More People Are Talking to Computers as Speech Recognition Enters the Real World

Machines with limited capabilities have a big future

Forecasts of the impending computerization of society tend to focus on the ever decreasing cost of computer power that makes new applications economically feasible where once the expense was too great. Just as important in deepening the computer's penetration into our daily lives, however, is the prospect of communicating with a computer through human speech. All computer users agree it would be highly advantageous to be able to talk to one's computer in a conversational way. A hotly debated question is: How natural does voice communication with a computer have to be to be useful? Surprisingly, the answer is not very.

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

The issue is being resolved by researchers and engineers in numerous companies who are producing speech recognition machines with limited capabilities that can be bought off the shelf today for a slew of applications, mainly industrial and military so far. All projections are that computer speech processing is on the verge of becoming a booming business, even if the personable computer epitomized by Hal in *2001* remains a dream that is decades or more away.

One reason that development of a voice communication capability may be imperative for expanding the purview of the seemingly already all-pervasive computer is in part psychological. Many people will prefer—and thus be willing to purchase—a machine that they can talk to (and that can talk back) to one that must be communicated with by way of an inflexible, artificial computer language and a typewriter-like keyboard.

Another reason is efficiency. Experiments by Alphonse Chapanis and his colleagues at Johns Hopkins University, for example, have shown that when two persons work together to solve problems, they find solutions roughly twice as fast when they communicate with speech as compared to the time taken when they communicate by other means, such as visual signaling, typing, and handwriting.

A second source of efficiency, at least in the industrial and military worlds now inhabited by speech recognition machines, is that the ability to convey information by talking frees a person to do other tasks simultaneously, as in sorting or inspection operations. Because of the now widespread use of remote computer terminals to gather information for record keeping or for direction of computer-controlled machinery, a workman has to determine such information as the destination of a sack of mail by inspection and then has to turn to enter the information on a keyboard. But he can perform both operations without moving by reading into the microphone of a speech recognition machine.

Computer speech processing actually encompasses several categories of activities. One is the use of digital electronics to compress voice signals so that many more of them can be sent over a single communications channel, such as a telephone line. To actually communicate with computers requires speech synthesis or generation typified by the highly regarded Texas Instruments hand-held learning aid, Speak and Spell (see box), and speech recognition, which is a much more difficult problem than speech synthesis and for which results are not as far along.

The source of the difficulty lies in the complex acoustic wave patterns associated with the sounds of the human voice. The pattern for any given sound consists of a superposition of many waves with different frequencies and amplitudes. Moreover, the pattern is stable only over periods of about 10 milliseconds, so that the frequencies and amplitudes making up the pattern vary constantly. An additional complication is that the sequence of patterns associated with a sound depends on such variables as where in a word it appears and what the sounds in the following or previous words are (Fig. 1). Finally, not only do physical differences between people cause differences in their speech, but a given person will speak differently when suffering from a cold, when in a highly emotional state, when very tired, and so on. Since the basic recognition act consists of a comparison between the acoustic wave patterns made by a speaker and reference patterns stored in the machine's memory to find the best match, to say the task is formidable is an understatement.

On top of the sound recognition task is the job of converting a string of sounds into words and the words into meaningful sentences. More than mere pattern matching is required to translate sounds into sentences. The machine must "understand" what every human more or less knows: the basic rules of grammar, syntax, and semantics. Incorporating such knowledge into computers is part of the domain of artificial intelligence, which draws upon the talents of linguists and psychologists, as well as computer scientists, and which has had limited success. (The speech processing side of artificial intelligence will be the subject of a subsequent article.)

An alternative course, followed by researchers and engineers more attuned to the electrical engineering discipline of signal processing, is to attempt to match the acoustic wave patterns of entire words spoken with well-defined pauses between them. Placing periods of silence between the words means that the machine can more accurately tell when a word begins and ends. Words spoken in isolation also do not exhibit the changes in the sounds that take place when words are spoken continuously, as in "Did" "You" versus "Dija." A disadvantage is that there are many more words (tens of thousands) than basic sounds (40 to 60, depending on the means of classifying them), and existing computer memory devices are too expensive and computer processing is too slow to permit matching patterns for such large vocabularies.

Although isolated word speech is somewhat unnatural and requires concentration on the part of the speaker, machines to recognize isolated words can be built with today's technology. The current excitement in speech recognition has been generated by those with an inclination to make the maximum use of existing technology and an orientation toward making salable products, and by agencies funding work in this area.

0036-8075/79/0216-0634$01.00/0

Speech Is Another Microelectronics Conquest

Christmas 1978 was widely touted as the season of the electronic toy. One of the more engaging of these is a handheld learning aid made by Texas Instruments called Speak and Spell. Perhaps the most intriguing aspect of Speak and Spell is that professionals in computer speech processing do not regard it as a toy—far from it. Speak and Spell is seen as a foot in the door for microelectronics in synthesis and recognition of speech by computers or man-machine communication by voice. Forecasters are saying the same potential for miniaturization and low costs that microelectronics brought to computers and consumer electronics, when applied to computer speech processing, will vastly increase the now somewhat limited range of applications for machines that can talk and that can listen. One fearless prognosticator foresees that every computer will be able to talk within 5 years.

Speak and Spell, which actually is more like a portable cassette tape recorder than a calculator in size, asks in a quite natural sounding voice that a specific word be spelled. The user "keys in" the answer by way of an array of push buttons labeled with the letters of the alphabet. Finally, the machine congratulates the speller for a correct answer or encourages another try after a wrong one.

Within the device are four microelectronic circuits. A microprocessor controls the operation, including selecting the word to be spelled and determining the correctness of the proffered answer. There are two memory chips that store the information needed for the synthesis of about 3.5 minutes of speech. But the most exciting component is the fourth chip which contains the circuitry for speech generation by the method of linear predictive coding (see main story). In linear predictive coding, the acoustic wave pattern associated with the sound of speech is represented by a set of parameters (ten in the case of Speak and Spell). The advantage of such a technique is that a rather modest amount of memory can store an impressive amount of speech. If the words used in Speak and Spell were represented simply by a digitized version of the appropriate wave patterns (pulse code modulation) the device could have stored only 4 seconds of speech, for example.

The words offered up for spelling were originally spoken by a professional actor, and, in some sense, the speech synthesis is only a very sophisticated form of playback of recorded speech. But the speech signal processing circuitry used in implementing the linear predictive coding, according to Robert Broderson of the University of California at Berkeley, is also central to other forms of speech processing, such as synthesis of speech from text (where no human speaker is involved) and speech recognition. This is the source of researchers' interest in Speak and Spell, which now retails for $50. The price of most commercial speech synthesizers is measured in thousands of dollars.

Texas Instruments' speech synthesis chip is actually part of a much larger development in microelectronics that is being stimulated in part by the conversion of communications systems, such as the telephone, to an all-digital format. Many microelectronic circuit manufacturers are coming out with a device, called a codec, that will one day appear in every push-button telephone for the purpose of converting the analog speech signal generated by a human being into a digital form. The next generation after that, although it may find use in more restricted applications, is a device called a vocoder. A vocoder, as originally developed a quarter of a century ago at Bell Laboratories, accomplishes the same task as a codec but with the added benefit of compressing the speech so that many fewer binary bits per second must be transmitted when a message is sent. Thus the same transmission line can carry many more messages. Vocoders are based on techniques like linear predictive coding, and Texas Instruments recently has reported implementing a vocoder on two chips, one for synthesis and one for analysis—that is, one for conversion of the compressed digital signal into an analog speech wave pattern and the other for the reverse task.

The miraculous benefits that miniaturization brings to microelectronics are most often expressed in terms of digital devices: microprocessors that place the computing power once requiring a roomful of electronics onto a single semiconductor chip and ultradense computer memory chips that store tens of thousands of binary bits in a rectangle a few millimeters on a side. But speech signal processing of the type performed in vocoders requires that at least some of the work be in an analog form. One strategy would be to convert the input speech pattern immediately into digital form and accomplish all the signal processing in a digital mode. At present, this approach is computationally inefficient because analog devices can perform operations such as Fourier transforms a thousand times faster than can digital ones (*Science*, 18 March 1977, p. 1216). And, points out Broderson, with existing fabrication techniques, engineers can make analog microelectronic chips, each with considerable processing power.

One solution, according to George White of the ITT Defense Communications Division, San Diego, is therefore to implement a speech processing system on more than one chip, the first for the analog signal processing stage and the second for the digital pattern classification stage. Since this is exactly the tack taken by Texas Instruments in its Speak and Spell, therein lies the reason for regarding this product as a forerunner of things to come.

What will all this lead to? Texas Instruments' George Doddington says that the main issue now is not what will be possible but how fast will prices come down to make new applications affordable. One thing that could limit the speed at which the price of the new speech processing chips drops is the high cost of developing them, several hundred thousand dollars for each circuit design. The large investment required necessarily limits developing such circuits to companies with the resources to swing the effort. Going somewhat less far out on a limb, ITT's White forecasts that the two-chip, analog-digital strategy will be the result of a slow evolution of capability culminating, in 20 years or so, in a microelectronics version of a speech recognition machine that can understand a moderate amount of continuously spoken speech and direct a computer to carry out the spoken commands.—A.L.R.

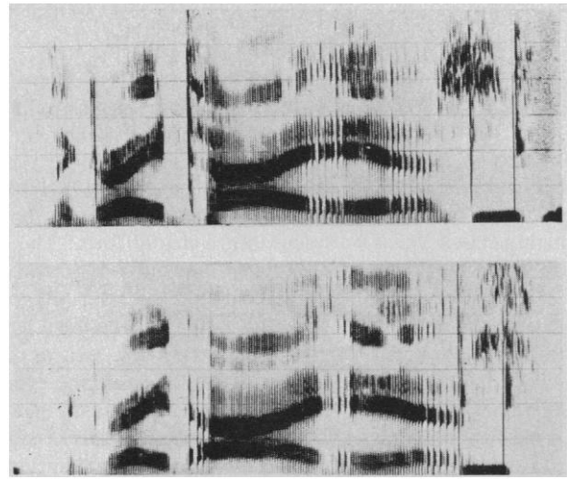

Fig. 1. Spectrograms for the phrases "gray tie is" (top) and "great eye is" (bottom). Spectrograms provide a display of the energy in the speech wave in terms of frequency (ordinate), time (abscissa), and intensity (darkness of the trace). The dark bars represent vocal tract resonances. The two spectrograms are similar except for the difference (center left) between the stressed t of "tie" and unstressed t of "great." [Source: C.H. Coker, in *Proceedings of the IEEE*]

J. Michael Nye of Marketing Consultants International, a Hagerstown, Maryland, consulting firm, estimates that the leading maker of isolated word recognition machines, Threshold Technology of Delran, New Jersey, has sold more than 250 units since opening for business several years ago, and another company, Interstate Electronics of Anaheim, California, may have marketed 40 or more word recognizers. These machines are priced from $10,000 to $80,000. Thomas Mandey of Quantum Science Corporation, a New York City marketing research organization, predicts that the price of isolated word recognizers could drop to about $5000 or less this year, which is, he says, the critical level below which sales could grow significantly. (Isolated word recognizers are available for as little as $300, but observers regard these as too limited in accuracy and flexibility for commercial applications.)

It is well known that large corporations, including Texas Instruments, ITT, Sperry Univac, and IBM, are looking hungrily at a perceived large demand for word recognition equipment. One prospect, unsettling for those small companies now dominating the speech recognition industry, is that the giants may take over as they did in hand-held calculators. Recently an advanced device based on an extension of currently available techniques and capable of a limited amount of continuous word recognition has been introduced, ominously enough for American manufacturers, by one of Japan's largest electronics companies, the Nippon Electric Company (NEC).

Numerous techniques exist for accomplishing the acoustic wave pattern matching needed for recognizing words. A simplistic one would be to directly compare the wave patterns point for point. In addition to being computationally inefficient, this approach can give a meaningless comparison because of such differences between the reference and sample speech patterns as the duration of a word. Nowadays, all word recognition machines make comparisons (statistical pattern recognition) between sets of parameters derived from reference and sample speech patterns.

Linear predictive coding (LPC) is the most widespread scheme of this type. It has been developed over the last 10 years, starting with the work of Bishnu Atal and his co-workers at Bell Laboratories. Later, it was extended by John Markel and Augustine Gray, Jr., formerly of the Speech Communications Research Laboratory, Los Angeles, and by John Makhoul of Bolt Beranek and Newman Inc., Cambridge, Massachusetts. These researchers were primarily interested in efficient ways of compressing speech for digital telecommunications. Linear predictive coding is based on a model of the human vocal tract consisting of a deformable acoustic tube. Parameters of the model include the fundamental voice frequencies, their intensities, and numbers specifying the cross-sectional area of the tube. To extract these parameters, the speech sample is electronically divided into segments 10 milliseconds long, and each of these is analyzed by one of numerous possible techniques. The analysis yields the set of LPC parameters.

Over the years numerous isolated word recognition systems have been put together, and the best of these can correctly identify words with an accuracy of 98 to 99 percent in a controlled laboratory setting for vocabularies up to 100 words. Operation in the field is an altogether different matter, and correct identifications can slip to 50 percent or less. A good portion of the error consists of instances in which the machine could not make an identification and therefore asked for a repeat speech sample, rather than misidentifying the word, a much more serious mistake. In a recent computer speech meeting, Stephen Moshier of Dialog Systems, Inc., Belmont, Massachusetts, judged that the major progress in isolated word recognition has not been in fundamental principles, but in field operation.

Numerous sources of error exist. In addition to the obvious problem of interference from background noise, machines are exceptionally sensitive to microphone placement. For example, microphone movement of a few centimeters can cause the machine to fail. Another is in the need for "training" the word recognizer. In training, the user must repeat each word in the vocabulary from five to ten times to the machine so that it can generate the reference patterns. If the user changes the way he talks or fails to observe the requirement for a pause between words, the word recognizer can botch an identification.

A third source of difficulty is psychological; people who do not think the word recognizer will work fail invariably to have their speech correctly identified. One story concerns employees whose job was inspecting glass tubes with stringent quality standards. For one or another reason, they were unsympathetic toward or distrusted a word recognition machine that was to replace a system in which inspection data were manually written down and later entered into a computer by way of punched cards. At first, workers complained they could not make the word recognizer understand them. In desperation, their supervisor spent an evening testing the machine and found little difficulty in being understood. The next day, the supervisor showed the workers how to make the machine understand spoken words and told them if they could not do as well, another job could be found for them. From then on the machine worked without error.

In many of the applications of isolated word recognition machines, the word recognizer simply replaces a keyboard terminal or some other input device to a computer. A major thrust of current activity is finding all-new applications and tailoring the properties of the word recognizers to these. One company taking this tack is Logicon, Inc., San Diego, which builds automated training systems of various types. Logicon is at present putting together a prototype air traffic controller training system for the Naval Training Equipment Center in Orlando, Florida. When completed this spring, the unit is to be tested at an air traffic con-

troller school in Memphis, according to Robert Breaux of the Navy center.

During training, the student sits before a simulated radar screen and is expected to make the vocal instruction appropriate to the indicated air traffic pattern. The same computer that generates the simulated radar display also interprets the student's instruction as analyzed by the word recognizer and provides feedback to the student by way of a voice synthesizer. The machine is capable of replacing a human instructor and other support personnel for a considerable portion of the training procedure.

Since the training system makes use of the proven isolated word recognition technique, the student's commands must fit the requirements of the machine. But, says Logicon's Mike Grady, the air traffic control terminology is rigidly defined, and the trainee must control his speech anyway so that the limited speech understanding capability is not necessarily a hindrance.

Another direction of current research is to extend the capability of isolated word recognition machines. Since training the machine requires considerable time and is not at all practical in easily envisioned applications involving the public, one goal is to reduce the need for training. What would be desirable is to find universal characteristics of all speech, so that once trained with one or a few speakers, the word recognizer could understand everyone. Recently, Lawrence Rabiner, Aaron Rosenberg, and Stephen Levinson of Bell Laboratories reported progress toward development of a so-called "speaker independent" isolated word recognizer for use in an automated telephone directory assistance system.

According to Rabiner, the researchers found that when the speech patterns of a group of 50 men and 50 women were analyzed, most of the patterns divided nicely into from 6 to 12 groups. Within each group there is little difference between the patterns of different individuals, but there are large differences between the groups. The result is that, having reference patterns for each group, the word recognizer can identify speech from a speaker not previously providing speech samples to the reference collection. The penalty, which entails much extra computation, is that speech samples must be compared with reference patterns from all of the 12 groups. At present, the machine can understand digits, letters of the alphabet (which are difficult because many sound alike), and certain control words and retrieves the correct phone number about 97 percent of the time when the questioner spells out the name of the person sought.

A major issue among speech recognition researchers is how close one can come to recognizing continuous speech without requiring extensive use of artificial intelligence-related techniques—that is, how well can machines recognize continuous speech from only the information contained in the acoustic wave pattern itself. One answer is that given by the NEC's newly announced machine, which has been praised by many as the new standard in the field.

Available in Japan since last April, the continuous speech recognition system will be officially introduced in the United States this month, although a few demonstration units have been shown to interested parties here. The machine has a vocabulary of up to 120 words, which are selected by the user during the training of the machine. Unlike previous word recognizers, only one training pass is needed (two passes for digits 0 through 9). After being trained, the system is said to be able to identify, with about a 98 percent accuracy in the laboratory, any combination of five words that together do not take more than 2.5 seconds to say. Response time is a fraction of a second. A major disappointment to many is the price, which ranges from $67,000 to $78,000, depending on the number of speech input channels. If the system becomes a big seller, however, the price could plummet. The U.S. Postal Service, for example, has just begun a test of the NEC system in one of its bulk-mail distribution centers. If the test were a success, a major market could open very quickly.

One of the secrets of the Japanese speech recognizer is a well-known technique called dynamic programming. In continuous speech, it is difficult to determine from the acoustic wave pattern when one word begins and another leaves off, in part because there may not be any pause between the words and in part because people speak words one way when talking naturally and another when saying words in isolation. Although originally developed for another purpose, dynamic programming provides a way of finding word boundaries that does not depend on there being a pause or a break in the wave pattern. According to George Doddington of Texas Instruments, the technique is not anything like a final answer, however, because considerable computation is required, which limits the size of the vocabulary. But consultant Nye estimates that 99 percent of the current types of tasks do not require more than a 200-word vocabulary. Having seen both continuous and discrete speech recognizers in action, *Science* can testify that there is a world of difference.

With an obviously bright future ahead for speech recognition, observers argue about what new applications will come forth and how fast. The conservatives talk mainly in terms of expansion of the existing industrial and military uses. Texas Instruments, for example, is working on an advanced version, for the Air Force, of a speaker verification system that it has had in daily operation for more than 4 years to control access to its central computer facility. In the new system, a speaker orally gives his identification number, which the machine checks. If the number is valid, it then compares the voice pattern with a reference to determine if the speaker is the individual who should have the number given. A similar system could be used, and Doddington predicts it will be in less than 5 years, for automatic financial transactions over the telephone. Ultimately, the largest markets are in the home, and futurists forecast that eventually such fancies as voice-actuated appliances will be commonplace. One can just imagine: "Television, please turn to channel 7. 'Wonder Woman' is on tonight and they have this fantastic computer you can talk to."—ARTHUR L. ROBINSON

With an obviously bright future ahead for speech recognition, observers argue about what new applications will come forth and how fast.

Additional Reading

1. *Proceedings of the IEEE* **64**, 405-558 (1976). A special issue devoted to man-machine communication by voice.
2. *Proceedings of a Workshop on Voice Technology for Interactive Real-Time Command/Control Application*, 6 to 8 December 1977, Ames Research Center, Moffett Field, Calif. Available from R. Breaux, Code N-71, NAVTRAEQUIPCEN, Orlando, Fla. 32813.
3. W. A. Lea, Ed., *Trends in Speech Recognition* (Prentice-Hall, Englewood Cliffs, N.J., in press).