解　説

ハードウェアにおける人間的要素

音　声　入　出　力　機　器†

千　葉　成　美††

1. はじめに

オフィスオートメーション（OA）システムに代表される最近の計算機システムの大衆化と共に，人間との接点，すなわちヒューマンインタフェースが重要視されるようになってきた．すなわち，人間にとって，より覚え易く，使い易いインタフェースが求められているわけである．このような望ましいヒューマンインタフェースとしての資格を有しているものの一つに，音声入出力がある．

音声は，相手とのコミュニケーションのために人間が最初に覚える手段であり，すでに2，3才の幼児でも，基本的な話し方言葉の運用能力を身につけている．また，文字文化を持たないような未開社会でも，音声言語だけは例外なく存在している．このように，音声は人間同志の間での最も自然なコミュニケーションの手段である．これをそのまま計算機のインタフェースとして用いることができれば，人間にとって好都合なことであろう．

このような音声入出力を実現するのが音声認識及び合成技術である．これらは，人間固有の無意識のうちに実現される機能であり，人間におけるその処理アルゴリズムはブラックボックスの中にある．

音声認識合成技術の研究は，30年以上の歴史を有するが，現在，あるレベルまでの実用化に成功しており，いろいろな分野で音声入出力装置が実際に用いられている．しかし，現在の技術段階は，単純なアルゴリズムで実現可能な範囲に限られている．本格的で広範な音声入出力の実用化は，今後の技術開発の成果にまたなければならない．

2. 音声によるヒューマンインタフェースの特徴

音声入出力は，従来からのキーボードやCRTなどの入出力方式に比較して独自の特徴を有している．これらを入力，出力別に調べてみよう．

音声入力の長所としては，理想的な音声認識が実現されたとした場合，次のようなものをあげることができる．

① 人間にとって最も自然な入力手段である．
② 特別の訓練なしに誰でも簡単に使用できる．
③ 一般に入力速度が大きい．
④ 即時性があり，会話的な入力に適する．
⑤ 手や視覚が占有されない．
⑥ 普通の電話を入力に利用できる．
⑦ 話者認識技術による本人の確認ができる．

ここで，現状の音声認識技術のレベルでは⑤，⑥以外はある程度割引いて考える必要がある．また，語彙数や発声方法に対する制限があり，認識精度にも限界がある．

一方，音声入力には，次にあげるような本質的な短所がある．

① 周囲雑音に影響され易い．
② 逆に，発声が周囲に対して雑音源となる．
③ 機密保持性が悪い．
④ 入力の中断ができない．

したがって，音声入力を実際に使用する場合には，これらの長所，短所をよく検討した上でシステム設計を行う必要がある．

音声出力の長所としては，次のようなものをあげることができる．

① 人間にとって親しみ易い出力手段である．
② 人間の注意を引きやすく，間をおいた情報出力に適する．
③ 視覚情報出力と組合せた場合の効果が大きい．

† Speech Input and Output Equipments by Seibi CHIBA (C & C Systems Research Laboratories, NEC Corporation).

†† 日本電気(株) C & C システム研究所

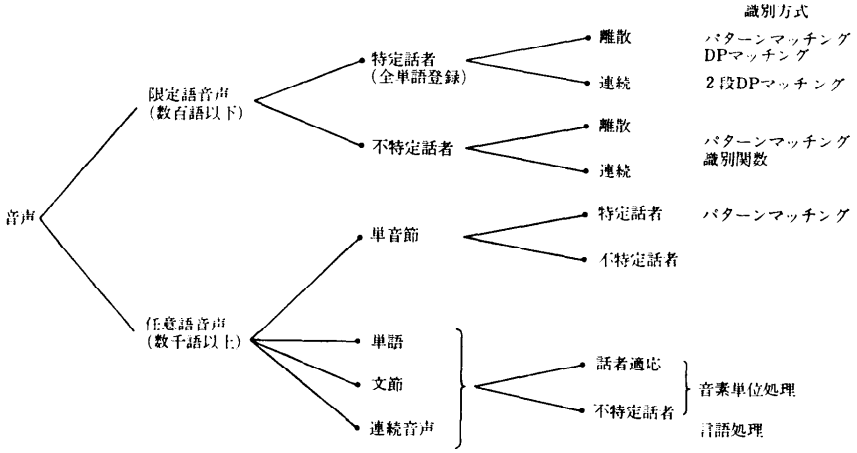

図-1　音声認識技術の分類

④利用者の位置，姿勢等に関する自由度が大きい．

⑤電話を出力に利用できる．

これに対して，音声出力にも次のような短所があり，実際に使用する場合に注意する必要がある．

①スピーカから出力する場合，周囲に対して雑音源となる．

②1次元的出力であり，前後の情報の参照が難しい．

③記録を残すのが容易ではない．

④注意を集中していないと聞きもらすことがある．

3. 音声認識合成技術の現状

3.1 音声認識技術[9)]

音声認識技術の究極の目標は，人間と同レベルの認識能力を実現することであるが，現状のレベルはこの目標に対してはるか手前にある．そこで，いくつかの制限条件を導入することによって必要な認識率を確保し，いろいろな中間段階のものが順次実用化されてきている．

このような制限条件の主なものとして，次のようなものがあげられる．

(1) 語　彙

a 限定語音声（数百語以下）

b 任意語音声（数千語以上）

(2) 話　者

a 特定話者（標準パターンとして登録）

b 不特定話者（登録不要）

(3) 発　声

a 離散発声（認識単位の間に休止をおく）

b 連続発声（休止不要）

ここで，当然aの方がbよりも制限がきびしいため技術的に容易であり，実用化が進んでいる．

図-1に，このような見方により，音声認識技術の分類を試みた結果を示す．

限定語音声認識では，単語単位の標準パターンとのマッチングを行う認識方式により，多くの場合，現実的なハードウェア規模で高い認識率が得られるため，実用化が進んでいる．

ここで，最も技術的に容易な特定話者離散単語認識システムの代表的な構成を**図-2**に示す．認識処理は，特徴抽出処理と識別処理とに大別される．

特徴抽出処理では，音声信号のスペクトル分析と，必要に応じてその分析結果に対するデータ圧縮が行われ，また，単語の始端，終端を決定する音声検出が行われる．

識別処理では，特徴抽出結果と，前もって登録された各単語の標準パターンとの間でパターンマッチングを行い，単語間距離を計算し，その最小値を検出して

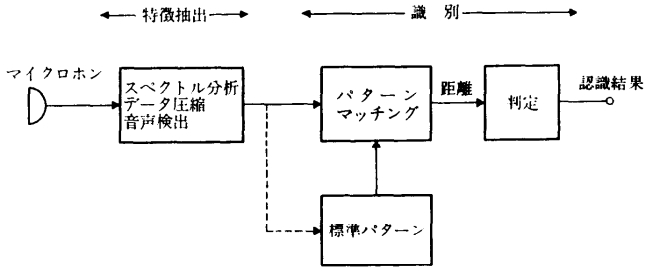

図-2　特定話者離散単語認識システムの代表的な構成

対応する単語の種類を認識結果とする．この場合，もし最小値がある閾値以下ならば，入力はリジェクトされる．

パターンマッチングにおいて問題となるのは，発声速度の変動による音声パターンの時間軸上の伸縮である．従来は近似的な正規化によっていたため，マッチングの精度が上がらなかったが，DPマッチング法[1),2)]の出現により，最適な時間正規化のもとでのマッチングが可能となった．

不特定話者を対象とする場合には，個人差による音声パターンのばらつきを吸収するために，各単語につき複数個の標準パターンを使用することが行われている．この場合には，多数話者の学習サンプルの分布からその代表点として標準パターンを選択するために，クラスタリングの手法が用いられる場合が多い．また，各単語の学習サンプルの分布の境界を識別関数として設定する方式も実用化されている[3)]．

連続単語認識に関しては，特定話者用として，DPマッチング法を発展させた2段 DP マッチング法[4)]が実用化されている．最近では，さらにそれを一般化し，オートマトンモデルによるシンタックス制御機能を導入した方式[5)]も開発されている．

任意語音声認識では，単音節を単位として，発声，認識を行う方式が，特定話者用として実用化されている．この場合には，すべての音素間の弁別が要求されるため，単語認識の場合よりも高精度の音声分析が行われる．識別処理としては，子音部と母音部に分けて，それぞれパターンマッチングを行う場合が多い．単音節はほぼかな文字に対応しており，その組合わせによって任意の日本語を表わすことができる．

単語，文節，あるいは文単位で発声された任意語音声を認識するためには，連続的な音素系列を分離して正しく認識する必要がある．これは非常に難しい問題であり，これまで種々の方式が試みられてきたが，十分な成果は得られていない．また，近接した音素相互間が干渉する調音結合と呼ばれる現象があるため，音響信号レベルの処理だけで，このような音素の分離と認識を完全に行うことは不可能と考えられている．そこで，単語辞書やシンタックス，セマンティックスなどの言語情報を援用し，全体としての文脈の中で音声をとらえることが必要となる．現状では，音響レベル処理，言語レベル処理のいずれもまだ基礎研究の段階にあり，今後の研究開発の進展にまつところが大きい．

3.2 音声合成技術[9)]

音声合成技術の目標は，自然音声と同程度の明瞭度と自然性，及び場合によってはいろいろな個人性を持った任意の合成音声を作り出すことである．しかし，現状ではこの目標はまだ達成されていないので，音声認識の場合と同様に，使用目的に応じていくつかの中間段階のものが実用化されている．この場合の分類の基準の主なものは，次の2つである．

（1） 語　彙

a 限定語音声⟶編集合成

b 任意語発声⟶規則合成

（2） 音声生成方式

a 波形符号化

b パラメータ符号化

語彙が限定されている場合は，単語や文節を単位としてアナウンサが発声した自然音声を，波形，またはパラメータの形で記憶しておき，それを適当に編集しながら読み出すことによって，出力音声を作ることができる．これに対して，任意語音声を合成するためには，音素（即ち子音C又は母音V）や音節（CV，VC）レベルの細かい基本単位を元にして，種々の合成規則を適用しながら，滑らかな音声を作り上げていく必要がある．このため，規則合成方式と呼ばれている．

音声波形の生成方式として，波形符号化は，自然音声波形をそのまま PCM（パルス符号変調）などの各種符号化方式で符号化して記憶しておき，必要に応じて再生して使用する．波形の情報量は1秒当り数十キロビットあるため，メモリの負担が大きいが，音声品質は優れている．これに対して，パラメータ符号化は，音声を音源パラメータと，PARCOR，ホルマントなどのスペクトルパラメータに分解して符号化する．このため，波形の場合の 1/10 以下に情報圧縮を行うこ

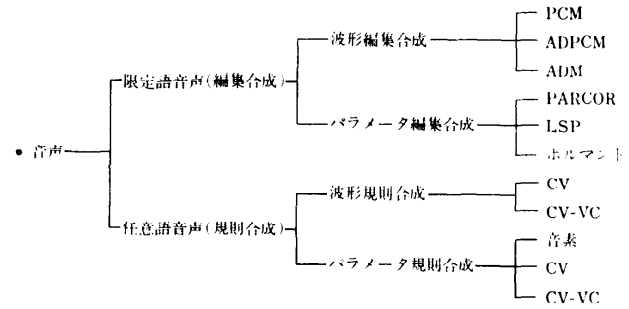

図-3 音声合成技術の分類

とができるので，コスト面で有利となる他，規則合成との適合性も良い．しかし，音声をパラメータに分解する過程で細部の情報が失われるため，一般に音声品質は低下する．

図-3 にこのような見地から試みた音声合成技術の分類を示す．今後の主要なテーマは，パラメータ規則合成の高品質化であり，現在，多面的な研究が行われている．

4. 音声入出力機器における人間工学的要素

4.1 音声入力機器

将来，音声認識技術のレベルが人間の能力なみに向上した時点では，音声入力に関する人間工学的な問題はほとんどなくなると考えられる．しかし，現在実用化されている音声認識装置では，性能の限界や使用法の制限などのため，信号入力系，レスポンス系，認識性能，誤り修正方式，発声単位，登録方式などに関して，ヒューマンインタフェースの観点から検討する必要がある．

（1） 信号入力系：

音響信号としての音声を認識処理のために電気信号に変換する部分である．基本的な電気的特性として，周波数特性（帯域），雑音特性などが問題となる他，マイクロホンのセッティング方式が人間工学の面から重要である．

周囲雑音に対して，信号対雑音比を向上させるため，従来，ノイズキャンセル機能を持つ接話形マイクが広く使用されてきた．接話形マイクでは，口との距離を一定に保つ必要があるため，ヘッドセット形として使われるのが普通である．この形のマイクは，85 dB (A) 程度までの騒音下でも音声認識が可能となること，及び頭に固定するために身体の移動が容易になることにより，仕分け装置制御のような産業用音声入力には非常に適合性がよい．しかし，今後一般化するオフィス環境での音声入力のためには，ヘッドセット形は扱いにくいため，これに代る形式のマイクロホンに対して強いニーズがある．この場合，電気信号のレベルで何らかのノイズキャンセル機能を実現する必要が生じる．

（2） レスポンス系：

入力された音声に対するシステムからのレスポンスであり，レスポンスタイムとレスポンス方式，レスポンスメディアをどうするかが問題となる．レスポンスタイムは 0.3 秒程度以内ならば多くの場合十分である．レスポンス方式としては，認識とリジェクトの区別のみを表わす簡単な方式と，認識の結果をフィードバックする方式とがあり，用途により使いわけられている．レスポンスメディアは，視覚によるものと聴覚によるものに大別される．視覚によるものでは，近距離の場合は CRT ディスプレイが使用されるが，遠くから見る場合には大形のプラズマディスプレイなどが用いられる．また，ランプの点滅で行う場合もある．聴覚によるものでは，単純な信号音で認識／リジェクトの区別のみを行う場合と，音声応答により認識結果をフィードバックする場合とがある．また，各種レスポンスが併用される場合もある．

（3） 認識性能：

これには，絶対的な認識性能，すなわち，リジェクトなしとした強制判定の場合の認識率と，リジェクトレベルと認識レベルとのトレードオフをどうするかの二つの問題がある．これらにも用途によって許容範囲がある．現状の技術で，実用性能として，リジェクト 1％程度に対して，ほぼ無視できるエラー率が実現されている．

（4） 誤り修正方式：

フィードバックされた認識結果から，認識誤りあるいは発声の誤りに気付いた場合の修正方式である．キーボードから取消しの指示をして発声をやり直すのが一般的であるが，電話からの入力のようにキーボードが使用できない場合は，音声によるコマンドで修正を行うことになる．この場合は，そのコマンド自体の認識を誤ることがありうるので，修正手順の設定には注意を要する．また，場合によっては，再発声をなるべく要求せずに，認識結果の第 2 位以下を順次フィードバックして，yes/no で答える方式が有効である．

（5） 発声単位：

単音節，単語，文節，文に分けられる．当然，より大きい単位で発声した方がより自然であるが，技術レベルにより制約され，任意語音声では，単音節単位の入力のみが実現されている．

（6） 登録方式：

特定話者用の認識装置では標準パターン作成のために登録が必要である．最初の登録に何回の発声を要するか，認識モードに入ってからの標準パターンの部分的な入れ換え，あるいは修正をどのような手順でやるか，などが問題となる．

4.2 音声出力機器

音声出力についても，明瞭度，音声品質，レスポン

スタイム，出力単位，信号出力系などについて，人間工学面からチェックする必要がある．

（1） 明 瞭 度：

パラメータ合成方式では，十分な明瞭度が得られない場合が多いので，適用領域が制限される．

（2） 音 声 品 質：

波形編集合成では，符号化方式，ビットレートによっては，明瞭度はあっても雑音に汚れた音質となる場合があり，やはり応用範囲が限られる．

（3） レスポンスタイム：

出力指示が与えられてから，実際に出力音声が出てくるまでの時間であり，磁気ディスクのような回転形メモリを使う方式等で問題になる場合がある．

（4） 出 力 単 位：

音声出力の場合には，入力の場合のように認識技術上の制約はあまりないので問題は少ないが，単音節単位の出力は非常にききとりにくいので，なるべく使わない方がよい．

（5） 信号出力系：

通常はスピーカが用いられるのであまり問題はないが，イヤホンが用いられる場合には，装着方法によっては違和感を与えるので注意を要する．

5. 音声入出力機器の応用分野と利用形態

5.1 音声入力機器

これまでに実用化された音声入力の主な応用分野として，産業用音声入力，音声ワードプロセッサなどをあげることができる．

（1） 産業用音声入力：

仕分装置制御や検査データ入力で代表される産業用音声入力では，一人のオペレータが継続的に作業をするため特定話者用の認識方式でよく，語数も100語程度ですむ場合が多い．このように技術的には最も容易な応用分野であり，また，音声入力による省力化の効果が大きいため，この分野から約10年前に音声入力の実用化が始まっている．

仕分装置制御[6]では，オペレータはヘッドセット形の接話マイクを使用し，両手で荷物に書かれた行先をさがして読み上げていく．行先が地名の場合には離散発声で十分であるが，郵便番号のような数字コードの場合には，連続発声でなければ，音声入力による作業効率の向上は十分でない．音声入力による疲労は特に問題にならないとされている．

検査データ入力でも，両手を使ったり，体を動かしたりしながら検査を行い，結果をヘッドセットマイクから入力する．具体的な検査対象としては，自動車，圧延鋼板や鋼管のきず，ICマスク（顕微鏡下）などが知られている．

（2） 音声ワードプロセッサ

日本語ワードプロセッサの入力方式としては，当初はある程度の訓練により高速入力が可能となる連想コード方式がよいとされた．しかし，ワードプロセッサの普及が本格化するにつれて，より簡単におぼえられる全文字配列漢字タブレットを経て，最近ではカナ漢字変換方式が主流となってきた．

このような日本語ワードプロセッサの大衆化時代における素人向の入力方式として期待されているのが音声入力である．ワードプロセッサ用の音声入力のためには任意語音声認識が必要であるが，当面は単音節認識に頼らざるを得ない．

日本語テキスト入力において，単音節認識による音声入力と，他の入力手段とを比較する評価実験[7]の結果では，カナキーボードの経験のない5名の被験者が，かな文字100字のテキストを入力するのに要した時間は，音声入力（日電のSR-200プロトタイプを使用）ローマ字キー入力，あいうえお順キー入力がほぼ同じ160秒前後で最も短く，オンライン手書ひらがな認識が約220秒でこれに続き，JISカナキー入力が約330秒で最も長い時間を要した．また，この場合に，各入力方式に対して，原稿の見易さと使い易さについて主観評価による順位付けを行ったところ，**表-1**のような結果となり，いずれも音声入力が最も高い評価を得た．この結果から，現状の単音節入力は音声入力の方法として理想的なものではないが，現在利用可能な他の日本語入力手段との相対的な比較では，素人向きの日本語入力方式として十分実用性があるといえる．**図-4**に音声ワードプロセッサVWP-100シリーズ（日電）の外観を示す．

（3） その他の音声入力

CAD用グラフィック端末のコマンド入力に音声入力（日電の連続単語認識装置DP-200プロトタイプを

表-1 各種入力方式の主観評価結果

入 力 方 式	原稿の見易さ	使い易さ
① JIS キー	5.0	4.8
② 手書きひらがな	2.6	2.8
③ 単音節音声	1.2	1.6
④ あいうえお順キー	3.0	3.0
⑤ ローマ字キー	2.6	2.8

（5人の平均）

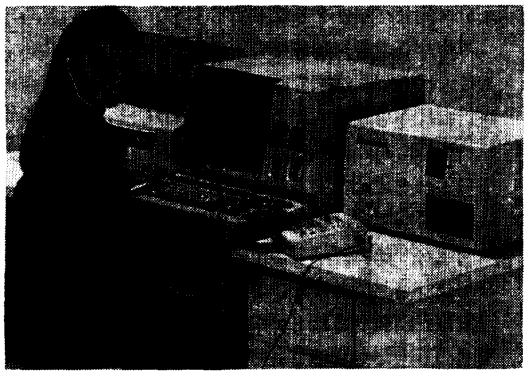

図-4　音声ワードプロセッサ VWP-100 シリーズ（日電）

使用）を実験的に適用した結果[8]では，キーボード入力に対して平均50%程度の作業時間短縮が実現されている．なお，米国ではカルマ社の CAD システムのオプションとして実用化されている．

その他，航空管制訓練シミュレータ入力用など，音声入力の特徴を生かした応用例は少なくない．

5.2　音声出力機器

これまでに実用化された音声出力の応用分野の代表例として，構内アナウンスシステム，盲人用読書器などがあげられる．

（1）　構内アナウンスシステム

駅，空港等において，発着案内などを行う構内アナウンスシステムでは，明瞭度や音声品質の重要性が高いが，語彙数は数百語以下に限定され，文章も定形化されている場合が多い．このような場合には，アナウンサが発声した音声波形を記憶しておいて使用する，波形編集合成方式が主に用いられる．記憶媒体としては，従来は主に磁気ディスクが用いられたが，最近ではビット単価の低下が著しい IC メモリ（DRAM）が用いられるようになり，レスポンスタイムが改善された．

（2）　盲人用読書器

書籍を OCR で読み取り，音声で出力するもので，イントネーションを自然につけるために，簡単な文章解析も行われる．任意語音声を出力する必要があるため，パラメータ規則合成方式が適している．英語の場合には OCR 部の負担が軽いため，米国では数年前から実用化されている．国内では通産省のプロジェクトとして，昭和57年度から5年間の計画でプロトタイプの開発が進められている．

（3）　その他の音声出力

電子翻訳機などの一般民生用機器に主にパラメータ編集合成方式により広く応用されているが，当初予想された程には応用分野は拡がっていない．この理由として，パラメータ抽出のための分析処理がメーカ側でかなりの時間をかけて行われるため，音声内容の変更が困難なことがあげられている．そこで，最近ではユーザ側でパラメータ抽出ができる分析システムが市販されるようになった．

5.3　音声入出力システム

音声入力と出力が同時に用いられる典型的な応用は電話から入出力を行うシステムである．銀行のテレホンサービスが代表的であるが，最近ではそれ以外の応用も本格化してきている．

（1）　銀行テレホンサービス

銀行では，従来から電話による顧客への振込通知や，顧客からの残高照会への対応を人手で行ってきたが，これらは銀行業務の中でも最も機械化が遅れた部分であった．そこで，まず，顧客からのプッシュホン入力と，銀行からの音声出力により，これらの業務を自動化するシステムが実用化されたが，プッシュホンの普及率が都市部でも10%程度と低いため，適用範囲は限られていた．

音声入力による本格的なテレホンサービスとしては，昭和55年3月からサービスを開始した住友銀行（大阪）の振込通知システムが第1号であり，続いて同年11月には三菱銀行が残高照会を含むテレホンサービスを開始している．これらは日電の不特定話者

表-2　残高照会のサービス手順

顧客側	システム側
発信	→応答
	「××銀行テレホンサービスです．サービスコードをどうぞ」
POPO	
	「残高照会ですね．店番・口座番号をどうぞ」
POPO…POPO	
	「暗証番号をどうぞ」
POPOPOPO	
	「そのままでお待ち下さい．」
	⋮
	「お待たせいたしました」「×××様の×日の残高は×××円でございます」
	「確認コードをどうぞ」
PO	
	「ありがとうございました」

（Pは発声合図のピー音を示す）

音声認識装置 SR-1000 シリーズ[3]により実現されたものである．**表-2** に残高照会業務の利用手順の一例を示す．

その後，電電公社により，共同利用形システムが商用化され，地方銀行を中心に利用されている．

テレホンサービスにおける利用手順は，これまで各行が独自に決めてきたが，利用者の立場からは，今後統一化されることが望ましい．

（2） その他のシステム

銀行系クレジットカード会社が最近テレホンサービスを開始した．通信販売の注文や，ローンの申し込みなどを自動化しようとするものである．

東京ガスは日電と共同で電話による緊急時社員一斉呼出しシステム「ワン・コール・システム」を開発し，本年3月から同社の第一線社員約7000人を対象に稼動させている．このシステムは，端末機から動員規模を指示すると，所定のリストに基づいて一括して対象人員の呼出しを自動的に行うと共に，音声入出力により出動の可否を確認し，端末機に動員可能者のリストを打出すもので，これにより緊急時体制は格段に充実する．

6. 人間工学面からみた音声入出力の課題

これまで実用化されてきた音声入出力機器における人間工学面からの問題点として，どのようなものがあるであろうか．

特定話者限定単語認識では，DP マッチング法によりほとんど技術的な問題は解決されたといえるが，人間工学面で残された問題として，マイクロホンの問題と，発声の許容度の問題がある．前者は，ヘッドセット形の接話マイクを使わずに，通常のオフィス環境程度の騒音のもとで音声入力を行いたいというものであり，このためには雑音のキャンセル法を新しく工夫する必要がある．後者は，パターンマッチングによる認識では，一般に認識範囲が狭すぎるため，人間の耳でははっきり聴き取れる発声でも，たまたま登録されている標準パターンとずれていると認識できない場合があることで，解決法としては，認識装置に学習機能を持たせることが考えられる．

単音節認識では，音響信号として類似したカテゴリが多いため，ある程度の誤認識は本質的に避けられない．このため，誤り修正の効率化，単語辞書やシンタックスなどの言語情報を用いた自動誤り訂正などのシステム的なバックアップを行う必要がある．

不特定話者音声認識では，通常，電話系で使用されることにより，認識率が劣化する傾向があるが，誤り修正の操作も音声入出力のみで行わなければならない．このため，認識結果の確認方式を含めて，誤り訂正の手順を最適化することが必要である．

音声合成では，規則合成における明瞭度と音声品質の向上が今後の課題である．明瞭度については定量的な評価が容易であるが，音声品質については定量的な尺度はまだ確立されていない．これは音声合成の研究を進める上でも重要であり，今後さらに検討されるべき人間工学的課題である．

以上あげたような人間工学面でのいくつかの課題が近い将来に解決されれば，ヒューマンインタフェースの代表的手段としての音声入出力機器の応用分野は，さらに大きく拡大していくものと期待される．

参 考 文 献

1) 迫江，千葉：動的計画法を利用した音声の時間正規化に基づく連続単語認識，日本音響学会誌，Vol. 27, No. 9, pp. 483-490 (1971).
2) Sakoe, H. and Chiba, S.: Dynamic Programming Optimization for Spoken Word Recognition, IEEE Trans., Vol. ASSP-26, No. 1, pp. 43-49 (Feb. 1978).
3) 渡辺，亘理，千葉他：不特定話者用音声認識装置 SR-1000 シリーズ，日本音響学会講演論文集，pp. 567-568 (昭和56年5月).
4) Sakoe, H.: Two-Level DP-Matching-A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition, IEEE Trans. Vol. ASSP-27, No. 6, pp. 588-595 (Dec. 1979).
5) Sakoe, H.: A Generalized Two-Level DP-Matching Algorithm for Continuous Speech Recognition, Trans. IECE Japan, Vol. E 65, No. 11, pp. 649-656 (Nov. 1982).
6) 江袋，磯野他：自動仕分け装置への応用，情報処理，Vol. 22, No. 4, pp. 318-326 (1981).
7) 吉田和永：日本語ワードプロセッサに対する各種入力手段の比較評価，情報処理学会昭和57年前期全国大会，2G-3, pp. 1019-1020 (1982).
8) 鶴田，首藤他：音声入力による CAD システムの評価実験，情報処理学会昭和56年後期全国大会，2 B-5, pp. 693-694 (1981).
9) 斉藤，中田：音声情報処理の基礎，オーム社 (昭和56年).

(昭和58年4月26日受付)