パソコンソフト連続音声認識

7R－5

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

近年、マルチメディア機能をもつパーソナルコンピュータ (PC) が普及し、音声・動画などを扱うアプリケーションが増加している。このような状況下で、キーボード・マウスの他に音声を入力インタフェースとして用いることができれば、使いやすさがより向上し、新しい形態のアプリケーションが可能になると考えられる。先に、筆者らは、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]。このシステムは単語ごとの発声を入力とするものであったが、今回、さらに、文、句などの複数単語の系列 (連続音声) の認識を可能にした。より自然な発声に近い入力形態で音声を入力することができる。また、いくつかの機能を追加し、音声入力インターフェースの改良を図った。

2. 特徴

本システムの特徴を以下に述べる。＊印のあるものが今回新たに追加されたものである。

1. 誰の声でも認識可能な不特定話者認識システムである。ユーザーの声の事前登録は不要である。
2. 50 単語程度の発声を用いてユーザーの声の特徴を学習し、さらに認識性能を向上させることができる。
3. 新規に登録する認識対象単語の読みをかな表記で定義できる。発声が不要で、認識辞書の作成が容易である。
4. 認識の結果、あらかじめ定義されたキーストロークをアプリケーションに送るキーエミュレーションの機能をもつ。これによって既存のアプリケーションを音声で操作することができる。
5. ＊文、句などの複数単語からなる発声を認識することができる (**連続音声認識**)。各々の単語に対しキーストロークを定義でき、一つの発声で複数の操作の組み合わせを実行することができる。
6. ＊連続音声の対象となる文を表形式で表示・編集するグラフィカルなインターフェースが用意されている (図 1)。単語を行列の形に並べて書くことにより、認識可能な文をユーザーが容易に定義できる。
7. 同時に認識できる単語数は 200 単語程度である。アプリケーションごとあるいは場面ごとに認識対象を切替えることができ、全体的により多くの単語を認識対象とすることができる。
8. インテル i486™ 程度の CPU 上で高速に動作する。入力と同期して認識処理を行ない、音声入力の終了と同時に認識結果を出力する。
9. ＊音声認識機能の起動・終了を音声入力を用いて行なう音声スイッチの使用が可能。認識を使用していないときの誤動作を防止する。
10. ＊リジェクト機能をさらに強化した。認識対象外の発声が入力された場合の誤認識を防止する。

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

3. システムの概要

本システムはサウンド取り込み機能をもつパソコンにおいて、Microsoft®Windows™3.1 上で動作する。システム構成を図 2 に示す。

3.1. 分析部

PC に入力後、AD 変換された音声信号に対し、そのパワー情報を用いて音声検出を行なう。音声の始端を検出すると、それ以降の音声信号に対して 16 ミリ秒フレーム周期のメルケプストラム分析を行なう。このとき、スペクトルサブトラクション処理によって、背景雑音の影響を低減している。

3.2. 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 を用いた不特定話者音声認識 [2] を用いている。不特定話者の標準パターンは、男女 43 名による音韻のバランスを考慮した 250 単語の 1 回発声を用いて作成されている。連続音声の認識においては、認識対象は有限状態オートマトンで定義され、フレーム同期 DP を用いて認識処理を行なう。

連続音声の認識処理は、主に、HMM の各状態における特徴ベクトル出力確率の計算と、入力音声と各単語

の半音節モデル列との時間軸整合のための漸化式計算、および、有限状態オートマトンの単語間ノードにおいて始状態からの単語列に対する累積出力確率を計算する単語間処理、の3つに分けられる。これらの計算が入力フレームごとに繰り返される。認識処理の高速化のためにはこれらの計算量の削減が必要である。

そこで、まず、出力確率計算については、確率分布の木構造を作成し、出現確率の小さい分布の出力確率をその上位ノードの分布の値で代用することにより、計算量を従来の1/10以下に削減している[3]。また、漸化式計算については、有限状態ネットワークの別々のアークに出現する読みの同じ単語を束ね、漸化式計算を複数の読みの同じ単語に対し1回のみ行なう処理(バンドルサーチ[4])を用いる。N桁連続数字認識の場合には計算量は$1/N$になる。

3.3. ユーザー学習部

ユーザーの音声を用いて、スペクトル内挿写像に用いた話者適応化[5]を行ない、ユーザー用の標準パターンを作成する。学習を途中まで行ない、後日追加の学習を行なうことも可能である。学習を行なうことにより誤り率が4分の1になることが実験で確認されている[1]。

4. 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。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。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に示すように、平均の文認識率が95.5 %と良好な結果を得た。

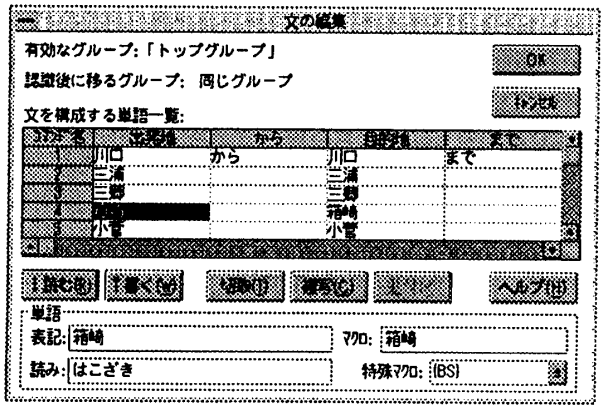

図1: 文法定義インターフェース

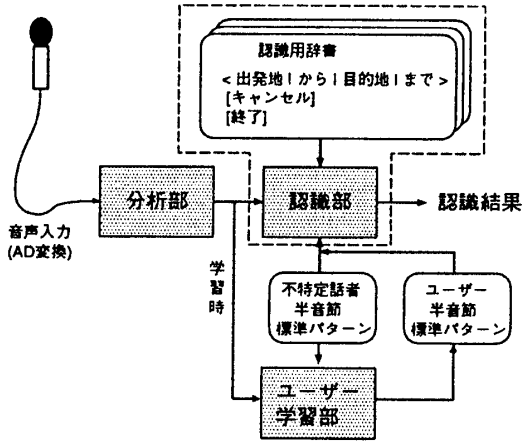

図2: システム構成

表1: 認識性能評価実験結果(%)

M1	M2	F1	F2	平均
95	100	97	90	95.5

5. API

キーボード互換に加えて音声認識機能を直接操作するアプリケーションを作成できるように、アプリケーション・プログラミング・インターフェース(API)を作成した。このAPIを用いて、「音声ダイヤル」[6]、「音声空席案内システム」[7]などのアプリケーションが試作されている。

6. おわりに

パソコン上でソフトウェアのみによる連続音声認識を実現した。音声入力プラットフォームとして、様々なアプリケーションでの利用を進めている。

参考文献

[1] 畑崎他: 第47回情処全国大会, 5V-5 (1993.10).
[2] 磯谷他: 日本音響学会講演論文集, 1-8-19 (1990.9).
[3] 渡辺他: 日本音響学会講演論文集, 1-8-7 (1993.10).
[4] 渡辺他: 電子情報通信学会論文誌, J75-D-II,11(1992).
[5] 篠田他: 電子情報通信学会論文誌, J77-A,2(1994).
[6] 野口他: 信学技報, SP94-54 (1994.11).
[7] 畑崎他: 第50回情処全国大会 (1995.3).