音声認識システムの最前線

放送大学学園理事長
早稲田大学名誉教授
白　井　克　彦

人間が音を使ってコミュニケーションするということは、どのようなメカニズムなのかは古くからの関心であった。音声科学の研究が盛んになって、音声生成の物理、発声のメカニズム、聴覚機構の解明、音声の知覚などの基礎的研究が進められてから、おそらく70年以上の時間が経過した。勿論、言語学に連なる音韻論、音声学などは、さらに古い研究の歴史がある。また、最近の認知科学や脳科学の発達も音声言語の研究に新しい発展をもたらしている。

しかし、情報通信技術が発展をはじめると、すぐに電子的手段で人間の音声を生成する音声合成と人間の声を自動的に認識する音声認識技術が研究されるようになった。ここでは、音声科学の知見が工学に結びつくと同時に、新しい問題提起や音声科学に新しい研究手段をもたらすことになった。

この分野の研究の特徴は、人間が音声をどう使うかという生理学的機構と言語が社会的にどう用いられるのかという、二つのいずれも人間の極めて本質的な部分に深くかかわる広がりの大きな技術であるということである。

他方で、様々な機械が日常生活に入り込むにつれて、人間と機械の情報交換をよりスムーズに人間にとって自然なものにしたいという欲求が強まってきた。その結果、1960年代には電子的手段を用いた音声認識や音声合成の研究が開始された。ところが、当時の真空管を用いたアナログ回路でできることは本当に単純なものしかできず、日本語でいえば、「あ」「い」「う」「え」「お」５母音の識別が、ようやくできる程度であった。つまり、人間が日常的に用いる音声言語の音声認識を実時間（リアルタイム）で処理できるようになるには、コンピュータの圧倒的な大規模化、高速化が必要であった。

もう一つ音声認識を実用的にするには、大きな困難性がある。それは、人間はあまり意識もせずに行っていることであるが、人間が音声コミュニケーションであつかっている、音声言語空間の大きさである。日本語音声を認識して文字表記にしたいと考えたとすると、まず、同じ母音の「あ」という音でも、発声する人によって色々異なるし、同じ人でも、発話文の中で大きく変化する。また、使う単語や文表現には大変多くの利用の幅があり、変化がある。これらの多種多様な変動は、統計的に把握する以外ないものである。したがって、音声認識を実用的にするまでには、大量の様々な音声データが必要であった。今日、世界の様々な言語の音声データがクラウド・サーバー上に出現することになった。そのデータを統計的に分析することで、まず基幹の音声認識システムを構成することができ、音声言語の広い変動に対処できるようになってきた。さらに、最初に考慮されなかった、未知の音声データもクラウド上を検索することで正しく認識される可能性もでてきた。

本特集では、日常的な話し言葉に対して、音声認識がどの程度可能で、応用がどのように広がっているかがわかる。近年の技術の進歩は、確かに著しいが、人間の音声認識能力に比べれば、まだまだ多くの技術的課題も指摘されている。