に展開しておく．発声終端が確定すると，このマッチングの結果を最終認識結果として出力する．先の終端候補点がキャンセルされると，詳細マッチングを中止し，次の発声終端候補点を待つ．

単語境界検出部で，第 2 の始端候補点が得られた場合には，ビームサーチの枝刈りの閾値設定用に求めている入力フレームごとの最大尤度を，フェノニック単語モデルの先頭の状態における尤度として与える．このようにして，入力側の始端を疑似的に開放する．

4. 特定話者認識実験

4.1 実験データ

大語彙音声認識のアプリケーションとして，病院における薬品のオーダーシステムを想定し，ある病院で実際に使用している薬品から無作為に 1,000 個を選択し，この語彙を使って本システムの性能を評価した．被験者は男性 2 名と女性 1 名，この語彙の平均モーラ長は 5.3，平均発声速度は 6.8 モーラ/秒であった．

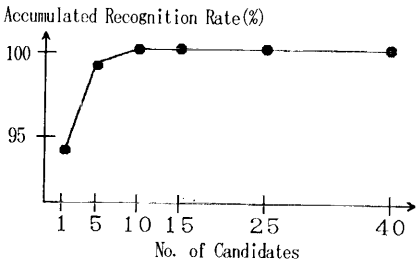


図 5 予備選択の累積認識率
Fig. 5 Accumulated recognition rates for word preselection.

表 1 1,000 単語認識率 (%)
Table 1 Recognition rate of 1,000 words (%).

話 者	男性A	男性B	女性A
予備選択率	100	100	100
最終認識率	99.7	99.7	99.6

モデル訓練用発声数：各単語 3 回
予備選択候補単語数：25
評価用入力単語総数：各話者 2,000 語

4.2 認識結果

各単語 3 回分の発声を使って，予備選択モデルの訓練を行った．また，このうち 1 回目をフェノニック単語ベースフォームとして登録し，2，3 回目をフェノニックマルコフモデルの最尤推定に用いた．まず，予備選択数と予備選択の累積認識率との関係を**図 5** に示す．10 位までとれば 100% の精度が得られており，この予備選択手法によって，候補数を 1% にまで絞り込めることがわかる．ただし，本システムでは，対象語彙，発声様式等の要因による劣化を見越して，25 位までを詳細マッチング処理するものとした．1,000 単語音声認識の実験結果を**表 1** に示す．いずれの話者に対しても 99.6% 以上の高い認識率が得られている．なお，詳細マッチング時の総計算点数はビームサーチによって全探索の約 12% にまで削減されているが，認識精度の劣化はほとんどなかった．また，平均応答時間は発声終了後約 0.4 秒であった．

5. おわりに

PC 用汎用音声処理カード上に，大語彙音声認識を実現することを試みた．システム構成の概略を報告した後，認識方式およびインプリメンテーション上の工夫について，計算量および記憶量の削減方法を中心に述べた．計算量については，まず，単語予備選択モデ

ルにより，対象語彙 1,000 語の場合，候補単語数を10〜25程度に絞り込めることを示した．さらに，ビームサーチの利用により，詳細マッチングの総計算点数も全探索の約 12% となり，合計で，計算量を約 0.3%(=(25/1000)×12%)にまで削減した．また，カード上に必要とされる記憶量も，フェノニックマルコフモデルの使用と，DMA 転送によるデータ参照によって大幅に削減することができた．これらにより，低コストの汎用音声処理カードを使って，応用プログラムの実行に影響を与えることなく，大語彙の高精度実時間認識が可能となった．

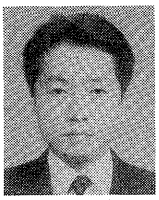
今後は，連続音声，不特定話者等，さらに高機能の音声認識を，同様の条件の下でどこまで実現可能か検討していきたい．

謝辞 研究の機会を与えて頂いた大河内正明音声情報処理担当部長，熱心にご討論頂いた菅原一秀主任研究員，年岡晃一主任研究員，橋本泰秀研究員，阪本正治研究員に深謝します．また，音声処理カード作成にご協力頂いた M. Ware 氏，D. Carmon 氏ほか IBM ローリー，大和両研究所の皆様に感謝します．

参考文献

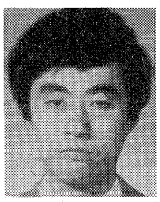
1) 中津良平：音声認識・合成技術の製品化及び市場動向，信学技報，SP 89-103 (1989).
2) 金澤博史，坪井宏之，竹林洋一：不特定話者音声対話システム TOSBURG におけるキーワードスポッティング処理，音響講演論文集，1-P-18 (1992).
3) 有吉 敬，松下 貢，藤本潤一郎：2入力による騒音下の単語音声認識方式，音響講演論文集，1-8-5 (1990).
4) Baker, J.: Large Vocabulary Speaker-Adaptive Continuous Speech Recognition Overview at Dragon Systems, *EUROSPEECH '91*, pp. 29-32 (1991).
5) 磯 健一，高木啓三郎，篠田浩一，山田栄子，服部浩明，Ehsani, F.，野口 淳，古賀真二，畑崎香一郎，渡辺隆夫：パソコン向けソフトウェア音声認識，音響講演論文集，2-Q-21 (1993).
6) Bahl, L., Brown, P., deSouza, P., Mercer, R. and Picheny, M.: Acoustic Markov Models Used in the TANGORA Speech Recognition System, *Proc. ICASSP '88*, S 11.3, pp. 497-500 (1988).
7) Nishimura, M.: HMM-Based Speech Recognition Using Dynamic Spectral Feature, *Proc. ICASSP '89*, S 6.12, pp. 298-301 (1989).
8) 西村雅史：フェノニックマルコフモデルに基づく音声認識のための話者適応化方法，信学論 (D-II), Vol. J 73-D-II, No. 10, pp. 1630-1638 (1989).
9) 大河内正明，黒田明裕，斎藤 隆，阪本正治，菅原一秀，鈴木和洋，年岡晃一，西村雅史，橋本泰秀：大語彙音声認識・日本語テキスト音声合成を実現した「日本語音声認識・合成アダプター」の開発，音響講演論文集，1-P-12 (1992).
10) Rausch, N. and Kline, D.: Mwave TMS 320 M 500 Signal Processor—Single Chip Solution for Multimedia, *Proc. ICSPAT '92*, Vol. 1, pp. 198-205 (1992).
11) Hinzelman, E.: The Mwave Operating System: Support for Hard Real-Time Multitasking, *Proc. ICSPAT '92*, Vol. 1, pp. 308-311 (1992).
12) IBM 日本語音声認識・合成プログラム (OS/2 版) 音声認識リファレンス・マニュアル，N: GC 180388-00, IBM Corp. (1992).
13) Rabiner, L. and Sambur, M.: An Algorithm for Determining the Endpoints of Isolated Utterances, *Bell Syst. J.*, Vol. 54, No. 2, pp. 297-315 (1975).
14) 西村雅史，大河内正明：状態の継続長を反映したマルコフ・モデルによる音声認識，音響講演論文集，1-4-18 (1985).
15) Bridle, J., Brown, M. and Chamberlain, R.: An Algorithm for Connected Word Recognition, *Proc. ICASSP '82*, pp. 899-902 (1982).

(平成5年11月26日受付)
(平成6年 4 月21日採録)



黒田 明裕 (正会員)

1957 年生．1980 年東京大学工学部電子工学科卒業．同年日本アイ・ビー・エム(株)入社．1983 年より同社東京基礎研究所において，主に音声認識，日本語テキスト音声合成システムの研究開発に従事．同研究所音声情報処理担当課長を経て，現在，マルチメディアパーソナルシステム担当課長．電子情報通信学会会員．



西村 雅史 (正会員)

1958 年生．1981 年大阪大学基礎工学部生物工学科卒業．1983 年同大学院物理系修士課程修了．同年日本アイ・ビー・エム(株)入社．以来，東京基礎研究所において，HMM による音声認識の研究に従事．電子情報通信学会，日本音響学会各会員．