＜報文＞

不特定話者に対する音声認識システムの認識率評価

近畿職業能力開発大学校
附属京都職業能力開発短期大学校　　殿　村　正　延

Evaluation of the recognition rate of the speech recognition system for unspecified speakers

Masanobu TONOMURA

要約　近年、従来のアプリケーションソフトウエアの付加価値の向上を目指し、ヒューマンインターフェースの１つである音声認識、音声合成の技術が取り入れられつつある。

本論文では、まず音声認識技術の動向について述べ、現状と今後を把握する。次にIBMの開発した音声認識技術を取り入れた「舞鶴観光ガイドシステム(プロトタイプ)」の構築を実際に行い、現在の技術を用いた場合の実用化の可能性を検証する。具体的には雑音の影響、不特定話者による認識率への影響、認識させたい適切な文章の長さ等について行う。

Ⅰ　はじめに

音声認識とは、マイクなどから採取した音声、主に人間の話し声をコンピュータが分析して理解することである[(1),(2)]。

音声認識を行うには、発音に対応する音の振幅スペクトルのパターンをデータとして保存しておき、そのパターンと音声を比較することによって、どういう発音がなされたのかを推測する。日本語であれば、例えば、「あ」なら「あ」に固有の、「わ」なら「わ」に固有の音声パターンを認識する。発音を認識した後は、言葉として理解する必要があるが、この段階では自然言語処理と呼ばれる技術が用いられる。基本的には、単語、文法や語法などに関するデータをコンピュータに保存した上で、それらを元にして意味を解釈する。例えば「ケンサクケッカヲインサツセヨ」と抽出された音声を、「検索結果を印刷せよ」というように、言葉としてコンピュータが認識できるようになる。この技術を用いることによって、キーボードで文字を入力することなくコンピュータを操作することが可能となる。

人間の言葉を機械に理解させる研究は、数十年前から進められてきた。しかし、当初は音声認識技術自体が未熟で、また、ハードウエアの性能も十分でなかったため、十分な成果をあげることは困難であった。しかし、ソフトウエア、ハードウエアの両面における近年のめざましい進化により、音声認識技術の利用が急速に広がってきている。業務向けの利用はもちろんのこと、個人向けにも比較的安価な音声認識ソフトが販売され始め、一般のユーザにも身近な存在になりつつある。

本論文では、まず、実用段階に入った音声認識技術の動向を把握する。次に、この技術の検証を行い、現時点での実用化可能な範囲を探る。具体的には、現在主流である連続音声認識を可能とし高い認識率を誇るIBMのViaVoice[(3)]を使用し、「舞鶴観光ガイドシステム（プロトタイプ)」を構築し実用に近い段階での検証を行う。このシステムは話者の音声を認識し、観光地情報をテキストデータ（画面表示）と音声合成を用いた音声で提供するシステムである。

論文受付日　H16.12.2

Ⅱ　音声認識技術の動向

1　音声認識技術の発展の歴史

音声の認識は、観測された音声信号から調音フィルタの振幅伝達特性（離散振幅スペクトル、あるいは離散パワースペクトル）を抽出し、音素の標準的な音響特徴(標準パターン)と比較照合することで行われる(4)。

1995年より以前は、これらの比較データを蓄積していくことにより音声認識技術が実用化されると考えられていた。しかし、実際には、

① 人によって音声データがかなり異なるため、システムを使用する個人ごとに比較データを準備する必要がある（音声特徴抽出の問題)。
② 認識する単語は事前に認識させたいだけ準備しなければならない（労力の問題)。
③ 同音異義語の識別（自然言語処理の問題）

等の問題があり、簡単な音声認識システムは実現できたが、自然発声による十分な数の語彙の認識システムの実用化を阻んでいた。

しかし、1995年に隠れマルコフモデル(Hidden Markov Model 以下、HMM）が登場し、音声認識技術が飛躍的に向上した。HMMの重要な点は学習機能があることである。以前は蓄積したデータを使用して正しく認識されるようにシステム設計者がマルコフモデルを調整する必要があったが、HMMにより自動化することが可能となり、上記問題点を克服可能とした。この機能をいち早く取り入れ製品化したのがIBMのViaVoiceである(当時Voice Type：1997)。

音声認識技術は、配送先ごとに商品を仕分ける物流や、社名や金額などを入力する伝票処理の分野においては比較的古くから利用されてきた。近年活用されている主な分野としては、携帯電話、携帯端末、カーナビゲーション、ＰＣディクテーション(書き取り)ソフトウエアなどがある。

2　音声認識製品

ＰＣ用ディクテーションソフトウエアは既に各メーカより提供されている。主な製品を表１に示す。音声認識ツールの競争時代に入り、性能と価格においてしのぎを削っている。適用例により優位性の違いはあるが、①～⑤の中では現時点では性能・価格で日本IBM社のViaVoiceが優れているといわれている。これらは共に特定話者による音声認識を想定しており、エンロールを必要とする。エンロールとは、自分の声の特徴を記憶させる作業のことである。加えて認識を行う音響環境の特性も記憶する。これにより特定環境における特定話者の認識率を向上できる。それに対し、⑥、⑦は事前に特定の環境での使用（例えば、自動車内、携帯電話等）を想定し、あらかじめボイスモデルを構築するため、使用者はエンロールを必要としない。つまり、特定環境における不特定話者の音声認識を可能とする。

⑥のAmiVoiceは、日・米・独共同で開発された音声認識エンジンであり、提供元のアドバンスド・メディアは1997年に設立され、比較的新しい。主に医療・テレフォニー分野等の大きなマーケットを対象としており高価である。⑦のJuliusは情報処理研究会連続音声認識コンソーシアムにおいて開発された日本語専用のエンジンであり、他のメーカ提供のものと性質が違うが、ソースがオープンであり、使用・内容の変更においてほとんど制限がないという特徴がある。今後、これを改変し組み込んだ製品が業界の牽引役となることが期待されている。

Ⅲ　ViaVoiceの認識率の検証

1　検証環境

音声認識システムの応用は、Ⅱ．1で述べたように多岐に渡る。今回ターゲットとしているのは、「音声認識観光ガイドシステム」の実用化であり、特定の音響環境における不特定話者による音声認識である。エンロールという作業を考えると、不特定話者用の認識エンジンを使用することが一見適当と思われるが、これらはあらかじめ導入環境の音響モデルを想定しており、必ずしも観光ガイドシステムを設置する場所に適しているとはいえない。よって、本論文では特定話者

表１　主なディクテーションソフトウエア製品

製品名	メーカ	エンロール
① ViaVoice	日本IBM社	有
② SmartVoice	NEC	有
③ DragonSpeech	アスキーソリューションズ	有
④ LaLaVoice	東芝	有
⑤ VoiceStyle	ソニー	有
⑥ AmiVoice	アドバンスド・メディア	無
⑦ Julius	連続音声認識コンソーシアム	無

用の音声認識エンジンの中で最も評価の高いViaVoiceを使用し、特定音響環境における不特定話者の認識率を評価することとする。

評価には、ViaVoice v9.0 Pro USB for Windowsのボイスモデル、観光ガイドシステム(プロトタイプ)の構築には、Microsoft VC++ v6.0, IBM ViaVoice v8.0 SDK for Windowsを使用する。なお、本論文は認識率の検証を主目的とするため、観光ガイドシステムの詳細については省略する。ViaVoice音声認識エンジンのアーキテクチャを図1に示す。Ⅰ.で述べたとおり、マイクから拾った音声は最終的に自然言語処理されたテキストとして出力される。

2　認識率の測定

2.1　認識対象センテンス

観光地名称：20語、一般会話：20語をベースとし、接頭語、接尾語：25語を音声グラマーに登録する。ViaVoiceは正式な言語理論として多くの言語に使用されているバッカス正規形式(BNF)の構文を採用した音声認識制御言語のサークル(SRCL)を用いている。SRCLを使用して音声グラマーを構成すると、共通句、任意選択句、および反復句を識別するための表記があらかじめ準備されているため、音声グラマーの一部である単語や句の組み合わせを簡単に作成することができる。これにより登録数はわずか65句であるが、グラマーの機能により4000余りのセンテンスが表現可能となる。例えば、

- 市役所はどこですか
- すみませんが、舞鶴市役所の場所を教えてください
- あの、市役所の場所を教えてください
- えー、舞鶴の市役所はどこですか

は、市役所の場所を尋ねる同一質問となる。

認識率評価のために、この組み合わせの中から適当に長文・短文、観光地についての質問・一般会話をランダムに15文選び出した。

〈センテンス〉

1. すみませんが、舞鶴の海水浴場について少し教えてください
2. あの、ポリテクカレッジ京都の場所を教えてください
3. 安寿姫塚についてもう少し詳しく教えてください
4. えー、海軍記念館はどこにありますか
5. 五老スカイタワーについて教えてください
6. もう一度お願いします
7. 市役所はどこですか
8. 今日は何日ですか
9. 今　何時なの
10. さようなら
11. ありがとう
12. こんにちは
13. あほか
14. すごい
15. 元気

2.2　測定の条件・結果

無作為に選んだ情報技術科の学生、男女各一人がクイックエンロールを行い、男女それぞれ10名を話者とした場合の認識率を調べる。クイックエンロールとは、ViaVoiceが準備した短い例文を1分程度読み上げ、音声の特徴を学習させる処理のことである。1回目で認識されなかった場合は、もう一度同じセンテンスを

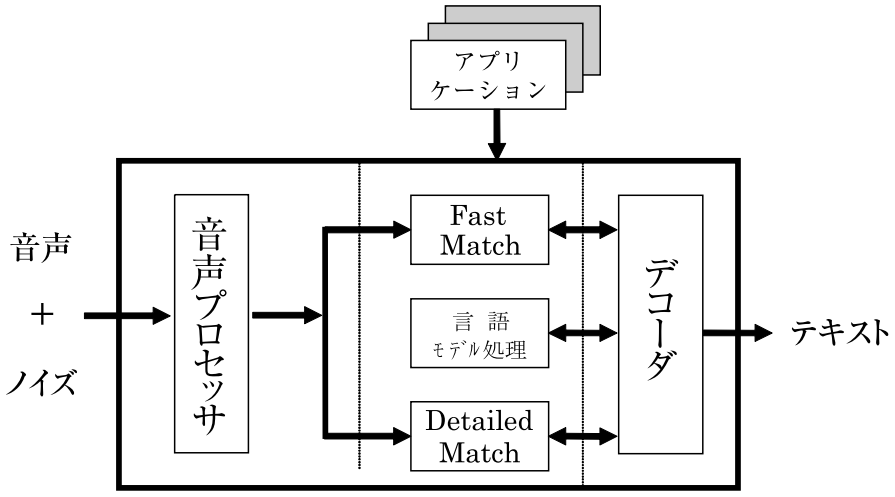


図1　ViaVoice音声認識エンジンのアーキテクチャ

読み上げる。最高3回まで行い、それ以降は認識できなかったとする。

結果を図2～5に示す。エンロールを男性が行った場合は、明らかに男性のほうの認識率が高く、女性がエンロールを行った場合は、女性のほうの認識率が高いという結果になっている。また、グラフが全体的に右上がりになっていることから、長い文章よりも短い文のほうが音声を認識しやすい傾向がある。

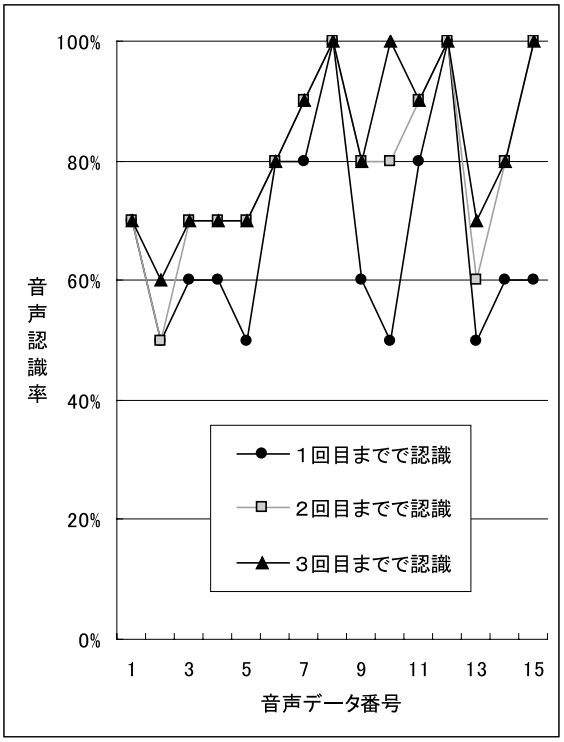


図2　女性がエンロールを行った場合の男性の認識率

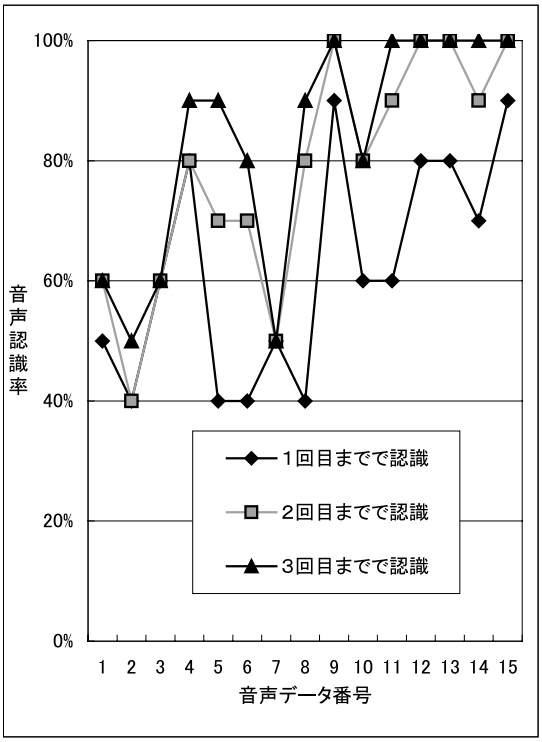


図4　男性がエンロールを行った場合の女性の認識率

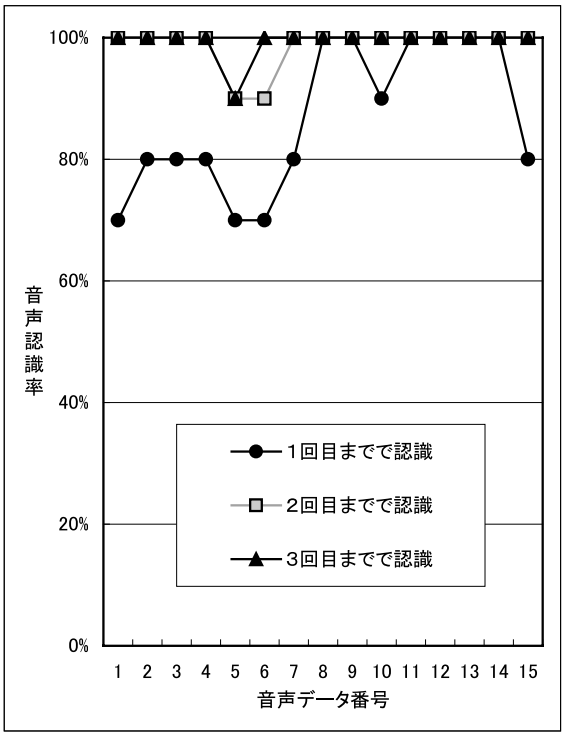


図3　男性がエンロールを行った場合の男性の認識率

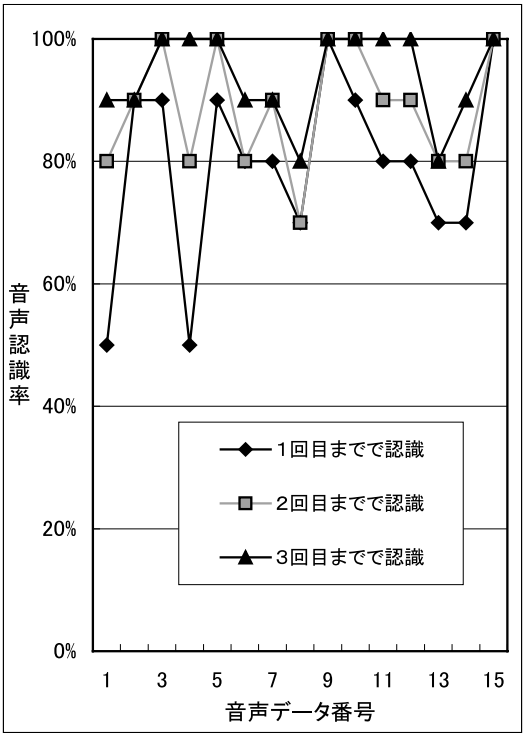
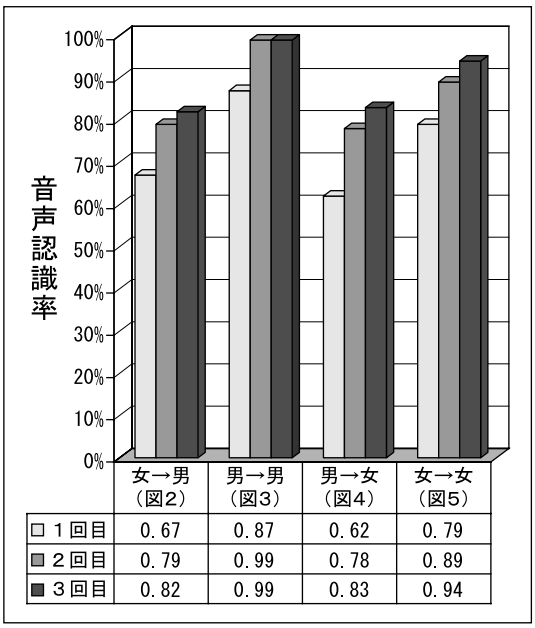


図5　女性がエンロールを行った場合の女性の認識率



	女→男（図2）	男→男（図3）	男→女（図4）	女→女（図5）
□１回目	0.67	0.87	0.62	0.79
▒２回目	0.79	0.99	0.78	0.89
■３回目	0.82	0.99	0.83	0.94

図６　各回の平均認識率の比較

図６に各回の平均認識率の推移を示す。１回目の認識に失敗した場合でも２回目、３回目にはかなり高い認識率になっている。この原因として、話者の発声（カクゼツ）が改善し認識されやすくなったこと、もう一つはHMMが確率的な振る舞いをすることが考えられる。

測定環境はかなり多くの学生が遠近で話している粗悪な環境（実用で想定している環境に近い）で行っているが、雑音による影響は認識率にそれほど影響していないようである。これはエンロール、及び測定を指向性の高いマイクを用いて行っているためと考えられる。ただし、マイクに向かって話している人と同方向近距離の場合は認識に影響を与えるのを測定から確認している。

ViaVoiceは特定音響環境における特定話者に対して使用されるものであるが、男女それぞれのエンロールで作成したボイスモデルを話者の性別に合わせて切り替えることで、特定環境における不特定話者に対して適用可能であるといえる。特に、今回ターゲットとしている観光ガイドシステムとして必要としている、表音文字に直して40文字程度のセンテンスは十分認識可能である。しかし、以下のような制約もある。

〈音声入力時のポイント〉

1. まとまった長さの文章を、朗読調で発声する
2. 単語をひとつひとつ区切らない
3. 声の大きさやマイクの位置を一定に保つ
4. 先頭の単語をはっきり発声する
5. 登録されていない単語は認識されない

これは、現在の音声認識エンジンが自由発声を行えないためであり、限界を示している。

学生に認識測定を行わせたときに、自分の声が認識されなかった場合、発声が機械的（抑揚がなくなる）のは興味深い現象であった。機械に認識してもらうために自らを機械に近づける（適応する）行動を取るのである。また、４回以上認識されない場合にはかなりのストレスを感じるようである。

このことから、ユーザに活用される音声認識システムを構築する場合には、

① システムが何でも認識してくれるという錯覚を事前に取り除くこと

② 事前検証を十分に行い、認識されるセンテンスの範囲内で構築すること

が重要となる。

Ⅳ．むすび

近年情報処理技術において、音声認識技術が注目されている。これは機能に制限があるものの技術レベルが実用段階に入ってきたこと、さらに、きたるユビキタス社会の一端を担うことが期待されているからである。音声認識の精度は、認識エンジンの性能とボキャブラリの質と量で決まり、現在も自由発声も視野に入れたより多くのボキャブラリを実時間で処理できる仕組みが研究されている。現在は読み上げ音声レベルであるが、ある制限を設けた中での実用化は十分可能である。

本論文では、具体的な応用例を想定しその有効性を検証したが、この中で得られたノウハウを今後、セミナーや企業との共同研究開発につなげて行くことを考えている。なお、本内容は平成１５年１２月１９日に行われた「近畿北部・まいづる地域　産学連携セミナー」で発表したものをまとめたものである。

[参考文献]

(1) 中川聖一、パターン情報処理、丸善、1999年11月、P5-57

(2) 石井健一郎 他、わかりやすいパターン認識、オーム社、1998年8月、P1-12

(3) 日本IBM、SMAPI開発者の手引き(IBM ViaVoiceTM SDK for Windows V1.5)

(4) 鹿野清宏 他、音声認識システム、オーム社、2001年5月、P1-51