Communicating with Computers by Voice

Researchers are finding that getting computers to talk is easier than getting them to listen

Computers would seem much friendlier if it were possible to verbally ask them for information or command them to carry out a task, and if they could respond by talking in natural, conversational English. Considerable commercial success has already been achieved with machines that recognize individual words spoken in isolation (that is, with pauses between them), and devices that can convert an arbitrary text into speech are appearing on the market. However, the prospects for automatic recognition of fluent, continuous speech in the near future are not encouraging, in part because of the complexity of the task and in part because of a dry spell in research support.

This is the second of two articles on the status of man-machine communication by voice.

A fundamental question in automatic speech recognition by computers is how much information is contained in the acoustic wave patterns that make up the words and sentences used in communicating by voice. In other words, can the machine decipher a voice command and act on it simply by decoding the wave pattern or does it need to "know" certain additional information? In all cases, the answer seems to be that extra knowledge must be programmed into the computer.

This knowledge falls into one of two categories. In the case of machines that can recognize words spoken in isolation, the knowledge is highly artificial: the machine knows that the acoustic wave pattern corresponds to only a single word, which can be decoded by a statistical pattern matching technique (*Science*, 16 February, p. 634). But for understanding the natural, continuous, fluid speech of humans, considerable knowledge about the structure of language and about the context of the task the computer is to perform is mandatory. It is the difficulty of devising effective strategies for imparting this type of information to the computer that makes speech recognition so hard.

Speech recognition by a computer comprises two operations. The first is the decomposition of acoustic wave patterns into the basic sounds or elements of speech, called phonemes. Existing speech recognition machines do this job with no more than 65 percent accuracy, in part because the wave patterns corresponding to neighboring phonemes tend to overlap. The second operation is to arrange the phonemes into words and the words into sentences, taking into account that a substantial fraction of the phonemes may be misidentified, and it is here that the need for linguistic and contextual knowledge arises.

Learning how to incorporate these capabilities into speech recognition machines is the province of a group of computer scientists, linguists, speech scientists, and psychologists working within the wider field of artificial intelligence. The kinds of knowledge they deal with include syntactic (so that the sentences arrived at are grammatically correct), semantic (so that the sentences have a logical meaning), and pragmatic (which includes considerations such as context so that other than strictly literal interpretations are made). The cartoon illustrates one of the pitfalls awaiting those who would talk to computers, which tend to take things literally.

If computer understanding of continuous speech is a dream yet to be realized, the inverse process of computer generation of speech is much closer to being realized. There is, in fact, a commercially available system made by Kurzweil Computer Products, Inc., Cambridge, Massachusetts, that combines an optical character recognition device with a voice synthesizer to make a reading machine for the blind. A book is simply laid flat over a glass plate as with a copying machine, and the system converts the text into spoken words. The first version of the Kurzweil blind reader became available 2 years ago, was expensive (about $50,000), and had a sound quality judged by some observers to be just adequate for a motivated person to be able to understand what was said. A new model with a much improved sound and a lower price ($19,000) is just now coming onto the market. A glimpse of the future is surely captured in Kurzweil's blind reader.

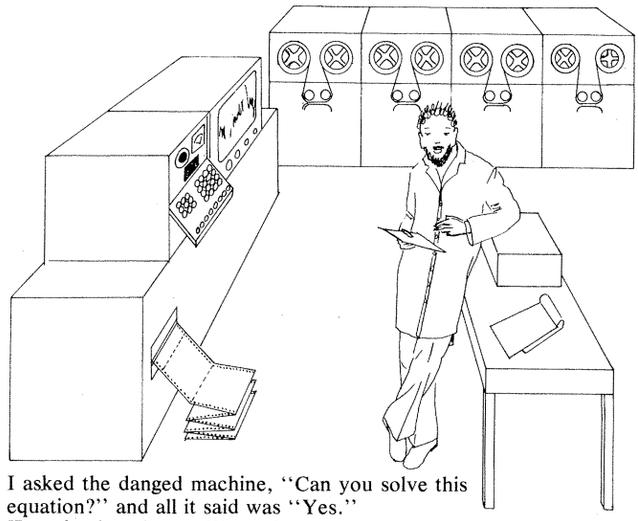
The difficulty in producing natural sounding speech stems from many of the same factors that make continuous speech recognition a problem. One could, for example, store phonetic representations of all the words in the dictionary in the computer. When the machine needed to generate a sentence, it could call up the pronunciation for each word and string the words together into a spoken sentence. However, not only would the sentence sound jerky, but the numerous factors that change the sounds associated with words or segments of words when speech is continuous, as compared to the pronunciation of the isolated word, would all come into play to wreak havoc with such a scheme. Moreover, according to Jonathan Allen of the Massachusetts Institute of Technology (MIT), in order to construct a good speech synthesis system, one must in effect develop a model of how a human reads aloud. Thus, knowledge about syntax, semantics, and context becomes important as well.

A way of limiting the complexity of speech synthesis systems due to the need for linguistic and contextual knowledge is to limit the scope of the task. Peter Denes, Mark Liberman, and Joseph Olive at Bell Laboratories are taking this tack in their work on an automated directory assistance system for telephone users. An experimental system now in operation at Bell uses a prerecorded voice to give the numbers and locations of laboratory employees, but the computer memory required to adopt such an approach for the directory of a large city is so huge that speech synthesis is necessary. In this application, the speech generated would be restricted to the form, "The number of Arthur L. Robinson of 1515 Massachusetts Avenue is 467-4326." Only the name, address, and number will be synthesized; the remainder of the reply, since it never changes, will be prerecorded and stored in the computer for playback as needed.

Allen of MIT has had success in unrestricted text-to-speech synthesis with a system he has been developing for sever-

0036-8075/79/0223-0734$00.50/0

al years. Allen's method relies primarily on rules for decomposing words into elements called morphs, which include word roots, prefixes, and suffixes. The MIT system incorporates a dictionary of pronunciations that now encompasses some 11,000 morphs. Allen says that about 95 percent of the words in an arbitrary text can be analyzed in this way. For the other 5 percent, a set of letter-to-sound rules are needed to convert words that cannot be decomposed into morphs directly into phonemes.



I asked the danged machine, "Can you solve this equation?" and all it said was "Yes."
[Drawing by Eleanor Warner]

In order to obtain the correct stress, pitch, and timing of words in the speech output, a syntactic analysis of phrases (not necessarily whole sentences) is carried out. The same algorithm that computes the pitch and timing also computes a set of parameters that control the output of a digital speech synthesizer. The synthesizer converts the parameters into an acoustic wave pattern—that is, sound.

One way to generate a set of parameters is to use a model that attempts to duplicate the motions of parts of the human vocal tract, such as the tongue, the lips, and the soft palate, as has been done by Cecil Coker of Bell Laboratories. The approach taken at MIT, however, is more abstract, and the parameters correspond to the properties of an electrical circuit model of the vocal tract. Nonetheless, the method is effective, and the sound quality of the MIT system is said to be so good that some listeners have thought they were hearing a human speaker.

Allen's system consists (as do almost all research systems for speech recognition and synthesis) of a set of computer programs to be run on a general purpose computer, as opposed to a special piece of hardware designed for speech processing. Researchers at Telesensory Systems, Inc., a Palo Alto, California, firm that specializes in devices to aid the blind, are hard at work reducing the MIT speech synthesis system to a practical form. The end product, which may be available in about a year at a price of "less than $10,000," is a blind reader somewhat like the Kurzweil product.

The Telesensory System blind reader will be an accessory for its now widely used Optacon, a device developed at Stanford University that enables the blind to read a text. The Optacon employs a set of vertically moving pins that form patterns on the user's index finger corresponding to the letters in the words sensed by a hand-held optical character recognizer. By use of the Optacon, a blind person will be able to select the portion of a page that is of interest without having to hear the entire page read.

The heart of the new product will be a microelectronic circuit specially designed to handle the speech processing part of the text-to-speech conversion—that is, the generation of the acoustic wave patterns from the phonemes. According to James Bliss of Telesensory Systems, by simplifying the MIT system it has been possible to considerably speed up its operation so that "real time" response is now possible without a great reduction in the quality of the speech output.

The more difficult problem of understanding continuously spoken words and sentences has a much less certain future and has had a checkered past. At one time speech understanding was criticized as impractical. One of the more influential knocks came from John Pierce (then at Bell Laboratories) in a letter to the *Journal of the Acoustical Society of America* 10 years ago in which he questioned the need for speech recognition machines and mused about the domination of the field of speech recognition by "mad scientists and untrustworthy engineers."

On the upswing of what is taking on the appearance of a boom or bust cycle, continuous speech recognition got a big boost in 1971 when the Defense Department's Advanced Research Projects Agency (ARPA) began a 5-year, $15-million program to develop a speech understanding system that could handle, with 90 percent accuracy, a multiplicity of speakers (men and women) talking from a 1000-word vocabulary and using an artificial syntax. Artificial syntax means that only certain combinations of words that would be appropriate for a specific task, such as making airline reservations, are allowed.

Wayne Lea of the Speech Communications Research Laboratory, Los Angeles, has pointed out that the ARPA project marked the first major attempt to link earlier speech recognition work, which was done primarily by electrical engineers versed in signal processing, with the idea of incorporating linguistic and contextual information into a system. Observers have commented that not a little bad feeling was generated when the agency seemed to give most of its support to those whom old timers in speech processing considered to be fancy dan buttinskies from the artificial intelligence community. However, says D. Raj Reddy of Carnegie-Mellon University, more than half of the ARPA funds actually went to researchers more properly classified as speech scientists than as artificial intelligencers.

The ARPA project proceeded in two stages. Those contractors judged to have made the most progress in the 2-year-long first phase were selected to complete systems to be ready for testing by the program deadline near the end of 1976. Four groups were chosen for the final phase: Carnegie-Mellon University; Bolt Beranek and Newman Inc., Cambridge, Massachusetts; and (jointly) System Development Corporation, Santa Monica, California, and SRI International, Menlo Park, California. Only one speech recognition system, called HARPY by its creators Bruce Lowerre (now at Systems Control, Inc., Palo Alto, California) and Reddy, met all the project goals. Most observers consider, however, that the project was successful in that considerable basic speech science progress was made and that participants learned how to integrate previously widely scattered bits and pieces of computer science, speech science, and linguistics into a working system.

The deleterious effect of the termination of the project by ARPA is less debatable than the degree of its success. Carnegie-Mellon has had other sources of support and has continued research, although at a lower level of activity, whereas the other three groups, contractors that primarily serve federal agencies, could not continue their projects. The largest continuous speech recognition effort is now at IBM's York-

town Heights laboratory, which has been a double maverick because it did not participate in the ARPA project and because researchers there have eschewed the artificial intelligence approach in favor of a method based on statistical communication or information theory.

Where the various speech recognition systems diverge is in their strategies for incorporating the use of linguistic and contextual information in deciphering a speech sample. Following the derivation of the phonemes making up the spoken words by a statistical pattern matching technique, the HARPY system applied the knowledge in a particularly simple way. For its limited task, which was document retrieval, HARPY determined in advance the strings of phonemes corresponding to all the possible sentences it might be asked to understand. Beginning at the left end of the sentence, the system compared the degree to which the phonemes stored in its memory as references matched those it "heard." As the analysis proceeded through the sentence, word by word, HARPY selected as candidate sentences for continuation only a set of those with the best matching scores up to that point in the analysis.

It would be possible to miss the correct sentence if, for some reason, the right word at some point in the sentence were to receive a very low score, but the computational task was reduced to a manageable one (not all possible sentences had to be evaluated) and the method was effective. HARPY correctly identified 91 percent of sentences from three male and two female speakers using a 1011-word vocabulary.

A second speech understanding system at Carnegie-Mellon was designed to be more readily adaptable to tasks requiring large vocabularies and tasks other than information retrieval. But the system, called HEARSAY II by its builders Lee Erman, Rick Hayes-Roth, Victor Lesser, Reddy, and their coworkers, was less accurate than HARPY and took much longer to accomplish the speech processing.

Perhaps the most ambitious of the speech recognition systems coming out of the ARPA project was the HWIM (Hear What I Mean) system built by William Woods and his colleagues at Bolt Beranek and Newman. HWIM had the least constrained syntax of any of the systems. Constraint is sometimes measured by the average number of words that are allowed to follow any given word in the spoken sentence. HWIM also relied most heavily on the use of linguistic and contextual information of all the systems and can almost be said to have reversed the usual recognition procedure by returning to the acoustic signal to verify hypotheses made on the basis of the linguistic sources of knowledge. This approach is called analysis by synthesis by its originator, Dennis Klatt of MIT.

But HWIM also performed less well, understanding only 44 percent of the sentences in its travel budget management task and running about 100 times more slowly than HARPY. How good or bad a system HWIM was was never completely evaluated, however, because its creators were making major changes in it up to the day before testing took place. As a consequence, none of the fine tuning and bug removing that was possible on the simpler HARPY was carried out.

The fourth ARPA speech understanding system, that from System Development Corporation and SRI International, was never fully evaluated because the former organization lost part of its computer facilities before the contributions of the two parties were melded into one system.

IBM's interest in speech recognition has been with an eye toward automating business offices with products such as a typewriter/computer that could accept dictation and turn it into a draft of a letter, perhaps even a final version, thus eliminating the need for someone to take shorthand or transcribe a recorded message. The group of researchers at the Yorktown Heights laboratory, headed by Frederick Jelinek, has concentrated on a statistical method that, although not directly using linguistic knowledge sources, incorporates a similar kind of information. For example, if a word is a particular part of speech, such as a verb, then there is a certain probability that the following word is another part of speech, such as an indefinite article and so on. By analyzing a large number of test sentences, these and other kinds of probabilities can be estimated. The selected sentence is the one that has the highest total probability as measured in some appropriate manner.

Jelinek and his colleagues chose this statistical method in order to be able to recognize arbitrarily constructed sentences—that is, sentences that do not need to fit into an artificial syntax designed for a specific task. The IBM group has enjoyed some success with this approach, and their system is now able to recognize about 75 percent of the words in test sentences made of words found in a lengthy U.S. laser patent. In a much easier test with a constrained syntax that Jelinek estimates is half again as complex as HARPY's, 100 percent recognition has been achieved.

Most researchers do not now believe that, even if they were given a roomful of supercomputers and an unlimited budget for using them, it would be possible to make a system that could understand unconstrained, natural speech with high accuracy.

Clearly, much more remains to be done. Most researchers do not now believe that, even if they were given a roomful of supercomputers and an unlimited budget for using them, it would be possible to make a system that could understand unconstrained, natural speech with high accuracy. What speech understanding researchers would like is another ARPA project, but one with a firmer prospect for stable, long-term support in place of a one-shot extravaganza.

Unfortunately, there is no clear sign of revival of funding of that type, although ARPA is far from having lost interest in speech understanding. The agency is said to still have speech understanding in mind, but in the context of an advanced access system to computer networks such as ARPANET. Such a system would incorporate voice as well as graphics and keyboard inputs. Moreover, the system would be very "knowledgeable" about the operation of the network because the expertise of computer scientists would be programmed into the system, and it would be able to use that expertise to respond to queries made in natural English. In this way, access to the network would be made easier for non-computer-oriented users.

—ARTHUR L. ROBINSON

Additional Reading

1. D. H. Klatt [*J. Acoust. Soc. Am.* **62**, 1345 (1977)] reviews the ARPA speech understanding project.
2. See additional readings at the end of the first article in this series [*Science* **203**, 634 (1979)].