

A Voice for the Laboratory Computer*

F.S. Cooper, T.C. Rand, R.S. Music, and I.G. Mattingly[†]
Haskins Laboratories, New Haven

It seems clear, even from the titles for this session, that voice answer-back from computers can be used in diverse ways. But responding by voice is by no means all that the ability to speak will enable a computer to do. One of these other things is to read aloud for the blind. This possibility--even most of the basic methods for realizing it--was clearly foreseen well before the technology was ready (Cooper, 1963). It is a challenging problem for several reasons: first, because it is so completely open-ended linguistically, with almost no limits on either vocabulary or sentence structure; second, because it is almost as open-ended with respect to the type fonts and page formats that must be accommodated; and third, because it looks so easy--but has proved so difficult--to find any cheap and simple substitute for spoken language as a means of conveying the printed message to the blind reader. The reading machine problem has been described in greater detail elsewhere (Studdert-Kennedy and Cooper, 1966; Cooper et al., 1969); moreover, its solution is beginning to be realized in a practical way. Field trials of a blind reading service are now under way with the aid of the Prosthetic and Sensory Aids Service of the Veterans Administration and the cooperation of a group of blind veterans at the West Haven Veterans Hospital.

Another potential use for talking computers is in shaping computer-assisted instruction to the needs of the lower grades. The youngsters could learn more easily from a talking informant than from conventional teletype or CRT displays, simply because reading is a skill they have yet to learn. A different use that the mature scholar will one day make of speech from computers is easy reference by telephone to a library's holdings, or at least the portion that is stored in machine-readable form. These, and other, applications of computer-generated speech are now beginning to get the attention they have long deserved. (See Flanagan et al., 1970, for a recent review.)

How do today's uses of voice response differ from the familiar weather and time announcements that telephone companies provide? And why employ computers when magnetic tape and sound track are so much cheaper? An obvious reason is that the simpler devices work well only with restricted message sets and do not extrapolate to more demanding problems. But in the background there is a more subtle reason, namely, that spoken language is

*This paper appeared in the 1971 IEEE International Convention Digest (Institute of Electrical and Electronics Engineers, Inc., 1971) 104-105.

[†]Also, University of Connecticut, Storrs.

by no means simple. This is why even a modest voice-response system needs a great deal of storage, or sophisticated synthesis, or some of each. The nature of the requirement can be seen most clearly in a comparison of the difficulties of outputting a stored message by writing it or by speaking it. Thus, printing a message that is stored in alphanumeric form is a completely straightforward operation and requires only a Teletype machine or lineprinter; by contrast, outputting this same message as speech requires only a little hardware but a lot of know-how and sophisticated programming. To be sure, one can trade memory space for sophistication, but only at a price that is prohibitive for many applications. (It is interesting to note that the inverse problem that computers have in getting an intelligible response from a human is even more difficult. Again, the explanation lies in the inherent complexity of the speech signal.)

THE TALKING COMPUTER AS A LABORATORY HELPER

Let us turn from these general considerations to some specific examples of how one laboratory uses, and plans to use, the voice-response capability that its computer has acquired while engaged in research on reading machines for the blind. We will mention some of the ways in which computer-generated speech is used, then indicate how that speech was produced and what trade-offs were involved in the choice of method.

The preparation of stimulus tapes for experiments on speech perception has been for us a most important use. The computer serves as a versatile informant and can respond appropriately to many variants of the question, "How would you say that?" For experiments on dichotic competition, two messages are needed that are accurately timed with respect to each other. The computer's ability to speak out of both sides of its mouth at once is a great asset. Sometimes the stimulus pairs consist of carefully tailored utterances from a human speaker; sometimes fully synthetic speech is used to get even closer control of intensities, durations, etc. We have made other uses also of the computer's ability to speak in ways that human beings cannot, e.g., more rapidly or slowly, in absolute monotone, adding or omitting formants, and other variants that were needed to test particular hypotheses. Stretched speech has been particularly useful in providing synchronized sound tracks for slow-motion movies of the articulators in action.

Monitoring type-in of information to the computer is another role for voice answer-back. One phase of the work on reading machines required that extensive texts be fed into the computer (for later conversion to speech recordings) in the form of phonetic transcriptions. A keyboard and storage oscilloscope used on-line provide the equivalent of a phonetic typewriter. Typing errors that might otherwise go unnoticed become immediately apparent when automatic voice feedback is provided for each utterance segment or sentence as it is typed.

A cheap remote terminal for inputting programs or data is an obvious variant. If one is copying handwritten programs, there is no real need for a visual display; only a keyboard and loud speaker are needed for type-in and verification on a line-by-line basis.

Confirmation (or correction) of responses in psychological experiments can best be given by voice when the task is a visual one. Thus, in combined dichoptic and dichotic perceptual tests (which need computer control of the stimuli in any case), we expect to use voice responses from the computer to alert and inform the subject.

Scoring test sheets on which subjects have written what they heard in a perceptual test and then entering the data into the computer for processing is an irksome but necessary part of our research. We now have a simple way of hand-scoring the test forms. In one pass, the entries on the form are identified and entered directly into the computer. The scorer has an immediate visual check on his part of the job, and voice answer-back from the computer serves to let him know that the computer correctly understood his written entries.

METHODS FOR GENERATING VOICE RESPONSES; TRADE-OFFS

There are only a few basic methods for getting speech from a computer, but a wide variety of ways to combine them for the best match to a particular task. Usually, the key factor is the form in which the information is stored. One method stores messages as speech wave-forms. This is a simple solution, but it makes heavy demands on memory capacity--roughly 50,000 bits per second of stored speech. The method is most useful when a very limited set of words will suffice for all messages. Thus, for example, the confirmation of subjects' responses or of hand-printed entries needs only a few different responses; we chose waveform storage for these tasks because it is so simple to enter the responses into storage and to use them as output.

Another method suitable for whole messages, or for messages composed of smaller fragments, is to store the control parameters that will enable a synthesizer to generate the speech waveforms. The advantages of this method are considerable: storage requirements are reduced by a factor in the range of 10-50, and there is more flexibility in composing realistic sentences since stress and intonation can be imposed on a composite spectral "skeleton." Disadvantages are that special hardware is needed to synthesize the speech, that the parameters are not related in a simple way to either the speech waveform or the message in its normal written form, and that, in consequence, a substantial amount of processing is required to derive the parameters or to revise them when the message set is to be changed. This is a method of storage we have used in compiling stimulus tapes for psychological experiments.

A third method is synthesis by rule. It is a two-stage process: the first converts a written English sentence into a phonetic transcription, and the second converts the transcription into control parameters for a synthesizer. In many applications, the first conversion will be bypassed since messages in phonetic form require even less storage than in alphabetic form and there are further substantial savings in processing time and memory space. The synthetic speech that can now be generated is still far from perfect, even after more than twenty years of research; nevertheless, it is good enough to be useful to the blind and is the spoken output used in our reading machine research. The advantages of synthesis by rule go well beyond those of synthesis from stored parameters: there

COMPARISON OF METHODS
FOR GENERATING VOICE RESPONSES

<u>Speech Generation</u>	<u>Message Storage</u>	<u>Bit Rate from Message Store</u>
1) Compilation	Waveforms, i.e., voice recordings	40,000 to 60,000
2) Synthesis	Control parameters	1,000 to 4,000
3a) Synthesis by rule	Phonetic symbols	60-100
3b) Synthesis by rule	Alphabetic Text	100-150

COMPARISON OF METHODS
FOR GENERATING VOICE RESPONSES

<u>METHOD</u>	<u>ADVANTAGES</u>	<u>DISADVANTAGES</u>
1) Compilation of voice waveforms (40-60,000 bps)	Simple hardware and software Easy to update messages	Needs very large random-access memory Slow, stilted speech Inflexible sentence structure
2) Synthesis from stored parameters (1-4,000 bps)	Moderate memory requirements Moderate flexibility Fair-to-good speech	Needs hardware synthesizer Parameters not easily derived Messages not easily updated
3a) Synthesis-by-Rule: phonetic symbols (60-100 bps)	Modest memory requirements Flexible Easy to update messages Better speech to be expected as rules are improved	Needs hardware synthesizer CPU time to compute parameters Sophisticated programs
3b) Synthesis-by-Rule: Alphabetic Text (100-150 bps)	SAME AS ABOVE, AND: Even easier to update	SAME AS ABOVE, BUT: More computation More memory (for word lists)

is a further reduction in storage requirements by a factor of 10-20; it is possible to deal with words that are not in the stored vocabulary; and the message set is easy to make or revise, since the entries stored in memory are just the messages themselves, in written form. The disadvantages of synthesis by rule are comparable with those of synthesis from stored parameters: a hardware synthesizer is needed and the programming--after one knows how to do it--is no more complex. However, the demands for processing time are greater, and one cannot hand-tailor the parameters (as one can when they comprise the stored message) to improve the naturalness of the responses. Clearly, though, the future lies with synthesis by rule, as the rules are improved and as more demanding applications are attempted.

SUMMARY

We have tried to explain that the gift of speech--now rare with computers--will one day become very common and very useful. We are only beginning to master the synthesis-by-rule techniques that will make this promise come true, but talking computers can already do a variety of jobs, including the sophisticated one of reading to the blind and the simpler but useful tasks around the laboratory that have been described and illustrated.

REFERENCES

- Cooper, F.S., 1963. Speech from stored data. IEEE Convention Record 7, 137-149.
- Cooper, F.S., J.H. Gaitenby, I.G. Mattingly, N. Umeda, 1969. Reading aids for the blind. IEEE Trans. Audio and Electroacoustics AU-16, 266-270.
- Flanagan, J.L., et al. 1970. Synthetic voices for computers. IEEE Spectrum 7, No. 10, 22-45.
- Studdert-Kennedy, M., and F.S. Cooper. 1966. High-performance reading machines for the blind. Proc. Internatl. Conf. on Sensory Devices for the Blind. (London: St. Dunstan's) 317-342.

[The oral version of this paper presented at the 1971 IEEE International Convention (March 22-25, 1971) included a film demonstration of 1) the hand-scoring of answer sheets from psychological tests, 2) the type-in of a phonetic text, with computer monitoring, and 3) on-line modifications being made to synthetic speech. In addition, tape recordings were used to demonstrate page-length passages of text synthesized by rule. In some passages the text was supplied to the computer as a phonetic transcription; in others, the entire process of converting printed text into spoken form had been simulated.]