Reprinted from THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA, Vol. 46, No. 4, (Part 2), 1049–1051, October 1969

Whither Speech Recognition?

J.R. PIERCE

Bell Telephone Laboratories, Inc., Murray Hill, New Jersey 07971

Speech recognition has glamor. Funds have been available. Results have been less glamorous. "When we listen to a person speaking—much of what we think we—hear is supplied from our memory. [W. James, *Talks to Teachers on Psychology and to Students on Some of Life's Ideals* (Holt, New York, 1889), p. 159]. General-purpose speech recognition seems far away Special-purpose speech recognition is severely limited. It would seem appropriate for people to ask themselves why they are working in the field and what they can expect to accomplish.

THE PURPOSE OF THIS LETTER IS TO EXAMINE BOTH MOTIVATIONS and progress in the area of speech recognition (that is, word recognition). There is a reason for this beyond the fun and insight that such an investigation affords. The country is supporting such work. Is this wise? Are we getting our money's worth?

The sort of human behavior encouraged by the lush funding of science and engineering after World War II, and especially after Sputnik, has been more imitated and elaborated than commented on or analyzed. Some of the strangest aspects of post-Sputnik behavior are exhibited in work on speech recognition as well as in other intriguing fields such as space, artificial intelligence, and cybernetics. One of the most fascinating aspects of behavior in such fields is its motivation.

It would be too simple to say that work in speech recognition is carried out simply because one can get money for it. That is a necessary but not a sufficient condition. We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn't attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%. To sell suckers, one uses deceit and offers glamor.

It is clear that glamor and any deceit in the field of speech recognition blind the takers of funds as much as they blind the givers of funds. Thus, we may pity workers whom we cannot respect. People who work in the field are full of innocent (in their own view) enthusiasm. What particular considerations have led to this enthusiasm?

Perhaps the soundest is a philosophical problem expressed in Turing's game.[1] Turing asked, On what basis can we say that a machine thinks? His perfectly rational answer was that if, in conversing with a machine, we cannot tell whether it is a human being or a machine, then we can scarcely deny that the machine thinks.

I believe that this sensible philosophical problem has been a considerable inspiration to workers in speech recognition (and also, for that matter, in speech synthesis). In science fiction, computers listen to (and talk to) human beings, as Hal does in the movie *2001*.

We should consider, however, that in deception, studied and artful deceit is apt to succeed better and more quickly than science. Experience teaches us that pretended telepathy or pretended communication with the dead is more convincing than the real thing if, indeed, the real thing exists. And, Weizenbaum's doctor program[2] has shown us that, via teletypewriter, a computer can "sound" much like a psychoanalyst (to the novice, at least) without really thinking at all. Indeed, a wag has proposed that computers are becoming so nearly human that they can act without thinking.

It is reasonable to believe that Turing's game may best be approached with the sort of artful deceit that men use in conversation when they want to hide their ignorance and make the most of their knowledge. It would perhaps be easier for a speaking computer to convince the listener that it was so deep as to be unintelligible, and had a cold or spoke with a Merovingian accent, than it would be for a speaking computer to understand what was said and reply sensibly in good, lifelike general American.

It seems dubious, anyway, whether Turing's interesting game is sufficient justification for spending much money on speech recognition. When we look further for reasons, we encounter that of communication with computers.

In this general form, the reason is as specious as insisting that an automobile should respond to *gee*, *haw*, *giddap*, *whoa*, and slaps or tugs of the reins. We communicate with children by words, coos, embraces, and slaps. We communicate with people by these means and by nods, winks, and smiles. It is not clear that we should resort to the same means with computers. In fact, we do very well with keyboards, cards, tapes, and cathode-ray tubes.

What about the possibility of directing a machine by spoken instructions? In any practical way, this art seems to have gone downhill ever since the limited commercial success of Radio Rex, a toy dog that jumped from his house when his name was spoken. Researchers are hard at it, however. Will they succeed?

There are strong reasons for believing that spoken English is, in general, simply not recognizable phoneme by phoneme or word by word, and that people recognize utterances, not because they hear the phonetic features or the words distinctly, but because they have a general sense of what a conversation is about and are able to guess what has been said. William James wrote, in 1899[4]:

> When we listen to a person speaking or read a page of print, much of what we think we see or hear is supplied from our memory. We overlook misprints, imagining the right letters, though we see the wrong ones; and how little we actually hear, when we listen to speech, we realize when we go to a foreign theatre; for there what troubles us is not so much that we cannot understand what the actors say as that we cannot hear their words. The fact is that we hear quite as little under similar conditions at home, only our mind, being fuller of English verbal associations, supplies the requisite material for comprehension upon a much slighter auditory hint.

This wisdom is confirmed by various anecdotes. It is said that a native speaker can understand a conversation on a noisy streetcar where a foreigner very fluent in the language cannot. In persistent efforts to understand noisy or indistinct speech, we continually try to guess what the utterance might be, and a conviction as to its content, even a false conviction, is catching.[5] Totally deaf people can understand speech by reading lips, and yet the clues they follow cannot be sufficient for deciphering all phonemes or even all words. A stenotypist can transcribe a stenotype record despite the fact that not all words are represented unambiguously.

These considerations lead us to believe that a general phonetic typewriter is simply impossible unless the typewriter has an intelligence and a knowledge of language comparable to those of a native speaker of English. This leaves a narrower question open. Are more limited, satisfactory economic applications of voice control realizable? Radio Rex was certainly a limited success. But, would something more, say, response to the 10 digits, or to some other limited vocabulary, be successful and worthwhile?

Another justification of speech recognition is sometimes given: that through such work we learn something about speech. This no one can deny. We do learn something. The argument would be more impressive, however, if those who engage in recognition showed more signs of an effective effort to learn about speech and

fewer signs of rapture for computers and for unproven schemes for, and theories of, recognition.

We all believe that a science of speech is possible, despite the scarcity in the field of people who behave like scientists and of results that look like science.

Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers. The typical recognizer gets it into his head that he can solve "the problem." The basis for this is either individual inspiration (the "mad inventor" source of knowledge) or acceptance of untested rules, schemes, or information (the untrustworthy engineer approach).

We would expect a scientist to check the literature concerning ideas, schemes, or information. Perhaps some point is old and has been established or confuted by a clear, simple, definitive experiment. If there is no clear experimental evidence, it might be possible for a scientist to devise a clear, simple, definitive experiment. So, a science of speech might grow, certain step by certain step.

The typical recognizer will have none of this. He builds or programs an elaborate system that either does very little or flops in an obscure way. A lot of money and time are spent. No simple, clear, sure knowledge is gained. The work has been an experience, not an experiment.

Those who are interested can trace in detail the uncertain course (if not progress) of speech recognition through the literature. Lindgren,[6] Marill,[7] David,[8] David and Selfridge,[9] Sections 11.2, 11.4, and 12.2 of a book by Saposhkov,[10] and Section 5.5 of Flanagan's book[11] cover much early work. Relevant Russian work is referred to by Kozhevnikov and Chistovich.[12] References 8–46 cover a wide variety of approaches.

It is hard to gauge the success of an attempt at speech recognition even when statistics are given. In general, it appears that recognition around 95% correct can be achieved for clearly pronounced, isolated words from a chosen small vocabulary (the digits, for instance) spoken by a few chosen talkers. Better results have been attained for one talker. Performance has gone down drastically as the vocabulary was expanded, and appreciably as the number and variety of talkers were increased. It is not easy to see a practical, economically sound application for speech recognition with this capability.

The arguments given earlier may lead us to believe that performance will continue to be very limited unless the recognizing device understands what is being said with something of the facility of a native speaker (that is, better than a foreigner who is fluent in the language). If this is so, should people continue work toward speech recognition by machine? Perhaps this is for people in the field to decide. Certainly, it would seem appropriate that before embarking upon such work, the worker should candidly ask and answer the following questions:

- Why am I working in this field?
- What particular thing do I hope to accomplish?
- Why is it worthwhile?
- Am I likely to succeed?
- How will I know whether or not I have succeeded?
- Where will success take or leave me?

Any application of the foregoing discussion to work in the general area of pattern recognition is left as an exercise for the reader.

The writer was fortunate in not having to seek out Refs. 8–46 for himself; he found them in a list of references compiled by S. R. Hyde of the Joint Speech Research Unit, Ruislip, Middlesex, England.

[1] A. M. Turing, "Computing Machinery and Intelligence," MIND **59**, New Ser. 236, 433–460 (1950).
[2] J. Weizenbaum, "Contextual Understanding by Computers," *Recognizing Patterns* P. A. Kolers and M. Eden, Eds. (The MIT Press, Cambridge, Mass., 1968), pp. 170–193.
[3] Sir R. A. Surtees Paget, *Human Speech* (Kegan Paul, Trench, Trubner and Co., Ltd., London; Harcourt, Brace & Co., New York, 1930), pp. 79–80.
[4] W. James, *Talks to Teachers on Psychology and to Students on Some of Life's Ideals* (Holt, New York, 1899), p. 159.
[5] E. H. Gombrich, *Art and Illusion*, Bollington Foundation (Pantheon Books, a division of Random House, Inc., New York, 1961), 2nd ed., p. 204.
[6] N. Lindgren, "Machine Recognition of Human Language," IEEE Spectrum **2**, Nos. 3 and 4 (1965).
[7] T. Marill, "Automatic Recognition of Speech," IRE Trans. Human Factors Electron HFE-2, 34–38 (1961).
[8] E. E. David, Jr., "Artificial Auditory Recognition in Telephony," IBM J. Res. Develop., **2**, 294–309 (1958).
[9] E. E. David, Jr., and O. G. Selfridge, "Eyes and Ears for Computers," Proc. IRE **50**, 1093–1101 (1962).
[10] M. A. Saposhkov, "The Speech Signal in Cybernetics and Communications," transl. by Joint Publications Res. Service, JPRS **28**, 117 (1965).
[11] J. L. Flanagan, *Speech Analysis Synthesis and Perception* (Springer-Verlag, Berlin, 1965).
[12] V. A. Kozhevnikov and L. A. Chistovich, "Speech: Articulation and Perception" transl. by Joint Publications Res. Service, JPRS **30**, 543 (1965).
[13] H. K. Davis, R. Biddulph, and S. Balashek, "Automatic Recognition of Spoken Digits," J. Acoust. Soc. Amer. **24**, 637–642 (1952).
[14] H. Dudley and S. Balashek, "Automatic Recognition of Phonetic Patterns in Speech," J. Acoust. Soc. Amer. **30**, 721–733 (1958).
[15] J. Wiren and H. L. Stubbs, "Electronic Binary Selection System for Phoneme Classification," J. Acoust. Soc. Amer. **28**, 1082–1091 (1956).
[16] H. F. Olson and H. Belar, "Phonetic Typewriter III," J. Acoust. Soc. Amer. **33**, 1610–1615 (1961).
[17] D. B. Fry, "Theoretical Aspects of Mechanical Speech Recognition" J. British Inst. Radio Eng. **19**, 211–219 (1959). Also, P. Denes, "The Design and Operation of the Mechanical Speech Recognizer at University College London," *ibid.*, pp. 219–229.
[18] J. W. Forgie and C. D. Forgie, "Results Obtained from a Vowel Recognition Computer Program," J. Acoust. Soc. Amer. **31**, 1480–1489 (1959).
[19] J. Suzuki and K. Nakata, "Recognition of Japanese Vowels—Preliminary to the Recognition of Speech," J. Radio Res. Lab., Tokyo **8**, No. 37, 193–212 (1961).
[20] T. Sakai and S. Doshita, "The Phonetic Typewriter," Proc. IFIP Congr. Munich, Infor. Processing Aug.–Sept. (1962).
[21] K. Nagata, Y. Kato, and S. Chiba, "Spoken Digit Recognizer for Japanese Language," Nippon Elec. Co., Res. Develop. No. 6 (1963).
[22] T. B. Martin, A. L. Nelson, and H. J. Zadell, "Speech Recognition by Feature Abstraction Techniques," Tech. Rep. No. AL-TDR-64-176 (AD604526), Air Force Avionics Lab. (1964).
[23] J. W. Falter, "Feature Abstraction: An Approach to Speech Recognition," Proc. Nat. Aerospace Elec. Conf., IEEE. 192–198 (1965).
[24] J. Gazdag, "A Method of Decoding Speech," Tech. Rep. No. 9, AFOSR-66-2385 (AD641132), Univ. of Illinois, June (1966).
[25] L. Gilli and A. R. Meo, "Sequential System for Recognizing Spoken Digits in Real Time," Acustica **19**, (1967/68).
[26] P. W. Ross, "A Limited-Vocabulary Adaptive Speech-Recognition System," J. Audio Eng. Soc. **15**, 414–419 (1967).
[27] P. B. Denes and M. V. Mathews, "Spoken Digit Recognition Using Time-Frequency Pattern Matching," J. Acoust. Soc. Amer. **32**, 1450–1455 (1960).
[28] G. Sebestyen, "Automatic Recognition of Spoken Numerals," J. Acoust. Soc. Amer. **32**, 1516 (A) (1960).
[29] W. F. Meeker, A. L. Nelson, and P. B. Scott, "Voice to Teletype Code Converter Research Program, Part II, Experimental Verification of a Method to Recognize Phonetic Sounds," Tech. Rep. No. ASD-TR 61–666, Wright-Patterson Air Force Base, Ohio (CRB 63/3885) (1962).
[30] B. Gold, "Word-Recognition Computer Program," Res. Lab. Electron, MIT Rep. No. 452, June (1966).
[31] P. N. Sholtz and R. Bakis, "Spoken Digit Recognition using Vowel-Consonant Segmentation," J. Acoust. Soc. Amer. **34**, 1–5 (1962).
[32] G. W. Hughes, "The Recognition of Speech by Machine," MIT Res. Lab. Electron, Tech. Rep. No. 395 (1961).
[33] G. W. Hughes and J. F. Hemdal, "Speech Analysis," Rep. AFCRL-65–681, (P137552), Purdue Univ. (1965).
[34] L. R. Talbert, G. F. Groner, J. S. Koford, R. J. Brown, P. R. Low, and C. H. Mays, "A Real-Time Adaptive Speech Recognition System," Tech. Rep. No. 6760-1 (ASD-TDR-63-660) (P133441), prepared by Stanford Electron Lab.
[35] J. A. Dammann, "Application of Adaptive Threshold Elements to the Recognition of Acoustic–Phonetic States," J. Acoust. Soc. Amer. **38**, 213–223 (1965).
[36] J. H. King and C. J. Tunis, "Some Experiments in Spoken Word Recognition," IBM J. Res. Develop. **10**, 65–79 (1966).
[37] D. Fraipont, "Voice Actuated Address Mechanism," Elec. Ass., Inc., Rep. No. 3 (AD 633711) (1966).
[38] C. F. Teacher and C. F. Piotrowski, "Voice Sound Recognition," Tech. Rep. No. RADC-TR-65-184, Rome Air Development Ctr. (AD 619964) July (1965).
[39] C. F. Teacher, H. Kellett, and L. Focht, "Experimental, Limited Vocabulary, Speech Recognizer," IEEE Int. Conv. Record, Part III, 169–173 (1967).
[40] M. Weiss, "A Study of Critical Instant Sampling of Speech Parameters for Automatic Recognition of Spoken Words," Rome Air Develop. Ctr., Rep. No. RADC-TR-65-371 (AD 38380) July (1966).
[41] M. W. Cannon, "A Method of Analysis and Recognition for Voiced Vowels," IEEE Trans. Audio Electroacoust. **AU-16**, 154–158 (1968).
[42] D. R. Reddy, "An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave," Computer Sci. Dep., Stanford Univ., Tech. Rep. No. CS49, Sept. (1966). [Paper based on this report in J. Acoust. Soc. Amer. **42**, 329–347 (1967).]
[43] J. N. Shearme and P. E. Leach, "Some Experiments with a Simple Word Recognition System," IEEE Trans. Audio Electroacoust. **AU-16**, 256–261 (1968).
[44] R. F. Purton, "Speech Recognition using AutoCorrelation Analysis," IEEE Trans. Audio Electroacoust. **AU-16**, 235–239 (1968).
[45] S. H. Lavington and L. E. Rosenthal, "Some Facilities for Speech Processing by Computer," Computer J. (British Computer Society, London, NW 1) **9**, 330–339 (1967).
[46] W. Bezdel, "Discriminators of Sound Classes for Speech Recognition Purposes," Conf. Speech Commun. Proc., AFCRL, Sec. B8, 104–108 (1967).