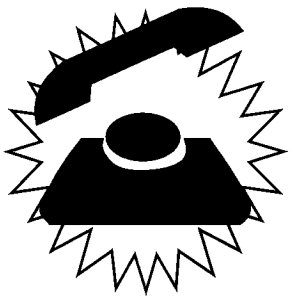
解説

電話の相手はコンピュータ？

－電話系音声認識とその応用－



石川　泰　yasushi@isl.melco.co.jp
岩崎 知弘　tiwasaki@isl.melco.co.jp
中島 邦男　kunio@isl.melco.co.jp
三菱電機（株）情報技術総合研究所

電話の相手はコンピュータ？

「...をご希望の方は1＃を，...をご希望の方は2＃を，詳しい説明をお聞きになりたい方は9＃を押してください」こんなメッセージをいらいらして聞きながら，電話器のプッシュボタンを操作された方も多いのではないだろうか．もちろん人間のオペレータが対応してくれるに越したことはないが，音声認識技術により，24時間話し中になることもなく，電話で予約ができたり，必要な情報が得られたらどれほど便利になるかは，改めて説明する必要もないだろう．

音声認識については，電話系に限らず，高い期待が寄せられ，1970年代から多くの研究開発が行われてきたが，最近の技術の進歩により，PCにおけるテキスト入力，いわゆるディクテーションや，音声認識機能を有するカーナビゲーションが製品として出回ったこともあり，いよいよ本格的な実用の時代に入ったといえる．この中でも電話系の音声認識は，古くから実用化が行われてきた分野であり，最近では応用分野も広がり，種々の音声認識を利用した電話サービスが運用されはじめ，「電話の相手がコンピュータ」である時代を迎えつつある．ここでは，研究開発の歴史を概観し，ニーズの高まり，現状技術とその応用例を紹介し，さらに，今後電話を切ったときユーザが「相手はコンピュータだったかな？」と悩むような日を迎えるための課題と展望を述べる．

電話系音声認識の研究開発

電話系の音声認識の研究開発の本格化は，音声認識研究の初期に，音声分析手法としての線形予測分析，音声の時間変化を考慮したパタンマッチング手法である「DPマッチング」（あるいはDTW（Dynamic Time Warping））の提案により単語認識の代表的なアルゴリズムが確立した1970年代後半に遡ることができる．電話系の音声認識の問題は，

(1) 不特定話者音声認識：利用者を限定できないため，発声者による音声の変形は，年齢層，方言，生理状態なども含め多種多様にわたる．さらに，通常使い慣れている電話での発話であるため，不明瞭な発声も多い．これらにいかに対処するか，

(2) 電話帯域，歪みへの対処：電話は伝送効率を向上させるため，300～3,400Hzに帯域が制限されている．また，電話端末や回線による歪み，雑音が音声に重畳する．これにどう対処するか，

が大きな課題となる．研究の初期では，不特定話者の音声の特徴を表現するため，1つの単語に対して複数の標準的な音声パタンを大量の発声データから抽出するマルチテンプレート法を基本方式として，マルチテンプレートの構成法，電話系の雑音や歪みに強い音響パラメータを抽出する分析方法に主眼が置かれた．これらの研究成果により1980年代の初頭には，数字と「はい」「いいえ」などの機能語，合わせて

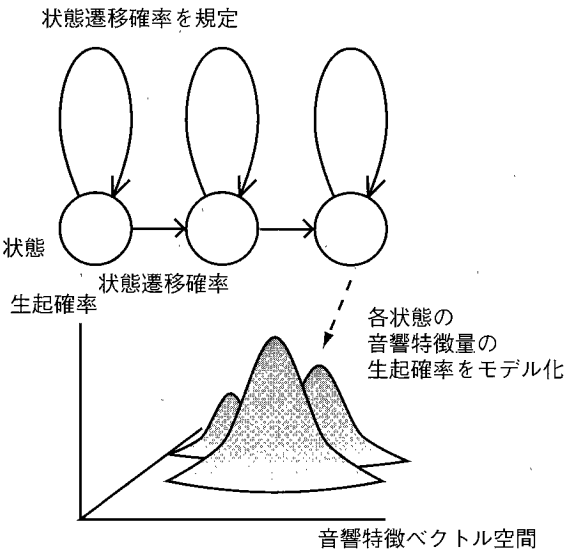


図-1　隠れマルコフモデルの概念

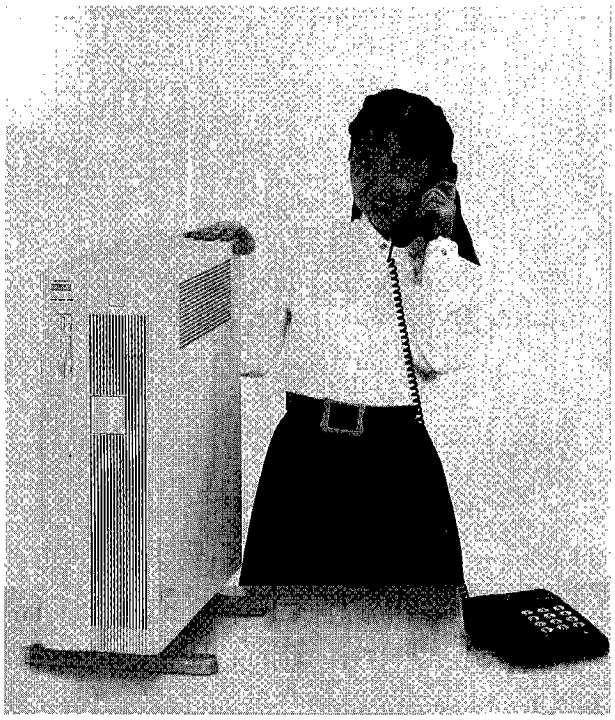


図-2　音声認識システムMELAVIS

10数単語認識を認識するシステムが開発された．特に当時の電電公社で開発された電話系単語認識による情報サービスシステムANSER (Automatic answer Network System for Electrical Request) が有名で，銀行や証券情報の提供サービスに適用された．10数字と機能語の認識は結局プッシュボタンと同じ機能を提供するに過ぎないが，当時は国内のプッシュホンの普及が遅れていたこともあり，我が国の音声認識の初期段階で最も大きな成果をあげたシステムとして評価されている．ANSERシステムは，プッシュホンの普及により，音声認識の利用という点ではその価値を失っていったものの，多くの銀行や証券会社が共同で利用できるアウトソーシングサービスの先鞭を切ったシステムであり，エレクトロニックバンキング・ホームバンキングサービスとして発展し，現在のNTTデータによるANSER-WEBではインターネットバンキングへの展開を示している．

しかし，電話系音声認識については，1980年代の初頭のANSERシステム以降，若干の語彙数の拡大や，システムの小型化が行われたものの，大きな進展は，1990年代後半を待つことになる．この大きな進展には，1980年代以降本格的な研究が進んだ隠れマルコフモデル (Hidden Markov Model, HMM) に基づく音声認識技術によるところが大きい．

ここで，簡単に隠れマルコフモデルを紹介しよう．従来は，音声の特徴をある特定の観測パタンで記憶しておき，そのパタンとの時間軸整合を行うパタンマッチングにより認識を行っていたため，音声認識の性能は，記憶しておく参照パタンに大きく依存するものであった．これに対して隠れマルコフモデルは，観測される音声の特徴量の時系列を確率現象としてモデル化する方法で，発声者などの違いによる音響特徴量の変動をモデルに表現させるとともに，時間軸の変動も，モデルの状態遷移の相違として表せるため，音声の種々の変動に対して精度のよいモデル化が可能となった（**図-1**）．観測系列を最も高い確率で出力するモデルを認識結果とするこの方法は，適切なモデルの学習が行われていれば，高い認識率が得られ，音声認識の性能向上に大きく貢献した[1]．

このような確率モデルに基づく音声認識手法により，不特定話者音声認識で十分に実用的な認識性能が得られるようになり，それまでの少数語彙認識から，数百語以上の大語彙化と認識率の向上が果たされた．さらには，基本的に音素モデルに基づく認識を実現したことで，認識対象語彙の大量音声データを収集することなくテキストにより自由に設定できるようになったのである[2]．1995年3月に発売された三菱電機のMELAVISは，業界で初めて1,000単語（総語彙10万単語）の認識を電話系で実現したシステムである（**図-2**）．このシステムでは，音素モデルについて，前後の音素環境を考慮し，前の音素からの入り渡り区間，音素特徴を有する定常区間，次の音素への出渡り区間に相当する音響的な特徴を考慮した細かい単位を用いることで，モデルの表現性と演算効率の向上を両立させ高い認識率を実現した．また，話者適応化により例外的

話者に対する耐性も強化している．さらに，ガイダンス音声出力中に発声があった場合，回線上で重畳してしまうガイダンス音声をキャンセルして入力音声のみを抽出認識可能とするバージイン機能（図-3），「あの」，「えー」などの不要語に対処する機能などを有することが特徴となっている．

発売当時は，音声分析や，モデルとの照合を実時間で処理するため，DSPを用いた並列処理を行っていたため，図-2のような，やや大きな筐体に専用の音声処理ボードを搭載したH/W構成をとっていたが，その後のCPU性能の向上などにより現在では，電話回線制御ボードを搭載したPCサーバ上で動作するソフトウェア音声認識エンジンとして動作している．

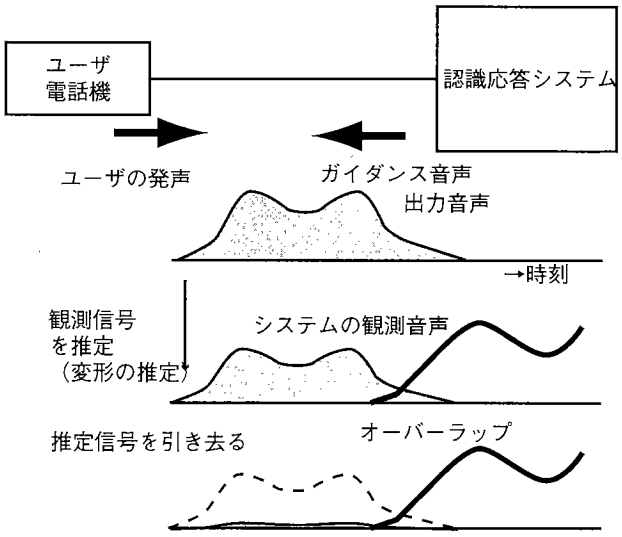


図-3　バージイン

高まる要求

さて，電話系の音声認識が今日脚光を浴びているのは，単に技術の向上により実現性が高まったというだけではない．そこには，市場における要求の高まりも見逃せない．従来，生産者・販売事業者は，市場の大多数の要求を把握し，それに適合する製品を大量に生産，安価に販売することで利益を追求してきたが，そのような経営スタイルが行き渡り，情報の流通，生産から市場投入への速度が速くなった状況では，単に一般的な消費者動向を把握しているだけでは，激化する競争の中で企業の存続が難しい時代を迎えたのである．また，市場のニーズ自体も個人のライフスタイルの多様化などを背景に画一的なものではなくなってきた．そのため，企業の経営は，個々の顧客のニーズを的確に把握し，獲得した顧客に厚いサービスを提供することで，顧客との関係をいかに継続・維持するかが重要視されるようになった．このような顧客重視の経営戦略は一般にCRM（Customer Relationship Management）と呼ばれ，1990年代後半から浸透し始めた．このCRMを実現するうえで，さらに注目を浴びるようになったのが，コールセンターである．コールセンターは従来から，電話による相談や故障の受け付け，さらにはテレフォンショッピングの窓口として，企業にとって重要な位置づけではあったものの，それを企業の経営戦略に活用しようという考え方はなかった．しかし，CRMを実現するうえでは，企業にとり直接顧客と接するきわめて重要なシステムであり，

- いかに個々の顧客の情報を蓄積活用し，サービスを充実させるか，
- コールセンターを通じて得られる顧客の要求や苦情をいかに素早く的確に経営戦略に反映させるか，

が企業の経営を左右すると考えられるようになったのである．そのため，電話とコンピュータとの統合によるシステムであるCTI（Computer Telephony Integration）が注目されるようになった．CTI自体は，1980年代のPBXの交換機能をコンピュータに公開するオープン化にその始まりをみることもできるが，1990年代後半に本格的にシステムの開発導入が進み始めた．特に，1997年から始まったNTTの発信電話番号表示サービス（ナンバー・ディスプレイ）は電話とコンピュータとの統合の可能性を大きく広げるきっかけとなった．

電話系音声認識は，このCTIにおいて，

- コールセンターの自動化，24時間化，
- 従来のプッシュボタンによる情報提供や1次受付（業務の振り分け）の効率化，サービスの向上，

を実現するものとして大きな期待を受けるようになったのである．

さらに，「音声認識というインタフェース」の導入において，他のインタフェースが存在する場合には，「使いやすさ」などの導入効果が定量化しにくく，企業向けのシステムではその投資効果が評価しにくい．しかし，電話系音声認識では，導入による効果を，同様のサービスを提

供するのにすべて人間のオペレータを雇った場合との運用コストの比較や，利用者の通話時間が短縮することによる話し中となる率の減少，用意する回線数と処理できる呼数の関係など，定量的に考えやすいという点もあり，音声認識システムへの企業の注目を大きくしたことも見逃せない．

実用化例

このような期待を受けて，現在，電話系音声認識装置は各社から販売されている．ここでは個々の製品の紹介は他の資料に任せることとし[3]，一般的な認識装置，認識S/Wの機能概要を説明する．

現在販売されている電話系の音声認識システムは，WindowsNTをプラットフォームとするS/W製品が多く，認識対象は，多くは離散単語発声であるが，要求の多い数字の連続発声に対応している製品も多い．また，連続音声認識を機能として有する製品も発売されているが，実用的なシステムを構築するうえでは，ユーザにまったく自由な発声を許すと意味を抽出することが困難となることなどから，離散単語音声認識システムとして利用されることが多い．認識語彙数は，カタログ上は数百単語から20万単語程度まで，あるいはS/W上の制約がない製品までとその幅は大きい．しかし，地名や人名を入力対象とする場合以外では，コマンドやメニュー選択として利用されるため，実システムでの認識語彙は数百単語以下の場合が多い．その他の機能としては，上に説明したバージイン機能，話者適応化，不要語除去，エコーキャンセルなどがそれぞれの製品で実現されている．

これらの装置を利用したシステムがすでにいくつか稼働している．

■情報提供システム

地方自治体のサービスや，業務の案内に利用されている．たとえば，1996年から運用されている岡崎市情報ネットワークセンターでは，行政分野の情報を1元的に管理提供するための情報拠点作りの一貫として，電話音声認識システムを導入し，「だれでも，いつでも，どこでも，はやく」情報が入手できるシステムを実現している．行政サービスの提供は他の自治体でも積極的に進めており，大阪柏原・羽曳野・藤井寺市の消防組合では地域の医療情報の提供システムに音声認識を導入し，注目を浴びている．ほかにも，交通情報サービス，駐車場案内や，天気予報などに実用システムがある．

■予約システム

チケット予約，ホテル予約などは電話系音声認識の最も期待の高い応用分野であるが，予約業務においては，座席や部屋などについてのユーザのやや複雑な希望を処理する必要があり，簡単なフローでは処理が困難であること，誤認識が重大な問題となり得ることなどから現時点では，実用例は少ない．今後の技術の進展が必要な分野である．

■受発注システム

予約システム同様，まだ本格的な実用システムは少ないが，音声認識の大きな利点である24時間の運用が可能であることを利用した化粧品のオーダーエントリーシステムの運用例がある．

■1次受付ほか

コールセンターへの電話を用件により振り分けるために，音声認識が用いられている例があるほか，KDDでは海外で「ジャパンダイレクト」の受付にかけられる大量のいたずら電話を防止するため，日本語によるキーワードを要求するメッセージを出力し，利用者の復唱を認識することで，正当な利用者かどうかを判断するシステムを導入している．このシステムによりいたずら電話のほとんどを自動切断することに成功している．また，企業内の利用としては，情報サービスのほか，内線交換に利用し，部署と名前を言うことで相手に内線電話を接続するサービスを運用している例がある．

■海外動向

音声認識についての技術については，我が国はトップレベルにあるものの，実用化では特に米国がやや先行している状況にある．これには，

- ヨーロッパ言語では単語という概念が明確で，単語認識によっても複雑なタスクを処理しやすく，複雑な内容でなければ，単文の表現の多様性も日本語ほどは大きくはないため，連続音声認識によるシステムも実現しやすい，
- 我が国では信頼性，確実性，きめ細かいサービスが重要視されるのに対し，効率などの観点で音声認識を容認する傾向が強い，

ことが背景にある．1997年に行われたEurospeech'97では，電話系の対話システムのコンテストが行われ，列車の情報検索サービスなど種々のシステムが会議参加者により評価されている．また米国においては，電話会社のコレクトコールの接続などに古くから実用化例があり，電話会社への通話の用件の分類，各種の情報サービスなどでの実用システム例は多い．

課題と展望

さて，このように今まさに電話系の音声認識の実用化の時代を迎えたわけであるが，今後さらに利用しやすいものとし，利用分野とユーザを拡大するための課題について最後に述べる．

■頑健性

電話系音声認識は，電話帯域であること，不特定話者が基本であることという音声認識にとって最も厳しい条件のもとに研究が進められてきた技術であるが，現在十分な性能が得られているわけではない．特に最近では携帯電話が急速に普及し，音声認識にとり厳しい品質の端末が増えたことなど，従来よりも検討課題が増えている状況にある．音声認識の基本アルゴリズムについては，種々の制約を緩和し，さらに頑健な認識システムを実現していく必要がある．主な課題は，

- 端末，回線の多様化への対処：携帯電話，インターネットプロトコル上の音声(VoIP)などについては，今後，音声通話品質の向上が図られていくことも予想されるが，多様化は進み，低品質な音声への対処は重要な課題となる．
- 騒音下音声認識：携帯電話の普及により，電話系音声認識の利用環境が拡大し，騒音下での認識性能を向上させることが重要．
- 発声の制限の緩和：不要語の除去や，発声から必要な単語だけを抽出するワードスポッティング，さらには連続音声，自由発話認識へと発展することが必要．
- ユーザの拡大：現在の不特定話者認識技術では，どのような話者でも平均的な認識率が得られるというわけではなく，大多数の話者は高精度に認識が可能であるが，音響モデルとの整合性が悪い話者では著しい認識率の低下が起こる．これを回避し，さらに認識が困難な老人や子供への対処も重要である．

これらについて，さらに研究が進められる必要があることはいうまでもないが，加えて米国で整備が進んでいる電話系の音声コーパスの整備も大きな課題である[4]．

■対話処理技術

電話系の音声認識でユーザが1回の発話で，目的を達成できるような業務はきわめてまれであろう．ユーザが複数の入力（発話）で目的を達成する対話型のシステムでは，インタフェースの設計により，音声入力としての利便性が得られないものとなったり，かえってインタラクションが増え不便なシステムとなることさえある．

- 誤認識時のリカバーなど，音声インタフェース設計技術の研究，
- インタフェース評価技術の研究に基づく設計基準・ガイドラインの作成，
- 簡便に音声対話システムを構築，修正ができる音声インタフェース構築ツールの開発，

が音声認識の基本的な認識率と同様きわめて重要な課題である．音声認識の認識率よりも，対話設計すなわち業務フロー設計がその効率，効果を決めるといっても過言ではない．

インタフェースの設計において最も重要な要求事項に「自己記述性：システム・対話の状況が正しくユーザに理解されるか」と，「可制御性：ユーザがシステムを制御できるか」がある．しかし，特に音声だけのチャネルによるインタフェースである電話系の対話システムではその実現が困難であり，さらに誤認識時にはきわめて厳しい状況になるのである．また，音声出力についてもその特徴を理解した設計が重要である．ディスプレイに表示されるガイダンスなどは，不要であれば無視をするか「斜め読み」をすればよいが，音声によるガイダンスは無視できず，時間もかかるものとなる．これらの特徴を考慮した音声インタフェース設計技術の向上が，電話系システムに限らず，システムの効率を向上させ，音声認識の普及の鍵となると考えられる．

技術的な研究，進展に加えて，メーカや設計者，対象業務によらない共通の利用方法や用語の統一化なども，インタフェースの「学習容易性」の向上に必要であろう．

電話系の音声認識には，インターネットが普及すれば必要性も薄まるのではないかという意見もあろう．しかし，通信手段が変化しても，音声インタフェースの有効性が低下するものではなく，インタラクションの多い情報の検索や予約業務などでは，音声認識の有効性が大きいことは疑う余地がないであろう．さらに携帯電話の普及で，「いつでも，どこからでも，だれでも」を実現する電話音声認識によるシステムはさらに要求も高まり，発展普及するであろう．21世紀，電話の相手がコンピュータであることがごく当然になっていくものと予想される．

参考文献，参考URL
1) 中川聖一：確率モデルによる音声認識，電子情報通信学会 (1988)．
2) Iwasaki, T. and Nakajima, K.: A Real Time Speaker-Independent Continuous Speech Recognition System, Proc. ICPR, pp.663-666 (1992)．
3) 音声入出力方式に関する調査研究報告書，00-標-2，(社) 日本電子工業振興協会 (Mar. 2000)．
4) http://morph.ldc.upenn.edu/

(平成12年7月6日受付)