READING AIDS FOR THE BLIND: A SPECIAL CASE OF MACHINE-TO-MAN COMMUNICATION

BY

F. S. COOPER, J. H. GAITENBY, I. G. MATTINGLY, AND N. UMEDA

Reprinted from IEEE TRANSACTIONS
ON *AUDIO AND ELECTROACOUSTICS*
Volume AU-17, Number 4, December, 1969
pp. 266-270

COPYRIGHT © 1969—THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.
PRINTED IN THE U.S.A.

Reading Aids for the Blind: A Special Case of Machine-to-Man Communication

FRANKLIN S. COOPER, Fellow, IEEE
JANE H. GAITENBY
IGNATIUS G. MATTINGLY
NORIKO UMEDA
Haskins Laboratories
New York, N. Y.

Abstract

The development of a reading machine for the blind offers insight into current problems of computer-to-man communications and poses a technical and humanitarian challenge. Approaches to the problem include compiled speech, reformed speech, and synthesis by rule. Of these methods, synthesis by rule may offer the best long-term trade-off between quality of the speech and cost and complexity of its production. Implementation of a high-performance reading machine will involve a central service facility that can generate tape recordings or provide voice responses to remote print scanners. Technical problems, especially in providing remote on-line service, seem formidable, but the organizational problems of matching central facilities to the blind user's needs may prove to be even more so.

Manuscript received August 1, 1969.

This work was supported in large part by the Prosthetic and Sensory Aids Service, Veterans Administration, under Contract V1005 M1253.

This paper was presented at the IEEE International Conference on Communications, Boulder, Colo., June 9–11, 1969.

I. G. Mattingly is also with the University of Connecticut, Storrs, Conn.

N. Umeda is also with Bell Telephone Laboratories, Murray Hill, N. J., on leave of absence from the Electrotechnical Laboratory, Ministry of International Trade and Industry, Japan.

Introduction

Communications engineers are, of course, familiar with machines that talk to men, since that is what the telephone and radio do best. They may be less familiar with reading machines for the blind, but will, we believe, find these devices interesting not only for humanitarian reasons but also because of the technical challenges they pose as a special case of communication *by* machines *with* people.

Reading Machines for Personal Use

Perhaps we should first explain what a reading machine is, and sketch the history of attempts to build a successful one. The primary function of a reading machine is to make ordinary printed or typewritten materials available to the blind man in audible or tactile form. The oldest of these devices was the Optophone [1], invented in 1912, and "reinvented" every ten years or so since then [2], [3]. Recent variants developed in the United States have been the Battelle Aural Reading Device [4] and the Mauch Laboratories Visotoner [5]. In operating one of these devices, the user scans the printed line with a hand-held stylus and hears sound patterns that represent the shapes of the letters, one complex sound for each letter. The sound-to-letter correspondences are, therefore, essentially arbitrary, and must be learned. This is possible with a few weeks of training, though reading rates remain disappointingly low even after months or years of practice. Potentially, at least, such a device can be small enough and cheap enough to be carried and owned by the individual blind user; also it can be used to read the denomination of money, the markings on packages, and so forth. Much the same can be said of devices that deliver their information as tactile patterns [6], [7].

The Price of Performance

Since the major limitation of optophone-type reading machines is poor performance, it is not surprising that much research has been undertaken in the hope that a more ingenious method of converting letter shapes into sound will improve the performance enough to match the pace of normal speech or sighted reading. These hopes have proved illusory; moreover, there is good reason to suppose that *no* arbitrary letter-by-letter code will ever approach the efficiency of spoken language. If high performance is required, then the machine must *talk*, preferably in the user's own language [8]–[10]. One might well have made a similar assertion a century ago about the future of the telegraph: it will be enormously more useful when it learns to talk.

But if a reading machine is to talk, its speech must be based on the *identities* of the printed letters and not merely on the letters as *shapes*. Hence, optical character recognition is an essential part of the total process, though one that we can safely put aside in this discussion since most of the technical problems have already been

solved, and also because different solutions affect primarily the cost [11] and not the nature of the speech that can be generated after the graphic characters have been identified. An automatic method for generating speech, once the letters are known, is the other half of the total process and the concern of this paper.

Speech from Compiled Voice Recordings

Three of the many possible methods of generating speech from printed text will be considered. The simplest of these is to record isolated spoken words, then reassemble them in the sequence that appears on the printed page. Let us refer to this method as compiled speech. The method is less simple than it may appear; the voice recordings of the individual words must be very carefully "tailored" if the *same* recorded word is to sound reasonably natural in a wide variety of contexts [12]. The problem is much simpler when the grammatical context is constant, as in reporting the time of day, and voice response devices of this kind are now quite common. But for reading machines there is the further problem that a very large memory is required. An average word will need about 40 to 50 thousand bits of digital storage, and the vocabulary should be some 20 000 words or more. Even then, some alternative procedure must be used, since a few words in each paragraph will be missing from the dictionary. These words can, of course, be spelled, though this is very disruptive to the reader's train of thought. Perhaps better solutions can be found; for the moment let us say merely that a "spelling problem" exists, and that it is *not* trivial.

The clear advantages of compiled speech are that the single words are spoken in a pleasant human voice and that the method is simple and straightforward. The disadvantages are: that the individual words have *unchangeable* stress and intonation; that a very large random-access memory is required; and that there is a spelling problem, however large the dictionary may be.

Speech from Stored Parameters

A second method also uses information about individual words stored in a dictionary and then assembled into the sequence called for by the printed text. The stored information, though, is in the form of control parameters that drive a formant synthesizer and cause it to "speak" the word. This is, then, re-formed speech in a very literal sense.

The method has several desirable features; since the control parameters are based on actual spoken words, they can recreate those words quite accurately. Also, the synthetic process allows independent control of the speech spectrum—the part contributed by the stored parameters—and of speech intensity, voice pitch, and relative durations. Thus, phrases and sentences can be reshaped as to their stress and intonation contours in ways that make the speech much more natural than it can ever be with compiled speech. However, the spelling problem remains, and we have not mentioned the magnitude of the task that is involved in hand-tailoring the parametric descriptions for a vocabulary of, say, 20 000 words. Moreover, the extraction of ideal control parameters—whether the extraction is done by hand or automatically from a spectrum analysis—can often be surprisingly difficult. We have done some work with reformed speech, but not enough to convince ourselves of its relative virtues and shortcomings.

Speech Synthesized by Rule

A third method leads to speech synthesized by rule—synthetic speech, in a strict sense. It requires much more processing than does compiled speech, but much less memory. In fact, the only memory requirements are those for the computer program—a few thousand instructions at most. The steps in the process are to convert the spelling into a broad phonetic (or phonemic) transcription, compute the control parameters using synthesis-by-rule procedures, modify the parameters to whatever extent is feasible, and then use them to drive a speech synthesizer [13].

An attractive alternative for languages such as English, in which spelling and phonemic transcription often fail to correspond, is to use a dictionary of phonemic equivalents of a printed lexicon. The disadvantages of doing so are the additional random-access memory that is required—and the spelling problem again. There are advantages, though, in addition to the primary one of having the "correct" phonemic equivalent for words, regardless of their spelling. For example, the dictionary can include grammatical information to serve as a basis for modifications of stress and phrasing, and this can do much to make the synthetic speech sound natural [14].

How does synthetic speech of this kind compare with compiled speech? In making this comparison, one must consider the fact that synthesis by rule can be greatly improved as more is learned about the rules and the use of syntactic information in modifying synthesis procedures; on the other hand, little can ever be done to make compiled speech better than it now is. Another consideration is that some blind people find normal speech distressingly slow, and would much prefer a very substantial increase in delivery rate. Now, synthetic speech can easily meet this need, since the speech rate can be made much faster than normal—or slower—with very little loss in quality or intelligibility. Thus, the future would seem to lie with synthesis by rule.

Fig. 1 illustrates in broad outline a scheme for a reading machine using synthesis by rule and a dictionary. After a character-recognition device has converted the printed text to a machine-readable text, the words in the text are matched with their equivalent phonemic transcriptions. Each transcription includes a marking of potential word stress. The dictionary also supplies syn-

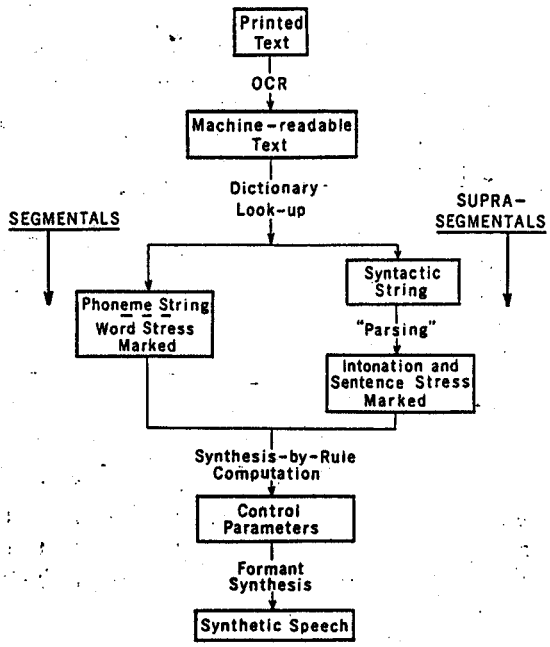


Fig. 1. Functional diagram for a reading machine that uses a dictionary, pseudo-parsing rules, and synthesis-by-rule computations to generate synthetic speech from printed text.

tactic information about the part-of-speech role that is usual for each word. This information, along with word order and the word-stress marking, is used by a parsing procedure developed by Umeda to determine which syllables in a sentence actually are marked as prominent (i.e., having "sentence stress") for purposes of the synthesis-by-rule computation. Word order, syntactic information, and punctuation are also used to mark the intonation contours and pauses. The phonemic transcription of the sentence, marked for prominence, intonation, and pause, is the input to the synthesis-by-rule computation developed by Mattingly which calculates the moment-by-moment values of the various parameters that control the generation of the synthetic speech. Unfortunately, Umeda's algorithm and Mattingly's rules for synthesis do not quite match in the functional significance they assign, for example, to stress markings. This, we think, is one of the reasons the speech falls a little short of our expectations.

Other Methods for Generating Speech

There are a number of variants that could be introduced into one or another of these general methods of making speech. We have described the processes as if words were the only possible units in which the dictionaries could be organized, but obviously a syllabary would involve far less storage and could even be complete enough to eliminate the need to spell [15]. Syllable synthesis may, indeed, be the best way to deal with the spelling problem, regardless of how most of the text is handled. However, complete dependence on a syllabary would introduce new problems, for example, breaking printed words into syllables—complete hyphenation—and inferring word stress. Also, the grammatical information that is potentially so valuable in assigning pauses and sentence stress would have to be abandoned. Other possible approaches are the use of half syllables, dyads [16], or diphones [17]—though experience thus far with these and other methods of synthesis [18] is not encouraging, and they seem to us to offer no special advantage here.

Reading Service Centers for the Blind

If we assume that good synthetic speech can be generated by one or more of these methods, we must still ask how it can be made available, in practical terms, to blind users. At present, optical character recognizers and computer-driven synthesizers with large random-access memories can be made available only through some kind of large-scale service center. One such possibility is an installation, situated within a general library, that would make tape recordings on request, even for a single individual—or more cheaply, of course, for a group. This would be most useful for the person who is reading extensively for instruction or pleasure, but inadequate for the individual who needs quick, auditory access to personal papers, letters, and books in the course of his daily work. For these latter individuals, it may be possible to provide hand-operated scanners that can transmit graphic information via telephone into an on-line service center, which would then send spoken messages back to the user. In a technical sense, the on-line service center must have essentially the same equipment that would be needed to make recordings, and in addition, must have time-sharing capabilities to serve a number of users in real time.

In these situations, requirements and system design interact strongly, especially for the on-line service center. For example, the user may wish to have the information returned word-by-word as he scans, or he may wish it phrase-by-phrase or sentence-by-sentence for easier comprehension. This preference determines the turn-around time within which the central processor must locate items in its dictionary and may determine whether that dictionary must be organized for true random access, or can scan its memory for all the words wanted by all the users over a span of a few seconds. There are engineering tradeoffs as well; for example, how much of the character recognition or speech synthesis should be done at the user's terminal, and how much at the central computer? The data load—hence, the line requirements and costs—are much affected by these choices. A quick calculation will make it clear that data loads are, indeed, a problem.

The operating costs of a service center will certainly be high, at least for the foreseeable future. However, the cost of recordings, on the basis of dollars per hour of speech, can be kept in the same range as the cost of using human readers who are paid for their time [8]. A service

center could supply book-length recordings promptly—as human readers could not—and it could also provide speeded speech, when wanted. The primary factor in determining cost is how well the demands for service match the capacity of the system. Thus, although the technical problems of providing a high-performance reading service for the blind are scarcely trivial, they may well be nearer solution than the practical problems of organizing such a service on a realistic basis.

References

[1] F. d'Albe, "On a type reading Optophone," *Proc. Roy. Soc.* (London), vol. 90, ser. A, no. A619, 1914.
——, *The Moon Element*. London: T. Fisher Unwin, 1924.

[2] F. S. Cooper, "Research on reading machines for the blind," in *Blindness*, P. A. Zahl, Ed. Princeton, N. J.: Princeton University Press, 1950, pp. 512–543.

[3] H. Freiberger and E. F. Murphy, "Reading devices for the blind: an overview," in *Human Factors in Technology*, E. Bennett, J. Degan, and J. Spiegel, Eds. New York: McGraw-Hill, 1963, pp. 299–314.
——, "Reading machines for the blind," *IRE Trans. Human Factors in Electronics*, vol. HFE-2, pp. 8–19, March 1961.

[4] J. L. Coffey, "The development and evaluation of the Battelle aural reading device," *Proc. Internatl. Congress on Technology and Blindness*, vol. 1, L. L. Clark, Ed. New York: American Foundation for the Blind, 1963, pp. 343–360.
J. S. Abma, "The Batelle aural reading device for the blind," in *Human Factors in Technology*, E. Bennett, J. Degan, and J. Spiegel, Eds. New York: McGraw-Hill, 1963, pp. 315–325.

[5] G. C. Smith, "The development of recognition and direct translation reading machines for the blind," *Proc. Internatl. Conf. on Sensory Devices for the Blind*, R. Dufton, Ed. London: St. Dunstan's, 1966, pp. 367–387.

[6] The Mauch Laboratories Visotactor is discussed in [5], p. 370 ff.

[7] J. Linvill, "Development progress on a microelectronic tactile facsimile reading aid for the blind," this issue, pp. 271–274.
J. C. Bliss, "A relatively high-resolution reading aid for the blind," *IEEE Trans. Man-Machine Systems*, vol. MMS-10, pp. 1–9, March 1969.

[8] M. Studdert-Kennedy and F. S. Cooper, "High-performance reading machines for the blind: psychological problems, technological problems and status," *Proc. Internatl. Conf. on Sensory Devices for the Blind*, R. Dufton, Ed. London: St. Dunstan's, 1966, pp. 317–342.

[9] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code," *Psychol. Rev.*, vol. 74, pp. 432–461, 1967.

[10] F. F. Lee, "Reading machine: from text to speech," this issue, pp. 275–282.

[11] See [5] and [10] for descriptions of optical character recognizers intended for use in blind reading aids.
E. A. Parrish, Jr., J. W. Moore, and E. S. McVey, "An experimental personal reading machine for the blind," *IEEE Trans. Bio-Medical Engineering*, to be published.

[12] J. H. Gaitenby, "Problems in machine conversion of print to 'speech'," *Bull. Prosthetics Research*, vol. 10, pp. 37–44, Spring 1966 (note in particular pp. 40–41).

[13] I. G. Mattingly, "Synthesis by rule of general American English," Ph.D. dissertation, Yale University, New Haven, Conn., 1968; also *Supplement to Status Report on Speech Research*, Haskins Labs., New York, N. Y., April 1968.
M. P. Haggard and I. G. Mattingly, "A simple program for synthesizing British English," *IEEE Trans. Audio and Electroacoustics*, vol. AU-16, pp. 95–99, March 1968.
References to other research on speech synthesis methods adaptable to blind reading aids will be found in Mattingly (above); in *Reports of the 6th Internatl. Congress on Acoustics* (Tokyo, 1968), vol. 2; and in *IEEE Trans. Audio and Electroacoustics* (Special Issues on Speech Communication and Processing), vol. AU-16, March and June 1968.

[14] R. Teranishi and N. Umeda, "Use of pronouncing dictionary in speech synthesis experiments," *Proc. 6th Internatl. Congress of Acoustics* (Tokyo, Japan), vol. 2, p. 155, August 1968.

[15] G. Dewey, *Relative Frequency of English Speech Sounds*. Cambridge, Mass.: Harvard University Press, 1923.

[16] G. E. Peterson, W. S.-Y. Wang, and E. Sivertson, "Segmentation techniques in speech synthesis," *J. Acoust. Soc. Am.*, vol. 30, pp. 739–742, 1958.

[17] N. R. Dixon and H. D. Maxey, "Terminal analog synthesis of continuous speech using the diphone method of segment assembly," *IEEE Trans. Audio and Electroacoustics*, vol. AU-16, pp. 40–50, March 1968.

[18] See [13] for sources.

Franklin S. Cooper (A'46–M'55–F'69) was born in Robinson, Ill., on April 29, 1908. He received the B.S. degree in engineering physics from the University of Illinois, Urbana, in 1931, and the Ph.D. degree in physics from the Massachusetts Institute of Technology, Cambridge, in 1936.

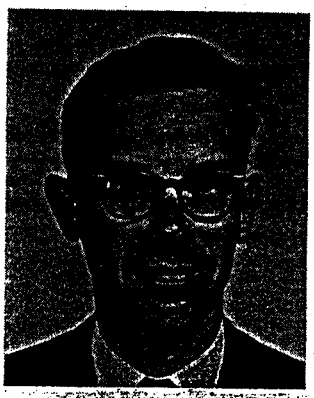
He was with the Research Laboratory of the General Electric Company from 1936 to 1939, where he worked on high-voltage conduction and breakdown processes in liquids and compressed gasses. He has been associated with Haskins Laboratories, New York, N. Y., since its founding in 1935, serving as Associate Research Director from 1939 to 1955, and as President and Research Director since 1955. During World War II, he was a Senior Liaison Officer of the Office of Scientific Research and Development, and was awarded the President's Certificate of Merit in 1948. At Haskins Laboratories, he has done interdisciplinary research on radiation biophysics, physiological effects of drugs and endotoxins, and currently, on sensory aids for the blind and the acoustic and articulatory nature of speech and the perceptual processes involved in its reception.

Dr. Cooper is a member of Tau Beta Pi, Sigma Xi, and the American Speech and Hearing Association, and a fellow of the Acoustical Society of America.

Jane H. Gaitenby (Mrs. Joseph J. Newman) was born in Niagara Falls, N. Y., on May 10, 1923. She received the B.A. degree in anthropology from Hunter College, New York, N. Y., in 1953, and did graduate work in linguistics and anthropology at Columbia University, New York, from 1953 to 1956, as an Atran Scholar.

Before and during her college years she was successively an artist, a social worker at Madison House Settlement, New York, a cartographer at the Language and Communications Research Center at Columbia University, and a volunteer in the archeology section at the American Museum of Natural History. In the fall semesters of 1958, 1960, and 1962, she taught linguistics at Hunter College. Since 1957 she has been a member of the research staff of Haskins Laboratories, New York, where her primary interest has been in spoken and written English phrase structure (the interrelations of morphology, syntax, stress, and intonation), especially for application to the development of a spoken output for reading machines for the blind.

She is a member of Phi Beta Kappa and the Acoustical Society of America.



Ignatius G. Mattingly was born in Detroit, Mich., on November 22, 1927. He received the B.A. degree from Yale University, New Haven, Conn., in 1947, the M.A. degree in linguistics frim Harvard University, Cambridge, Mass., in 1959, and the Ph.D. degree in English frim Yale University, in 1968.

He taught at Groton School from 1947 to 1948, and at Yale University from 1950 to 1951. From 1951 to 1966, he was employed by the U.S. Department of Defense. Since 1966, he has been associated with Haskins Laboratories, New York, N. Y., and has taught at the University of Connecticut, Storrs. As a Guest Researcher at the Joint Speech Research Unit, Eastcote, England, from 1963 to 1964, he became interested in speech synthesis by rule, and has published a number of papers on this subject.

Dr. Mattingly is a member of the Technical Committee on Speech of the Acoustical Society of America, the Linguistic Society of America, the Linguistic Circle of New York, and Phi Beta Kappa.

Noriko Umeda was born in Kobe, Japan, on January 10, 1933. She received the B.A. degree in 1957 and the M.A. degree in 1959, both in linguistics, from the University of Tokyo, Japan.

In 1962 she joined the research staff of the Electrotechnical Laboratory, Ministry of International Trade and Industry, Japan, working on speech analysis and synthesis. In February, 1969, she joined Bell Telephone Laboratories, Inc., Murray Hill, N. J., on a leave of absence from the Electrotechnical Laboratory. She has been working on rules for speech synthesis. She is also associated with Haskins Laboratories, New York, N. Y.

Mrs. Umeda is a member of the Acoustical Society of America.