Current Comments

From Vocoder to Vocalock— Speech Recognition Machines Still Have a Long Way to Go

Number 34 August 25, 1980

Ever since computers were first built, people have been fascinated with the idea of someday talking with them. Such a capability has long been a staple of science fiction. For example, Arthur C. Clarke's *2001: A Space Odyssey* features casual conversation between the crew of a spaceship and their computer, Hal.[1]

This sort of easy idiomatic communication between humans and machines will continue to exist strictly within the realm of science fiction for many years to come. But there are simpler machines that can respond appropriately to a few voice commands. Some of them are at work on production lines.[2] Back in the 1940s the Vocoder, developed at Bell Laboratories, was thought to be the key to the voice-activated typewriter.[3] Now futurists predict that the "understanding typewriter" is not far away.

Imagine your office equipped with an understanding typewriter. To write your colleague, you simply dictate into a microphone connected to the typewriter. The machine instantly types out your critique of his last paper, corrects your grammar, and eliminates the "ahs" and "uhs."

Speech recognition devices are not to be confused with optical character recognition (OCR) machines, which I have discussed previously.[4] The Kurzweil reader is an example of an OCR. It can read aloud from a printed page in a synthesized electronic voice. I might add, however, that in spite of its other successes, the Kurzweil OCR reader cannot yet be adapted to ISI®'s data input needs.

Speech recognition systems must accept voice commands instead of printed characters as an input, and they must correctly identify each word. I noted that there were still significant problems with OCR technology. But as we shall see, the problems associated with speech recognition are far greater. Both computers and humans find it easier to talk than to listen.

I recall reading about the Vocoder in a collection of essays entitled *Bibliography in an Age of Science*.[5] Then in the early 1960s, the Sperry Gyroscope Company invented the Sceptron, a device that identifies sound waves by their frequency content. ISI's Irv Sher developed an application. In 1965, he patented a door lock that would open only in response to an individual's voice.[6] The door mechanism was called Vocalock. To operate Vocalock, you first pushed a button to activate the system, and then spoke into a microphone in the door. The system analyzed the sound and opened the lock if it recognized the voice. Vocalock could be programmed to recognize any number of individual voices. Interestingly enough, Robert Heinlein in his 1961 book *Stranger in a Strange Land* described a future in which voice-operated locks are commonplace.[7]

It should be noted at this point that the term "speech recognition" can apply to several types of machines. The Vocalock, for example, was a device that identified and responded to a particular individual's voice. But the semantic content of what that voice said was irrelevant. Vocalock would perform its job whether the voice said "open sesame" or "shaboom." On the other hand, the speech recognition machines now in use on production lines do consider semantic content. It matters very much whether the human operator utters "start" or "stop." These machines have a very limited vocabulary, and the commands they accept must be pronounced distinctly, with clear pauses between words. This brings us to yet another type of machine, one that can respond to a human language of unlimited vocabulary spoken in a natural manner. Progress toward building such a machine is the chief concern of this essay.

There are several good reasons for pursuing research into voice recognition. For one thing, accurate voice recognition would facilitate direct communication between people and machines. As T. B. Martin of Threshold Technology, Delran, New Jersey, observes in a review article,[2] such communication has always been tailored to the operational requirements of the machine. But speech recognition systems would finally begin to allow machines to adapt to the requirements of people. This assumes that use goes beyond audio recognition to semantic comprehension. It is one thing to recognize words. It is another to understand speech.

Another advantage of speech recognition machines is an economic one. As computers become cheaper, most of the cost of data processing is involved in data preparation and entry.[8] Speech recognition systems may eliminate these high costs. At ISI, our editors might be able to read citations, titles, addresses, and other information aloud onto magnetic tape. Then the tapes could feed the data directly into our computer system. Ironically, one of the first data entry methods recommended to us back in 1962 was a system involving this same first step, to be followed by having operators key the data from dictaphone machine headphones. But just imagine the difficulties of dealing with homonymic names and words. Spelling out the last name of an author to avoid ambiguity and error would be a tiresome task compared to the present procedure.

Research into speech recognition machines began some 30 years ago. In 1952, scientists at Bell Laboratories reported that they had "taught" a computer to recognize the spoken digits "one" through "oh." These digits were spoken into a microphone connected to the computer, and the acoustic spectrum of each word was stored in the computer's memory. Thereafter, the machine compared the spectrum of a user's spoken word against the ten acoustic patterns stored. Bell's computer was able to identify the correct digits with a high degree of accuracy. However, the system had to be adjusted to accommodate different speakers.[9]

The early history of speech recognition research was described in a 1965 review by Nilo Lindgren, a staff writer for the *IEEE Spectrum*.[10] According to Lindgren, the development of limited-capability speech recognition systems such as the one at Bell created much optimism within the research community. It was not apparent at first just how difficult it would be to design a system that could respond to natural human speech of a fairly large vocabulary. Most experts thought it was primarily a question of having a computer compare incoming speech signals against acoustical patterns stored in memory. As Lindgren puts it, "It was the considered view of researchers that once they had found some method of analyzing acoustic signals into their basic component parts, the automation of speech recognition would quickly follow.... But extensive research on speech at the acoustic level...increasingly revealed the complexity of the speech process and forced the realization that this viewpoint was far too simple."[10]

It is one thing to teach a computer to recognize a limited number of words with distinct pauses between them. It is quite another matter to develop a system that can respond to a continuous stream of human speech. The problems involved are numerous and complex. Only a few of them can be discussed here.

If you are going to try to anticipate every possible pattern of connected speech utterances, you are going to need a lot of computer memory. Consider a vocabulary of just ten words. If you wanted to form word strings from one to seven words long, there are more than ten million ways the ten words can be combined. If each variation were treated as a single pattern, as they would be in a *connected*-speech recognizer, you would have to store more than 10 million reference patterns.[8] Clearly then, a simple pattern recognition scheme such as that employed in the early Bell Labs' machine could never accommodate a sizeable vocabulary of natural speech.

If a machine is to understand continuous speech, it must be able to determine where one word ends and another begins. This is even more complicated than you might assume. It's no surprise that humans can have trouble distinguishing one word from another. Just think of the old song "Mares Eat Oats." D. Raj Reddy, Carnegie-Mellon University, reports an experiment in which four subjects were asked to listen to a sentence containing the phrase "in mud eels are." None of the listeners correctly repeated what they heard. One listener thought the phrase was "in muddies sar," another "in model sar."[8]

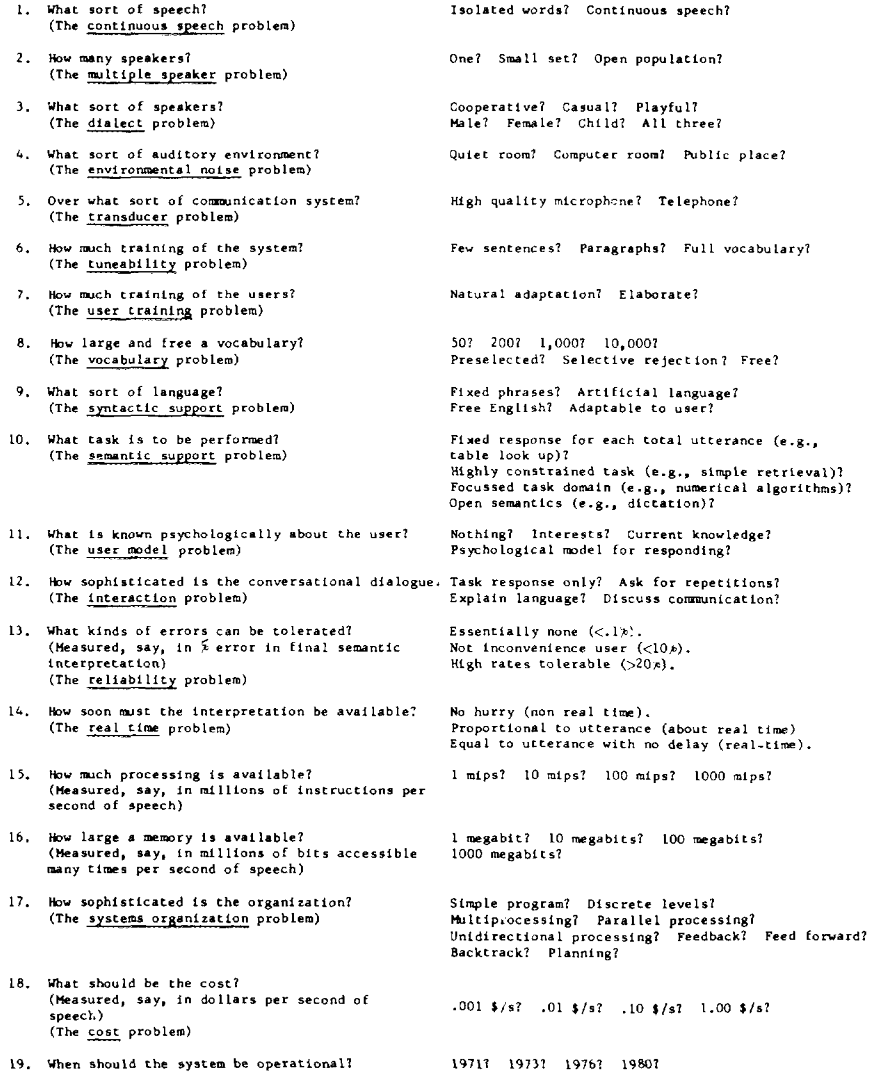
Another problem concerns the number of people who can use any speech recognition device. In order to teach a vocabulary to a computer, words must be read into it with a microphone. The early speech recognition models could only respond to the people who "trained" them.[10] Everyone's voice has its own distinctive acoustical spectrum. Linguists involved in speech recognition research have been looking for characteristics that are common to all human voices in order to enable machines to recognize most speakers.

There's also the problem of the imprecise nature of human speech. In written language, communication is unidirectional, without feedback. Therefore, all the information must be included in the writing. But as J.J. Mariani and colleagues note, "The spoken message implies that both interlocutors generally share the same environment and the same knowledge of the situation.... So speech may be fuzzy, poorly defined, noisy or ambiguous for external observers."[11]

Other aspects of natural speech pose problems. People often intone the same words in different ways, depending on the context.[12] People don't always use correct grammatical constructions when speaking. They sometimes pause in mid-sentence and clutter their speech with extraneous sounds such as "uh" or "well." They might say "dija" instead of "did you."[12] Related to this problem is the matter of background noise. Any speech recognition machine must be able to distinguish the true speech signal from noises in the surrounding environment. Lindgren observed that once the enormousness of these problems became clear, many researchers in the 1960s gave up on speech recognition and pursued other work.[10]

In the early 1970s, speech recognition research experienced a revival, largely through the infusion of federal money. In 1970, the Advanced Research Projects Agency (ARPA) of the US Department of Defense created a panel to review the state of the art and to develop a set of reasonable goals for research. Over the next five years,

Figure 1: *Problems considered by the ARPA study group for development of a speech-understanding system.*



1.	What sort of speech? (The continuous speech problem)	Isolated words? Continuous speech?
2.	How many speakers? (The multiple speaker problem)	One? Small set? Open population?
3.	What sort of speakers? (The dialect problem)	Cooperative? Casual? Playful? Male? Female? Child? All three?
4.	What sort of auditory environment? (The environmental noise problem)	Quiet room? Computer room? Public place?
5.	Over what sort of communication system? (The transducer problem)	High quality microphone? Telephone?
6.	How much training of the system? (The tuneability problem)	Few sentences? Paragraphs? Full vocabulary?
7.	How much training of the users? (The user training problem)	Natural adaptation? Elaborate?
8.	How large and free a vocabulary? (The vocabulary problem)	50? 200? 1,000? 10,000? Preselected? Selective rejection? Free?
9.	What sort of language? (The syntactic support problem)	Fixed phrases? Artificial language? Free English? Adaptable to user?
10.	What task is to be performed? (The semantic support problem)	Fixed response for each total utterance (e.g., table look up)? Highly constrained task (e.g., simple retrieval)? Focussed task domain (e.g., numerical algorithms)? Open semantics (e.g., dictation)?
11.	What is known psychologically about the user? (The user model problem)	Nothing? Interests? Current knowledge? Psychological model for responding?
12.	How sophisticated is the conversational dialogue. (The interaction problem)	Task response only? Ask for repetitions? Explain language? Discuss communication?
13.	What kinds of errors can be tolerated? (Measured, say, in % error in final semantic interpretation) (The reliability problem)	Essentially none (<.1%). Not inconvenience user (<10%). High rates tolerable (>20%).
14.	How soon must the interpretation be available? (The real time problem)	No hurry (non real time). Proportional to utterance (about real time) Equal to utterance with no delay (real-time).
15.	How much processing is available? (Measured, say, in millions of instructions per second of speech)	1 mips? 10 mips? 100 mips? 1000 mips?
16.	How large a memory is available? (Measured, say, in millions of bits accessible many times per second of speech)	1 megabit? 10 megabits? 100 megabits? 1000 megabits?
17.	How sophisticated is the organization? (The systems organization problem)	Simple program? Discrete levels? Multip.ocessing? Parallel processing? Unidirectional processing? Feedback? Feed forward? Backtrack? Planning?
18.	What should be the cost? (Measured, say, in dollars per second of speech) (The cost problem)	.001 $/s? .01 $/s? .10 $/s? 1.00 $/s?
19.	When should the system be operational?	1971? 1973? 1976? 1980?

Reprinted by permission of the publisher from **Newell A.**, *et al.*
Speech Understanding Systems. Copyright 1973 by Elsevier-North Holland.

ARPA spent $15 million on the development of a "best possible" speech recognition system.[12]

The ARPA study group's assessment of the state of the art and its recommendations for research are contained in a highly readable report which was published in 1973.[13] The report also addressed the problems associated with speech recognition, some of which I have already discussed. Figure 1 is taken from the ARPA report. It provides a good illustration of the complexity of these problems.

One of the ARPA study group members, D.H. Klatt of MIT, summarized the goals of the project in a 1977 review article.[14] ARPA was to develop several prototype speech recognition machines to accept continuous speech of a general American dialect. The systems were to be able to accept new speakers with only slight tuning. They were to have a vocabulary of 1,000 words and a syntax appropriate for whatever specific task they were designed to perform. An error rate of no more than 10% would be tolerated. These goals were to be met by November 1976.[14]

ARPA provided funding to a number of contractors for research into speech understanding systems. After two years, four contractors judged to have made promising developments were selected to complete their work. All four contractors actually developed systems. The one that came closest to meeting ARPA's goals was HARPY, developed at Carnegie-Mellon University.

HARPY was designed by Bruce Lowerre and D. Raj Reddy. Like its competitors in the ARPA project, HARPY improved upon an innovation that had come into use during the late 1950s. Instead of matching the acoustic patterns of whole words, as the Bell Labs' machine had done, researchers began to consider breaking down words into their constituent phonemes. Phonemes are "the basic linguistic units which have the property that if one replaces another in an utterance, the meaning is changed."[15] The English language consists of about 40 phonemes. Employing this method lowers the risk that the system will mistake one word for another which sounds similar. The contractors in the ARPA project attained further specificity by breaking down spoken English into even more basic units called phones. The HARPY machine was programmed to recognize 96 phones.[16]

Allen Newell, Carnegie-Mellon, who served on the ARPA study group, explains that HARPY had a grammar that functioned as a generator of sentences. The grammar determined which combinations of words in HARPY's vocabulary were permissible. The grammar could conceivably generate about five billion different sentences.[16]

HARPY employed an innovative search strategy, which was described in a recent paper by A.L. Robinson in the IEEE *Transactions on Professional Communication*. The system determined in advance all of the possible sentences it might be asked to understand to perform its task, which was document retrieval. When addressed by a speaker, HARPY "compared the degree to which the (phones) stored in its memory 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."[12] By this process of eliminating improbable sentences, HARPY was able to cut down on the search time it required.

HARPY was "trained" to accept speech from three male and two female users, and it did so with 91% accuracy. It could also accept speech from people it was not trained for, although accuracy then diminished somewhat. HARPY had a vocabulary of 1,011 words.[12]

In 1976, funding for the ARPA speech project was greatly reduced, although Carnegie-Mellon and other institutions involved in the project continued research activity. A glance at papers presented at the 1979 International Conference on Acoustics, Speech, and Signal Processing shows that many organizations are still very interested in speech recognition research.[17] The interest in speech recognition is worldwide. Recent papers on the topic have been written by researchers in a number of countries, including the Soviet Union,[18] India,[19] Italy,[20] and the Federal Republic of Germany.[21]

HARPY represented an improvement over the classical pattern-matching machines. But developers continue to search for other methods. At present, the most concentrated effort to develop a speech understanding system is taking place at IBM's Yorktown Heights, New York, facility. The giant of the computer industry hopes to apply voice recognition to office machines. This includes development of the previously described understanding typewriter.

IBM did not take part in the ARPA project, and its systems follow a different conceptual strategy for understanding speech. The IBM group, which is headed by Frederick Jelinek, uses a statistical method that assigns probabilities to word sequences.[22] For example, if a speech recognition machine hears a verb, then there's a certain probability that the next word will not be a verb, but some other part of speech. The IBM researchers analyzed a number of test sentences to devise probabilities that can be applied to a great number of possible sentences.

Recently, Jelinek and colleagues announced the successful testing of a speech recognition device using an IBM 360/168 mainframe computer.[23] The system uses a sophisticated acoustic processor that digitizes incoming speech signals and matches them against prototypes stored in the computer's memory. The IBM system is very accurate and has a 1,000 word vocabulary. But it is slow. It may take the computer 100 minutes to recognize a single sentence. Moreover, this system too must be "trained" by each individual user.

The IBM system is a promising development, but we are still a long way from speech recognition systems of unlimited vocabulary. Perhaps the obstacles facing speech recognition research will be solved with the next generation of computer software. One of several approaches under investigation is a concept called "fuzzy sets."[24] In most speech recognition systems, utterances must be classified and put into reference sets. Sounds that fall outside the sets will not be recognized. But human speech is not so precise. And as S. Rivoira and P. Torasso of the University of Turin, Italy, put it, "Fuzzy languages have potential for imprecise patterns, and the basically subjective concept of a fuzzy set makes the fuzzy membership assignment to the pattern segment a heuristic (self-educating) choice."[24]

Speech understanding systems have the potential to provide a wealth of opportunity for people in the coming information society.[25] Speaking to a computer is a lot less threatening than communicating with it via codes or data entry keyboards. As Joshua Lederberg recently told *Research Resources Reporter*, speech input will make computers in general more acceptable. Right now, people are "just not going to sit at a terminal that they don't know how to operate, or don't have time to use.... Voice entry of data would make a very big difference."[26]

While it is not difficult to imagine some of the future social consequences of a universal voice recognizer, it would be an important breakthrough just to have a machine that could recognize one person's voice input. At ISI we could certainly afford the time to "educate" the computer to recognize each indexer's voice. And I'm sure that executives could learn to speak more precisely if they knew they could eliminate a lot of the headaches of preparing manuscripts and letters. However, we need more research, both basic and applied. We certainly need more basic research on how we perceive and understand speech. But I suspect this problem also needs the attention of a few innovative engineers who are able to combine basic knowledge with unique technological skill. Speech recognition has come a long way but it still has a long, long way to go.

* * * * *

My thanks to Thomas Di Julia, Patricia Heller, and Esther Surden for their help in the preparation of this essay.

REFERENCES

1. **Clarke A C.** *2001: a space odyssey.* New York: NAL, 1968. 221 p.
2. **Martin T B.** Practical applications of voice input to machines. *Proc. Inst. Elec. Electron. Eng.* 64:487-500, 1976.
3. **Dudley H.** Remaking speech. *J. Acoust. Soc. Amer.* 11:169-77, 1939.
4. **Garfield E.** Has OCR finally arrived? Or is it a technology with a lot more problems than meet the eye? *Current Contents* (19):5-13, 7 May 1979.
5. **Hill A G.** The storage, processing and communication of information. (Ridenour L N, Shaw R R & Hill A G, eds.) *Bibliography in an age of science.* Urbana, IL: University of Illinois Press, 1951. p. 73-88.
6. **Sher I H.** US Patent 3,184,937. 25 May 1965.
7. **Heinlein R A.** *Stranger in a strange land.* New York: Berkeley, 1961. 414 p.
8. **Reddy D R.** Speech recognition by machine: a review. *Proc. Inst. Elec. Electron. Eng.* 64:501-31, 1976.
9. **Davis K H, Buddulph R & Balashek S.** Automatic recognition of spoken digits. *J. Acoust. Soc. Amer.* 24:637-42, 1952.
10. **Lindgren N.** Machine recognition of human language. Part 1—automatic speech recognition. *IEEE Spectrum* 2:114-36, 1965.
11. **Mariani J J, Lienard J S & Renard G.** Speech recognition in the context of two-way immediate person-machine interaction. *1979 IEEE international conference on acoustics, speech, and signal processing,* 2-4 April 1979, Washington, DC. New York: IEEE, 1979. p. 269-72.
12. **Robinson A L.** Communicating with computers by voice. *IEEE Trans. Prof. Commun.* 22:159-65, 1979.
13. **Newell A, Barnett J, Forgie J W, Green C, Klatt D, Licklider J C R, Munson J, Reddy D R & Woods W A.** *Speech understanding systems.* New York: Elsevier, 1973. 137 p.
14. **Klatt D H.** Review of the ARPA speech understanding project. *J. Acoust. Soc. Amer.* 62:1345-66, 1977.
15. **Flanagan J L.** *Speech analysis, synthesis and perception.* New York: Academic Press, 1965. 317 p.
16. **Newell A.** Telephone communication. 24 July 1980.
17. **IEEE Acoustics, Speech, and Signal Processing Society.** *1979 IEEE international conference on acoustics, speech, and signal processing,* 2-4 April 1979, Washington, DC. New York: IEEE, 1979. 993 p.
18. **Velichko V M & Zagoruiko N G.** Synthesis of speech-understanding systems. *Sov. Physics Acoust.* 24:87-8, 1978.
19. **Sarma V V S, Yegnanarayana B & Ananthapadmanabha T V.** A speaker recognition scheme on a minicomputer based on a signal processing facility. *Acustica* 41:117-21, 1978.
20. **Bernorio M, Bertoni M, Dabbene A & Somalvico M.** Quasi—natural language understanding in the semantic domain of robotics. *Cybernetica* 22:159-72, 1979.
21. **Zwicker E, Terhardt E & Paulus E.** Automatic speech recognition using psychoacoustic models. *J. Acoust. Soc. Amer.* 65:487-98, 1979.
22. **Jelinek F.** *Self-organized continuous speech recognition.* Warrendale, PA: Society of Automotive Engineers Congress, 25-29 February 1980, Detroit, MI. SAE Technical Paper 800198, 8 p.
23. **Jelinek F, Mercer R L & Bahl L R.** Continuous speech recognition: statistical methods. (Unpublished paper), 1980. 35 p.
24. **Rivoira S & Torasso P.** An isolated-word recognizer based on grammar-controlled classification processes. *Patt. Recog.* 10:73-84, 1978.
25. **Garfield E.** 2001: an information society? *J. Inform. Sci.* 1:209-15, 1979.
26. **Freiherr G.** The problems and promises of artificial intelligence. *Res. Resour. Rep.* 3(9):1-6, September 1979.