論文 / 著書情報
Article / Book Information

論題(和文)	パソコン向けソフトウェア音声認識
Title(English)	
著者(和文)	磯 健一, 高木啓三郎, 篠田浩一, 山田栄子, 服部浩明, F. Ehsani, 野口 淳, 古賀真二, 畑崎香一郎, 渡辺隆夫
Authors(English)	Koichi Shinoda, atsushi noguchi
出典(和文)	日本音響学会平成5年度秋季研究発表会講演論文集, Vol. 2-Q-21, No. , pp.
Citation(English)	, Vol. 2-Q-21, No. , pp.
発行日 / Pub. date	1993,

2—Q—21　パソコン向けソフトウェア音声認識*

○磯健一 高木啓三郎 篠田浩一 △山田栄子 服部浩明 △Farzad Ehsani
野口淳 古賀真二 畑崎香一郎 渡辺隆夫（日本電気(株) 情報メディア研究所）

1　はじめに

従来のパソコン向け音声認識システムはＡＤ変換用ＬＳＩとＤＳＰなどのアクセラレータチップを搭載した専用ハードウェアを必要とするものが多かった。近年ＣＰＵ性能の向上に加えて、マルチメディア向けにＡＤ変換機能が標準装備化されるようになり、ソフトウェアのみで音声認識機能を付加することが可能になりつつある。

しかし高度な音声認識処理を実装するためにはＣＰＵの処理能力は依然として不足しており、これまで実現されたソフトウェアシステムは単語単位の標準パターンに基づく小語彙認識に留るものが多かった。

我々は新しい高速アルゴリズムに基づいて、パソコン上にソフトウェアのみで半音節認識単位による不特定話者中語彙単語認識システムを実現したのでここに報告する。

2　システム

本システムが前提としているハードウェアは、インテルｉ４８６™程度のＣＰＵとＡＤ変換機能を搭載したパソコンである。システム構成を図１に示す。マイクから入力された音声は標本化周波数１１ｋＨｚ、精度１６ビットでＡＤ変換され、分析・認識・ユーザー学習部でそれぞれ処理される。不特定話者単語認識、および５０単語程度の発声による話者適応化機能を有している。同時に識別可能な単語数は最大２５０単語であるが、複数の辞書を場面に応じて切り替えることによって、全体としてはより多くの単語を扱うことができる。システムはキーボード互換に動作し、認識結果はあらかじめ辞書に定義したキーボード入力として他のアプリケーションプログラムへ送られる。認識対象単語は発声による登録不要で、単語の読みと対応するキーボード入力だけを登録することにより任意に定義・変更が可能である。

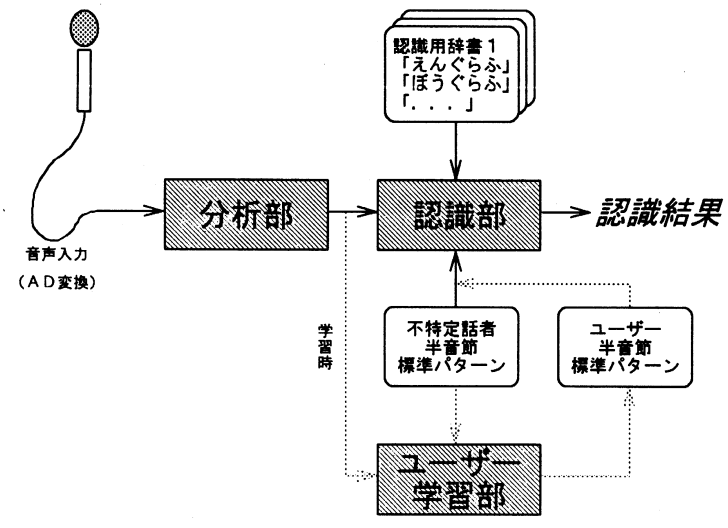

図１　システム構成

2.1 分析部

分析部では１６ミリ秒間隔でメルケプストラム分析処理を行っている。ＦＦＴパワースペクトルをメルケプストラムに変換する処理においては、はじめに帯域制限（１５０～５０００Ｈｚ）した対数パワースペクトルをケプストラム領域でのリフタリングにより平滑化（コサイン変換＋逆コサイン変換）してから、メル変換（周波数軸→メル軸）とコサイン変換（メルスペクトル→メルケプストラム）を行う。ここでメル変換を行列近似することにより、ＦＦＴ対数パワースペクトルをメルケプストラムに変換する一連の処理を行列乗算１回に縮約している。また背景雑音の影響を除去する目的で、スペクトルサブトラクション処理を行っている。

2.2 認識部

認識部では不特定話者半音節認識単位の混合ガウス分布ＨＭＭ［１］を標準パターンとして照合処理を行っている。入力音声の各時刻における主な処理は、ＨＭＭの各状態における特徴ベクトル出力確率の計算と、ビタビアルゴリズムによる漸化式計算である。この内、出力確率計算は約１５００個のガウス分布確率計算を含んでおり、照合処理の大半を占めている。ここではあらかじめ全ガウス分布を木構造にクラスタリングして、認識時の照合演算

*Software-only speech recognition for personal computers, by Ken-ichi ISO, Keizaburo TAKAGI, Koichi SHINODA, Eiko YAMADA, Hiroaki HATTORI, Farzad EHSANI, Jun NOGUCHI, Shinji KOGA, Kaichiro HATAZAKI, Takao WATANABE, NEC Corporation

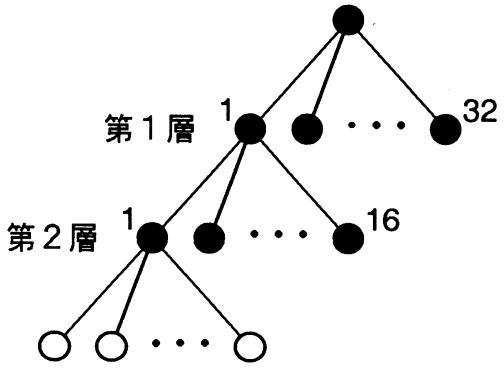

図２ 木構造標準パターン

量の低減を図っている［２］。まずガウス分布（要素分布）を３２個にクラスタリング（第１層）し、次に各クラスタ内をさらに１６個にクラスタリング（第２層）する（図２）。各層のノードにはそのクラスタを近似するガウス分布（代表分布）を付与する。第２層のノード下に要素分布が結ばれる。認識時には第１層および第２層の代表分布のみを保持し、はじめに第１層の３２個の代表分布の出力確率を計算し、その内の確率値の大きい上位５ノードについて第２層の代表分布を計算する。確率値の計算されなかった第２層の代表分布およびその下に結ばれた要素分布については、その上位ノードの確率値で代用する。以上の処理により、確率値の大きな要素分布については正確に、小さな要素分布については粗い精度で計算が行われる。演算量は１５００回の計算が３２＋５×１６＝１１２回に低減される（１／１０以下）。

一方、漸化式計算においては認識対象外語彙のリジェクト機能を安定させるために、尤度補正に基づくリジェクション方式［３］を組み込んでいる。補正のための参照尤度としては各時刻におけるガウス分布確率値の最大値を累積したものを用いた。

2.3 ユーザー学習部

不特定話者認識に加えて、システムはユーザー学習機能を有している。ユーザー学習部ではスペクトル内挿写像に基づく話者適応化処理［４］により、不特定話者標準パターンからユーザー標準パターンを作成する。本システムでは話者適応化時の演算量低減のため、木構造標準パターンの再クラスタリングは行わず、各ノードのガウス分布のパラメータのみを適応化している。すなわち木構造標準パターン第２層のガウス分布だけをスペクトル内挿写像によって話者適応化し、その上位ノードの代表ガウス分布を再計算している。

3 評価実験

実オフィス環境（比較的静か）において本システムのオンライン評価実験を行った。不特定話者標準パターンの学習には男性２３名、女性２０名による音素バランス２５０単語１回発声を用いた。男女混合不特定話者標準パターンに加えて、男性、女性別々の性別不特定話者標準パターンも作成した。各標準パターンを構成する半音節HMMの状態数は３（例外的に１状態の半音節もある）、各状態のガウス分布混合数は２とした。評価用の発声は学習とは異なる話者（男性３名、女性３名）による学習とは異なる２５０単語発声を用いた。ユーザー学習には評価用話者の５０単語１回発声を用いた。

表１ 認識性能評価結果 （単位%）

		標準パターン		
		男女混合	性別	学習後
評価話者	男性A	92.8	95.2	97.6
	男性B	92.8	89.2	94.8
	男性C	88.4	93.2	98.4
	女性A	92.8	97.6	99.6
	女性B	90.4	93.2	98.0
	女性C	92.4	95.2	99.2
平均		91.6	93.9	97.9

実験結果を表１に示す。少数話者の実環境におけるオンライン評価のため、不特定話者（男女混合、性別）認識性能では若干話者によるばらつきが見られるが、話者適応化後は安定した性能が得られている。

各標準パターンのメモリ使用量は約１２０ｋバイト、システム全体の必要メモリ量は約７００ｋバイトであった。認識時の応答はＣＰＵにインテルｉ４８６™ＤＸ２（６６ＭＨｚ）を用いた場合、遅延なしの実時間応答が実現された。

4 おわりに

高速アルゴリズムの開発により、特別な付加装置なしでパソコン上にソフトウェア音声認識システムを構築した。評価の結果、不特定話者２５０単語認識および話者適応化により良好な結果が得られることを確認した。

謝辞

システム開発にご協力いただいた坂井信輔氏、水野正典氏に感謝致します。

参考文献

[1] 磯谷他，音学講論，1-8-19 (1990.9).
[2] 渡辺他，音学講論，(1993.10).
[3] 渡辺他，信学論，J75-D-II (1992.12).
[4] 篠田他，音学講論，1-8-12 (1990.9).