論文 / 著書情報
Article / Book Information

論題(和文)	音声認識技術の現状
Title	
著者(和文)	古井貞熙
Authors	SADAOKI FURUI
出典 / Citation	The journal of Information Science and Technology Association, Vol. 46, No. 1, pp. 19-25
Citation(English)	The journal of Information Science and Technology Association, Vol. 46, No. 1, pp. 19-25
発行日 / Pub. date	1996, 1
権利情報 / Copyright	本著作物の著作権は情報科学技術協会に帰属します。 Copyright (c) 1996 Information Science and Technology Association.

音声認識技術の現状

古井貞熙*

音声認識は，コンピュータに音声の意味内容を理解させる狭義の音声認識と，誰が話しているかを判断する話者認識に分けることができる。前者はさらに（孤立）単語音声認識と連続音声認識に分けられる。隠れマルコフモデル(HMM)，統計的言語モデルなどの技術的進展により，近年，音声認識や話者認識の種々の実験システムや実用システムが作られるようになってきた。今後の課題としては，個人差，雑音などへの適応化法，自然な対話音声の認識法などがある。

キーワード：音声認識，孤立単語音声認識，連続音声認識，連続単語音声認識，文音声認識，会話音声認識，音声理解，音声分析，HMM，統計的言語モデル，話者認識

1．まえがき

音声認識とは，人が発声した音声をコンピュータで知識処理することである[7]。音声には，発声者が意図した言葉の意味内容の他，誰が話しているかという話者情報も含まれている。どちらも我々の日常の対話において，重要な役割を果たしている。音声による対話は，人と人との最も自然で，容易かつ効率的な情報交換手段である。人とコンピュータの間でも音声を使って対話ができるようになれば，極めて便利になると期待される。

音声認識技術の研究は，1950年代から今日まで，世界中の多数の研究者によって，40年以上にわたって行なわれてきた[1)7)]。著者自身がこの研究分野に足を踏み入れてからも，もうすぐ四半世紀が経過しようとしている。その間，Stanley Kubrick の有名な映画「2001年宇宙の旅」の中のコンピュータ HAL や，George Lucas の「スターウォーズ」のシリーズの映画の中のロボット R2D2を始めとして，音声認識のできる数多くのロボットや装置が，空想科学の中で人々の関心を集めてきた。話し言葉を認識し，その意味を理解し，誰が話しているかを判定することができる知的機械を作りたいという願望は，多くの研究者の夢であり続けてきた。

音声は目で見ることができないので，音声認識の難しさを直感的に把握するのは難しいが，いわば毛筆の続け字で書いた行書あるいは草書をコンピュータで読み取ったり，誰の筆跡であるかを判定するようなものである[1)]。個人差や雑音（よごれ）によるパターンの変形が大きく，書き間違いに相当する言い間違いや言い直しも，日常の会話では頻繁に起こる。これらの問題にどの程度対応できるかが，重要なポイントである。これまでの40年間の技術的進歩により，音声認識のできるコンピュータの夢はやっと部分的に実現されつつあるが，任意の話題についての，あらゆる人による，あらゆる環境での音声を理解できるという目標からは，まだ程遠い状況にある。

コンピュータに，人が話した言葉を理解させる技術が狭義の音声認識技術，誰が話しているかを判断する技術が話者認識技術である。本解説では，主に前者を対象とするが，後者についても触れることにする。このため以下では，「話者認識」と区別して「音声認識」を狭義の意味で用いる。

2．音声認識の分類と基本的構成

音声認識は，図1に示すように，区切って発声した単語を認識する孤立単語音声認識と，単語を連続して発声した音声を認識する連続音声認識の二通りに分けることができる[8)]。後者はさらに，比較的多数の語彙を対象とし，文法などの言語的知識を用いて，その意味内容を理解しようとする文音声認識と，比較的少数の語彙を対象とし，単語間の言語的知識を用いずに，音響的特徴のみによって認識する連続単語音声認識に分けることができる。文音声認識は，会話音声認識あるいは音声理解ともよばれる。音声認識の形態はさらに，誰の声でも認識できる不特定話者型の音声認識と，学習を行なった特定の話者の音声のみを認識する特定話

* ふるい さだおき
NTT ヒューマンインタフェース研究所古井特別研究室
〒180 東京都武蔵野市緑町3-9-11
Tel. 0422-59-3910
東京工業大学大学院情報理工学研究科
〒152 東京都目黒区大岡山2-12-1

（原稿受領 1995.11.01）

者型の音声認識に分けることができる。さらにその中間として，後に述べる話者適応型の方法が，最近活発に研究されている。

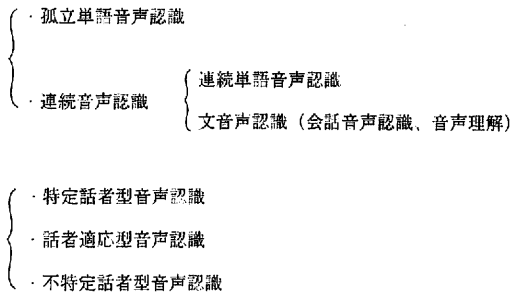


図1　音声認識の分類

認識を行なうためには，図2に示すように，音素などの音声の基本的な単位の標準パターンあるいはモデルを蓄えておき，認識すべき音声をそれらの単位と音響的に照合して，どの単位がその音声に含まれているかを判定する。その結果とその他の種々の言語的知識，すなわち語彙(単語辞書)，文法などを用いて，文としての認識の判定を行なう。

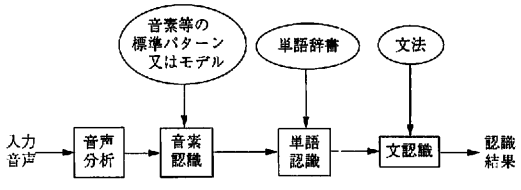


図2　音声認識の構成

音声の基本単位としては，音素から単語まで種々のレベルが考えられる。音素は，音声の最も基本的な単位であり，日本語には，/a/, /i/, /u/, /e/, /o/の5種類の母音と，/p/, /t/, /k/などの約20種類の子音がある[8]。英語など，もっと多数の母音と子音をもつ言語も沢山あるが，音素の数が50を越える言語はほとんどない。すべての音声は，これらの音素の連続によって表すことができるので，音素を単位とすれば少数の基本単位と音声との照合によって音声認識ができそうに思えるが，実はそうではない。

その主たる理由は調音結合，すなわち各音素の物理的特性が，前後の音素の影響によって変形する現象である[8]。このため，同じ音素でも，その前後の音素に依存した複数のパターンあるいはモデルを用意する必要がある。音素の数が数十程度であっても，当該音素の直前あるいは直後のいずれかを考慮する場合（このような音素単位をダイフォンと呼ぶ）で数百，前後両方を考慮する場合（このような音素単位をトライフォンと呼ぶ）には千種類以上の単位が必要となる。図3に，トライフォンによって単語を表現した例を示す。

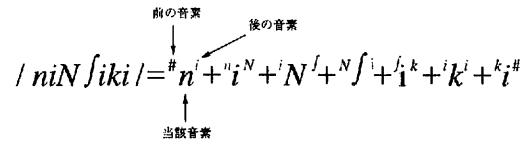


図3　「認識」という言葉をトライフォンの系列で表わした例（#は無音を表わす）

人が音声を聞いて理解しようとするとき，音としての情報だけでなく，語彙に関する情報（どのような単語があるかという情報）を始め，種々の情報を組み合わせて音声を聞き取っている。多くの場合，状況や話題の展開から無意識に単語や句を予測したり，文脈から推し量って認識している。その一つの理由は，音素情報のあいまい性である。すなわち，通常の音声の中では各音素は必ずしもはっきりと発声されず，その部分だけを聞いたのでは何の音素か分からないことがめずらしくないということである。このため，コンピュータによる音声認識においても，語彙，文法などの情報を組み合わせて用いることが不可欠である[5]。

3．音声分析の方法

音声認識の第一歩は，音声の物理的特徴を，コンピュータで扱える形に変換することである。これを音声分析という。そのステップは次のようである[4][8]。

(1)　フィルタによる周波数帯域制限

次のステップの標本化において，標本化定理が満たされるように，音声波を周波数帯域を制限する低域通過フィルタに通す。4～8 kHz 以下の周波数帯域に制限することが多い。

(2)　標本化，量子化

ディジタル処理ができるように，音声波を標本化し，さらに量子化する。

(3)　スペクトル分析

音声波から，10ms 程度の時間ごとに30ms 程度の区間を切り出し，各区間から短時間スペクトルを計算する。スペクトル分析法としては，線形予測分析法を用いることが多い。

(4)　特徴パラメータ抽出

音声認識に最も重要な情報は，スペクトルの全体的な形状にある。これを安定かつ効率よく表現する方法として，通常，ケプストラム分析法を用いる。ケプストラムは，対数スペクトルの逆フーリエ変換である。

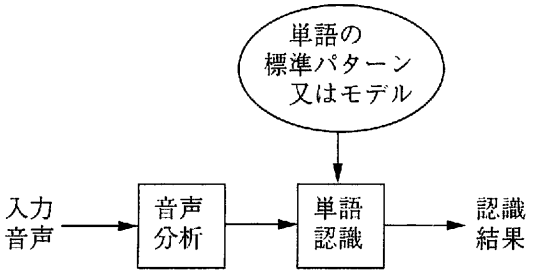
さらに，スペクトルの時間変化に，聴覚的に重要な情報が含まれていることが分かっているため，ケプストラムの時間変化を表すΔケプストラムというパラメータを抽出する。

(5) 音声区間検出

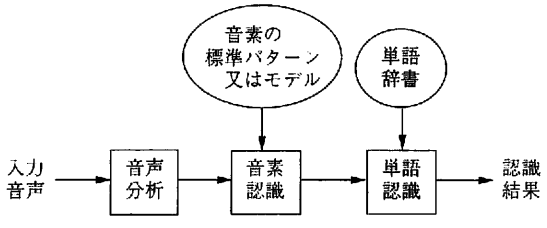
音声認識をするためには，音声が存在するか否かを検出することが必要である。この判定には，音声の振幅の大きさやその継続時間長などが用いられる。

4．単語音声認識の方法

音声認識の技術の中で，区切って発声された単語を認識する技術は，最も基本的なものであり，長い研究の歴史がある。典型的な二種類の単語音声認識の方法を，図4に示す。一つは基本的単位として単語を用いる方法であり，もう一つは音素を用いる方法である[8]。前節で述べたように調音結合現象があるため，単語に含まれる各音素の変形を忠実に表現するためには，各単語をそのまま単位として用いるのがよい。しかし認識すべき語彙の数が数千というように大きくなると，すべての語彙を蓄えるための記憶容量が膨大になるだけでなく，認識をするための計算量も膨大になってしまう。このため，語彙数が大きいときには，(トライフォンなどの調音結合を考慮した) 音素単位の結合として各単語を表現する方法を用いる。なお，音素と単語の中間である音節（日本語の「かな」に対応）や，それを半分に割った半音節と呼ばれる単位を用いる方法も研究されている[3]。



(a) 単語を単位とする方法



(b) 音素を単位とする方法

図4　単語音声認識の構成

音声の特徴パラメータ系列と，各音素や単語の基本単位との類似の度合いを計算する際に最も重要な課題は，音声の時間的な伸縮にどう対処するかである。例えば，同じ音素でも，それが含まれる言葉，発声者，発声速度などによって長さが変化する。しかも，その変化はゴム紐のように線形（一様）ではなく，部分的に伸びたり縮んだりする非線形な伸縮である。このように非線形に伸縮するパターンを，そのずれを調整しながら照合する方法として，動的計画法（DP；dynamic programming）を用いた方法（DPマッチングと呼ばれる）が提案され，広く用いられている[4]。

音声には，時間軸だけでなく，スペクトルの形そのものが，個人差を始めとする多様な要因によって変化する性質がある。この問題に対処するため，近年，隠れマルコフモデル（HMM；hidden Markov model）による方法が広く用いられている[4]。図5に，HMMの例を示す。HMMは，確率状態遷移機械であり，各状態遷移確率と，状態遷移に伴う特徴パラメータの観測(出現)確率によって特徴付けられる。これにより，音声の確率的変動が，極めて効率よく表現できる。

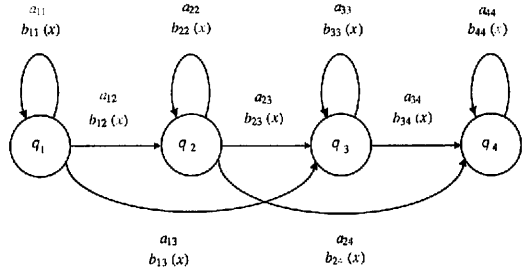


図5　HMM（隠れマルコフモデル）の例
(q_i：状態，a_{ij}：遷移確率，$b_{ij}(x)$：スペクトルパターンxの出現確率)

HMMによって音声を認識する際は，音声の特徴パラメータ系列が，音声の各基本単位に対応するHMMから生成される確率（尤度）を計算する。この計算も，DPを用いることによって，能率よく行なうことができる。その結果として，最も確率の高い基本単位を選び，認識結果とする。

5．単語音声認識の実用システム

単語を認識する装置は，すでにかなり実用化されている。世界で最初の大規模なシステムは，NTTが1981年に初めて実現した，音声認識を用いたバンキングサービスシステム「ANSERシステム」である[2]。このシステムが認識できるのは，10数字と6種類のコマン

ドで，しかも一語ずつ区切って発声することが必要である。このように機能は限定されているが，極めて多数の話者の音声を分析して，不特定話者が発声した音声が認識できるように，各単語の標準パターンが作られている。このシステムにより，通常の電話を通じて，銀行口座の残高照会，振込情報などを音声認識と音声合成によって自動的に得ることができるようになった。

比較的最近実用化された大規模な音声認識システムに，米国AT＆Tの電話オペレータの自動化システムがある。このシステムが認識できる語彙は，「コレクトコール」，「指名通話」などの6種類のキーワードであるが，不特定話者型であることの他に，「お願いします」などの余分な言葉が付加しても認識できる特徴がある。これは，文音声中からキーワードだけを検出する「キーワード・スポッティング」という技術を用いているためである。音声に関する制約としては，その文音声の中に，キーワードが必ず一つ含まれているということだけである。このシステムは全米に広がりつつある。

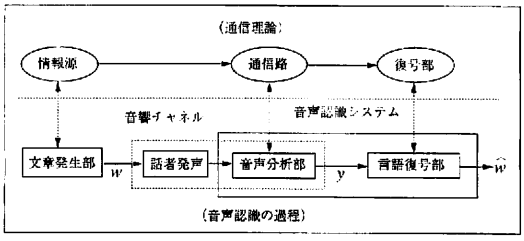
その他の単語音声認識システムとしては，音声ダイヤルサービスがある。受話器を上げて，番号をダイヤルする代わりに，あらかじめ自分の声で登録してある相手の名前や，「自宅」「会社」などの言葉を発声するだけで接続してくれるサービスである。これも米国の一部で用いられている。日本では，音声ダイヤル機能を持つ電話機も販売されているが，あまり売れていない。その他には，単語音声認識装置を，株価案内（カナダ），コンサート等の案内（フランス），病院や運転免許教習所の自動予約受付（日本）に用いている例などがある。

6．連続音声認識の方法

2章で述べたように，連続音声認識には，連続単語音声認識（文法なし）と文音声認識（文法あり）がある。前者の代表的なものは，連続数字音声認識である。このためには，あらゆる種類の数字の組み合わせ（並び）の候補と入力音声とを照合する必要がある。例えば7桁の数字の組み合わせは，10^7の膨大な組み合わせがあり，「なな」「しち」のような読みの変化まで考慮するとさらに増え，まともに実行すると膨大な計算が必要となる。しかしこの場合も，DPを用いることによって，大幅に計算量を減らすことができる。数字の発音も，その前後の数字によって変化するので，それを考慮して各数字に複数の標準パターンあるいはモデルを用いる方法が研究され，これにより認識性能が大きく向上することが確認されている。

近年の音声認識の基本的な研究は，文音声認識の方向へと進んでいる。その研究の流れは，ディクテーションと対話音声の認識・理解を対象とした研究に分けられる。前者はいわゆる音声ワープロに相当し，書き言葉を発声したものを文字に変換するものである。そこで認識の対象とされる言葉は，書き言葉であるから，文法的にはほぼ正確であると想定することができる。そのかわり，認識しなければならない語彙や文体が極めて多様であるという点に難しさがある。後者の対話の場合には，対象（タスク）をある分野の情報案内などに限定すれば，語彙を制限することができるが，話し言葉特有のあいまいな文章や，文法的に誤った文章も対象としなければならない難しさがある。

近年では，連続音声認識システムを，図6に示すように，音声分析部とそれに続く言語復号（処理）部から構成し，情報処理論・通信理論の問題として，確率モデルを用いて定式化することが広く行なわれている[5)7)8)]。この際，話者による発話と音声分析部を，一つの音響チャネルとしてモデル化する。話者は情報源に対応する文章wから，その話者の発声のくせに従って音声波形を生成する。音声分析部はスペクトル分析・特徴抽出を行なって，時系列データ（特徴パラメータ系列）yを得る。音響チャネルは雑音を含んだ通信路に対応し，言語復号部にyを送る。言語復号部はyから元の文章wを推定する。wは，ベイズ則によって展開した図中の式を最大化するように推定される。条件つき確率$P(y \mid w)$は音素モデルとの照合によって得られ，具体的には，HMM，DPマッチング，ニューラルネットなどが用いられる。文章wが発声される事前確率$P(w)$は文法によって得られ，近年では，単語の連鎖確率に代表される統計的言語モデルがよく用いられる。言語モデルとして，文脈自由文法，有限状態ネットワーク文法などを用いる方法も研究されている。



$$P(\hat{w}|y) = \max_{w} \frac{P(y|w)\,P(w)}{P(y)}$$

図6　確率モデルに基づく音声認識システムのモデル化

7．連続音声認識システムの例

連続音声認識の内，連続数字音声認識はすでに実用化レベルにかなり近づいている。その用途は，音声ダイヤル，照会のためのクレジットカード番号の入力などである。

文音声認識はまだ研究段階にあるが，世界で最大のプロジェクトは，米国のARPA（Advanced Research Projects Agency）によるものである[3]。その内，ディクテーションについては，Wall Street Journal新聞の記事を読み上げた文章を認識するプロジェクトが強力に進められた。音素認識にはトライフォンのHMMを用い，文法には統計的言語モデルを用いて，65,000語の語彙を対象にした場合，93%の単語認識率が得られている。同じくARPAの対話音声の認識・理解プロジェクトとしては，ATISシステム（Air Travel Information System）が，やはり強力に進められてきている。これは，米国内の航空路線を用いて旅行をすることを想定して，コンピュータによる案内システムに，音声で問い合わせや予約をするシステムである。自由に発声した話し言葉を対象とするため，言い間違い，繰り返しなど難しい問題が多数含まれている。どこまで難しい音声を対象とするかによって異なるが，現在90～95%程度の単語認識率が得られている。

日本では，電話番号案内システムを想定した研究などが行なわれている。極めて多数の人の名前や住所を，かなり自由な文体で発声した音声の認識ができるように，約80,000語を語彙とし，文脈自由文法を用いた認識システムの実験が進められている[2]。

8．話者認識の分類と基本的構成

話者認識とは，音声に含まれる発声者の特徴（話者情報）を用いて，コンピュータなどにより，誰の声であるかを自動的に判定することである[6][8]。これが実現できれば，(1)車や建物の入口の鍵として音声が使える他，(2)特定の会員や利用者を対象とした電話によるバンキング・買い物・情報案内サービス，(3)音声メール・コンピュータのリモートアクセス，などにおいて音声による本人確認ができるようになり，極めて便利になると期待される。カードや暗証番号は，他人に盗まれたり落としたりするおそれがあるが，音声にはそのような心配がない。ただし，同じ人の声が多少変化しても正しく動作し，他人が真似た声では誤動作しないようにすることが必要である。

話者認識の一般的方法としては，図7に示すように，まず登録すべき各話者の声の特徴を学習する。具体的には，各話者にいくつかの単語や文章を発声してもらい，その音声波形から話者の特徴を抽出する。各話者に対して，その代表的なスペクトル時系列を標準パターンとして蓄えるか，スペクトルの変動を考慮した確率的なモデルを蓄える。次に，認識をする場合には，未知音声について，学習の時と同じ方法で話者の特徴を抽出する。その特徴と，登録されている各話者の標準パターンあるいはモデルとの類似度を調べ，その値によって話者の判定を行う。

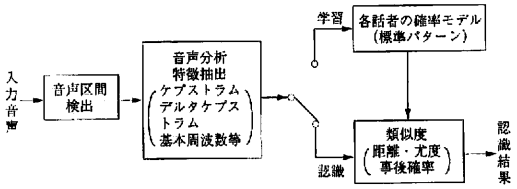


図7　話者認識の基本的構成

話者認識の方法は，話者識別と話者照合に分けることができる。音声が，あらかじめ登録されている多数の人の内の，誰の声であるかを判定するのが話者識別である。自分が誰であるかを音声，カード，登録番号などで名乗り，音声が，名乗った本人の声であるか否かを判定するのが話者照合である。音声を本人確認に用いる応用のほとんどは，話者照合に該当する。話者識別の場合は，多数の登録話者の中から，未知音声と最も類似度の高い話者を選ぶ。話者照合の場合は，類似度が一定のしきい値よりも大きければ名乗った本人の音声であるとして受理し，そうでない場合は他人の音声と判定して棄却する。

話者識別の性能は，登録話者の数が大きくなると，それだけ選択肢が増えるので低下するが，話者照合の性能は，常に本人か他人かの二者択一の判定を行なうので，その数には依存しない。

話者認識の方法はさらに，あらかじめ決まっている言葉（キーワード）を発声するテキスト（発声内容）依存（限定）型と，任意の言葉を発声してよいテキスト独立型に分類できる。前者の場合は，個々の音素に依存した個人性情報を直接的に用いることができるので，一般にテキスト独立型の方法よりも高い認識性能を得ることができる。しかし応用によっては，決まった言葉を用いるのが難しい場合もあるため，テキスト独立型の方法も，最近活発に研究されている。

9．話者認識技術の動向

話者認識の実用化システムで最も大規模な例は，米国のUSスプリント社のVoice Phone Cardサービスである[6]。これは，電話用クレジットカードの安全性を高めるため，そのカードを用いる際に，本人の社会保

障番号を発声することを求め，その音声の話者照合を行なう。その結果本人と確認できると，音声ダイヤルによって電話をかけることができるようになる。すでにこのカードは100万人のユーザによって使われている。

従来の話者認識の方法には，本人のキーワードの音声を録音して装置の前で再生すれば，本人になりすまして装置を容易にだますことができるという問題があった。このため最近，テキスト指定型話者認識法が提案された[6][7]。この新しい話者認識法では，認識装置を用いるたびに，装置の側から新しいテキストを，文字あるいは合成音声でユーザに提示し，ユーザが本人の声でそのテキストを正しく発声したと判定できる時のみ，受理判定をする。そのテキストとしては，初めてでも比較的発声しやすい，短い文章を用いる。指定される言葉を予測することはできないので，テープレコーダによってだますことは極めて難しくなる。

この基本的方法を図8に示す。あらかじめ各登録話者ごとに，学習用音声を用いて，各音素のスペクトルを表すモデル（音素モデル）を作成しておく。未知音声を，各登録話者が多種多様なテキストを発声したときの文章（音声）モデルと比較できるようにするためである。認識時には，指定したテキストに対応して，登録話者の音素モデルを接続して，そのテキストの文章（音声）モデルを作成する。入力音声をその文章モデルと比較し，類似性が十分に大きいときのみ，本人の音声として受理する。この方法により，安全性高く話者が認識できることが確認された。

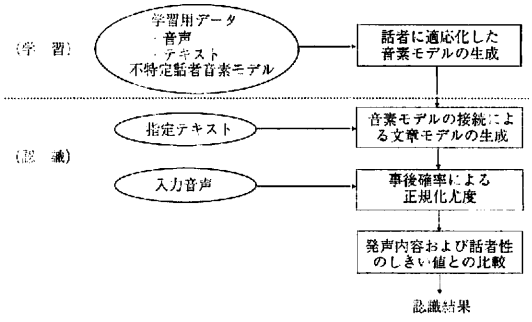


図8　テキスト指定型話者認識法

10. 今後の課題

音声認識技術は，近年実用化システムも徐々に増えつつあり，周囲の雑音や電話網での歪み，さらに声の個人差への対処など，実環境を意識した研究も活発に行なわれている。人が音声を認識する際には，個人差を含むこれらの実環境での音声の変化に，実に巧みに適応していることが分かっている。コンピュータにも，このような適応性を持たせることが極めて重要であり，この観点から，話者適応型の方法，雑音や歪みへの適応化法などの研究が活発に行なわれている[4]。

長期的な音声認識の研究の中心的な課題は，音声による人とコンピュータとの対話の実現である。対話音声の認識に関しては，まだわかっていないことばかりであり，書き言葉と大きく異なる対話音声を認識するために解決しなければならない問題が極めて多い。この中には，「あのう…」「お願いします」などの余計な言葉が付加した場合などへの対処法の研究も含まれる。

近年，携帯用のポータブル端末の小型化が進み，キーボードを持たず，手書き文字で入力する型も広く用いられているが，これが音声で入力できるようになれば，入力速度は大幅に向上すると思われる。電話による情報サービスなど，今後一層音声認識を用いた装置が増えてくることであろう。音声認識の究極の技術に，自動通訳電話がある。この実現への道のりはまだ遠いが，着実な進歩が期待されている。

話者認識技術も，これからの情報化社会でのセキュリティ確保の手段として，需要が増してくると予想される。

参照文献

1）古井貞熙，音声認識の過去・現在・未来，NTT R & D, Vol.43, No.10, p.3-8(1994)
2）鹿野清宏，音声認識技術の現状と展望，NTT R & D, Vol.43, No.10, p.9-16(1994)
3）Lee, C.-H.; Rabiner, L.R., 音声認識の将来, NTT R & D, Vol.43, No.10, p.17-28(1994)
4）南泰浩, 松岡達雄, 音声認識の音響処理技術, NTT R & D, Vol.43, No.10, p.29-38(1994)
5）松岡達雄, 南泰浩, 音声認識の言語処理技術, NTT R & D, Vol.43, No.10, p.39-48(1994)
6）松井知子，古井貞熙，話者認識技術，NTT R & D, Vol.43, No.10, p.49-56(1994)
7）古井貞熙，音声認識，電子情報通信学会誌，Vol.78, No.11(1995)
8）古井貞熙，音響・音声工学，東京，近代科学社，1992, 254p.