論文 / 著書情報
Article / Book Information

論題	標準語化すすめる音声認識
著者	古井 貞熙
出典	科学朝日, 1987, 4月号, pp. 41-45
発行日	1987, 4
承諾番号	25-2865
Note	朝日新聞社に無断で転載することを禁じる。

特集　機械は日本語を変えるか 5

標準語化すすめる音声認識

古井貞熙
NTT基礎研究所室長／音声情報処理／ふるい・さだおき

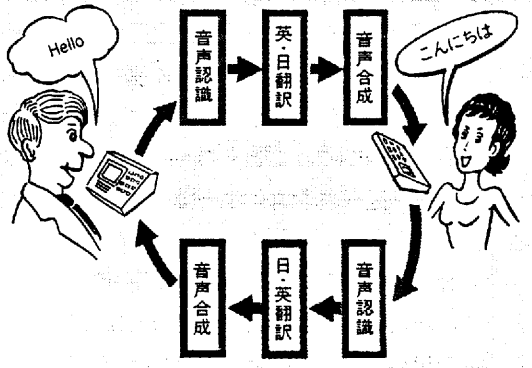
音声は最も手軽なメディア

音声は人間同士でコミュニケーションを行う時に，最も手軽にかつ最も頻繁に使われるメディアである。音声がもしなかったら，いかに不便であるか，容易に想像することができる。文字として書かれた言語も，もちろん知識の交換方法として効果的かつ永続的であり，書物・雑誌などは一方向的情報伝達手段としては極めて有効である。しかし，情報の相互交換の面では，音声に勝るものはない。文字を持たない文化はあるが，音声を持たない文化はない。音声はそれだけ私たちの生活の基本であるということができる。

人間同士が音声で会話をするように，人間と機械（コンピューター）とが音声で対話をすることができたら，どんなに便利だろう。しゃべった音声が文字になるワープロ（音声タイプライター）があったら，ワープロがもっと手軽になるのではないだろうか。キーボードをたたいてコンピューターに命令やデータを入力するのに比べて，音声なら短時間にたくさん入力できる。さらに，電話によって遠隔地から命令することもでき，手や目が自由になるので，移動したり作業をしながら入力できるという長所もある。

究極の夢としては，音声を認識して日本語を英語に，英語を日本語に翻訳し，音声合成して出してくれる通訳電話というシステムも

❶夢の通訳電話。



考えられている（❶）。これが実現できれば，世界中のだれとでも自由に会話ができるようになるだろう。ただし，このように便利に使えるためには，音声認識装置がたくさんの言葉を，しかも文章として続けて発声しても認識できることが必要であり，現在よりも格段に技術が進歩する必要がある。

話し言葉の認識は難しい

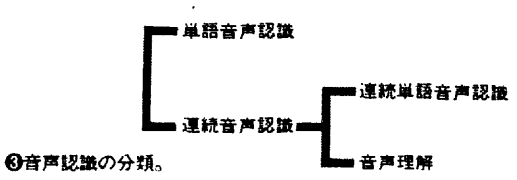
音声認識の基本的難しさは，音声自体の言語としての性質にある。私たちの話し言葉は通常，書かれた文章よりもはるかに文法的に不完全で省略が多く，あいまいさを多く含んでいる。私たちは無意識のうちに，現在の話題，文章全体としての文法的，意味的制約などの言語的情報を考慮して発声し，聞く側も次に話される言葉を予測しながら聞いている。そのような音声を音声認識装置が受け入

れられるようにするためには，音声認識装置に高度の人工知能としての機能を与えることが必要になる。

第２の難しさは，音声の物理的性質にある。私たちの耳は，音声に含まれる周波数成分の移り変わりを聞き取っているといわれている。❷は，自然に「爆音が……」と発声したときの周波数成分（周波数スペクトル）と音声パワー（波形の振幅）の時間変化を示したものである。山の尾根のところ，すなわち周波数成分が特に強いところはフォルマントと呼ばれ，聴覚的に感じられる音色と密接な関係がある。このように私たちが「ば」「く」「お」「ん」「が」という５つの音（音節：日本語の仮名に対応）をしゃべっているつもりでも，物理的にはこれらの明確な時間的境界はなく，連続的なスペクトル変化として現れる。

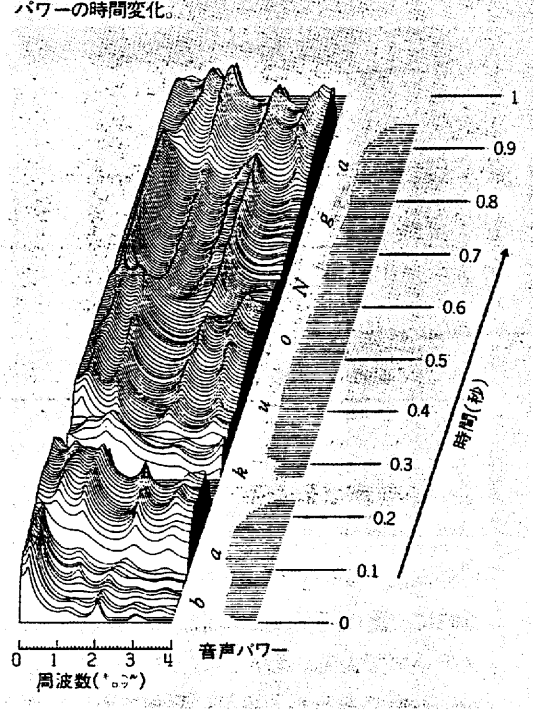
音声は，文字にたとえれば，草書のようなものといえる。われわれが❷のようなスペクトルの流れから，５つの音節を聞き取れるのは，必ずしも物理的区切りがあるからではなく，私たち自身に日本語の音節のきまりが身についていて，それにあてはめて聞き取るからであり，外国の人が聞いたら，これが５つの音節として聞き取られるとは限らない。

音節を連続して発声すると，口の形が連続的に変化するため，隣同士の音の影響を受けてそのスペクトルは大きく変化し，あいまいになる。このような現象は一般に調音結合と呼ばれる。このため，連続して発声した音声の波形の一部を切り出して聞いてみると，その音に聞こえない場合もある。仮に「ステクトル」と発声されても，私たちは「スペクトル」という言葉しかないことを知っているので，だれもが何の抵抗もなく「スペクトル」と聞くであろう。このようなことは，日常の会話では頻繁に起こっている。



❸音声認識の分類。

❷「爆音が…」と発声した時の周波数成分（スペクトル）と音声パワーの時間変化。



音声認識を難しくしているもう１つの要因は，音声の個人差である。同じ言葉を発声しても，発声器官の大きさと構造，その動かし方などの個人差のため，生成される音声の物理的特徴（スペクトル）は発声者によって大きく異なる。自由に発声した音声では，違う人の違う音素（ａ，ｉなどの発音記号に対応する音声の基本単位）が同じスペクトルを持つということも起こる。このため，すべての人に共通に含まれる音素や音節の物理的特徴は，極めてあいまいである。

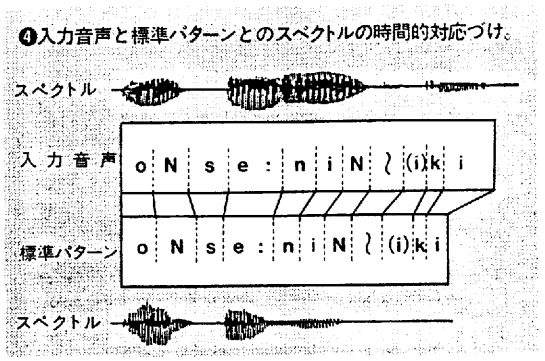
標準との一致度調べが基本

音声認識の形態は，❸に示すように分類することができる。単語音声認識は，区切って発声された単語音声を認識するもの，連続音声認識は，単語を連続して発声した音声を認識するものである。連続音声認識はさらに，比較的少数の用語と限られた文型を対象と

し，発声された言葉をすべて認識しようとする連続単語音声認識と，多数の用語と多様な文型を対象とし，発声者が意図した意味内容を聞き取ることを主眼とする音声理解に分類することができる。人間同士が会話をするときは，あいまいな文章から相手の意図を聞き取るという機能が中心になっていると考えられ，これを実現するためには，高度な音声理解システムの構築が必要である。

音声認識に関する最初の学術論文は，1952年の米国ベル研究所の数字音声認識装置「オードレイ」に関するもので，その後世界各国で盛んに研究が行われるようになった。これまでに実用化された音声認識システムでは，限定された発声者があらかじめ定められた単語を，1つひとつ区切って発声した音声を対象とするもの（単語音声認識）が多い。このように制限すれば，あらかじめ発声者の各単語のスペクトル変化パターンを「標準パターン」として蓄えておき，入力音声とこれらを重ね合わせて比較することによって，どの単語が発声されたかを判定することができる。

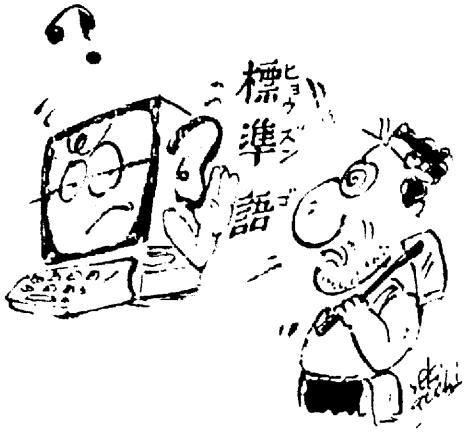
この場合でも，発声のたびにスペクトルは微妙に変化し，言葉の長さは不均一に伸縮するので，蓄えられている標準パターンと入力音声とを❹に示すように時間的に適切に対応づけてから，一致の度合いを調べることが必要である。どの時点とどの時点を対応づけたら最もよく合うかを，すべての可能性について調べたのでは膨大な処理量になってしまうが，動的計画法という手法を使えば，能率よく最適な対応づけが見つかることがわかっている。このようにすれば，数百単語の音声を99%程度の精度で認識することができる。



❹入力音声と標準パターンとのスペクトルの時間的対応づけ。

しかし，銀行における残高照会などのサービスのように，不特定多数の人を対象として音声認識をしたいとの希望も多い。このためには現在では，発声者による音声のバリエーションを調べて，同じ単語について標準的なものを複数用意しておき，入力音声とこれらのすべてとを比較して最も類似しているものを見つけ出すという方法がとられている。

この方法では，多数の標準パターンとの比較を行わなければならないので，認識のための処理量が非常に大きくなる。このため声のバリエーションを確率的に表現するなど，いろいろな工夫がなされているが，いずれにしても高い認識性能を維持するためには，認識する単語の種類を限定することが必要で，16〜20単語程度に限定することが多い。このようにすれば，電話音声の場合に不特定の話者に対して，95%程度の認識率が得られる。



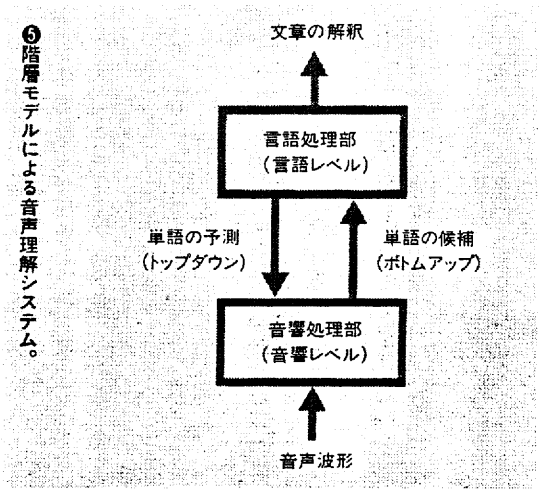
内容を理解するシステムへ

まだ実用化されていないが，最近世界各国の研究機関で音声理解システムの研究が盛んに行われるようになってきた。これは，人間が自然に発声した文章音声を解析して，その意図する内容を理解するシステムを作ろうというものである。このためには，人間が相手

の音声を聞き取る際のように，種々の知識を組み合わせることが必要である。これらの知識の中には，用語，文法，意味，さらに今の話題は何かという文脈に関する知識などがある。人間の会話に生じやすい主語の省略などに関する常識も，システムに教えておく必要がある。

連続した文章として発声すると，単語を単独で発声するときよりも一層音素の特徴があいまいになるので，そのようなあいまいな物理的特徴に基づいて文法的に正しく，意味が通り，話題にあった正しい文章を探索する技術が，システム実現のポイントとなる。あいまいな知識に基づいて，すべての可能性を探索していたのでは膨大な処理量になってしまうので，確からしい情報を手がかりに，いかに能率よく探索範囲を絞り，速く結論に達するかが重要である。

種々の知識をどのように組み合わせるかによって，種々のシステム構成が考えられているが，❺はそのうちの階層モデルと呼ばれる構成である。音響処理部において，音声波形から取り出した音素や音節の性質によって単語の候補を作り出し，言語処理部へ送る。言語処理部では，逆に文法，意味，文脈などの知識に基づいて，可能性の高い単語を予測して音響処理部へ送る。この両者の処理の相互作用によって，最終的に最も可能性の高い文章が出力される。



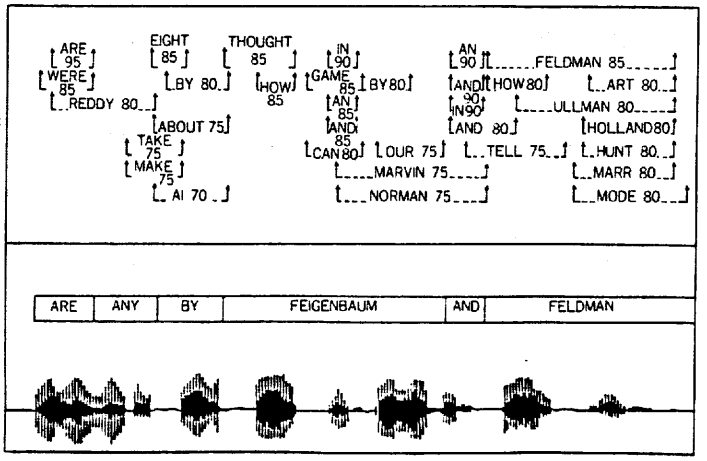
❺階層モデルによる音声理解システム。

❻は「…… are any by Feigenbaum and Feldman……」と発声したときの音声波形と，音響処理部で検出された種々の候補単語を示したものである。音としての物理的性質があいまいなため，たくさんの候補が検出されており，これらに種々の知識を組み合わせて，発声された文章が推定される。

音声理解システムは，人間が持っている極めて複雑かつ高度な機能を工学的に実現しようとするものなので，まだ基礎研究段階にある。しかし，話題を限定することによっていくつかの試作機が作られつつある。

機械には標準語ではっきりと

高度な音声認識システムの実現のためには，音声の持つ３つのあいまいさへの対処法を中心に，今後，研究の大きな進展が必要である。しかし，すでに銀行の残高照会サービスなどのように，音声認識を実際に使いたいというニーズが各方面にあるので，研究者の努力と並行して，大きな不便のない範囲で使用する人間の側がシステムに合わせるという



❻❺のシステムの動作例。aは音声波形、bは入力した文章、cは音響処理部によって検出された候補単語。単語の脇の数字は、確からしさの大きさを表す。

形で，用途が広がってくるものと予想される。例えば音声認識装置に対しては，ある程度はっきり，標準語で，文法的にきちんと発声する，という譲歩である。

これに多くを期待するのは，技術者として怠慢のそしりを免れないが，すでにテレビなどの普及で方言は徐々に少なくなりつつあり，音声認識装置の普及が標準語化をこれから一層進めることも予想される。特に通訳電話で，日本語音声から英語音声に自動的に変換するためには，その日本語がかなり論理的に明確であることが必要だろう。この意味で，日本語がそのよい特徴を失わない範囲で，論理的により明確な言語に変わる必要があるかもしれない。

同時にシステムの側に，自動的に発声者に適応する機能を持たせる研究も重要である。人間には，相手の声の質の変化や話し方のくせ，話題の移り変わりなどにうまく適応する機能があり，これが会話をスムーズに進める上で重要な役割を果たしていると思われる。

話し手の感情を察する研究も

音声認識の別の形態としては，だれがしゃべっているのかの認識（話者認識），発声者の感情の認識なども研究されている。特に前者は，いわゆる声紋鑑定のコンピューター化で，音声を本人か否かの一種のカギとして使うことを目的として研究されている。音声を使って自動的にだれの声かが判定できれば，カードやカギを使うより簡単であり，両手がふさがっていても使うことができ，盗まれる心配がないなど，メリットが多い。情報化社会の進展とともに今後このようなニーズも高まってくると思われる。

後者の感情の認識は非常に難しい課題で，まだほとんど手がつけられていない。将来通訳電話の夢が実現されるためにはこのような研究も不可欠であろう。◀