論文 / 著書情報
Article / Book Information

論題(和文)	音声認識と知識の体系化
Title	
著者(和文)	古井 貞熙
Author	SADAOKI FURUI
出典(和文)	日本音響学会誌, Vol. 59, No. 10, pp. 589-590
Journal/Book name	, Vol. 59, No. 10, pp. 589-590
発行日 / Issue date	2003, 10

音叉

音声認識と知識の体系化*

古井貞熙 (東京工業大学)**

1. 音声認識による知識の体系化

最近，電子メールでの情報交換が日常的になってきて，外国出張でホテルに着いたら，まずインタネット接続ができるかどうかを確認するのが習慣になってしまった。しかし，やはり人と人とのコミュニケーションの基本は音声である。自分の考えを人にすばやく伝える方法としては，多くの人にとって音声が最も容易な方法であると言えよう。テレビのニュース，学会や一般の講演，大学の講義，インタビューなども，音声による情報の伝達が基本である。毎日大量に生まれているこれらの音声ドキュメントを，録音しておいただけでは，後で利用するのが難しい。このため，自動的に音声認識して，文字化し，内容のインデックスをつけることが期待されている。これによって，誰が何と言ったかといった情報や，いろいろな人が持っている知識を，即座に流通させ，活用することができるようになる。

第2次世界大戦中のユダヤ人虐殺の歴史を風化させないために，膨大な被害者がその体験を語ったインタビューを，音声認識して文字化する NSF の「MALACH (Multilingual Access to Large Spoken Archives)」プロジェクトが，米国を中心に進められている。欧州では，過去の種々の分野における，膨大な放送記録を，音声認識によって文字化する「ALERT」プロジェクトが行われている。

これらを正確に文字化するだけでも，大変な努力を要するが，それだけでは不十分で，その内容を要約し，分類し，関連付け，体系化することが求められている。音声は人の思考プロセスと密接な関連を持っている。言葉を覚えるときには，音を伴って覚えるのが普通である。覚えた言葉を混同するときは，音が似ているものに間違えることが多い。新しい言葉や知識は，長期記憶にある意味的概念と結びついて，覚えられる。人の思考と密接に関連した音声を中心に，大規模な知識をいかにして構築し，体系化し，活用し易くするかが，21 世紀の重要な研究課題となるであろう。

2. 近年の音声認識の進歩

現在のコンピュータによる音声認識では，音声が与えられたときの，事後確率を最大化する単語列（形態素列）が選ばれる。この背景には統計的パターン認識理論があり，事後確率の最大化によって，音声認識誤りを最小化することができるからである。事後確率を直接計算するのは通常難しいが，ベイズの定理によって，音響モデル（音のモデル）に基づく音声の尤度と，言語モデル（語彙と文法のモデル）に基づく単語列の確率の積に変換することができるので，この積が最大となる単語列を選べばよい。音響モデルとしては隠れマルコフモデル（HMM），言語モデルとしては単語の連鎖確率を用いるのが普通である。これらの統計モデルは，多数の人が発声した音声や，大量のテキストから学習する。

アナウンサが原稿を読み上げているような音声であれば，このような方法で，95%程度の精度で音声認識できるが，考えながら自由に発声した話し言葉の音声認識精度は，まだ 70%程度である。原稿を読んでいる音声や，語

* Speech recognition and knowledge systematization.

** Sadaoki Furui (Department of Computer Science, Tokyo Institute of Technology, Tokyo, 152-8552) e-mail: furui@cs.titech.ac.jp

彙が限られた一問一答型の音声に対してだけ音声認識できたのでは，音声認識の用途は広がらない。そこで，話し言葉の音声認識精度を上げようとすると，当然ながら，話し言葉に合った音響モデルと言語モデルが必要になる。話し言葉には，間投詞，言い直し，繰り返し，言い間違い，あいまいな発声などが頻繁に含まれるので，これらに対処できる言語モデルをいかに作るかが問題である。

言語モデルつまり単語の連鎖確率を計算するためには，話し言葉を忠実に単語列で表したデータベースが不可欠である。書き言葉のテキストは，新聞をはじめとして，電子的なものが大量に入手できるが，話し言葉に対しては，大勢の人の音声を録音して，人手で書き起こし，単語の区切りを入れる必要がある。これには大変な労力を要するが，1999 年から進められている国の「話し言葉工学」プロジェクトで，そのデータベース作りが進んでいる。2004 年の春までには，世界最大の話し言葉データベース（CSJ; Corpus of Spontaneous Japanese）ができる予定である。

3. 音声認識のための知識の体系化

しかし，話し言葉のデータベースが構築されても，それを用いて，これまでの方法で音響モデルと言語モデルを作成するだけで，話し言葉の音声認識が完璧にできるようになるとは考えにくい。それには種々の理由があるが，最も根本的な理由は，それが，人が音声を認識するときの原理とは大きく異なるためである。人が音声を認識するプロセスは，その人が持っている知識との対応付けである。外国語の音声を聞き取ろうとする場合に経験するように，理解できないことを認識することはできない。

これまでの音声認識では，上で述べたように，発声内容の意味に立ち入ることなく，音響的特徴と，語彙と文法だけで，音声を文字に変換し，その結果から意味を抽出するというプロセスがとられてきた。一方，人が音声を聞き取るときには，その人が持っている知識を総動員し，その知識との関連において，音声の認識が行われる。まずどんな分野や話題に関連した話なのかを判断し，それに関する知識との結び付けが行われる。聞いたことがない新しいことが話されても，それ以外の理解できる部分との関わりにおいて，その未知の部分が処理される。想像をたくましくして理解あるいは誤解に至ることもあるし，話し手に質問することによって理解に至ることや，しばらくほうっておかれることもある。結果として，分かっているつもりだったのに，しばらくしてから，実はとんでもない思い込みで，全く誤解していたということも起こる。

聞き手が予想できる内容であれば，相手の音声を認識するのは簡単なので，耳を澄ます必要もないが，知らないことであれば，発音，文法などの情報のみならず，多少なりとも関係すると思われる知識が，注意深く使われる。コンピュータによる音声認識でも，このような知識との関連付けとしてのアプローチをとる必要がある。このためには，その場の状況を含めた多様な関連知識を体系的に蓄積し，活用できるようにしておくことが不可欠である。このようなアプローチをとることにより，言い直し，繰り返し，言いよどみなどの影響を受けにくい，従来法よりもフレキシビリティのある，音声認識が実現できると期待される。これができれば，日本語に限らず，あらゆる言語の音声認識に共通に適用できるであろう。

これまでいろいろな分野で多様な知識（データベース，コンテンツ）が作られてきたが，それらは個別知識にとどまっているものが多く，互いに関連付けることが難しい。人が持っている共通の知識を体系化するためには，意味概念を体系化することが必要で，人の能力に近づくには多くの困難が考えられる。なかなか超えられない壁があるかもしれない。人が簡単にやっているように見えることでも，それをコンピュータで実現するのは，しばしば難しい。しかしこのような研究が，21 世紀においては，極めて重要になると考えられる。