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連続音声認識
Title(English)	
著者(和文)	篠田浩一, 坂井信輔, 磯健一, 畑崎香一郎, 渡辺隆夫, 水野正典
Authors(English)	Koichi Shinoda
出典(和文)	日本音響学会平成6年度秋季研究発表会講演論文集, Vol. 2-8-3, No. , pp.
Citation(English)	, Vol. 2-8-3, No. , pp.
発行日 / Pub. date	1994,

2—8—3 パソコン向けソフトウェア連続音声認識 *

⊙ 篠田 浩一　坂井信輔　磯 健一　畑崎香一郎　渡辺隆夫　△ 水野正典†
(NEC 情報メディア研究所　† NEC 情報システムズ)

1. はじめに

先に、パソコン上でソフトウェアのみで動作する音声認識システムを開発した[1]。今回、さらに連続音声認識を実装し、より良いインタフェースの実現をはかった。バンドルサーチ[2]を用いて計算量を削減し、異なり単語数が100程度の規模のタスクで実時間の動作を可能としている。

2. システムの概要

本システムはMicrosoft^{R}Windows^{TM}3.1で動作するWindows Sound System 1.0(A)の音声認識部に用いられ、様々なアプリケーションに対し音声入力機能を付加することが可能である。同時に認識可能な単語数は離散・連続あわせて200単語である。各場面ごとに単語セットを定義できる。システム構成を図1に示す。今回新たな機能として連続音声認識機能、音声による認識開始・終了機能(音声スイッチ)を加えた。

2.1. 分析部

分析周期は16ミリ秒、周波数帯域は150～5000Hzであり、パワー差分、メルケプストラム10次元、メルケプストラム差分10次元の計21次元の特徴量を計算する。また、背景雑音を除去するためスペクトルサブトラクション処理を行なっている。

2.2. 認識部

認識単位として半音節を用いた混合ガウス分布HMM[3]を標準パターンとして用いている。各状態のガウス分布混合数は2である。計算量の多くを占める出力確率計算の計算量を低減するため、以下のように、確率分布の木構造化を行なっている[4]。最初に全ガウス分布に対しクラスタリングを行い、さらにクラスタ内の分布集合に対しクラスタリングを行う処理を繰り返すことにより、木構造を作成する。各クラスタには、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。今回のシステムでは、総分布数は1500、第1層の分布数は32、第2層の分布数は16である(図2)。第1層の32個の確率分布に対し代表分布の出力確率計算をしたのち、確率値の大きい上位5ノードに対し、さらにその下の第2層の分布に対する確率計算を行なう。計算量が10分の1以下になる。また、次章で述べるように、今回新たに連続音声認識機能を付加した。

2.3. ユーザー学習部

ユーザー学習部では、スペクトル内挿写像に用いた話者適応化を行なっている[5]。演算量低減のため、確率分布の木構造の再構成は行なわれず、各クラスタの代表分布は、そのクラスタに属する各分布の適応化後に再計算される。50単語1回発声を用いたユーザー学習を行なうことにより、250単語認識で誤り率が8.3％から2.1％へと約4分の1に減少している[1]。

2.4. 音声スイッチ

音声認識の起動、終了を音声により行なうことを可能にし、インターフェースの改善をはかった。起動命令、終了命令に用いる語彙はユーザーが任意に定義することが可能である。

3. 連続音声認識

3.1. 方式

連続音声認識部は、文法として有限状態オートマトンを用い、認識アルゴリズムとしてフレーム同期アルゴリズムを用いている。バンドルサーチアルゴリズム[2]を適用することにより計算量を削減している。バンドルサーチでは、文法ネットワークの各アークのうち読みが同じ単語を出力するアークを束ね、それらのアークに関しては漸化式計算を共通化する。最適解は保証されないが、劣化の程度は小さいことが実験的に確認されている[2]。

今、簡単のため、1つの単語Wに対し有限状態ネットワークの複数のアーク$A_1, A_2, \ldots, A_n$が対応している場合について説明する。以下、対数尤度を尤度と書く。各時刻tにおいて以下の処理を行なう。単語Wの始端においては、各アークの始端の尤度$Ps(t, A_i)$のうち最も大きい尤度を単語始端の尤度$F(t)$として記憶する。

$$F(t) = \max_i Ps(t, A_i) \qquad (1)$$

単語Wの漸化式計算ではその始端尤度$F(t)$を用いる。単語Wの終端においては、終端尤度Gと単語開始時刻t'における単語始端の尤度$F(t')$との差をもとめ、それを各アークA_iの始端尤度$Ps(t', A_i)$に加算することにより、アーク終端の尤度$Pe(t, A_i)$とする。

$$Pe(t, A_i) = Ps(t', A_i) + G - F(t') \qquad (2)$$

ユーザーによる文法定義のインタフェースとして、図3に示すような辞書エディタを用意し、表形式での文法の入力を可能にした。

---

* Software-only continuous speech recognition for personal computers,
by Koichi SHINODA, Shinsuke SAKAI, Ken-ichi ISO, Kaichiro HATAZAKI, Takao WATANABE and Masanori MIZUNO†
(NEC Corporation　†NEC Informatec Systems)

表 1: 認識性能評価実験結果 (%)

M1	M2	F1	F2	平均
95	100	97	90	95.5

また、認識辞書に対する認識処理と並行して連続音節認識を行ない、その結果の尤度 L_0 を補正尤度として用いて認識結果尤度 L を補正する、尤度補正によるリジェクションを行なっている [6]。

$$L' = L - L_0 \tag{3}$$

ここで、L' は補正後の尤度であり、L' に対し適当な閾値を設定してリジェクションを行なう。

3.2. 評価実験

連続音声認識のオンライン評価実験を行なった。標準パターンとして、男性 23 名、女性 20 名による音素バランス 250 単語 1 回発声を用いて学習された、不特定話者標準パターンを用いた。首都高速のランプおよびインターチェンジ 99 地名の、「[出発地] から [目的地] まで」という文法を定義し認識対象とした。単語パープレキシティは 9.9 である。認識用データとして、男性 2 名女性 2 名の計 4 名の 100 文 1 回発声を用いた。表 1 に文認識率を示す。話者によるばらつきが見られるものの、平均認識率 95.5 % と良好な結果を得た。この実験条件で、連続認識処理用に必要な部分のメモリ量は、約 100KB であり、システム全体で約 800KB である。認識時の応答は CPU にインテル i486 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

パソコン上で連続音声認識を実現した。99 地名を用いた連続認識タスクによる評価の結果、良好な認識性能を得た。音声入力プラットフォームとして、今後、様々なアプリケーションに使用したい。

参考文献

[1]磯 他: 音学講論, 2-Q-21 (1993.10)
[2]渡辺, 吉田, 畑崎: 信学論 (D-II), J75-D-II,11(1992)
[3]磯谷他: 音学講論, 1-8-19 (1990.10)
[4]渡辺他: 音学講論, 1-8-7 (1993.10)
[5]篠田, 磯, 渡辺: 信学論 (A), J77-A,2(1994).
[6]渡辺, 磯谷, 塚田: 信学論 (D-II),J75-D-II,8(1992).

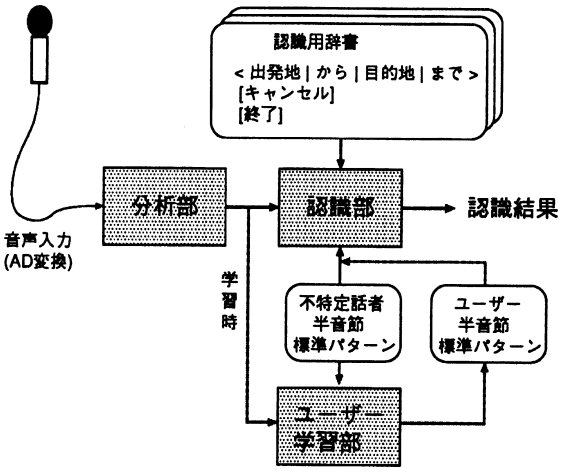

図 1: システム構成

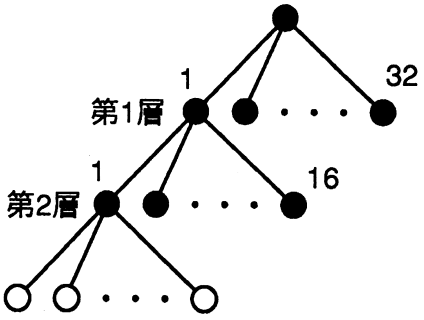

図 2: 木構造標準パターン

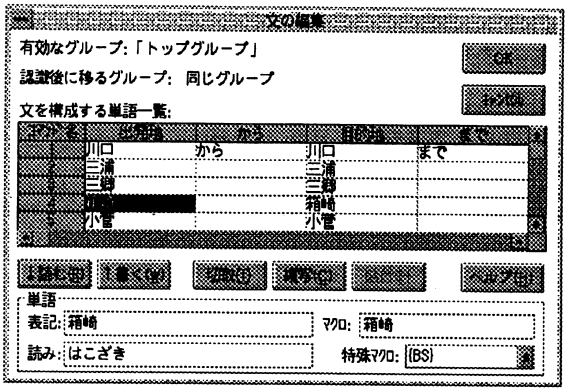

図 3: 連続音声認識用辞書定義エディタ