Kobe University Repository : Kernel

PDF issue: 2026-11-25

音声認識ソフト利用可能性に関する評価研究 : 発話に困難のある障害者、高齢者を対象として

李, 義昭
中林, 稔堯

(Citation)
神戸大学発達科学部研究紀要,14(2):147-156

(Issue Date)
2007-03

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81000694

(URL)
https://hdl.handle.net/20.500.14094/81000694

KOBE UNIVERSITY

神戸大学発達科学部研究紀要
第14巻第2号　2006

研究報告

音声認識ソフト利用可能性に関する評価研究
－発話に困難のある障害者、高齢者を対象として－

Evaluation research about the possibility of the speech recognition software
－ For the disabled person and elderly people with the difficulty utterance －

李　義昭*　　中林稔堯**
Li I-Chao*　　Nakabayashi Toshitaka**

要約：本研究は、すでに多くの分野で利用され始めている音声認識技術のこのままでの進展が、当技術が発話に困難のある人々（言語障害者）への新たな障壁（バリアー）となるのではないかという視点に立って行われた。具体的には、音声認識技術の確認から始まり、市販音声認識ソフトでの認識率評価を行い、その利用可能性を検討した。最期に、発話困難者について専用の音声モデルを作成することによって、音声認識技術の利用可能性を示した。音声認識ソフトの認識率評価においては、健常者に比べて言語障害者の認識率は低下すること、特に発話の困難な身体障害者は全く利用できないことが判っている。一方、利用可能性の検討では、多くの分野においてこの技術が有用であること、また、障害者の職業開発、就労に役立つことが示された。音声モデルの作成では、発話困難者の専用音声モデルと汎用の音声モデルによって、認識率の変化を対比した結果、語彙数ははずか9個であるが、専用の音声モデルにおいて、その認識率は大幅に向上した。今後の課題は、発話困難者の音声データを大量に収集することにある。このように集められた資産が開発の方向を定め、新たな障壁を防止し、音声認識技術の有用性をさらに高めるであろう。

Ⅰ　評価研究の目的

近年の音声認識技術の進展にはめざましいものがある。音声認識ソフトの開発は、「将来、キーボードは不要になるだろう」と言い、電話での問い合わせなどは、その多くが音声応答になっている。カーナビゲーションは音声で目的地を指示し、順路の説明や注意などはシステムが音声で回答するようになっている。携帯電話では特定の相手の番号を音声で検索し、通話が出来るようになった。警備やセキュリティー関係にも、音声技術を利用したものが現れている。

今後、多くのIT機器や制御機器、家電製品や介護機器などに、その技術は利用されようとしている。しかし、PCや通信機器の発達がコミニュケーションハンディキャップ或いはデジタルデバイトを生じさせたように、私たちは新たな技術進歩が、新たなバリアーを作る可能性があることを経験的に知っている。この評価研究は、音声認識技術の進歩においても起こりうるバリアーに配慮し、障害児（者）の音声認識技術利用に関する可能性を検討するものである。

Ⅱ　評価研究の方法

最新の音声認識技術の現況を文献、資料、その他によって調査・把握した。次いで、市販汎用音声認識ソフトの認識率、精度、仕様、その他を、ソフトメーカからの技術資料などの提供と連携研究の下で、障害児（者）、高齢者、健常者などの協力を得て評価、検討を行い、現況における音声認識技術および市販汎用音声認識ソフトの問題点と今後の課題について明らかにした。さらに、市販汎用音声認識ソフトの障害児（者）への利用可能性を実験・評価した。具体的には音声による、静止画・動画等の起動、制御を実験、評価し、障害者の音声認識ソフトの利用可能性を検討した。また、障害者の社会参加を促進するために、音声認識ソフトを利用した職業訓練および職域拡大の可能性について考察した。最後に、音声認識ソフトの課題である、発話に困難や訛りのある障害児（者）が音声認識技術を利用できない問題に関する解決策（利用の可能性）について、実験・評価を行い、その開発の方向性を提示した。

1　音声認識技術の概要

(1)　基本構造

音声認識技術とは、基本的には音響的処理、音素認識、言語処理の3つの段階をへて認識結果が出力される技術である。音響処理は入力された音声（図1）を認識しやすくするため、特徴を良く表わす時系列ベクトルである「特徴量ベクトル」に変換する。近年はLPCケプストラム（cepstrum）が良く使われる（LPCは線形予測符号化の略で音声スペクトル（図2・図3）の概形を推定したもの、

* 追手門学院大学経済学部講師
** 神戸大学発達科学部教授

（2006年9月1日　受付
2006年9月1日　受理）

またケプストラムは対数スペクトル（spectrum）に逆フーリエ変換したもの)。音素は日本語をローマ字であらわしたときのアルファベット1文字に相当するもので、音声を構成する最小単位である。音素認識は入力された音声に対して、あらかじめ用意してある標準の音素モデルと比較して最も「それらしい」音素を選びそれを結果とする。「それらしい」音素の選び方は、最近ではHMM（Hidden Markov Model: 隠れマルコフモデル）が最も一般的である。言語処理は音素認識で得られた結果とあらかじめ用意された辞書にある単語とを比較して、最もそれらしい単語を選び結果として出力する。

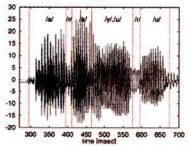


図1

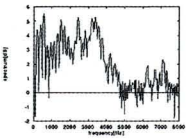


図2

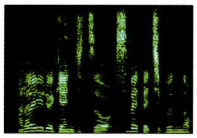


図3

(2) 代表的なシステム

1）登録型

登録型の音声認識方式は、認識させたい言葉を発声しあらかじめ登録しておく。認識時に入力された音声と登録されている音声を比較･照合し、最も音響的な特徴のそれらしい言葉が認識結果となる。この方式では話者に固有のパターンが登録されることから、特定話者の音声認識となる（図4)。この方式は1970年代に研究開発されたが、登録の不便さや認識率の問題などから限られた分野でのみ使用された。

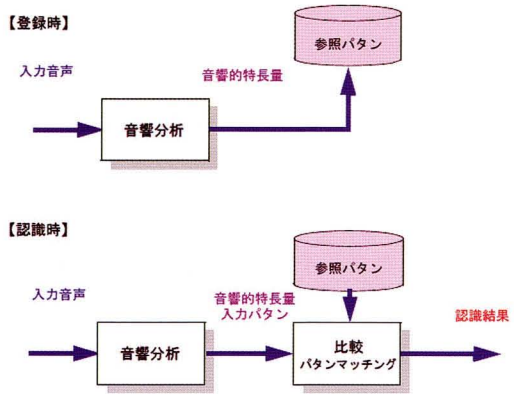


図4　登録型の音声認識

2）登録不要型

登録が不要で誰でもが使える認識方式（不特定話者音声認識）で、統計的な手法が用いられている。隠れマルコフモデル（HMM）に基づく方法が代表的である。この方式は不特定の話者 の音素モデル（音素は音声を識別する最小の単位で母音や子音など）と、音素で書かれた言葉の辞書を用いて音声認識を行うものである。辞書には、機器の操作であれば、その用語、カーナビゲーションの目的地設定では、地名や施設名を音素記号記述で書いておけば認識ができる。音素記号記述は、カナ記述から変換できるので、ユーザが認識用語を決めるようなアプリケーションでも、簡単に認識辞書を構築することができる。この音素モデルとして、音声の特徴や種々の変動を統計的に考慮・処理する隠れマルコフモデルが開発されたことで、精度の高い音声認識の実現が可能となった（図5)。この方式は、1980年代後半に開発され、現在は音声認識技術に広く利用されている。

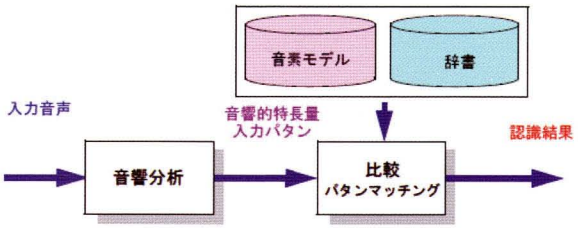


図5　不特定話者の音声認識

3）ディクテーション型

入力される文は任意で連続的に発声される。文は単語の切れ目さえなく、単に音響的に近い音素の系列やひらがな記述では、文章入力としての精度は得られないことから統計的な言語モデルが適用される（図6)。

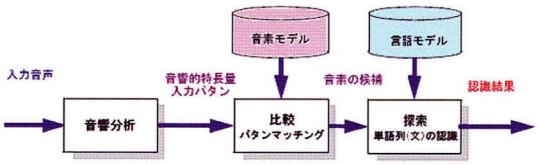


図6　ディクテーション

このモデルは、一般に言われる文法ではなく、単語の連鎖の確率が用いられる。たとえば､「東京」という単語のあとに「都」や「近郊」「大学」「駅」「横浜」などの単語や「は」「で」「へ」などの助詞が高い確率で起こる可能性があり「猫」「村」「応用」などの単語が連続する確率は低い。ディクテーションでは､膨大なテキストデータから、認識対象の語彙（通常数万）の連鎖確率を二単語連鎖、あるいは三単語連鎖で記憶しておき、この確率を用いて、認識の音素の候補からそれらしい単語の連鎖を、音素の比較で求め、その音素である確率とトータルで最も確率の高い単語列を検索し認識結果とする。通常、この言語モデルの作成には、テキストデータとして整備されている新聞記事などが用いられている。基本的に語彙として扱っていない単語や、言語モデルにない言い回し、一般的でない用語などが含まれると、認識率が極端に低下する（図7)。



図7　単語列の探索

2　登録型と不特定話者

音声認識技術は言語として識別する場合に同じ音（同じ音素）として認識する音の音響的な特徴をいかに表現し、入力された音声と

対比して識別をするかにある。同じ音であると識別している音も音響的特徴がさまざまに変化している。この変動は、次の要因が起因となる。

1) 話者（人ごとに発声器官が違う）
2) 発話スタイル、方言（人ごとに発声習慣が異なる）
3) 前後の音（発声器官は連続的に動くため、前後の音の影響を強く受ける）
4) 発話ごとの変動（発声するたびに、発声器官の動きは微妙に異なる。発話速度が変化しても、大きく特徴が変わる）
5) マイクとマイクの距離（音声を電気信号に変えるまでにも大きな変化が生じる）
6) 周囲の騒音（音声だけが入力されることはまれで、入力された信号には、音声以外の音が混入する）

不特定話者の音素モデルでは、一般にこれらの変動を確率現象として捕らえ音のモデルを作成する。音の特徴が二次元平面（実際は10～40次元程度のベクトル量で表現される）の位置で表されるとする。たとえば、種々の「あ」と「え」という音を分析収集した場合（図8）。

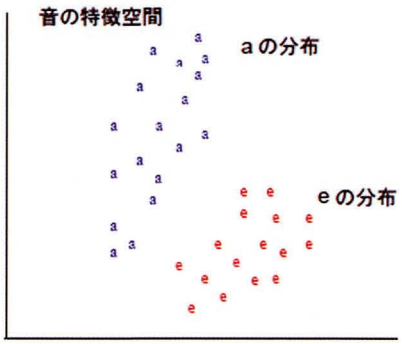


図8　音の特徴と分布

一般に隠れマルコフモデルは、この分布を確率モデルで表現し（図9）、音の特徴空間に対する分布をモデルが表現する。モデルは、ひとつの分布で表現しておくこともあるが、より精度の高いモデルでは（図10）、複数の分布で表現することもある。たとえば、男女でこの分布は大きく異なり、分布を複数で持つ方が高い精度が期待できる。また、モデルを複数持つこともある。

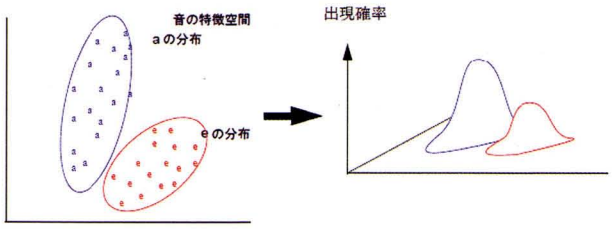


図9　音声の特徴の確率モデル

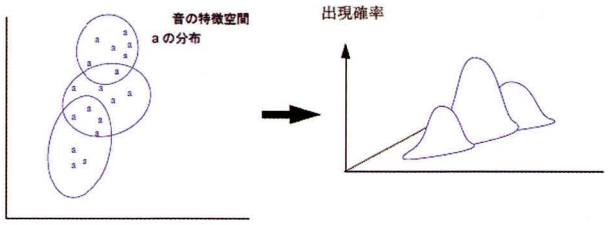


図10　複数の分布によるモデル化

個々の音素が前後の音素の影響を受けて大きく特徴が変化し、発声のたびに変動するため、その人の音声の特徴が記憶されている登録型の音声認識でも、この点を配慮しないと高い認識率は得られない。たとえば、登録時の発声が○、入力時の発声が×であれば、「a」は、登録パターンの「e」に近いことになり、誤認識が発生することになる（図11）。

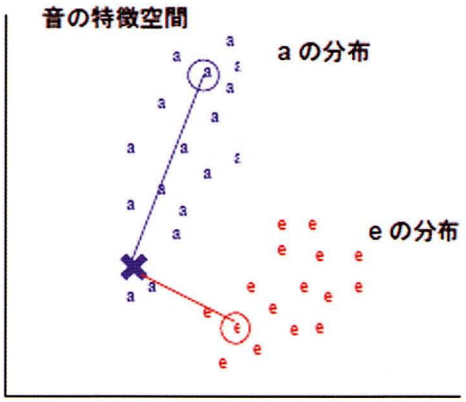


図11　登録型の音声認識における距離値判定

従って、登録型の認識では同一の話者が発声しているが、認識率は100%になる訳ではなく、不特定型の認識が平均的には、安定的な高い認識率が得られる場合がある。これは、不特定話者の音素モデルが単に、話者による変動をモデル化しているだけではなく、確率モデルの中に、話者以外の種々の要因による音の特徴の変動を組み込まれているからである。

3　話者適応

確率モデルは、一般的な発声をする大多数の人に高い認識率を提供できるが、多くの人の特徴からずれている発声、つまり確率的に低い発声の話者では極端に認識率が低下する。認識率が平均して95%であった場合、すべての人の認識率が95%程度ではなく、一般的には大多数は100%近く、一部が80%や70%の認識率ということになる。音声認識では、この問題を解決するため、高い精度が要求されるディクテーションでは、多くの人になるべく高い認識率を提供する目的で、話者適応という技術を用いる。「エンロール」「音の学習」などの機能がそれで、音素モデルを利用する人に向けてチューニングする技術である。ここでは、単に入ってきた音声で、モデルを書き換えるのではなく、統計的特徴量の学習を行い、他のモデルとの関係など、偶然に観測されたデータに強く依存せず、そのモデルの話者によらない特徴を維持することを考慮することが必要である（図9）。数回観測されたデータでモデルを書き換えると

登録型と同じく、認識率が低くなることがある。一般的に、適応化には、学習用のデータが多ければ多いほど性能が向上する。音素モデルの学習では、音素が前後の音に大きな影響を受ける性質から、モデルの作成では変動を考慮して大量（数100文章）の音声データを用いる。少数（数10文章）のデータでは、変動の特徴がカバーできず、もとのモデルの特徴を利用せずにモデルを作成すれば、認識率を下げることになる（図12）。

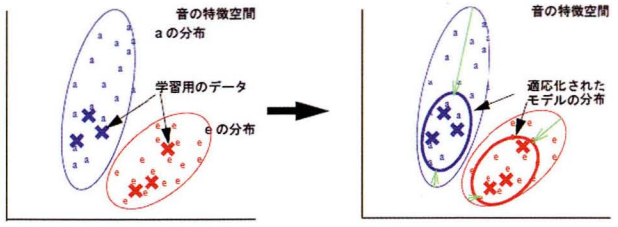


図12　話者適応の模式図

Ⅲ　市販汎用音声認識ソフトの市場における評価（注1）

音声認識ソフトは、「声で操作する」というインターフェイスに魅了される人、キーボード操作が苦手な人、大量の文章を入力しなければならない人にとっても非常に興味あるものである。近年の音声認識技術の進歩は、多くの安価な家庭・個人用の汎用音声認識ソフトを市場に送り出している。音声認識ソフトの導入には、マイクの音声レベルの調整や「エンロール」、「音の学習」が欠かせなかった。そのため、使用できるまでには時間が必要であったが、現在ではかなり短縮化され、省略できるものもある。PC（パーソナルコンピュータ）にとって言葉も音楽も雑音も単なる音の情報であり、入力された音は「音素モデル」によって音素を摘出し、その音素列から「言語モデル」によって単語・文章を推測する。そのため比較的正しい日本語で記述された新聞記事は認識精度が高く、主語／助詞が省略された、「あたしのだ」や「あたしんだ」などのくだけた会話文では認識精度は低くなっている（表1）。

表1　会話文

	簡易エンロール	フルエンロール
新聞記事	70～97%	88～99%
会話文	55～84%	56～85%

作動環境（Windows版）：CPU Pentium Ⅲ-800EBMHz、メモリ　256MB

なお、ミドルウェアの音声認識ソフトは、多くの中小メーカー（独自開発の音声認識技術よりは基本技術の購入による応用技術の販売が主である）から提供されている。また、研究・開発用のフリーソフトとして、日本では大語彙連続音声認識エンジンJuliusが提供されている。このプラットフォームは音響モデルと言語モデルの分離が可能で、それぞれのモデルを差し替えることにより、各種の実験、研究が行える。

1　市販汎用音声認識ソフトの選定方法

主な4社の市販汎用音声認識ソフトである「La La Voice2001」（株）東芝、「Via Voice Pro V8」日本アイ・ビー・エム（株）、「Smart Voice Ver.4.01」NEC、「ドランゴンスピーチVer.5」ドラゴンシステム（株）から、その認識率、使いやすさ、メーカーの協力度などの視点から、男性、女性、子供、知的障害、身体障害などのタイプ別認識率評価用音声認識ソフトと利用実験用音声認識ソフトの選定を行った。なお、追加的にCPUの性能（速さ）による認識率の対照実験を行った。

2　評価の結果

単語、ひらがな文、会話文、新聞記事の4種類の評価用テキストによって、その認識率を算定した。評価は同一のPCを用いて、基本的に「エンロール」「音の学習」などは行わず、第一次認識率（一回目の出現文字・文）に基づいて行った。なお、認識率の算出は、下記の式を用い、同音異義語は正解とした。

（出現文字－誤認識文字）÷（出現文字）×100＝
認識率（少数以下2位四捨五入）

表2　認識率対照表

市販汎用音声認識ソフト	単語	ひらがな文（絵本）	会話文（児童書）	新聞記事
La La Voice2001 ㈱東芝	64.8%	82.5%	80.9%	86.2%
Via Voice Pro V8 日本アイ・ビー・エム㈱	35.0%	92.3%	82.8%	88.5%
Smart Voice Ver.4.0i NEC	73.0%	91.7%	86.9%	91.1%
ドラゴンスピーチ Ver.5 ドラゴンシステムズ㈱	55.3%	73.3%	76.6%	80.7%

作動環境（Windows版）：CPU　Pentium Ⅲ-500、メモリ　128MB

総じて、単語および会話文の認識率が低く、市場での評価どおり新聞記事に関する認識率は各メーカーとも高くなっている。単語はメーカーによって、特にばらつきが目立ち、主語／述語のある、整理された文である絵本、新聞記事は各メーカーとも誤認識が少ない。これは、「言語モデル」の完成度によるものと考えられる。なお、後述のとおり、「エンロール」「音の学習」を長時間行い、誤認識の修正を繰り返しPCに学習させると、認識率は向上する。音声認識ソフトの使用方法は各メーカーで特徴があり、「La La Voice」は「エンロール」機能はなく、「La La Voice」および「Smart Voice」は声を男女別に登録する機能がある。また、「ドラゴンスピーチ」は他のメーカーに比べて、かなり長時間の「エンロール」機能をもっている。なお、CPUの差異による認識率の評価を新聞記事により追加的に行ったが、認識率はCPUの差異に各メーカーとも影響を受けるといえる結果となった（表3）。

表3　追加的認識率対照表

市販汎用音声認識ソフト	新聞記事
La La Voice2001　（株）東芝	90.5%
Via Voice Pro V8 日本アイ・ビー・エム（株）	95.1%
Smart Voice Ver.4.0i　NEC	93.7%
ドラゴンスピーチ Ver.5 ドラゴンシステムズ（株）	89.3%

作動環境（Windows版）：CPU　Pentium 4-2.4G、メモリ　1GMB

3　選定の結果

以上の実験結果をもとに、認識率、使いやすさ、メーカーの協力度などを踏まえ、平均的に認識率の高かった「Smart Voice」をタイプ別認識率評価用音声認識ソフトとして、また開発用のソフト（ＳＤＫ）を用いずにかなりの開発が行える機能を持つ「ドラゴンスピーチ」を利用実験用音声認識ソフトとして選定した。なお、評価研究に必要な技術的資料の提供を申し出たメーカーは「Smart Voice」「ドラゴンスピーチ」の２社である。また、「ドラゴンスピーチ」からは評価研究用として、新バージョンの「ドラゴンスピーチ６」および「ドラゴンスピーチ７」の提供を受けた。

Ⅳ　市販汎用音声認識ソフトの評価実験

1　評価の方法

タイプ別認識率評価用音声認識ソフト「Smart Voice」について、「エンロール」「音の学習」「同音異義語の修正・学習」などを行い、単語、ひらがな文、会話文、新聞記事、それぞれについて段階的な認識率の向上を実験・評価した。また、大人（A: 男性、B: 女性）子供（小学生）、知的障害（A：男性、B：女性）身体障害（男性）など、異なる発話を特徴にもつタイプ別に単語、ひらがな文、会話文、新聞記事の認識率を評価、検討した。

2　評価の結果

音声認識技術が不得意とされる、単語・会話文に関しても、「エンロール」「音の学習」を長時間行い「同音異義語の修正・学習」を繰り返し行うと、認識率は90%を超える結果となった。また、主語／述語のある整理された文である、いわゆる音声認識ソフトが得意な絵本、新聞記事は第一次の段階から、その認識率は90%を超えており、

表４　エンロール・修正・学習後の認識率

	単語	ひらがな文（絵本）	会話文（児童書）	新聞記事
第一次の認識率	30.0%	90.1%	75.3%	91.3%
第一次の同音異義語を修正した後の認識率	46.7%	89.8%	73.3%	95.4%
第一次の修正後ＰＣに学習させた後の認識率	76.7%	92.5%	91.7%	97.5%
第二次の同音意義語を修正した後の認識率	90.0%	96.2%	98.8%	100.0%

作動環境（Windows 版）：CPU　Pentium Ⅲ -850、メモリ　256MB

さらに「同音異義語の修正・学習」を繰り返し行うと、さらに認識率は向上した。特に新聞記事は同音異義語を正解とした場合、その認識率は100%に達した（表4）。

表５　対象者

事例	特記すべき事項
 子供（小学生）	声変わりが終わっていない小学生であるため、かなり高音の発話を行う。当初、声の登録を男性で行ったため、認識率はかなり低いものであった。しかし、声の登録を女性とした場合、その認識率は標準以上となった。音声認識ソフトが持つ、男性として予定される「音素モデル」に、この小学生の声は、多くが予定されていなかったものと考えられる。
 知的障害（A：男性）	軽度の知的障害を持つ男性であるが、発話はかなり正確である。日常の会話では、聞き取りに困難は無く、コミュニケーションに問題は無い。しかし、今回の実験・評価結果で、ひらがな文、会話文、新聞が他の被験者に比べて、認識率がかなり低下している。読点、句点の切り方、文章の構成力などに困難があるため、主語・述語のある整理された文を必要とする「言語モデル」に起因するものと考えられる。
 知的障害（B：女性）	軽度の知的障害を持つ女性で、発話にかなりの困難がある。日常の会話では、聞きとりにくいことがかなりあり、コミュニケーションは注意深く聞き、繰り返し確認するなど、日ごろ交流があれば問題は無い。しかし、あまり交流が無い、初対面である場合、かなり注意深く聞かないと意思の疎通が難しい。極端な認識率の低下は「音素モデル」および「言語モデル」の双方に起因するものと考えられる。
 身体障害（男性）	知的障害はないが、身体障害により発話にかなりの困難がある。日常の会話では、注意深く聞けば、断片的に聞きとることが出来るが、初対面である場合、ほとんど聞きとることは出来ない。日ごろから、親密な交流のある、母親や施設の職員はコミュニケーションに困難は無い。音声認識では「エンロール」段階から問題があり、「音素モデル」に予定された音素が無いと考えられ、ＰＣにとって彼の発話は未知の言語となっている。

異なる発話の特徴を持つと思われる、大人（A: 男性、B: 女性）子供（小学生）、知的障害（A：男性、B：女性）身体障害（男性）などの協力者に対して、実験・評価を行った結果、発話に困難・障害がある、訛りがあるなど特徴的な発話をする対象者に、極端な認識率の低下が起こった（表6）。これは、「音素モデル」に、その人たちの音素が登録されていないことに起因するものと考えられる。ＰＣや音声認識ソフトにとって、彼らの発話は「未知の音」や「理解できない外国語」であることを意味するものと思われる。特記すべき事項として表５の事例がある。

表6　話者タイプ別認識率

	単語	ひらがな文（絵本）	会話文（児童書）	新聞記事
大　　人（男性）	73.1%	85.8%	87.2%	89.5%
大　　人（女性）	64.5%	90.3%	76.4%	91.2%
子　　供（小学生）	80.7%	94.1%	86.8%	90.5%
知的障害A（男性）	85.2%	74.7%	62.2%	76.6%
知的障害B（女性）	40.0%	23.5%	26.1%	26.9%
身体障害（男性）	0.0%	実験不能	実験不能	実験不能

作動環境（Windows版）：CPU　Pentium Ⅲ-850、メモリ　256MB

3　今後の課題

音声認識ソフトは多くの可能性を秘めている。しかし、市販汎用音声認識ソフトの評価・実験で判るように、発話に障害や困難があったり、発話に訛りなどの特徴があったりすると、認識率は極端に低下し、時には認識がまったく不可能な状態となる。これは上記「音素モデル」が健常者の男女を標準に構築されているためであり、PCにとって障害のある人の発話は全く判別できない言語となる。また、主語/述語のある整理された文である、新聞記事などから構築された「言語モデル」の特徴から、読点、句点の切り方、文章の構成力などに困難のある知的障害者、幼児などの認識率は低下する。現在、開発の主流となっている音素型音声認識技術は、このタイプの障害のある人々には利用できないか、或いはその利用に困難が生じている。このままの形で音声認識技術が多くの機器に組み込まれていけば、障害のある人たちの前に新たなバリアーが出現し、近い将来、これらのバリアーを解消するために新たな社会的費用を負担する事態が生じることになる。

Ⅴ　市販汎用音声認識ソフトの利用実験

1　実験の方法

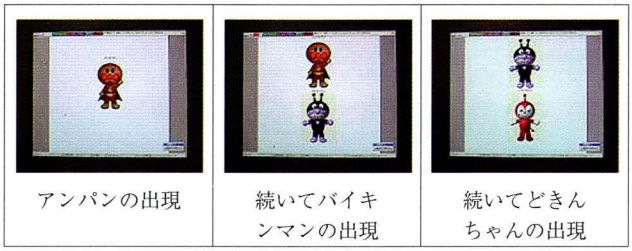
利用実験用音声認識ソフト「ドラゴンスピーチ」のディクテーション機能と各種の開発機能を用いて、障害児や幼児の理解・交流・参加などを促すツールの開発、実験を行った。具体的には、絵や画面の切り替え、動画の起動・制御などを利用して学習理解深めたり、障害児と健常児が共に遊んだり、交流することによって、共生の可能な環境、道具の開発実験を行った。さらに、中国語で話すと、日本語が出現し、日本語で話すと英語が出現する「おもちゃ翻訳機」に挑戦した。また、障害者や高齢者の職域拡大のための、音声認識ソフトの利用可能性の検討を行った。具体的には、障害や高齢でキーボード入力が苦手な人でも、データ入力（アンケート調査などの結果を数字列として入力し、それをもとに統計的な分析がなされる）やテープ起こし（会議や聞き取り調査のテープ音声を聞きとってテキスト入力して納品する）などを通じて在宅の仕事の開拓に関する実験を行い、その可能性を検討した。さらに、聴覚障害者の支援に関する音声認識ソフトの利用可能性の評価実験を行った。具体的には、大学で行われている聴覚障害学生に要約筆記者（ノートテイカー）を付ける支援の補助として、講義内容をディスプレイ上にテキスト表示し、ノートテイカーの補助的役割を音声認識ソフトの利用によって担わせることが可能であるかを検討した。

2　実験の結果

(1)　絵や動画の制御

市販汎用動画作成ソフトと音声認識ソフトとの組み合わせによって、絵や画面の切り替え、動画の起動・制御などの開発実験を行った。たとえば、「アンパンマン」という発話には「アンパンマン」の映像、「バイキンマン」という発話では「バイキンマン」の映像が出現する。また、「アンパンマン」と叫べば「アンパンマン」が飛んできたり、「キック」と叫べば「アンパンマン」がキックしたりするなどの音声認識ソフトの利用可能性を検討した。

写真1　画面の切り替え

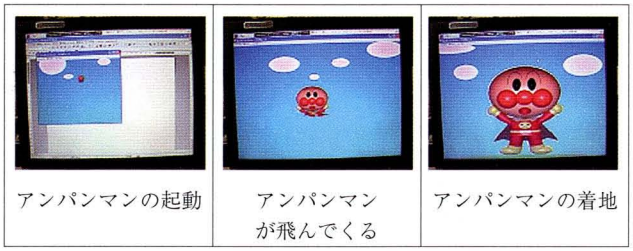


アンパンの出現　｜　続いてバイキンマンの出現　｜　続いてどきんちゃんの出現

作動環境（Windows版）：CPU　Pentium Ⅲ-850、メモリ　256MB

画面の切り替え実験では、音声による画面の切り替え機能を利用した、「アンパンマン」の発話に対して「アンパンマン」の絵、「バイキンマン」の発話に対して「バイキンマン」の絵、「どきんちゃん」の発話に対しては「どきんちゃん」の絵を対応させた。写真1は発話に対応して、絵が順次出現しているところ。

写真2　動画の起動

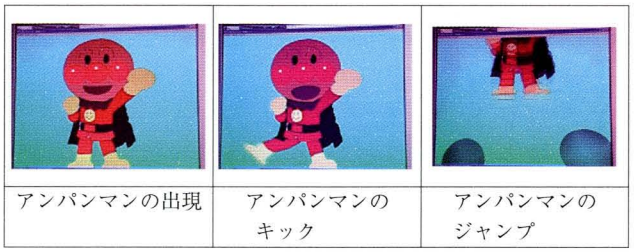


アンパンマンの起動　｜　アンパンマンが飛んでくる　｜　アンパンマンの着地

作動環境（Windows版）：CPU　Pentium Ⅲ-850、メモリ　256MB

動画の起動では、音声によるアプリケーションの起動機能を利用、「アンパンマン」の発話に対して、あらかじめ用意された動画を起動させ実行する実験を行った。具体的には一分ほどの「アンパンマン」がテーマ曲に乗って、飛来・着地する動画を用意、それを起動・実行させた。写真2は、「アンパンマン」の発話から起動・着地までの様子。

写真3　動画の制御

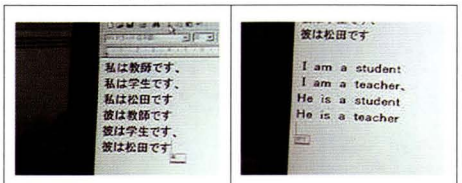


アンパンマンの出現　｜　アンパンマンのキック　｜　アンパンマンのジャンプ

作動環境（Windows版）：CPU　Pentium 4 2.4G、メモリ　1GB

動画の制御では、音声認識ソフトによるＰＣの命令・制御機能と動画作成ソフトのマウスによる画像制御機能を利用して、音声発話による動画の制御実験を行った。具体的には、「アンパンマン」「キック」「パンチ」「危ない」などの発話による命令語を用意し、動画にはその発話された命令語に対応した制御反応を用意し、それを起動・実行させた。写真３は、「アンパンマン」の起動・出現から「キック」の発話にたいする「キック」の動作、「危ない」の発話に対応した「ジャンプして危険を回避する」という動作をさせているところ。

写真４　おもちゃの翻訳機



中国語から日本文　　日本語から英文

おもちゃの翻訳機では、発話に対するテキスト（文字）の出現機能を利用して中国語の発話に対して、日本語の出現、日本語の発話に対して英語の文字を出現する簡単な翻訳機のようなものを実験した。具体的には表６のような対応表に基づき、「我是」の発話に対して「私は」、「他是」の発話に対して「彼は」、「老師」の発話に対して「教師です」などを対応させ、この組み合わせでテキスト（日本文・英文）を出現させることを行った。写真４は、ＰＣが色々な組み合わせによる発話を、テキスト（文）として出現しているところ。

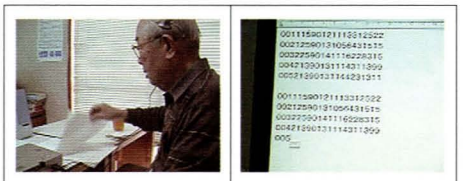
表６　対応表

中国語	日本語	英語
我是（ウォーシー）	私は	Ｉ　ａｍ
他是（ターシー）	彼は	Ｈｅ　ｉｓ
老師（ラォシー）	教師です	ａ　ｔｅａｃｈｅｒ
学生（シェーセン）	学生です	ａ　ｓｔｕｄｅｎｔ
松田（ソンテン）	松田です	

(2)　障害者及び高齢者の就労の開拓

市販汎用音声認識ソフトのディクテーション機能を用いて、障害者や高齢者のＰＣの利用に関する実験を行った。特に障害があるためにキーボードが利用できない人、高齢や知的障害のためにＰＣの利用やシステムの理解、キーボードの入力操作や構造的な理解が苦手な人に、音声入力による在宅就労の可能性を研究した。具体的には、データ入力（大学や研究所などの研究機関がアンケート調査を行った結果を、データ入力会社などに依頼して、単純な数字列データとして納品させ、それをもとに統計的な分析を行う）やテープ起こし（企業や大学などでの会議、アンケートなどの聞き取り調査の結果をテープ音声に記録しておいたものをテープ起こし会社などに依頼して、その記録した内容をテキストとして入力し、文字や文章の形で納品させる）などの在宅の仕事が出来るか、その可能性について検討した。

写真５　データ入力



音声での読み上げ　　数字列のテキスト出力

表７　認識結果

対象者	大人（男性）	高齢者（男性）	身体障害（女性）
認識率	100%	99%	99%

作動環境（Windows 版）：CPU　Pentium Ⅲ-850、メモリ　256MB

データ入力では、簡単なアンケート調査の結果（５人分）を用意し、大人（男性）、高齢者（男性）、障害者（女性）に対して、音声入力におる認識結果を評価・実験した。具体的には、アンケート調査において記入された文字番号の結果を数字で発話し、ＰＣ画面上に、その結果をテキスト出力させた。障害者（女性）は緊張すると、発話に少し困難が生じるが、認識結果は実用可能なものである。写真５は、音声による読み上げとその結果を出力しているとおころ。また表７は認識結果である。

写真６　テープ起こし



テープ内容の読み上げ　　テキスト文字・文での出力

表８　認識結果

対象者	大人（男性）	高齢者（男性）	身体障害（女性）
認識率	93.8%	97.8%	98.6%

作動環境（Windows 版）：CPU　Pentium Ⅲ-850、メモリ　256MB

テープ起こしでは、会議内容のテープ記録を用意し、それを短く聞き取りながら、内容を発話し、ＰＣ画面にテキスト文字・文章として出力させた。なお、高齢者（男性）に関しては、高齢のため軽度の難聴となっており、テープの内容が聞きとりにくいため、テープ記録の内容をテキスト文として出力された文章を読み上げてもらった。また、障害者（女性）に関しては、テープデッキの操作に困難があるため、高齢者と同様、文章を読み上げてもらった。写真６は、作業所（かなり話し声など雑音のある環境）でのテープ内容の読み上げ光景とＰＣ画面へのテキスト文字・文章の出力の様子。表８は認識結果である。

(3)　聴覚障害学生支援

市販汎用音声認識ソフトのディクテーション機能を用いて、聴覚障害学生の大学講義における授業理解支援の可能性について実験・評価を行った。具体的には、現在、大学で行われている聴覚障害学生にノートテイカーを付ける学習支援の補助として、講師が話す講

義内容を全90分間に渡ってＰＣに音声入力した。それと同時に、その出力結果をディスプレイ上（誤認識も含めて）にテキスト文字・文として表示し、ノートテイカーの補助的役割を担わせることが可能であるかを検討した。認識率の算出は、講師が話し言葉として発話した「会話文」と講師が講義レジュメを読み上げた「文章」部分に分けて算出した。なお、誤認識のうち同音異義語として出力されたものは正解とした。これは、「エンロール」「音の学習」を長時間行い「同音異義語の修正・学習」を繰り返えすと、認識率が向上する音声認識ソフトの性質を考慮したためである。認識結果は表9の

写真7　大学講義における利用実験


教授の講義風景


聴覚障害学生の様子


ノートテイカーに確認

表9　認識結果

	会話	文章
認識率	92.3%	100%

作動環境（Windows版）：CPU　Pentium Ⅲ-850、メモリ　256MB

表10　実験結果のコメント

	コ　メ　ン　ト
聴覚障害学生	80%ほどは変換されているようで支援にはなります。ディスプレイで見る場合かなり目が疲れるので、スクリーンに映し出すのが良いのではないか。
ノートテイカー	80%程度の認識では文意がとりにくい。文章を読み上げている場合は本文と対比できるが、会話文は一旦文意がずれると意味が取れなくなる。今回の実験からでは、ディスプレイがあるために作業が増える。また、ノートテイカーは要約筆記者であって、20%程度の文字表現でいかに全体を表現するかにその妙がある。

ようであり、対象者からは表10に示したコメントがある。認識結果と対象者のコメントにおける認識率の誤差は、同音異義語を正解としたことによるものと考えられる。また、写真7は、大学の講義で評価・実験を行っているところ。

3　今後の課題

市販汎用音声認識ソフトは、かなりの利用可能性を示している。絵や動画の制御では、知的障害児（者）などが興味を持ち・理解しやすいコテンツが、いかに用意されるか。また、誤認識の少ない命令語や、操作性の良いものが開発されるなど、多くの工夫が必要とされる。障害者・高齢者の就労開拓では、音声入力以外のＰＣ操作環境の改善が必要である。たとえば、高齢者の補聴器、アンケート調査表をめくる、テープデッキを操作する、身体的障害を補う器具の用意など、作業環境の開発・改善が課題となっている。大学での聴覚障害者支援のノートテイカー付与には、もともと大学生とその専門分野に関して同等の知識、あるいは、それ以上の知識を有する人材の確保が困難であるとうい問題がある。音声認識ソフトの利用においても、事前に講師がＰＣに対する「エンロール」「声の学習」などの時間を十分に確保する必要がある。また、大学講義でよく用いられる多くの専門用語を講師の膨大な書籍・論文などを、前もってＰＣに学習させておき専門用語の認識精度を向上させるなどの準備の課題がある。このような課題以上に、我々にとって最も大きな問題は、発話に障害があったり訛りがあったりする人々が、既に社会資源となっている音声認識技術を利用できない状態にあるということである。

Ⅵ　特定話者（特定の障害者）による音声認識技術の利用可能性に関する実験

1　実験の概要

発話に障害・訛りなどがある障害ある人を含めたすべての人が利用できる、音声認識技術の可能性を評価、検討した。具体的には、発話に困難のある特定の話者（障害児1名：男性）の音素モデルを作成し、企業向け汎用製品に搭載、その製品の標準音素モデル（汎用音響モデル）と新たに作成した特定話者の音素モデルによる認識率の比較評価を行った（図13）。最初に、発話に障害のある障害者1名から9単語（各5回）を音声データとして収録し、4回分を音素モデル作成用、1回分を評価テスト用として認識率の比較評価を行った。その結果、標準音素モデルでは22.23%の認識率を示し、特定話者用音素モデルでは100.00%の認識率を得た。なお、この評価実験は、インフォコム株式会社の協力を得て行われた。

2　実験の方法

(1)　音声データ採取作業（注2）

特定話者音響モデル作成のため、障害者1名の音声を録音採取した。発話内容は50音と認識対象語彙である、おかあさん、きんきゅう、たすけて、いたい、うれしい、あんぱんまん、どきんちゃん、ばたこさん、ばいきんまんの9単語である。なお、50音は、あ行、か行などそれぞれの行で一息に発話収録した。

(2)　音声データ編集作業（注3）

収録された特定話者の音声を、トレーニングおよび認識実験に用いるために、5回分の発話音声を各1ファイルに編集、個々の認識対象語彙である、おかあさん、きんきゅう、たすけて、いたい、う

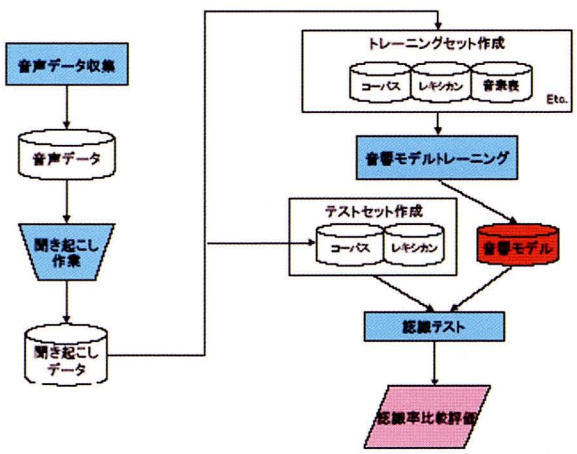


図13　作業工程

れしい、あんぱんまん、どきんちゃん、ばたこさん、ばいきんまんの発話ごとに、それぞれ5ファイルを作成した。

(3) 音響モデル作成作業（注4）

収録・編集した音声データをもとに、聞きおこし作業を行い、音響モデルを作成した。作成方法に関しては、インフォコム株式会社の関連会社である「ScanSoft社」の企業秘密であるため、開示されていない。

(4) 認識プログラム作成作業（注5）

プログラムは、音声ファイルを認識エンジンにて認識させ、認識結果を表示するものである。具体的には、音声認識リソースおよび録音音声ファイルを選択して、認識させる実行プログラムである。なお、プログラムのソースは「SpeechPearl」のAPIが記述されているためインフォコム株式会社の判断で開示されていない。

3 実験の結果（注6）

(1) トレーニング・テストセットの作成

特定話者（障害者1名）の発話は、清音すべてが規則性をもって、濁音になる特徴を有している。たとえば、「アンパンマン」→「あんばんまん」、「バタコサン」→「ばだござん」のごとくである。したがって、作成される特定話者用音素表は、清音のトレーニング、清音・濁音の対照認識などが出来ない特徴を持っている。また、レキシカン（認識させたい語句とその発音記号）は「アンパンマン」→「aNbaNmaN」、「バタコサン」→「badagOzaN」と表記される。コーパス（聞き起こし語句と音声ファイルの言語データベース）の作成では、収録・編集された各単語5ファイルのうち、4ファイルをトレーニング用コーパスに用い、残り1ファイルをテスト用コーパスに用いた。

(2) 認識テスト

認識評価テストは、テスト用に用意された残りのファイル9単語に対して、汎用製品に搭載された標準の音素モデル（汎用音響モデル）と今回作成した特定話者の音素モデルを差し替える方法で行った。認識の結果、汎用音響モデルでは認識された語は「いたい」と「きんきゅう」の2語であり、「アンパンマン」「うれしい」「おかあさん」「たすけて」「どきんちゃん」「バイキンマン」「バタコサン」の7語は認識されなかった（表11）。

表11 汎用音響モデル評価

発声語	誤認識語
アンパンマン	代用「きんきゅう」
うれしい	削　除
おかあさん	代用「きんきゅう」
たすけて	代用「うれしい」
どきんちゃん	代用「うれしい」
バイキンマン	代用「うれしい」
バタコサン	代用「うれしい」

これに対して、今回作成した特定話者用音響モデルは、9単語すべてを認識する結果となった（表12）。

表12 認識率比較評価

	特定話者用音響モデル	汎用音響モデル
ワードベース誤認識率（WER）	0%	77.78%
代用語数（sub）	0%	66.7%
挿入語数（ins）	0%	0%
削除語数（del）	0%	11.1%
正解語数	9語	2語

$$\text{ワードベース誤認識率（WER）} = \frac{\text{代用語数（sub）} + \text{挿入語数（ins）} + \text{削除語数（del）}}{\text{レキシカンの登録語彙数}}$$

(3) テスト結果

特定話者用音響モデルと汎用音響モデルでは、認識に大きな差が出ている。特定話者（障害者）の音声にはかなりの特徴があり、健常者用に構築された汎用の音響モデルでは対応できない。PCは、その発話を全く新しい言語と捕らえていると考えられる。対して、今回作成された、特定話者用音響モデル（障害者個人用音響モデル）は、用意された9語彙すべてを認識しており、発話に困難や訛りがある人が音声認識技術を利用できる可能性を示している。

4 今後の課題

音響モデルを特定の障害者専用にチューニングし、専用の音響モデルを用意することによって、発話に困難のある特定話者（特定の障害者1名）の9語彙に関して、そのすべてを認識させることに成功した。健常者用の語彙は、すでに10万語、100万語とうい膨大な量に達している。第一の課題は、特定話者用の語彙数を、健常者なみの実用可能な量に増やすことである。また、第二は、複数の障害者に、これらの評価・実験を試みることである。このような評価・実験の集積の結果から、次の二つの開発の方向性が示唆される。

①もし、障害のタイプ別に発話の特徴などが、統計的に一定の傾向を持つようであれば、現在進んでいる、幼児用・高齢者用音響モデルのように、障害A・障害B用の音響モデルを開発して行くという方向性が必要である。

②また、それぞれの障害特性で全く異なった音響モデルが必要だということならば、Aさん用、Bさん用のように個人専用の音響モデルを、いかに簡易・低価で構築出来るかのという開発の方向性が必要になってくる。

Ⅶ 結語

音声認識技術は既に多くの分野で利用されているが、これからもますます発展し、その技術は、あらゆる機器や経済社会システムに応用・組み込まれることになると思われる。

今回の特定話者用「音素モデル」の評価研究結果は、わずかな語彙数ながら障害者用「音素モデル」構築の可能性を示したものといえる。これは、カーナビゲーションなど汎用機器の操作や任意に連

続的に発声される音声を文章化する音声入力ワープロなどIT機器、インターネット利用へのアクセスの可能性を示し、障害のある人々の経済社会への参加の促進に貢献するものと考えられる。
近い未来において、音声認識技術はすべての社会システムや家庭における電子制御機器に応用されていると考えられる。そのとき、障害のある人を含めた、すべての人々は、それぞれの個人用の「音素モデル」を持ち、汎用の電子制御機器にキーを差し込むように、彼/彼女だけの制御に利用するようになっているものと思われる。

注：

1) 市販汎用音声認識ソフトの市場における評価
（ASCII24 http://ascii24.com/ より抜粋）
2) 「音声データ採取作業報告書」インフォコム株式会社　2003年
3) 「音声データ編集作業報告書」インフォコム株式会社　2003年
4) 「音響モデル作成作業報告書」インフォコム株式会社　2003年
5) 「認識プログラム作成作業報告書」インフォコム株式会社　2003年
6) 「神戸大学用音響モデル評価報告書」インフォコム株式会社　2003年

参考文献：

1) 石川泰「障害者の方向けの音声認識について」 三菱電機株式会社情報技術総合研究所　2003年
2) 情報処理学会「音声認識システム」株式会社オーム　平成14年
3) 古井貞熙「音響・音声工学」近代科学社　1992年
4) 中川聖一「確率モデルによる音声認識」コロナ社　1988年
5) 今野清水「知的障害者の言語とコミュニケーション」学苑社　1999年
6) Rena B. Lewis : "Assistive Technology and Learning Disabilities : Today's Realities and Tomorrow's Promises", Journal of Learning Disabilities, Vol.31, No.1, pp.16-26 (1998).
7) Keith Wetzel : "Speech - Recognizing Computers : A Written-Communication Tool for Students with Learning Disabilities ?" ,Journal of Learning Disabilities, Vol.29, No.4 pp.371-380 (1996).