ヒューマンインタフェース 97－25
音声言語情報処理 40－25
（2002. 2. 2）

音声認識技術の利用形態とその性能評価に関する一検討

李 琳　伊藤 憲三
岩手県立大学大学院ソフトウェア情報学研究科
〒020-0105 岩手県滝沢村巣子152-52
E-mail:g231y018@edu.soft.iwate-pu.ac.jp;itoh@soft.iwate-pu.ac.jp

あらまし　音声認識技術は、ハードウェアの進歩と共に発達し、最近では種々の分野で多くの商品が販売されるようになっている。また、現在のようなブロードバンド社会では、情報端末機器の小型化や携帯化を考慮して音声認識技術の利用形態も広まっている。しかし、音声認識技術の利用形態は、認識率、耐雑音性あるいは音声応用という特異性等から応用分野が限られているのも現状である。本報告では、音声認識技術のより最適な利用形態（出口）を模索するために次の検討を行った。(1)音声認識技術の利用実態調査、(2)電話系とマイク系の代表的な音声認識エンジンの基本性能評価、(3)実環境における評価、(4)音声認識技術の利用形態に関する考察。
キーワード　音声認識技術、利用形態（出口）、利用実態調査、基本性能評価、耐雑音性、文章理解度

A study on basic performance evaluation and find an outlet for application of speech recognition technology

Lin Li　Kenzo Itoh
Graduate School of Faculty of Software and Information Science
Iwate Prefectural University
Sugo152-52,Takizawa-mura, Iwate 152-52, Japan
E-mail:g231y018@edu.soft.iwate-pu.ac.jp, itoh@soft.iwate-pu.ac.jp

Abstract　The speech recognition technology (SRT) have been researching and developing, and many equipment are proposed and sold in wide application area. In addition, application of SRT increase for many area such as broadband or Web-internet communication systems. However, it is very difficult that look for a good match application of SRT, because, SRT can not use on every environment or condition. This paper studies and describes as following. (1) Fact-finding of SRT application. (2) Basic performance evaluation for two types speech recognition engine. (3) Influences of noise. (4) Considerations for application (Outlet) of SRT.

Keyword　Speech recognition technology (SRT), Application(Outlet), Fact-finding of application, Evaluation of basic performance, Performance for noise, Articulation for sentence

1. まえがき

音声認識技術は古くてまた新しい技術であり、音声符号化や音声合成技術と共に発達してきた。これらの技術は、近年ハードウェアの進歩やパーソナルコンピュータの普及と共に急速に発展し、種々のサービスが提案されてきている。例えば、電話系を用いた利用形態では、音声による質問回答システム(以下 Q/A システムと呼ぶ)である。また、音声ワープロは、コンピュータの個人利用の拡大と共に大きな市場となってきている。音声を用いたマン・マシンインタフェースでは、人間が本来持っているコミュニケーション手段として、自然で気軽に使えることが大きなメリットとして挙げられる。現在のようなブロードバント社会では、「オンラインはいつでもどこでも」という観点から、ネットワークに常時接続して利用できるインフラが整いつつある。このような情報社会では、情報端末機器の小型化や携帯化を考慮した音声認識技術の利用が益々重要になってくると考えられる。しかし、現状の音声認識技術では、どのような使用環境でも高い認識率を得ることは困難である。また、「発声する」という音声の持つ特異性(利便生と欠点)などから、利用形態が限られているのも現状である。すでに使われている音声処理技術を利用したサービスシステムでは、通常、既存のインターフェース(例えば電話のプッシュボタン、パソコンのキーボード等)と併用してオプション的な機能として活用されているのが現状である。

本報告では、音声認識技術の最適な利用形態(出口)を見つけ出すために以下の検討を行った。(1)現状の利用形態調査。(2)代表的な音声認識エンジンの基本性能評価。(3)実環境における性能評価。最後に、これらの検討結果をもとに、利用形態(出口)に関する考察を加えた。

2. 音声認識技術の利用形態調査

2−1 調査方法

現在、音声認識技術を利用したサービスは非常に多く、新聞や雑誌などに掲載されるようになってきている。音声認識の利用形態を明確に分類できないが、その使用環境から次の2種に分けられるであろう。(1)マイク入力を主とし、パーソナルコンピュータ(PC)上で使用する形態(オフライン型)。(2)電話系を利用し音声応答技術と共に利用する Q/A システム(オンライン型)[1]。PCを使用する形態では、音声入力の環境条件が良く、接話型マイクを使用することによって高い認識率を比較的容易に得ることができる。この場合には、マウスやキーボードなど他の入力モードと競合することになる。一方、電話系では信号帯域が制限され雑音や伝送歪も大きく、認識の条件としては悪くなるが、音声が主な信号であるために音声認識技術の利用形態としては重要な出口と考えることができる。しかし、実際のサービス状況を見てみると必ずしも有効に動作していないのが現状である。ここでは、日本電子工業振興会の「音声入出力方式に関する調査研究報告書」[2]、各種の学会誌や関連雑誌[3][4][5][6]や下記のホームページなどをアクセスして情報を収集した。

http://it.jeita.or.jp/jhistory/committee/humanmed/speech/index.html;
http://www.voistage.com/;
http://www.omron.co.jPCma/index.html;
http://www-6.ibm.com/jp/voiceland/

2−2 調査結果

表1は、現在の音声認識技術の応用分野を電話系とマイク系に分け、利用形態を簡単にまとめたものである。電話系では音声合成技術を併用して Q/A システムを構成しているが、規則合成音の品質が問題となる場合には、録音編集形が利用されるケースが多いようである。調査の中で、全日空のチケット予約サービスでは、自由発話が可能になっていてユーザーにとって使い勝手良いと感じられるシステムである(ノースウエスト航空のサービスも同様)。このシステムでは、オムロンのNuanceの認識エンジンが使用されている。ポータルサイドの「Voizi Lab !」(例えば、スポーツ、健康など)では、一つ一つの項目確認がユーザーにとって非常にわずらわしく感じられる。これは、認識率の低さを「確認」というステップで回避する為で、このような一問一答形式の Q/A システムは、ユーザーにとっては好ましくないと思われる。インターネット検索サービスは、接続不可能なものなどが目立った。習志野市役所の情報案内サービスは、出産や戸籍関連などの手続き情報などを24時間サービスしており、利用効率も高いと思われる。以下、そのシステムの概要を述べる。

習志野市は、市民の利便性を目的に、全国の自治体で初めて音声認識技術を活用する24時間対応の電話による市民の行政情報提供サービス「テレホンガイド習志野」を導入した。このサービスは、「耳寄り情報習志野」として平成5年にスタートした。しかし、サービスの選択にコード表を利用していたため、非常に使いにくいシステムであった。そこで、このシステムに音声認識技術を利用することで、利用者は各種サービスに対応するコードを覚える必要もなく、コード表を紛失して利用不能になることからも開放され、非常に使いやすくなった。また、同システムにより休日や夜間の情報提供も可能となり、緊急を要する情報を24時間いつでも得ることができるようになった。このシステムで用いている音声認識の性能は(認識率=80%程度、語彙数

表1 音声認識技術の応用

分類	利用形態(例)	サービス名の一例	特徴
電話系	案内 (テレホンガイド)	習志野市音声ガイド／大学合否照会／医療照会テレホンガイド／道路交通情報	・オンライン ・不特定話者 ・実環境の影響が大きい (雑音/周波数制御など) **メリット** ・24時間無人サービスが可能 ・電話応対の迅速化・標準化
	予約 (診療予約)	音声認識診療予約システム／会議室予約システム／CARアンサー(車検情報)	
	検索 (商品情報)	航空会社予約・照会システム／野村の株価ダイヤル(株価照会)	
	ポータルサイド (ウェーブ情報)	Voizi Lab !	
マイク系	制御 (機器制御)	カーナビ／AIBO	・オフライン ・特定話者 ・使用条件良い **メリット** ・ハンズフリー通話 ・障害者と高齢者のパソコン利用支援
	PCソフト (音声ワープロ)	ViaVoice(IBM) ／SmartVoice(NEC) ／宛名職人2000／NHK ニュース番組字幕／語学教育	

約300語程度)と、必ずしも十分とはいえないが、それをサービスの「水先案内人」として非常にうまく利用していると思われ、最適な利用形態の一例である。

一方、マイク系での利用形態は、荷物の仕分け(機器制御)などが音声認識利用の草分けであった。しかし、最近では音声ワープロに代表されように、PCソフトのオプション的な利用形態が主流である。

3. 音声認識技術の基本性能評価

3－1 実験方法

先に述べた調査結果から、音声認識技術の利用形態は、電話系に代表されるオンライン形式と、音声ワープロに代表されるオフライン形式に大別でき、その性能の範囲内で利用されていることが分かった。従って、具体的に利用形態を考える場合には、その基本となる認識エンジンの性能を的確に把握する必要がある。そこで、ここではA社の音声ワープロとB社の電話応答システムに使われている認識エンジンについて、その基本性能を確認することとした。

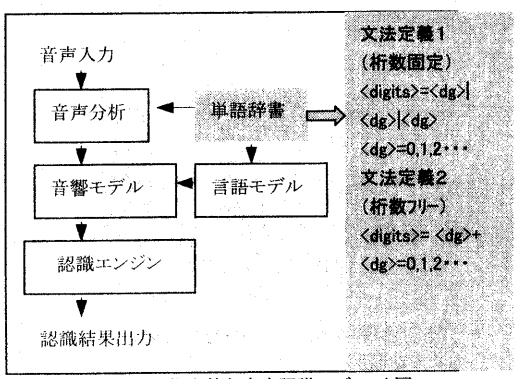


Fig. 1 代表的な音声認識のブロック図

基本的な音声認識エンジンは、HMM を基本とするものが主流である[7]。単語辞書は、認識対象の単語を文法ファイルを利用して定義され、この定義方法によって、応答時間や認識率に影響を与える(Fig.1参照)。同図の例では、数字を発声する場合に、文法定義1では桁数を固定した場合、文法定義2では桁数を自由にした場合を「単語辞書」でそれぞれ定義することになる。

表2に、実験に用いた評価用音声資料を示した。発声音声は、数字音声と単語である。

表2 評価用音声資料

入力系	マイク系 (ヘッドセット)	電話系 (公衆電話回線)
標本化 量子化	22.05kHz 16bit	8kHz 16bit
話者	男性 6 名	男女 160 名
タスク (データ数)	4桁数字(120) 7桁数字(120) 単語(300)	4桁数字(1020) 7桁数字(1195) 単語(2966)

認識結果の評価尺度としては、以下に示す式(1)～(3)とした。

$$PW = \{Nc/N\}*100 \quad (1)$$

$$PC = \{(N-D-S)/N\}*100 \quad (2)$$

$$PA = \{(N-D-S-I)/N\}*100 \quad (3)$$

ここで、Nは全単語数、Ncは正解の単語数、Dは脱落数、Sは置換数、Iは挿入数である。これらの認識率は、もっとも基本的な性能を表現すると思われる。すなわち、PWは単語の正解率、PCは数字毎あるいはカタカナ表記の場合には1文字毎の正解率、PAはPCの条件に挿入を考慮した正解率をそれぞれ示している。実際の認識実験では、認識辞書(文法定義)を、数字音声の認識の場合には桁数を固定した場合(固定)と桁

数を制限しない場合(フリー)の条件で行った。また、この辞書の語彙数を1000語彙、3000語彙、及び5000語彙の3種類とした。

3－2 実験結果

3－2－1 マイク系

Fig.2(a)(b)に、音声ワープロ(A社)の認識結果の一例を示した。Fig.2(a)は数字音声、Fig.2(b)は単語音声の結果である。同図には、発声終了から結果が返されるまでの時間をレスポンスタイム(RT)(秒)として示した。

実験結果から、以下の結果が得られた。(1)単語正解率(PA)は(PC)より劣化が大きい(挿入による誤認識の影響)。(2)辞書の大きさが1000語から3000, 5000語になることによって、認識率が10%近く劣化するが、しかし、レスポンス時間は0. 057秒とほとんど変化しない(Fig.2(b))。その他、音声長等の各種パラメータの設定を行わなくても、ある程度の認識率が得られ、機能がフレキシブルであるといえる。

一般に、音声ワープロの認識エンジンでは、言語モデルによる結果補正が可能なため、リジェックションの閾値を低い値に設定しており、できるだけ脱落による誤認識を避けるように設計されている。これは逆に、誤り挿入を許してもなるべく正解単語候補を湧き出すように設計する傾向がある。ここで実験対象にしたA社の認識エンジンの全体の設計は、ディクテーション機能に徹しているといえよう。ただし、被験者によっては20分～30分(最大60分以上)の事前学習が煩わしく感じる場合もある。

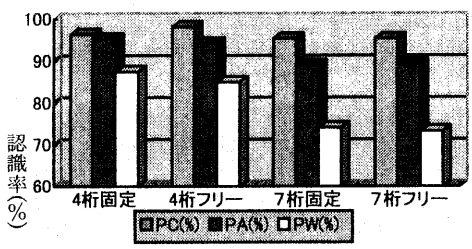
3－2－2 電話系

電話系の認識エンジンには、多数の話者の特徴をカバーするために、音響モデルにはバリエーションの多い音声データが必要とされる。また、用途に合わせた利用者の使用環境、雑音及び歪みなどについても考慮しなければならない。

(a).数字の認識結果

レスポンスタイム(RT)

	4桁固定	4桁フリー	7桁固定	7桁フリー
RT(秒)	0.078	0.08	0.156	0.151



(b).単語の認識結果

レスポンスタイム(RT)

	1000 語彙	3000 語彙	5000 語彙
RT(秒)	0.057	0.057	0.057

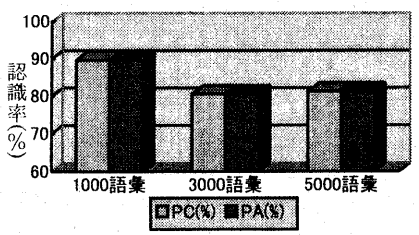
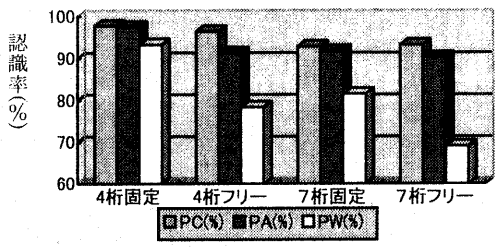


Fig.2 音声認識エンジン(A社)の実験結果

(a).数字の認識結果

レスポンスタイム(RT)

	4桁固定	4桁フリー	7桁固定	7桁フリー
RT(秒)	0.196	0.21	0.267	0.286



(b).単語の認識結果

レスポンスタイム(RT)

	1000 語彙	3000 語彙	5000 語彙
RT(秒)	1.819	1.928	3.487

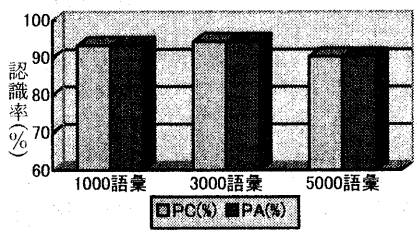


Fig.3 音声認識エンジン(B社)の実験結果

Fig.3(a)(b)に、B 社の音声認識エンジンの実験結果を示した。なお、音声長に関するパラメータなどは4桁数字に、また語彙数は3000語に設定して実験を行った。(a)は数字音声、(b)は単語音声の結果をそれぞれ示した。同図から以下のことが分かる。(1)4桁から7桁にすることで、認識率が低下する。(2)A社の場合と異なり、語彙数増加に対してはその影響は少ない。(3)レスポンス時間は、辞書の大きさに準じて遅くなる。これらの結果から、B社の認識エンジンは、各種パラメータを自由に設定できるために、レスポンス時間を犠牲にしても認識率を向上(確保)したい場合、また種々のサービスを構築する上では柔軟性のある設計となっているといえる。

以上の考察結果から、Q／Aシステムを実用性の高いサービスにするためには、対話の流れを決め、発声コマンドを明確にする(語彙数をサービスに合わせて制限するなど)必要があると考察できる。

4. マイク系における実環境評価

4－1 耐雑音性

4－1－1 実験方法

静かな録音ブース(暗騒音30dB)内で収録した音声資料(新聞記事:約500文字)に対し、雑音(計算機室)を付加し、前述した A 社の音声ワープロを用いて認識実験を行った。雑音レベルは、6段階(SN=40, 22, 17, 12, 7, 3dB)に設定した。認識結果は、文字毎の正解率(PC)で計算した。発声者は、学生10名(男女各5名)である。

4－1－2 実験結果

Fig. 4に、実験結果を示した。同図から、付加雑音の増加が認識率に大きく影響を及ぼすが、SN(17dB)以上では、平均的にほぼ80%の認識率が得られることが分かる。

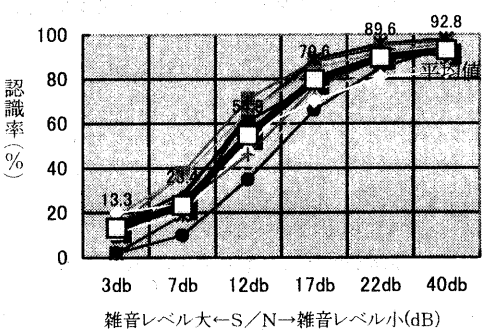


Fig. 4 雑音を付加した場合の認識結果の一例

このことから通常の使用状態(静かな研究室など)では、ハンズフリーでの使用で80%程度の認識率が得られるが、雑音を発する機器等がマイクの近傍にある場合には、接話マイクなどその使用によってSNを向上させる必要があることが分かる

また、SN が悪くなると個人差も大きくなる。なお、研究室(サーバー室)の実際の信号対雑音比を測定すると約15dB、また、同室でプロジェクターの電源をONにした場合は約5dBとなった。この実験から、音声ワープロの応用に当たっては実環境での評価が重要であること、また、通常の事務室等では説話マイクなどを使用してSNが20dB程度以上確保できれば、実用に耐え得る認識結果が得られることなどが分かった。

4－2 文字情報から得られる文章理解度

音声ワープロの利用形態として、福祉支援(聴覚障害者の口述筆記)がある。この場合の利用形態としては、100%の認識結果を要求しないことも予想される。そこで、ここでは、文章認識率とその文の理解度[8]との関係について若干の検討を加えた結果について述べる。すなわち、種々の認識率の文章を提示した場合、その文章の内容がどの程度理解できるかを調べた。

100名の学生を被験者に用いて実験した結果を表3に示した。実験に用いた認識対象文章は、前述(4-1-1)した新聞記事約500文字である。この結果から、文章の正解率が75%の条件では、7割以上が文章を理解するのが難しいことが分かる。また一方、文章の正解率が80%以上であると、8割以上の学生が文章の文法的な誤りを排除しながらも内容を理解できることが分かる。従って、講演や授業などの音声口述として、耳の不自由な人に内容が分かる有効な手段として、音声認識技術が利用できると考えられる。

表 3 文章理解度アンケート調査結果

認識率(%)	95	90	85	80	75
非常によく理解できる	2				
大体理解できる	9	6	8	3	3
考えながら、理解できる	6	12	11	13	3
理解するのが難しい	3	2	1	4	13
全然理解できない					1

(人数)

以上の結果は、文章の内容、難易度、被験者の知識などに影響されると考えられる。また、実際の授業などでは不要語や雑音の問題なども考慮する必要があり、今後より詳細な検討が必要と思われる。

5. 音声認識技術の新しい利用形態

前述したように、音声認識技術は認識率や音声応用

という特異性などの観点から応用分野が限られている。しかし、音声認識技術は人間にとって自然でかつ有効な情報伝達手段という面から、また社会環境への融合性や高齢者や障害者への思いやりなどの観点から、新しいビジネスチャンスが生まれる可能性を秘めている。これまで、既存の音声認識の応用状況及び音声認識エンジンの基本性能について調べた結果などから、以下の新たな利用形態の可能性があろう。

5－1電話系における利用形態

・無人化ホテル予約システム:最寄りの駅名などを発声することで、その周辺のホテルの空き部屋情報を提供し即予約できるシステム。このシステムの主な利用者は、ビジネスマンなどの一般の人であるとし、電話(携帯、公衆、家庭)の普及により、24時間(特に深夜時において)、電話しか使用出来ない状況における即答型の情報サービスシステムである。

・日常応急処置ガイド:電話で症状等を発声、軽い症状への処置方法、一般症状から当てあまる病気の関連情報を気軽に得ることができる。最終的には近くの病院を案内することも可能。24時間で日常生活の健康及び病気に関する情報を提供するサービスを目指すものである。

・タクシーお迎えサービス:携帯電話のGPS機能とタクシーのGPS機能を連動して、携帯電話から名前と台数を発声すると、無線配車システムによって最も近い場所にある車両で迎えに来ることができ、サービスの無人化と迅速化を目指すものである。

以上、いずれの利用形態も24時間対応サービスシステムである。

5－2　マイク系

・音声口述システム:講演や授業など、講演者や先生の話を音声ワープロによって文字情報に変換し、聴力障害者に提供する。(現状では人手による概要記述)。主な利用者は聴力障害者とライターとする。先述した主観評価結果によれば、文章正解率が80%以上あれば、文章の内容がほぼ理解できることから、実用性が高い応用と考えられる。

・バス運行情報案内システム:バス停では、行き先を音声入力するだけで、バスの運行状態と道路の混雑状態を音声合成などで知らせるシステムである。このシステムの主な利用者は、お年寄りや身体障害者である、既存のバス停の情報案内サービス(文字、音声、サイン音)を音声認識、合成技術によって一体化し、マイク付の小型端末より音声指示で情報を得ることができる。

以上2例のいずれも福祉支援応用であるが、不用語のリジェクションや戸外でのSN確保の問題を解決する必要がある。

6. 結論

以上の結果から、まだ完成度が十分でない音声認識技術の利用形態(出口)は、かなり制限されていることが分かった。しかし、企業や公共機関の付加サービスとしての「水先案内人」としては、十分にその役割を果たせるとの結論に達した。このような利用形態から、ユーザーに対する音声認識技術の存在性と有効性を認識してもらうことが重要である。今後は、あらゆる環境で使用できるような認識技術の確立が期待される。また、情報化と高齢化が益々加速する社会においては、高齢者や障害者支援への利用面でも音声認識や合成技術の利用促進が計られるであろう。音声認識技術の利用形態に対するキーワードは、以下のようであろう。インターネット、ブロードバンド社会、WEBアプリケーション、高齢社会、障害者支援

「謝辞」日頃頃ご指導頂く伊藤研究室の堀内助教授、伊藤(久)助手の先生方に感謝いたします。習志野市役所広報課の係りの方からは、サービスの説明などを丁重に説明頂いた、ここに深く感謝いたします。また、種々の実験に協力頂いた本学生に感謝いたします。

「参考文献」

[1] 河原達也:ここまできた音声認識、情報処理学会誌 VOL.41、pp.436～439(Apr.2000)
[2] 『音声入出力方式に関する調査研究報告書』日本電子工業振興協会、(Mar.2000)
[3] Computer TELEPHONY:ビジネスシーンでみる音声認識の可能性、pp.19～33、(Apr.2000)
[4] 壇辻正剛:IT時代の語学環境としてのCALL、情報処理学会誌 VOL.42、pp.1001～1005 (Oct.2001)
[5] デジタルバイヤ:最新音声認識ソフトで遊ぼう、(株)アスキー、pp.150～155、(May.2001)
[6] 朝日新聞:文化往来、テレビの生番組　字幕の導入広がる、pp.32、(28.Dec.2001)
[7] 鹿野清宏、伊藤克亘、河原達也、武田一哉、山本幹雄:『音声認識システム』、オーム社、(May.2001)
[8] 三浦種敏監修:新版『聴覚と音声』、電子情報通信学会、pp.400～416(Jun.1991)