Vol. 36 No. 11　　情　報　処　理　　Nov. 1995

解　説

音声言語情報処理の現状と研究課題

6. 音声認識技術実用への課題†

嵯峨山　茂　樹††

1. はじめに

いうまでもなく，音声認識はまだ100%認識できるような完全な技術ではない．それでも使い方によっては有用であると考えられてきた．音声認識のアルゴリズム[1),2)]は大きく進歩して性能が向上し，パソコンなどのハードウェアの飛躍的発達から計算コストも下がり，応用可能性はかつてに比べ格段に広がってきたかのように思える．しかし，期待したほど音声認識は使われていない，と感じている人は多いようである．

かつて音声認識の応用が少なかった頃は，高いコストや低い性能がおもな問題だった．「コストがもう少し安く性能がもう少し高かったら，応用は一挙に広がるのに」と多数の技術者が悩んでいた時代はもう10年以上も前のことである．当時に比べ，これらの点で大きく進歩してきているように思われるのに，かつて想像していたほど音声認識は使われていない．

なぜなのだろうか．どうすればよいのだろうか．

ひたすら研究をして音声認識性能を向上させれば，応用は自然に開かれてくるものなのだろうか？それとも，何か盲点のようなものがあって，それが問題の本質なのだろうか？何を解決すれば爆発的に実用になるのだろうか？実は人間は，機械に向かって話すなどということは，大嫌いなのだ，という極端な説もある．だから，音声認識はだれも決して使いたがらない絶望的な技術なのだと．「値段さえ安ければ音声認識を使いたいんですけどねえ」というシステム屋さんたちの言葉は本心だったのだろうか，それとも音声研究者へのお世辞だったのだろうか．

一体何が本当の問題なのだろうか．

このような疑問を抱いて，当学会の音声言語情報処理研究会（SIG-SLP）の初会合（1994年5月20日）の準備として「なぜ音声認識は使われないか・どうすれば使われるか？」と題して電子メール討論を行ったところ，多数の音声・言語研究者から返答を得た．研究会当日にはアンケート調査も行った．本稿では，討論内容を紹介するとともに，筆者の考えを加えて議論する．読者には音声認識の1つの側面を一緒に考えていただくことで，次の段階への前進につながれば幸いである．

2. 事前E-mail討論と投票調査

議論開始にあたって，音声研究者メール（onsei-mail@etl.go.jp）や言語研究者などのメールグループに投稿した内容を次に示す．いままでに筆者自身が考えたり，多くの人から公式・非公式に聞いていた22種の理由（(a)-(v)）を示し，どの項目への賛同が多いか，一種の投票の形の調査を試みたが，むしろ自由回答の形の発言が多く，結局自由討論となった．

そこで，その内容から新たに6項目（(A)-(F)）を抽出し項目に新たに加えた．第1回音声言語情報処理研究会においては，この事前討論内容を基に報告・議論[5)]するとともに，参加者を対象に，28種の項目の趣旨を筆者が説明し，ただちに挙手による賛同の票数を数えた．当日その場で集計されたもの（に筆者自身の票も追加した票数）を，各項目ごとに示す．挙手投票終了時点の参加者総数は83名であった．

事前討論：「なぜ音声認識は使われないか・どうすれば使われるか？」

† Discussion Toward the Practical Use of Speech Recognition Technology by Sigeki SAGAYAMA (NTT Human Interface Laboratiories).

†† NTTヒューマンインタフェース研究所

嵯峨山@NTTヒューマンインタフェース研究所です．

◆今年度，情報処理学会の「音声言語情報処理研究会」が発足し，5/20に東京で第1回の研究会が計画されておりますが，その中で「なぜ音声認識は使われないか・どうすれば使われるか？」というテーマで議論をしたいと考えて，そのための時間をいただいております．

◆いままで多数の研究者が，音声認識は有用な技術であると信じて研究開発に多大な努力を払ってきました．しかし，実際にはその実用化は思ったほどはかどってはいないようです．何が問題なのでしょうか？何を解決すれば爆発的に実用になるのでしょうか？不況の中で音声認識研究の立場をよくするためにも，この問題は一度討論する価値がありそうです．

◆そこで，以下の事前討論に皆様の返答をお寄せいただけないでしょうか．返答の中から私が取捨選択してまとめて当日に報告させていただき，議論の出発点にしたいと存じます．また，皆様がどのようにこの問題を考えておられるかを明らかにしてみたいと思います．この形式にかかわらず，フリーディスカッションの形でご意見をくださっても結構です．

◆質問「あなたは，なぜ，これまで音声認識技術が期待するほど使われて来なかったと思いますか？以下の選択肢の中に当てはまると思うものがあれば，複数回答でお答え下さい．」

(a)認識率が低い．音声認識は技術が未熟である．まだ使いものにならない．　(36票)

(b)自由発話音声認識ができるまでは本格的に使えない．いまの連続音声認識は使いものにならない．　(23票)

(c)話者の間で認識性能の差が大きすぎる．認識率が低い話者がいるのでは使えない．　(13票)

(d)扱える語彙サイズ（語数）がまだ小さすぎる．　(11票)

(e)語彙制約があるのでは使えない．任意の音声・文字変換できなければ不自由．　(4票)

(f)ワードスポッティング技術* が肝要．(18票)

(g)リジェクト力** が弱いことが最大の問題．　(40票)

(h)雑音や発声変形や回線変動などに対するロバストネスが不足．　(46票)

(i)ヒューマンインタフェースが未熟だから使いにくい．　(41票)

(j)うまい対話制御が重要．これがうまくいっていないから使われない．　(22票)

(k)人は機械に向かって話すのには抵抗がある．音声認識技術は本質的に嫌われる．　(7票)

(l)音声認識誤りがどのように起こるのか，どう発声すれば避けられるのか分からない．この不透明感がユーザにとって最も辛い．　(41票)

(m)キーボードなどの他手段に比べて入力効率が決してよくない．　(17票)

(n)音声認識機能とアプリケーションのインタフェースが確立していない．　(24票)

(o)コストの問題．現状では，音声認識技術は高価すぎる．　(11票)

(p)まだ速度が十分でない．　(6票)

(q)マイクロフォンが口の近くになければ動作しないのでは応用が限られる．　(10票)

(r)音声認識の応用について知恵が足らない．現在の技術レベルでも工夫すれば使えるはず．　(37票)

(s)言語処理が問題．音響処理はかなりのレベルに達しているが言語モデルあるいは言語処理（構文解析，意味理解．状況，知識，社会常識）が遅れている．　(18票)

(t)音声認識への要求条件が厳しすぎる．もっと育てるつもりで使えるところから使うべきだ．　(0票)

(u)音声認識を使う人間を訓練する必要がある．キーボードだって一日で使えるようにはならない．　(6票)

(v)音声認識の使い方の社会的コンセンサスがまだない．　(11票)

(A)音声研究者が実用化など考えていないからだ．この技術はいける，という確信がなければ実用化はできない．　(22票)

(B)提供されるサービスが十分に知的（intelligent）である必要がある．「対話」が知的であるかではなく「サービス」が知的であることが肝要．　(21票)

* あらかじめ決められているキーワードを，入力音声中から検出する技術である．読者も，外国語の音声から知っている単語だけ聴き取れた経験があるだろう．

** あらかじめ決められた音声認識対象の語彙や文法から外れた内容が発声された場合に，そのことを検知する能力である．

(C)次の3つの条件がまだうまく噛み合っていない. (34票)

Ⅰ. 音声認識理解の性能と機能がある程度のレベル(応用による)に達していること

Ⅱ. システムに組み込む人という意味でのユーザが容易に組み込めるためのサポートツールが用意されていること

Ⅲ. 音声認識理解を利用することを前提とした応用技術が開発されていること(たとえば,音声認識を前提としたCAI技術など)

(D)長年の研究成果があまり開かれたものになっていない. 市場を立ちあげるためには,基礎技術以外にも製品やサービスとして完成させるアイデアや技術や根性や資本が不可欠. (7票)

(E)音声認識というものはすでに人間が非常な夢を描き,理想が先行してしまった技術であるために,なかなか便利であると認めて貰えない. 発想の転換が必要. (4票)

(F)メンタルモデル(システムがユーザにどのように見えているか)とシステムイメージのギャップを埋める手段が確立していない. (36票)

以上のうち(a),(g),(h),(i),(l),(r),(C),(F)は研究会当日に30票以上の賛同を得ていることから,参加者の多くの問題認識は,

「音声認識に関してまだ基本性能が十分でなく,リジェクト力,雑音や回線変動に対するロバストネスを含めた基本性能がまだ不足である. 応用の可能性については楽観的だが,応用については知恵不足. また,応用する上でヒューマンインタフェースとメンタルモデルが現実的問題である. 応用開発ツールを用意することも重要. 」

とまとめることができるだろう.

3. なぜ音声認識は使われないか

メール討論への参加は大変活発であった. 多様な討論意見から僅かしか伝えられないのは残念であるが,筆者の独断で選択し要約したものを以下に示す.

「どう発声すれば認識してくれるか分からない,コストが高い」…「何と発声すれば装置は認識してくれるのか?」という部分の不透明さ. すなわち,発話を誘発するようなインタフェースが,現時点では必要だ. しかし,現状の音声認識装置の性能でも用途を限定すれば,利用できるアプリケーションは,必ず存在. 利用できる単語や文(認識装置が認識できる言葉)の発声を支援する環境をユーザに提供する必要がある. 認識装置の特性を最もよく理解している開発者が,アプリケーションの開発にも取り組まなければいけないのかもしれない.

「提供されるサービスが知的でなければならない」…音声認識(対話)が実際に役に立つためには,提供されるサービスが十分に知的(intelligent)である必要がある. 人間に(音声で)ものを頼むのが快適なのは,相手が十分に知的だからではないか.「対話」が知的であるかどうかではなく,「サービス」が知的であることが重要だ. まず,何かの知的なサービスが世の中で利用されるようになって,次にその最も快適なインタフェースとして,音声対話が注目されるようになるという順番.

「音声認識を使いこなしているのはまだ少数」…数年前の結論(音響学会誌48巻1号,1992)は,(1)ロバストネス不足(特に周囲の雑音や使用者の声の変動に対する)が,一般の人から見ると使えないという評価につながっている,(2)音声認識を用いて本当にメリットがあるというサービスが現われていない,の2点.

米国でも(1)実際に音声認識技術を使いこなしている例はきわめて少ない. (2)米国のベンチャ企業でも音声認識装置で商売できているところはほとんどない. DARPAの金で食っているか,資本金を食いつぶしながら生きているところが多い.

「信念なくして研究や技術が成功するはずはない」…技術進歩のスピードの違いが,日米の当事者の思い入れの深さの違いを反映している. 信念なくして,研究や技術が成功するはずがないということは,日米を問わず,あてはまる. 日本よりも,米国の方が,研究者・投資家・ユーザの信念が一般に強いように思うので,今後,米国の方が,研究のピッチが高まるだろう,と予測する. 日本独自の技術で,米国の研究者も心して見習わなければならない分野は,談話制御. ユーザの発話をたくみに誘導する日本の技術は,きわだっている.

「3つの条件(技術レベル,サポートツール,応

用技術の存在）が必要」…次の3つの条件がうまく嚙み合うことが必要．

Ⅰ．音声認識理解の性能と機能がある程度のレベル（応用による）に達していること

Ⅱ．システムに組み込む人という意味でのユーザが容易に組み込めるためのサポートツールが用意されていること

Ⅲ．音声認識理解を利用することを前提とした応用技術が開発されていること（たとえば，音声認識を前提としたCAI技術など）

この3つが徐々に回転して進むように配慮．始めは音声屋がⅠ，Ⅱ，Ⅲをカバーし，良い発展のサイクルに入れる努力が必要．

「中途半端な知的より単語認識でロバストネスを高める」…知的でさえあれば使われるだろうか．機械に，しかも他人の目の前で話しかけるのは抵抗がある．認識誤りが0にならない以上，確実なボタンの方を選ぶのではないか．（自動改札や銀行のATMでトラブると，使っている方が恥ずかしい思いをする．）

マルチモーダルの一環という考えもなるほどとは思うが，本質的に音声でないとだめという場面がないと，結局使われない．音声の決定的な利点は，

1．ポータブル（キーボードなどはサイズに限界がある）

2．触らなくてよい（手が使えない，装置がどこか分からない場合）

3．電　　話

4．多くの語彙にランダムアクセス可

だと思う．

中途半端に知的より単語認識でロバストネスを高める．数～数十の単語認識（スポッティング）でも使える用途は結構あるんじゃないか．

「メンタルモデルとシステムイメージのギャップを埋める手段が確立していない」…メンタルモデルとシステムイメージのギャップを埋める手段が確立していない．メンタルモデルとは「システムがユーザにどのように見えているか」を表わすもので，システムイメージとは実際のシステムのこと．両者にギャップがあると誤操作の原因になり「使いにくい」と評価されてしまうのではないか．システムの約束事を把握していないことからくる「とまどい」の解消，そして，認識誤りがあった場合でも最後には間違いなく目的を達成できるという信頼感が，音声認識システムの普及に必要．GUIの設計においては「アイコンとマウスによるWYSIWYG（What you see is what you get）」という解が定着．音声インタフェースの設計についても，音声の特性を活かしたうまいモデルはないものか．

4．音声認識はどのように使われるか

今後，音声認識が使われるとすればどのような利点で使われるだろうか．いわば音声認識が拠り所とできるのはどんな点だろうか．筆者なりに整理しておこう．

電話回線の利用：電話回線を通して情報のやりとりができるという利点は，音声認識の重要な効用である．1981年に開始されたNTT（当時，電電公社）のANSERシステムに始まる電話回線経由のサービスは，現在世界的に盛んに開発されている[6]．公衆サービス以外に，まず24時間無休止電話情報案内サービスなどの形で，官公庁，企業，個人で今後も最重要な応用形態として用いられ，やがて電話に自動応対する電子秘書に発展するだろう．

快適さ：慣れないキーボードを使わないで済む，という利点が長く考えられてきた．この点に加えて，巧みな会話やシステムの構成により，コンピュータと快適なコミュニケーションが可能になると考えられる．

他に手段がない場合：手が塞がっている作業などに音声認識合成は有効な情報入出力手段である．特にナビゲーションシステムなどの車載機器への入出力（運転中は目と手が塞がっている）が大きい市場となろう．続いて，福祉・高齢者関係が重要となろう．

効率：従来も音声はタイピングより情報速度が速いといわれてきた．しかし文書入力でのキーボードの置き換えより，むしろ音声コマンドや，図面作成[4]，住所入力[8]などのマルチモーダル入力として情報入力効率を高める方が有望である．

他入力手段との協調要素として：音声入力と他の入力モードとの協調動作（synergy）は重要な面である．「キーボードの代わりに音声認識」ではなく，「キーボードとともに音声認識」を使うことになろう．作業効率を上げるために「使いやす

い」というより「手も声も使え」になるかもしれない．

小型・軽量化：コンピュータはますます小型・軽量化し，究極には携帯電話と融合した携帯端末[2]として音声入出力コンピュータへと進むだろう．情報通信・処理機械を極限まで小型にするには，キーボードも表示パネルもない音声認識合成によるしかない．

低コスト化：将来の究極のコンピュータは，CPU，マイクロフォン，イヤフォン，電池だけからなる，きわめて安価なものになるだろう．むしろキーボードや表示パネルつきが高級品として残り，音声認識入力のコンピュータは「安くて小さいから」使われるだろう．

面白さ：音声認識は本質的に面白いのではないだろうか．ゲーム機などでは，自分で声を出すことですっかりその気になって没入することが考えられる．

ロボットとの対話：古典的なロボットより，むしろサイバスペースなどのバーチャルワールド中の仮想人物と対話をするために音声認識合成が不可欠な技術となるだろう．

5. 音声認識が使われるために何をすべきか

では，上にあげたような音声認識の応用が実際に開かれるために，何をすべきだろうか．以下は筆者の考えである．

◆ロバストネスの向上

まず第一の必要条件は当然ながら性能の向上である．しかし，従来多かった実験室でのトップデータの向上だけでは意味がない．広く現実の音声に対してロバストな認識能力を持つことが実用上の大きな課題である．項目a，b，c，e，f，g，h，qなどにあげたが，ロバストネスへの要求には次のような多様な面がある．

音響的なロバストネス：雑音，マイクロフォン特性，回線特性，反響，残響，ロンバード効果，ガイダンス音声との発話衝突（barge-in）などの実際的な使用環境に対するロバストネスは，最近の音声認識研究の中心的課題になっている．多入力マイクロフォンやエコーキャンセラを始めとする音響的な処理が期待されるとともに，音声認識アルゴリズムのもう一層の進歩が必要である．

話者差に対するロバストネス：話者の間の音声の特性の差はきわめて大きく，真に不特定話者音声認識を行うことは容易ではない．どうしても認識性能が低い話者ができてしまうことが問題である．いまのところ音響モデルの学習音声の話者を増やすこと以外に有効な手立ては見つかっていない．（ところで，不特定話者音声認識とテキスト独立話者認識の両方が可能であるというのは，一見不思議なことではないか．）

発話に対するロバストネス：利用者は，必ずしも想定したように話してくれないので，発話の仕方に対するロバストネスが望ましい．語彙外あるいは文法外の発話のリジェクションも，任意の発話中からキーワードを検出するword spottingも，易しい問題ではない．さらに話者が自由に発話した音声言語は，余剰語（「え－と」「あの－」など），言い淀み，言い直し，言い間違い，倒置，未知語（その場の造語，想定語彙外の語），非文法的な文などの現象を含む．このような自由発話の音声認識は近年の音声認識研究の新たな目標として進展中だが，実用段階は少し先である．

音声入力手段としてのロバストネス：観点を変えて，どのような内容でも最終的に音声入力することができるという意味でロバストな手段として，語彙制約なし，任意の音声・文字変換の技術に対する要望は強い．通常の音声認識では事前に認識語彙のリストを定義する必要があるが，そのためにキーボード入力せねばならないのでは，何のための音声認識か分からない，という意見もある．このような事前語彙定義を必要としない技術は現実的要求である．音声認識だけで高い精度を得るのは難しいが，ペンやカーソルによる修正を組み合わせれば使える可能性がある．

電話を通して情報入力を行うには，ロバストな音声対話技術が重要である．たとえば，「嵯」という字は「山扁に差し引きの差」などと説明することが多いが，このような対話を通して発声者から氏名の漢字を獲得するような，氏名獲得対話エージェントがあれば便利であろう．同様に，住所獲得エージェント，任意文章獲得エージェントなどの対話エージェントの研究が望まれる．

利用者から対話を通して任意の漢字文字列を聞き取ることのできるような対話エージェントが実現できれば，あらゆる音声対話の場面で最後の手段としてロバストな情報入力に用いることができ

よう．

◆マルチモーダル音声入力

音声認識において「マルチモーダル」という語は多義的に使われているが，音声認識と他の手段を組み合わせるものである．形態としては次のような種類がある．音声認識率を100%にすることは困難なので，他手段と組み合わせることは現実的に必要である．

順次使用：音声認識の複数結果候補からマウスクリックで正解を指示する．自由発話大語彙住所音声の認識（番号案内タスク）[3]など．

併存：音声認識，メニュー選択，キーボード入力のどれでも受け付ける．操作者の習熟に応じて使い分けることができる．たとえば，住所入力タスク[8]．画面上の情報入力スロットは，キーボードでもプルダウンメニュー選択でも音声認識でも同等に使える．

分担：たとえば，マウスポインタを頻繁に移動しがちな作図ツールやCAD作業では，マウスポインタは描画に専念し，色，線種，図形の種類，などの属性指定やコマンドなどは，音声で行う[4]ことができる．協調により作業効率が向上する．

協調：他の手段と同時に情報を獲得し，組み合わせることにより情報入力の総合精度を向上する．たとえば仮名漢字変換候補中から音声で選択したり，音声を発声しながら手書き文字を認識，などの可能性がある．音声認識と読唇の協調[9]などもある．

◆音声認識合成ビジュアルエージェント

人間と機械が音声認識と合成によって対話する場合に，機械の側にも人間のような顔がある方が対話がスムーズに運ぶだろうと考えるのは自然なことだろう．最近，そのような「音声認識合成ビジュアルエージェント」あるいはエージェント型マルチモーダルインタフェース（たとえば文献7））の試みが多くなされるようになった．

この考え方は，今後インターネットなどの通信と融合して重要だろう．たとえばオンラインネットワークショッピングで，百貨店を呼ぶと販売員エージェントが派遣され，音声合成と音声認識を介して顧客に商品の説明をして注文をとる．

バーチャルリアリティの世界などでは，ヴァーチュアルな世界の生き物が喋り，人間の声を理解するのが自然であろう．そうすると，音声認識と音声合成の必要性は必然的になる．このような世界の登場人物や生物はマルチモーダルである．触っても，話し掛けても反応する．これらはその世界で動き，話す．これが音声認識・合成の究極の利用形態の1つになり，ゲームや教育などで使われるようになるだろう．

この先にある研究課題は「人工人格」ではないだろうか．音声認識・合成から態度や反応までを複雑にかつ統一性をもって制御するには，何が基本原理なのか．興味の尽きない問題である．

6. おわりに

音声認識はなぜ期待されるほど使われていないか．どうすれば使われるようになるのか．問題点はどこか．このような，音声研究者が皆抱いていた疑問について議論し，音声認識の今後の課題について考えた．

しかし，音声認識は音声認識研究者だけの問題より，すでにそれを応用する側のビジネスコンセプトの知恵比べの時代に入っているといえるだろう．マルチメディアにしても，モーバイルオフィスにしても，PDAにしても，将来の技術構想には音声認識に期待するところが大きい．e-mail討論を行った1年前に比べても，現在は音声認識に対する応用意欲は広く急速に高まってきているように感じられる．「音声認識はなぜ使われないか」という論題は過去のものになりつつあることを願いたい．

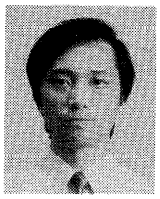
謝辞　音声言語情報処理研究会にてe-mail事前討論およびアンケート調査に参加された方々に深謝します．

参考文献

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) 嵯峨山茂樹：音声認識，日経バイト7月号（100号記念特集号），日経マグロウヒル社，pp. 212-221（1992）．
3) 吉岡　理，南　泰浩，山田智一，鹿野清宏：電話番号案内を対象としたマルチモーダル対話システムの作成，音学講論，1-8-19，pp. 37-38（Oct. 1993）．
4) 西本，志田，山岡，小林，白井：音声・マウス・キーボードを用いたマルチモーダル作図環境，音学講論，1-7-21，pp. 41-42（Apr. 1994）．
5) 嵯峨山茂樹：なぜ音声認識は使われないか・どうすれば使われるか？，情報処理学会研究報告，Vol. 94，No. 40，pp. 23-30（May 1994）．

6) Proceedings of IVTTA 95 (1995 IEEE Workshop on Interactive Voice Technology for Telecommunications Applications), IEEE Communications Society.
7) 伊藤克亘, 長谷川修, 栗田多喜夫, 速水 悟, 田中和世, 山本和彦, 大津展之: 音声・視覚・画像をもつインタラクションシステム, 情報処理学会研究報告 95-SLP-5, pp. 31-38 (May 1995).
8) 荒井和博, 吉岡 理, 嵯峨山茂樹, 山田智一, 野田喜昭, 井本貴之, 管村 昇: 音声認識機能を持つ住所入力システム, 電子情報通信学会 1995 年総合大会講演論文集, SD-9-7, 情報・システム 1, pp. 379-380 (Mar. 1995).
9) Duchnowski, P., Hunke, M., Busching, D., Meier, U. and Waibel, A.: Toward Movement-Invariant Automatic Lip-Reading and Speech Recognition, Proc. ICASSP 95, pp. 109-112 (1995).

(平成 7 年 9 月 14 日受付)



嵯峨山茂樹 (正会員)

1948 年生. 1972 年東京大学工学部計数工学科卒業. 1974 年同大学院修士課程修了. 同年日本電信電話公社に入社. 武蔵野電気通信研究所にて音声情報処理の研究に従事. 1990 年, ATR 自動翻訳電話研究所に移り, 自動翻訳電話プロジェクトを遂行. 1993 年より, NTT ヒューマンインタフェース研究所にて音声情報処理 (音声認識・音声合成) の研究に従事. 現在, 同研究所音声情報研究部音声情報処理方式研究グループリーダ. 著書「自動翻訳電話」(共著) など. 1990 年発明協会発明賞, 1994 年日本音響学会技術開発賞, 1995 年本会山下記念研究賞を受賞. 日本音響学会, 電子情報通信学会, IEEE, AVIRG 各会員.