Reprinted from:

International Conference on
Sensory Devices for the Blind
June 1966, St. Dunstans, London

HIGH-PERFORMANCE READING MACHINES FOR THE BLIND

Psychological Problems, Technological Problems and Status

by

Dr. MICHAEL STUDDERT-KENNEDY* and FRANKLIN S. COOPER

Haskins Laboratories, New York, N.Y.

The task of the reading machine is to code written text for easy and rapid assimilation. We have no wish to belittle low-rate reading machines, whether tactile or acoustic, but we will confine our attention here to high-performance acoustic devices that might be expected to yield listening rates close to that of normal speech.

Of the acoustic displays currently available for such a device there is little question that the most efficient is speech. The first part of this paper will attempt to show why this is so, by examining principles on which the speech code operates. We will also discuss the problems of applying these principles to non-speech or even to a special class of "speechlike" signals that might be produced by relatively simple mechanisms.

The second part of the paper will start from the point that speech is a satisfactory signal—indeed the only one presently available—and will canvas ways in which it can be generated by a reading machine, the relative advantages of the several methods, and the current status of research along these lines.

Normal speech may be comfortably followed at a rate of more than 200 words a minute. The listener handles some forty to fifty bits of information a second. This rate is an order of magnitude greater than listeners have achieved with any non-speech code so far developed. There seem to be two main reasons for this. First, speech signals are multi-dimensional and their dimensions are not arbitrary: they are determined and organized by characteristics of the articulatory apparatus that generates the signal. Second—a reason closely related to the first—speech signals form a complex pattern of overlapping, or shingled, cues, a flowing auditory display of which elements are provided to the listener in parallel rather than in series.

* Also at Barnard College, Columbia University, New York, N.Y.

We may illustrate the dimensions of speech with an example. Figure 1 presents a spectrogram of the syllable [ga]. It is not immediately obvious how the very large amount of information in the display may be reduced to humanly manageable levels. But enough has been learned for us to describe the important dimensions of the display and the articulatory events that give rise to them. First, the vocal tract is closed near the velum (a period of silence). Closure, during which air pressure builds up in the oral cavity, is followed by release (burst of noise centred around 2000 cps), by the onset of laryngeal vibration (harmonic structure and an intense first formant), and by movement of the articulators (formant transition) into a more open position for the following vowel (sustained formant pattern).

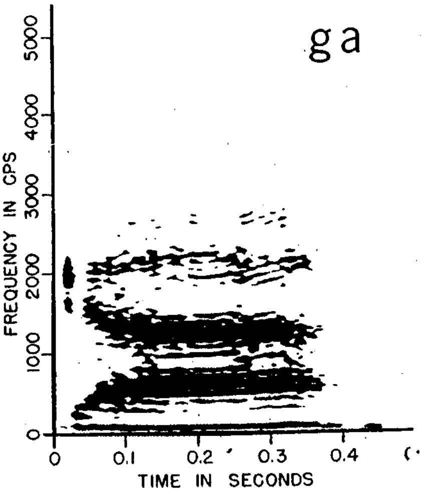


FIG. 1. Spectrogram of the syllable [ga].

The acoustic events of this description (noise, formant array, formant frequency, direction of formant transition) are known to be among those used by listeners to identify the syllable. But it is important to note that they differ from the usual acoustic dimensions of frequency, intensity and duration: their character and necessary sequence are determined by articulatory events. These events may also serve to define the phonemes. The frequency and duration of the noise burst, together with the frequency, direction and extent of the second formant transition, serve to identify the first phoneme as a velar stop consonant [k, g]. The timing of the onset of the first formant relative to the noise burst distinguishes a voiced from a voiceless consonant and permits the identification of [g].

Such a classification of speech sounds in articulatory terms follows not only conventional phonetic practice, but also the evidence of several perceptual studies. A number of these have dealt with the acoustic features that may serve as cues in the perception of the phonemes (for general reference, see Liberman, Cooper, Harris, MacNeilage and Studdert-Kennedy (in press); also, for example, Liberman, Delattre and Cooper, 1952; Liberman, Delattre, Cooper and Gerstman, 1955; Liberman, 1957). Our description of the spectrogram suggests that the listener might, in principle be able to check off, one by one, the identifying features, gradually narrowing his range of possible confusion. There is evidence from reaction time studies reported by Chistovich (1962) and by Kozhevnikov and Chistovich (1965) that for some phonemes—certain Russian voiced stops—the listener is able to do just this, correctly narrowing his choices to the class of voiced stops before he has yet received the information that permits him to identify the point of closure. Related to this is the evidence of Miller and Nicely (1955) that perceptual confusions among consonants, heard through noise, are systematic: certain features, defined by the authors in articulatory terms, tend to be lost while others are correctly heard. More recently, Wickelgren (1966) has shown similar systematic errors in short-term memory for consonants.

Speech is evidently organized along a characteristic set of dimensions. Whether we treat the dimensions as articulatory or attempt to define them in acoustic terms, the underlying organization remains articulatory. From it there flow advantages of parallel processing, reduction of error, and a consequent increase in rate, all suggested by the studies we have cited. The advantages may be inherent in the system, due to the fact that the listener is also speaker, so that his perceptual processes are facilitated by some form of articulatory reference (Liberman, Cooper, Studdert-Kennedy, Harris and Shankweiler, 1966), or they may stem from the listener's long practice with the code. In either event, we may doubt whether the advantages will accrue to a multi-dimensional signal unless the dimensions and organization of speech are used.

Before commenting further on this, we will discuss a second characteristic of speech, its overlapping cues. Perhaps we have given the impression in describing the spectrogram that cues for the identification of [g] are perfectly distinct from cues for the identification of [a]. This is certainly not so. We may show this most simply by attempting to cut a tape recording of the utterance in such a way as to separate [g] from [a]. If we do this, we will find it impossible to achieve a satisfactory [g] without at the same time getting the impression of a following vowel (cf. Harris, 1953; Peterson, Wang and Sivertsen, 1958). Given the brevity of the formant transition (approx. 50 msecs.), it is not surprising that the listener requires some information where the formant is going *to*, in order to assess where it is coming *from*. Similarly he may need information on where a formant is coming *from*, in order to decide where it is going *to*—particularly if it does not stay for long at its point of arrival.

We may illustrate this by describing an experiment recently carried out at Haskins Laboratories in collaboration with Dr. Björn Lindblom of the Royal Institute of Technology in Stockholm. A series of steady-state synthetic vowels was prepared on the Institutes synthesizer, OVE II. The second and third formants were varied in approximately equal steps along a continuum that ranged from a vowel heard as [ɪ] (as in "bit") to a vowel heard as [u] (as in "book"). The vowel series was then placed in dynamic context so as to provide a series ranging from [wɪw] to [wuw]. Figure 2 schematizes the two extreme stimuli of this dynamic series. Both series, the steady-state and the dynamic, were prepared at two durations (fast: 100 msec, and slow: 200 msec).

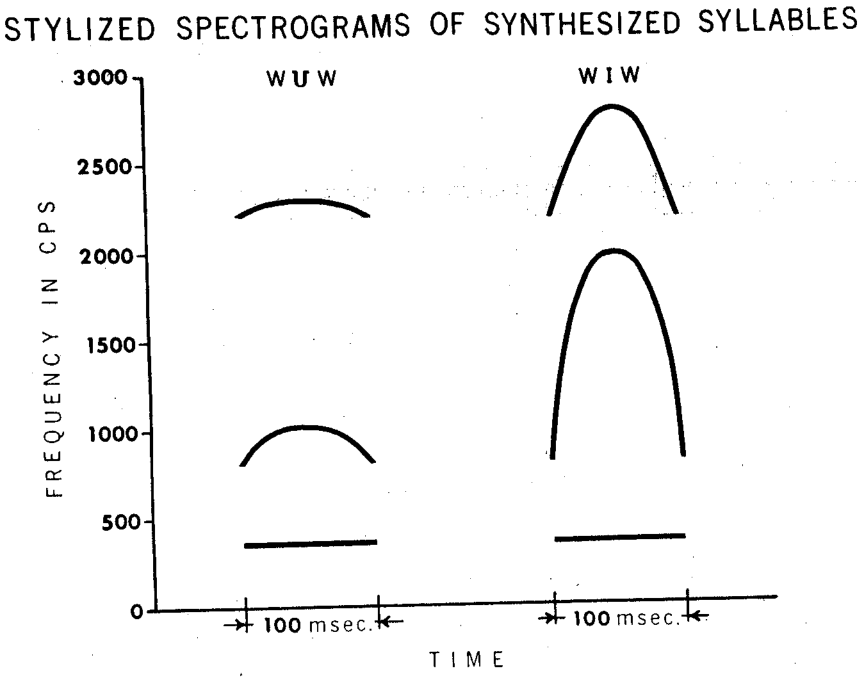


FIG. 2. Stylized spectrograms of synthetic syllables for experiment on the perception of vowels in dynamic context.

The purpose of the experiment was to determine whether there was a shift in the phoneme boundary between [ɪ] and [u] as a function of context rate of utterance. Tape-recordings of the stimuli were spliced into random orders and presented to a group of listeners. They were asked to judge of each stimulus whether its vowel was [ɪ] or [u].

Figure 3 displays the results. The percentage of responses in each vowel class is plotted as a function of stimulus number in the series. Above are results for the steady-state or null context, below for the consonantal context; on the left are results for the slow stimuli, on the right for the fast. Comparing the two plots on the left, we see that the effect of placing the

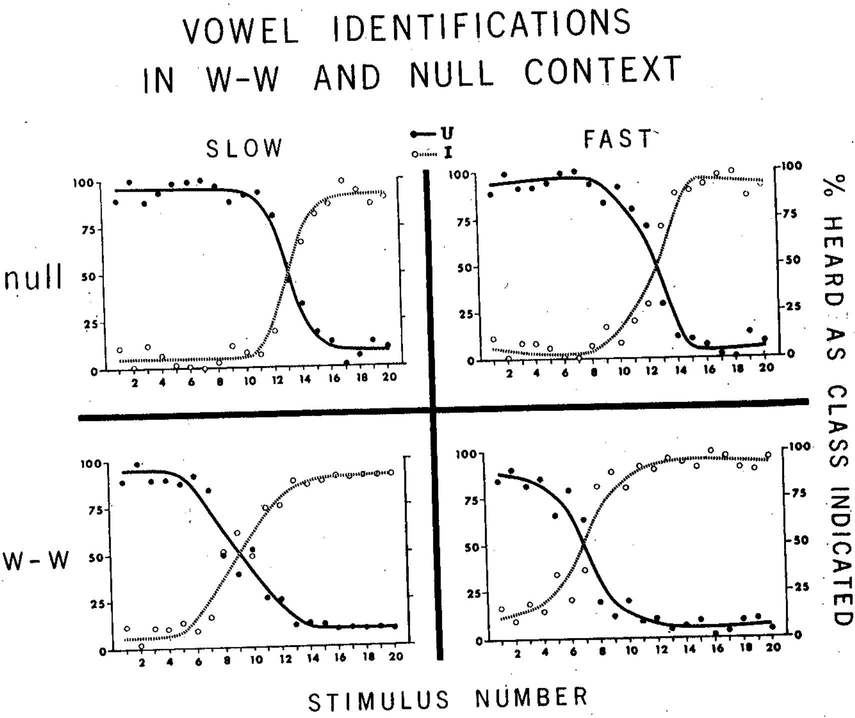


FIG. 3. Results of experiment comparing identification of vowels in null and dynamic contexts at two rates of utterance.

vowels in this dynamic context is to shift the phoneme boundary so that more of the stimuli are heard as [ɪ]. Since placing the vowels in dynamic context effectively reduces the time available to move from consonant to vowel and back, it is not surprising that a greater shift in phoneme boundary is observed where duration is further reduced, in the fast series (right hand plots). There is little effect of rate when the vowels are not in dynamic context (top row).

The results suggest that listeners somehow compensate for the information trimmed away by the rapid transitions into and out of the vowel: they extrapolate from the acoustic data to the formant "target" at which the speaker seemed to be aiming. But we need not speculate here on the perceptual mechanism involved. It is enough to see that listeners' identification of a rapidly articulated vowel may depend upon its neighbouring consonants—just as perception of a consonant may depend upon information from a neighbouring vowel.

Such overlap of cues for neighbouring phones is typical of speech. Its effect on the temporal economy is not hard to imagine. For example, in some of the synthetic stimuli of this experiment three phones were packaged into 100 msec, a rate, if sustained, of 30 phones per second or about 360 words per minute. Had the phonic elements been discrete, non-speech signals, each would have lasted little more than 30 msec, and the composite would presumably have been an unintelligible buzz.

But there is more than simple temporal overlap involved. The acoustic signals sufficient for identification of the vowels of the experiment were actually different in different contexts. This brings us back to our comments on the dimensions of speech. For these signal changes were not arbitrary. They were determined—or would have been determined in natural speech (Stevens and House, 1963; Lindblom, 1963)—by characteristics of the articulatory apparatus. The tongue, drawn back for the production of [w] both before and after the vowel, has little time to reach forward to form a vowel of the same formant pattern as the steady-state front vowel [ɪ]. The resulting pattern may be shifted substantially toward that of the steady-state back vowel [u]—and yet may be heard as [ɪ].

The overlap of cues in the speech signal is related to the rapidity of the articulatory movements that generate the signal. The rapidity of these movements may cause changes in the acoustic cues themselves. The final pattern is an intricate, peculiar and highly efficient code. Once again, we may doubt whether the efficiency of the system can be duplicated by some arbitrary non-speech code.

Finally, we should like to consider the application of the principles we have been discussing to the design of a reading machine. Presumably the output of the ideal machine should form a multi-dimensional pattern of overlapping cues. We may conceive of three possible types of device to fulfil these requirements: (1) a simple machine with a non-speech output; (2) a fairly simple machine with a "speechlike" output; (3) a more complex machine that generates speech.

The possibilities of various non-speech machines have been studied rather exhaustively over the years and there is no need to enumerate their difficulties again. One possible path of improvement—multi-dimensionality—has now been tried. The value of multi-dimensionality in reducing errors of identification had been shown by many studies (for example, Klemmer and Frick, 1953; Miller, 1956; Pollack, 1952; Pollack and Ficks, 1954). Recently, Nye (1965) modified the conventional optophone to permit variations in the number of dimensions in its display. He found than an increase in the number of dimensions could improve rate as well as accuracy of identification. But performance was still poor—less than 2 bits/second with the more successful display.

A "speechlike" device may seem more promising. There are two main requirements: (1) It should avoid all-out character recognition since this involves elaborate and expensive machinery. Yet the number of features that must be extracted from letter shapes, and used to control a suitably multi-dimensional output will be about half as many as would suffice for all-out character recognition. Thus, the reduction in machine complexity will not be great. (2) It must generate "speechlike", i.e. pronounceable, sequences from the letter-shape features. Nye (1965) has tested a display that was controlled by letter features and used acoustic dimensions available on a speech synthesizer, such as hiss, buzz frequency, formant frequency. His listeners' performance improved over their performance with clearly non-speech displays, but still fell far short of what would be expected with speech itself. One difficulty seems to have been that the sound sequences were not easily pronounceable. Perhaps a better choice of dimensions could have been found, but the reasons for the difficulty lie deeper. One is that the acoustic dimensions of speech, subject as they are to articulatory constraints, do not combine freely in arbitrary ways. Hence some combinations of the shape features of letters would lead to unspeakable combinations of the distinctive features of speech sounds. Similar difficulties would attend an attempt to introduce overlapping cues into the signal: the overlap would be arbitrary, since the shape features of adjoining letters are independent. Thus, the dimensional patterns within and across phonemes would not be linked by a common principle of organization, as in speech. The patterns would therefore be only marginally "speechlike", and the advantages of speech would not accrue.

To sum up, we have argued that the only known acoustic signal adequate to the coding of written text for easy and rapid assimilation is speech. The advantages of speech stem from its intricate, multi-dimensional pattern of overlapping cues, determined and organized by the articulatory apparatus. Some other satisfactory set of dimensions and principle of organizations might be found. But there is no rationale for the search and we seem little closer to the discovery today than we were fifty years ago. We are therefore inclined to accept our fate and to give our attention to the development of a reading machine that talks.

In the first part of this paper, a case has been made for speech as the *preferred*—if not the only available—acoustic output for a high-performance reading machine. If one accepts this conclusion, then certain other questions arise: What methods are available for generating speech from the printed page? What are the relative merits of these methods, in technical terms—in economic terms? In short, is a high-performance reading machine that talks ordinary English realizable, in any practical sense? In the remainder of this paper, we shall try to deal with these questions and, in the absence of a demonstrable working system, we shall tell you what we are trying to do at Haskins Laboratories about the evaluation of synthesis methods and preparations for pilot studies leading to a library-type installation.

How can one make the printed page talk? There are three general methods, though they permit a bewildering variety of combinations and modifications. Each of these methods has its special problems, and these too have many variants (Cooper, 1963). One of the general methods is to generate *synthetic speech* on the basis (essentially) of how the words are spelled, that is, in much the same way that a child "sounds out" unfamiliar words if he is learning to read by the phonic system. The second method is to produce *compiled speech* from a stored vocabulary of voice recordings of individual words. The third method, something of a hybrid between the other two, also makes use of stored data about all the individual words that are likely to be encountered, but now the stored information is a set of control signals that will enable a speech synthesizer to "say" the word. This kind of speech, reconstituted from stored control signals, might be called *re-formed speech.*

It is clear by hindsight that the three methods are not equally easy to realize, nor will they yield speech of equal acceptability, at least initially. We shall see that the third method, in one or another of its variants, seems best suited for use in initial trials. It may, nevertheless, be useful to examine all three methods in somewhat more detail, noting their merits and limitations.

Synthetic speech. The series of studies on acoustic cues for speech perception, mentioned earlier, had made it possible by the late fifties to consider the use of these cues in a set of rules that could be applied to a text (written in phonemic transcription) and that would generate synthetic speech from it. The rules for synthesis (Liberman, *et al.*, 1959; Cooper, *et al.*, 1963) were initially drawn up as a kind of cookbook for painting spectrographic patterns that could be converted into sound on a pattern playback device. The essential point, however, is that the conversion of a phoneme sequence into intelligible speech was done in a rigidly prescribed manner, and so in a way that could be automated. The use of a computer to synthesize speech by essentially these rules was demonstrated by Gerstman and Kelly (1961); related studies are under way in several laboratories (Holmes, Mattingly, Shearme, 1964). Figure 4 indicates schematically the steps in making synthetic speech for test purposes and the way in which such a reading machine would operate; the figure also shows some examples of the hand-drawn patterns used in the research.

The rules for synthesis are comparatively simple and it may be possible to keep them so even after additional rules are added to make the phrasing and intonation more lifelike. Thus, a comparatively small computer, or special-purpose circuitry, and a formant-type synthesizer would suffice to generate synthetic speech from a phoneme sequence.

There are, however, two other problems. One is to meet the need, common to all high-performance reading machines, for an optical character recognizer that can scan the printed page and identify the successive letters of the text. The other is to convert English spelling into pronounceable form—a formidable task if it is taken seriously, or a trivial one if the listener is prepared to accept some bizarre sound sequences in the place of familiar words. These substitution forms do not change, however, and it should be possible to infer the spelling for unfamiliar words.

Thus, synthetic speech, as a means of realizing a reading machine, poses a very real dilemma: It is potentially a simple method, but an "iffy" one—it will work *if* a simple letter recognizer can be built (Mauch and Smith, 1966),

Synthetic Speech:

1) Printed Page
2) Optical Character Recognition
3) Spelling → Phonemes

HL Research	Prospective Reading Machine
4) Apply Rules for Synthesis to get	4) Apply Rules for Synthesis to get
5) Hand-drawn Spectrogram.	5) Control Signals equivalent to a Spectrogram.
6) Use Pattern Playback to hear	6) Use C.S. with a Formant-type Synthesizer to hear
7) Synthetic Speech.	7) Synthetic Speech.

Examples of hand-drawn patterns for synthesis by rule:

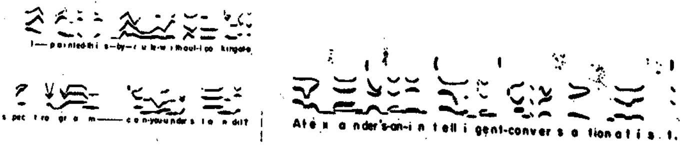
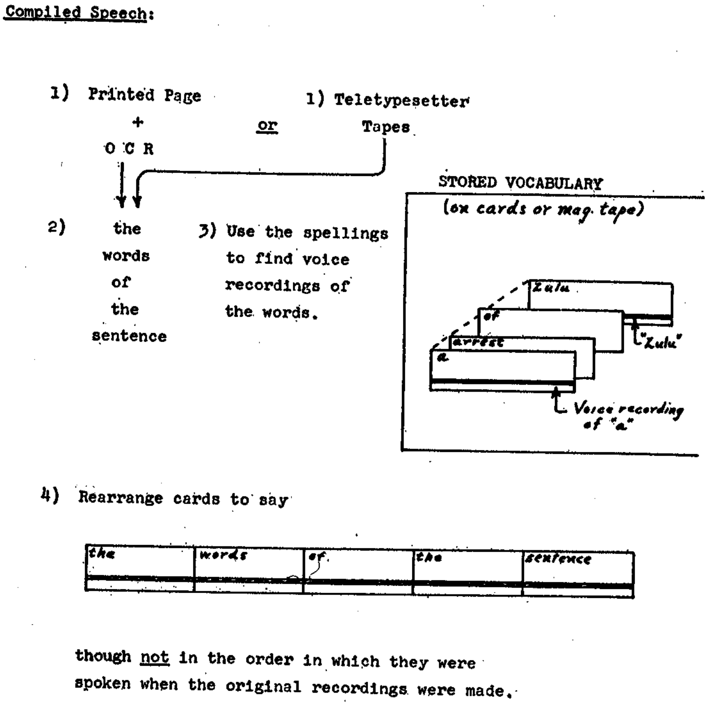

I—painted-this—by—rule-without-looking
spectrogram——can-you-understand-it
Alexander's-an-intelligent-conversationalist.

FIG. 4.

if special circuitry can be designed for implementing the rules, and *if* the listener will be satisfied with bizarre pronunciations and less-than-perfect intelligibility. None of these can be resolved without a great deal more research. The other horn of this dilemma is that really good performance will require much more work on rules for synthesis and on converting spelling to phonemes—and after these improvements, the equipment needs may be as great as those of other methods for generating speech.

Compiled speech. The simple and obvious way to get speech is to put it together, word-by-word, from recordings of the individual words of the language. This resolves immediately all the problems of correct pronunciation, clear articulation and even pleasant voice quality. Figure 5 shows schematically how a compiled sentence would be prepared from recordings of words stored on strips of magnetic tape that are attached to file cards. Automation of the method is straightforward: the spelling of successive words of the text are provided by an optical character recognizer (or by Teletypesetter tapes, when these are available), the spellings serve as addresses in finding the voice recordings, and the recordings are assembled by a start-stop tape recorder. (For the description of an experimental device, an Interim Word Reading Machine, see Cooper, 1963.)



Compiled Speech:

1) Printed Page + O C R

or

1) Teletypesetter Tapes

2) the words of the sentence

3) Use the spellings to find voice recordings of the words.

STORED VOCABULARY

4) Rearrange cards to say

though not in the order in which they were spoken when the original recordings were made.

FIG. 5.

The method has its problems, some obvious and some not. One might expect good intelligibility if the individual words are clearly spoken. Most of the commercial devices that give stock quotations or directory information by telephone use essentially this method. They can operate with a limited vocabulary tailored to a set context, and so produce quite a good speech. It comes as something of a shock to hear how unintelligible a sequence of words can be if those words were merely lifted out of other spoken sentences that happened to contain them. The method depends, of course, on havin g pronunciations that can serve acceptably in *any* context—not only as to intelligibility, but also without bizarre effects on the naturalness of the compiled speech. It is possible, by suitable methods, to get word recordings that will meet this need, and so to generate texts that are highly intelligible—and that even sound fairly natural—at rates of well over 100 words per minute. There is an occasional odd effect when the context leads one to expect a particular pronunciation, but the recording provides a different one.

A more disturbing feature of compiled speech is the interruptions caused by words that were *not* included in the recorded vocabulary. These can be spelled from recordings of the letter sounds, but this breaks the listener's train of thought and becomes annoying if it happens too frequently. This "spelling problem" is closely tied to another problem, that of the size of the storage device. If many thousands of words could be stored, then spelling would be infrequent and not troublesome, but this would be gained only at the cost of a large and expensive memory. As a very rough indication of the trade-off between frequency of spelling and size of vocabulary, one can expect that about one word in twenty will have to be spelled if the vocabulary is in the range of 5–10,000 words, or about one word in a hundred, if the vocabulary is 15–20,000 words. A desk-type dictionary such as Webster's Collegiate contains about 60,000 different words.

Whatever compromise is made between inconvenience to the listener and size and extent of the storage equipment, the method of compiled speech will require a large installation, and will have an inherent upper bound on the naturalness of its output, since very little can be done to change the acoustic shapes of words once they are in recorded form.

Re-formed speech. This is a hybrid method that compiles control signals on a word-by-word basis and then uses them to operate a speech synthesizer, following the steps outlined in Fig. 6. There are several advantages: (1) The

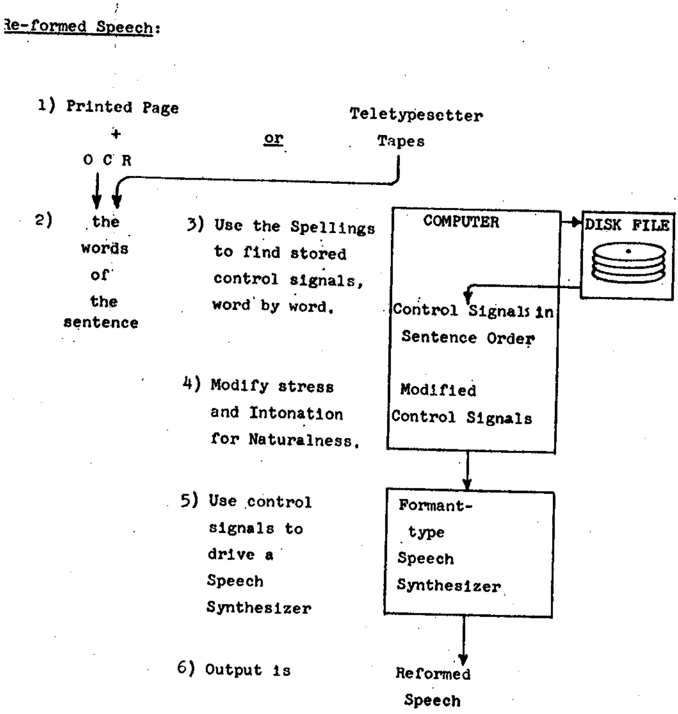


FIG. 6.

storage requirements are drastically reduced. Good speech can be generated from 2400 bits per second for the control signals as compared with 48,000 bits per second or more for the digital storage of speech waveforms. (2) Even though the speech is synthetic, the control signals preserve the correct pronunciation and syllable stresses of human speech, and do so regardless of how the printed word may have been spelled. (3) In additon, because the word is synthesized, it is possible to modify the pitch and intensity patterns of successive words in ways that will make the speech more lifelike. Commas and periods can serve as signals to introduce sustentions and terminal junctures that will further contribute to naturalness; moreover, the way is open to still further improvements as more is learned about how to infer the syntactic structure of sentences and to use this information, as a person does, in controlling stress and intonation.

There are disadvantages, too. The requirements for equipment are considerable, and it is unlikely that this method would ever be adaptable to anything but a library-type installation. A further difficulty is that the spelling problem remains, though in a less severe form since it will be feasible to store very large vocabularies. But vocabulary size alone is not an answer and some method better than letter-by-letter spelling needs to be worked out. Perhaps Spelled Speech of the type that Metfessel (1966) has described will serve, or perhaps synthesis-by-rule can be used for the non-vocabulary items. But a more promising method would appear to be syllable-by-syllable synthesis, using stored control signals for a syllabary that is stored as part of the main vocabulary. Certainly, there will be problems with this "syllable spelling", first in deciding where to introduce syllable breaks into the written form of the word, then in linking the control signals to get smooth acoustic transitions from one syllable to another, and again in assigning stressed or unstressed status to the individual syllables of polysyllabic words. These are difficult, but not unmanageable, problems. They do not require solution before a re-formed speech system could be put to use; rather, they offer improvements and eventually, perhaps, a simpler type of system in which syllable synthesis would be used exclusively. Such a system is described briefly in the Appendix, but we should remember that much research lies between this goal and where we stand today.

Overall System Requirements and Operating Costs

The emphasis thus far has been on methods for generating speech. This is, of course, an essential element in the kind of reading machine we are considering. What are the other elements? An obvious one is an optical character recognizer that will identify the letters on printed or typewritten pages. The need for the recognizer can be bypassed during pilot tests, and even afterward for certain types of material, by using Teletypesetter tapes that have been discarded by the printing industry. But this is only a partial solution, and an optical character recognizer must be considered an essential element in the system. Equipment of this type is under very intensive development (Feidelman, 1966) for use in processing business documents.

sorting mail, and other commercial applications, though the machines that are now in use fall a little short of our requirements: those that are available as production models operate only with special type fonts designed for the purpose (though they may read several such fonts); true multifont machines that could cope with most of the books and typewritten materials in a library are still in the experimental, or one-of-a-kind class. It is reasonable to expect, from present progress, that suitable optical character recognizers for multifont work will be commercially available within a very few years.

There is little prospect, however, that either these recognizers or the speech generating equipment that we have already discussed will be either simple or cheap for quite some time. This sets a third requirement on high-performance reading machines, namely that the "machine" will have to be a service establishment operating in the context of a library or comparable institution. For economic reasons, it will have to serve the blind reader by preparing tape recordings for him and for others who may wish to read the same material. Only in this way can the high operating speeds of character recognizers and computers serve to offset their high costs.

It is instructive to look at these costs, partly because they point out the problems that could make or break the entire undertaking, and partly because one would like some indication of whether high performance reading machines are a practical possibility or only a mirage, however alluring technologically. In the paragraphs that follow we shall summarize some estimates that are rough at best and involve weightings and choices of alternatives that may be strongly biased by our misconceptions and limited familiarity with some of the equipment. We shall try to indicate, in an Appendix, the bases for the estimates, but we are not prepared to defend the resulting cost figures, except in a very general way.

What is a reasonable way to compare costs between a large installation and other options that might be open to the blind individual, or to governments or organizations that provide services to him? In what follows, we have reduced all of the dollar figures to the cost per hour of what the blind man will hear. If this is the voice of a friend reading aloud, then there is no cost in dollars. If it is the time of someone who is hired to read to him, or to make recordings, then it is the hourly rate plus tape and mailing charges. This kind of situation is presumably the one against which comparisons should be made, especially if the reading machine is assumed to be used principally to meet individual requests. The cost per reader hour would, of course, be very much lower if the same tape-recording served a number of individuals.

The principal items we have considered in trying to guess the costs of a high-performance reading machine installation are (1) rental costs for the optical character recognizer, (2) rental costs for a computer and associated equipment for use in generating the speech, and (3) the labour and materials needed to operate a service centre in connection with a library.

The results of the estimates sketched out in the Appendix are summarized in Tables 1 and 2. The first three systems referred to in the tables

Table 1
Estimates of Quality and Speed

System	*Speech quality**	*Speed factor*	*Equipment*
1. Synthesis-by-Rule	3–4†	20×·RT	Medium to small computer
2. WRM/A, D	5†	$\frac{1}{30}$ to $\frac{1}{4}$	Experimental Basis with Interim WRM or Computer
3. PRM/1	10	4	Medium to small computer with disc files
4. PRM/2	8	30	Medium computer with 32 K core, 48 bits; disc file
	8	300	Medium to large computer, 100 K core, 48 bits; Mag. tape.
5. WRM/s	5	300	Medium computer with 32 K core, 48 bits; Mag. tape

* Estimates of relative quality are on a ten-point scale.
† Speech should be rather slow for good intelligibility.

Table 2
Estimate of Operating Costs

System	*Cost of generating speech*	*Overall costs*	*Comments*
1. Synthesis-by-Rule	$2/hr	$7/hr	
2. WRM/A, D	160	—	Experimental use only
3. PRM/1	10	15	
4. PRM/2	2–10	5–15	Depends on system size
5. WRM/s	1–2	4–5	Depends on system size

Note: Costs are per hour output speech.
Overall costs include the cost of generating speech (computer rental), optical character recognizer @ $2/hr, and labour and materials @ $1–4/hr.

are essentially the three we have already discussed. Systems 4 and 5 represent possible methods that would operate faster, with a better balance between the operating rates for character recognition and speech generation, and so at lower overall cost per hour. The five systems are shown in Fig. 7 by diagrams that indicate the principal differences in mode of operation.

The estimates of speech quality in Table 1 are mostly guesswork. There is, to be sure, some background of experience with speech synthesized by rule, or compiled by hand, or synthesized from carefully prepared control signals, but the degradations in quality that may result from the operating economies introduced in Systems 4 and 5 can only be guessed.

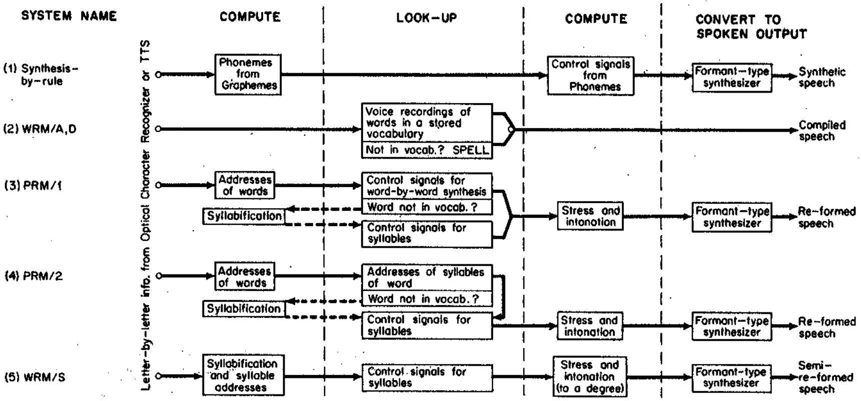


FIG. 7.

The cost figures in Table 2 were arrived at by dividing the rental cost per clock hour of a suitable computer by the speed factor at which it could be used, and then adding the cost per hour of speech for the optical character recognizer, labour, and magnetic tape. A number of special assumptions are spelled out in the Appendix, but it should be emphasized again that the figures given here are rough estimates at best.

One point that is not obvious in these tables, but emerges from the discussion in the Appendix, is that the cost of generating speech is a sensitive function of the access time to stored information, which puts a premium on computers with large core memories and on methods with small storage requirements. A more general point is that these cost estimates are for a service centre working at full capacity to serve individual subscribers. If the same recordings were to be used by, say, ten readers, costs per hour would drop to little more than a tenth of the estimates; conversely, if the service centre were to work at half capacity, the costs would be roughly twice as much, unless arrangements could be made to rent time on a large computer only to the extent that it was needed, thereby reducing the penalty for operating at a fraction of capacity.

The point of these considerations, and others that will readily occur to the reader, is that the practicality of a high-performance reading machine centre will depend more on the solutions of organizational and administrative problems than on the elegance of the technical methods. The latter are not yet fully worked out, but they can be, probably within the next couple of years or surely within this decade. The real problem lies in putting these possibilities to practical use.

Research Status and Plans

A substantial background of experience has been gained over the past several years from studies of synthetic speech and compiled speech. This includes much that cannot usefully be summarized here but the principal conclusions are (1) that speech synthesis by rule will require much more research on the rules before the speech will be fully acceptable, and so is not the best choice of method for the present, though it has long-term promise, (2) that compiled speech would serve quite adequately, at least up to 120 words per minute, (3) that the most noticeable defects of compiled speech (unevenness in intensity, pitch, and pacing of successive words) could be largely eliminated by simple smoothing operations, and (4) that reformed speech makes the smoothing operations easy and, besides, reduces the storage requirements to such an extent that a pilot experiment in the operation of a reader service centre becomes feasible.

The present directions of research on reading machines at Haskins Laboratories is toward field tests aimed at comparing compiled speech and reformed speech in terms of the reactions of blind users to extensive recordings made by these two methods. Following this, and assuming that reformed speech will prove to have the advantages we expect of it, we should like to go far enough with an experimental reader service centre to explore the problems and lay the groundwork for setting up an operating centre if some suitable group is interested in establishing and running it.

Compiled speech. There are several reasons for doing further work on compiled speech, in spite of the fact that it seems less promising than reformed speech. For one thing, we have a substantial investment in this area already in the form of experience with the method and a recorded vocabulary of some 7200 words. Recordings of compiled speech will provide a basis for finding out what problems are involved in carrying through an evaluation study of reading machine outputs with blind subjects. Undoubtedly, a number of practical problems will arise and we may as well get this experience with a system that is essentially ready to go.

Production of the compiled speech recordings could be done in either of two ways. We have built and successfully tested the Interim Word Reading Machine referred to in an earlier section and could use it, or we could use digital equipment that is now being installed. We think the latter is the better choice, and that we should invest the rather substantial effort of loading the 7200 words into computer memory rather than into a special-purpose analog device that would inevitably require special maintenance. The recordings will use punched paper tapes as the input and so by-pass the use of an optical character recognizer; this input can be either Teletypesetter tape, or tape that has been specially punched for the experiments. The general procedure for proceeding from punched tape to recorded speech is outlined in Fig. 8 The text can be processed in batches of 2–3 min of speaking time (about 500 words of text). The spelled versions of the text are read into the computer memory, re-arranged, and used to search for the digital waveform recordings on a magnetic tape. A 2400-ft reel of computer tape provides storage space for about 5000 words of waveform data. Each batch of input text will require one complete scan through the magnetic

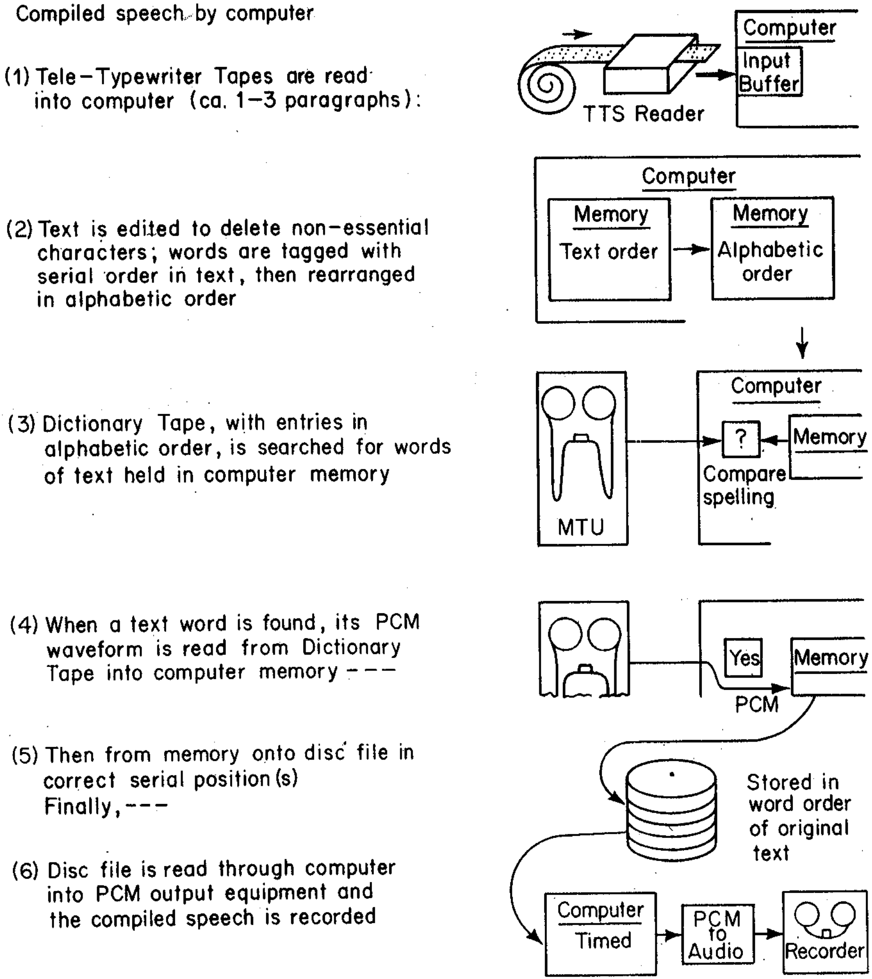

Compiled speech by computer
(1) Tele–Typewriter Tapes are read into computer (ca. 1–3 paragraphs):
TTS Reader
Computer
Input Buffer
(2) Text is edited to delete non–essential characters; words are tagged with serial order in text, then rearranged in alphabetic order
Computer
Memory
Text order
Memory
Alphabetic order
(3) Dictionary Tape, with entries in alphabetic order, is searched for words of text held in computer memory
MTU
Computer
?
Memory
Compare spelling
(4) When a text word is found, its PCM waveform is read from Dictionary Tape into computer memory – – –
Yes
Memory
PCM
(5) Then from memory onto disc file in correct serial position (s)
Finally, – – –
Stored in word order of original text
(6) Disc file is read through computer into PCM output equipment and the compiled speech is recorded
Computer
Timed
PCM to Audio
Recorder

FIG. 8.

tape, with each word that is found being transferred to its proper text position on a disc file. The final step is to read the disc file through the computer (for buffering and correct timing of the waveform samples), and then through a PCM demodulator to an ordinary quarter-inch tape recorder running at normal speed.

The waveform information is originally stored on the magnetic tape from the word recordings that we now have on cards. These are converted to digital form (ten bits) from which a six-bit version with logarithmically scaled steps is computed and stored on digital tape. The only special devices needed are the PCM encoder and decoder to convert from audio to digital and back to audio. These have been built and tested; for the present purpose, they operate at a sampling rate of 8 kc and use a ten-bit linear amplitude scale. Speech of telephone quality is to be expected.

Re-formed speech

The work on reformed speech for the next six months or so will be concerned with the detailed procedures for measuring and storing (on a disc file) the control signals for the 7200 word vocabulary, in adjusting the formant generator synthesizer, and in writing computer programmes to do the "smoothing" operations on intensity, duration, and pitch. Thus, the recordings of reformed speech for field trials should begin to be available by about the time trials of compiled speech can reasonably be completed.

Substantially more special equipment is needed for generating reformed speech, but it is all complete except for final tests and adjustment when operating under computer control. The interface between computer and formant generator synthesizer provides buffer storage for a forty-eight-bit time slice and D/A converters to generate the analog control signals. Table 3 shows the control functions and bit assignments that will be used, at least initially. The formant generator synthesizer is a parallel type with one

Table 3
Bit Allocations for Re-formed Speech

Function	*Number of Bits*	*Comments*
Formant Frequencies:		
F1	5	Linear over normal range
F2	5	Linear over normal range
F3	4	Linear over normal range
F4	0	Fixed freq., adjustable
Formant Amplitudes:		
A1	3	Decibel scale
A2	3	Decibel scale
A3,4	3	Ratio A3/A4 adjustable
Fricative Filter	2	Choice of filters
Fricative Ampl.	3	Level of hiss into filter selected
Overall Ampl.	4	An "override" control
Excitation (Buzz, Hiss, Buzz + Hiss, Silence)	2	Four choices
Pitch	6	Linear scale, range adj.
Nasality	2	Four degrees
Duration	3	An "override" control
Duration Enable	1	Only marked time slices can be modified
Odd/Even Word Identification	2	These could be salvaged for other functions
Total	48	Two computer words for each time slice

fixed, and three variable, formants. Overriding controls are provided for intensity and duration; they may also be added for pitch. Finally, a device for obtaining digital information about formant frequencies, intensities, etc. by merely tracing sound spectrograms is nearly complete. This will greatly simplify the task of measuring and storing the digital control signals.

In addition to these hardware-oriented parts of the research, studies are underway on how changes in duration, for example, should be made to the various parts of words and phrases in order that the result may yield a lifelike speeding or slowing of the connected speech. Other studies that will parallel the work on loading the computer memory with control signals will include experiments on "coupling" the control signals for separate syllables in order to blend those syllables into a single word, i.e., the basis and limitations of syllable spelling.

In summary, our research to this point has led us to put aside the method of synthesis by rule, and to concentrate on generating and testing user acceptability of compiled speech as an immediate objective while we work out the remaining problems of generating reformed speech for further field tests. The present prospect is that reformed speech will be the method of choice, and that it should permit full-scale testing of a high-performance reading machine within a very few years.

The research reported here on reading machines for the blind was supported by contracts with the Research and Development Division of the Prosthetic and Sensory Aids Service, Veterans Administration, Washington, D.C.

Appendix

Estimates of operating speeds, equipment, and costs for various methods of generating speech from printed text were summarized in Tables 1 and 2. Such estimates involve a number of assumptions and approximations. The following paragraphs will explain the principal points that were considered but without attempting to do so exhaustively.

Optical Character Recognition

Let us start with the optical character recognizer. It will illustrate some of the problems and the crudeness of the estimates. The nearest approximation to the required equipment that was commercially available on the American scene at the end of last year was the Farrington Page Reader with a speed of 300–400 characters per second and a rental cost of about $20 an hour on a lease basis (full-time, one shift). If we assume a final output to the blind user of 210 words a minute—a reasonably brisk pace—then the OCR operates at about twenty times real time. and so would cost about a dollar per hour of output text, if it operates at capacity. However, the device that will be needed must read the various typefaces that are found in ordinary books and so will be more complex and more expensive. The estimate should therefore be increased by 50 per cent, or even doubled. The cost entered in Table 2 was $2/hour of output speech.

Generating the Output Speech

Estimating the cost of generating speech necessarily involves looking at the details of several possible processes. Figure 7 shows some of the options in block diagram form. Each of the five systems has its own requirements for instrumentation, and these will be noted. Except for the final stage of conversion to spoken output, where conventional formant-type synthesizers can be used, it is probably desirable to employ general purpose digital computing equipment, since it is available without the heavy investment in engineering development that would be required to build special-purpose devices, whether analog or digital.

System 1. Synthesis by rule requires two types of computation. One is to convert phoneme strings into control signals. Holmes, Mattingly and Shearme (1964) in their procedures for computer synthesis, use about 300 instructions per phoneme which, in a small computer, would mean that the control signals are computed at about forty times real time. Bhimani and Mitchell (1965) in discussing the programme for grapheme to phoneme conversion, said that they processed 6000 words in 150 sec, which is roughly twenty times real time. They were using a large computer but since they were converting to phonetic representations in five different dialects, it may not be unreasonable to use their factor of twenty for a smaller problem on a smaller machine. Very roughly, then, we might expect the cost of generating synthetic speech by rule to be about one twentieth of the hourly rental charge on a small computer, i.e. in the range of $1–2·00 per hour of output speech.

System 2. A word reading machine—either analog or digital—will almost certainly not be competitive in a cost sense because of the enormous amounts of memory required to store a vocabulary in waveform terms. The experimental procedure described in the text uses a computer to generate compiled speech for listener evaluation tests, but uses the computer very inefficiently, i.e., about four hours of computer time are needed for every hour of output speech. This would be intolerably expensive for anything but experimental work.

System 3, referred to as a "phrase reading machine" because of the overriding controls on stress and intonation, is essentially the one already described for the production of a reformed speech. Here it is assumed that a word omitted from the vocabulary would be syllabified by methods like those used in computer-controlled typesetting, and that control signals for these syllables would then be drawn from storage and used to synthesize the missing word.

The controlling factor in the cost of operating such a system turns out to be the access time to the stored data. Even though the control signals are a much more compact description than the speech waveform, it is still necessary to use a disc file to store even a minimal vocabulary. For the particular disc file IBM 1311 that we shall be using in an experimental system, random access takes about 250 msec. Thus, a speaking rate of

210 words per minute, requiring three or four accesses per second, requires the disk file to work about as fast as it can. Two or three additional disk files, with overlapping access times would increase the speed by a factor of three or four, and also would provide plenty of storage space for at least a twenty thousand word vocabulary; the additional equipment would increase the rental cost of the system by less than 20 per cent.

A note about vocabulary size and storage requirements may be useful at this point. The minimal vocabulary referred to above would consist of the 1400 syllables that, according to Dewey (Relative Frequency of English Speech Sounds (Cambridge, Mass. and London, 1923)) account for 93 per cent of all occurrences plus about 600 high-frequency words, to give a basic set of words and syllables that, again according to Dewey, would account for 75 per cent or more of the words in running text. The basic set would then be supplemented by a 5000 word vocabulary to bring the total to 7000 words, roughly the capacity of a single disc file of the type considered here. A more nearly optimal vocabulary would include the 7000 word minimal vocabulary, and additional 3000 syllables to cover all those observed by Dewey, and another 10,000 words, for a total vocabulary of 20,000 items.

As to storage requirements, the control signals for the formant-type synthesizer we shall be using for experimental work require 48 bits to specify each time slice, or 2400 bits per second to specify the spectrum at 20 msec intervals. At a speaking rate of 210 words per min, i.e., 3½ words per sec or about 4 syllables per sec, this is 600 bits per syllable.

System 4, also a phrase reading machine, differs from System 3 in storing control signals for the syllable only. The reason for considering such a system is to see whether or not it might be feasible to reduce the total storage requirement to the point that core storage—with its very much shorter access times—could be used for the entire vocabulary. The process involves a two-stage look-up operation for polysyllabic words: first, a general vocabulary of word spellings (or of those spellings in a suitably reduced form) is searched for the specific word being processed; if it is found, the entry will give exact addresses for each syllable and additional information about relative stresses, grammatical functions of the word, etc. In the second stage, the control signals for successive syllables are fetched from memory, modified as to stress where this is necessary, and put together to form the control information for the entire word. If the text word is not in the general vocabulary, then syllabification and "syllable spelling" would be used as in System 3.

Certain consequences of this overall strategy are fairly evident. The quality of the speech will suffer to some degree (not fully predictable without further research) because of the errors in blending the syllable control signals to give control signals for the complete word; however, most of the merits of System 3 can be preserved since the normal spoken form of the word (with its usual pattern of syllable stresses and vowel reductions) is preserved. The speed of the prccess will depend almost entirely on how much of the working information can be stored in core memory, since the total number of references to memory is substantially larger than in System 3.

A more detailed examination shows that the storage requirements will be essentially the same as in System 3 for the first 5000 entries (i.e., all the syllables that are likely to be encountered plus a few hundred high-frequency bi-syllabic words). However, additional words that are stored as syllable addresses will require only 5–7 per cent of the space they would have needed in control signal form, and so an additional 15,000 words adds no more to the memory requirements than an additional 1000 words in control signal form, but significantly reduces the spelling problem.

Even so, the storage requirements are severe—roughly 100,000 words of core memory, assuming 48 bit words. The equipment configuration might be a computer with this large memory capacity and a high-speed magnetic tape unit to record the control signals for off-line conversion to speech at more convenient rates. The speed limits would probably be set by the magnetic tape unit, and could run up to 300 times real time. On this basis, the costs for computer time per hour of output speech might be about $2·00 per hour, perhaps less.

Another configuration would use core memory for only the high-frequency syllables and words (about 2000 of them) and disc files for the remainder. Since most of the data retrievals would be from core memory, the time penalty due to the disc file would be much less severe and an overall speed of perhaps thirty times real time seems feasible. The operating costs with this smaller configuration—32,000 words of forty-eight bits each—should be no more than about $10·00 per hour of output speech.

System 5. The final system considered here would generate all the output speech by "syllable spelling" i.e., the words of the printed text would be broken into syllables by computer programme and these fragments of spelled words would then be used as the addresses for retrieving control signals for the most probable pronounced syllable to match the spelling. Clearly, the quality of the speech will suffer, even though overall stress and intonation controls can be imposed as in some of the earlier methods. The principal thing that has been lost by relinquishing a stored vocabulary is information about the usual pronunciation of particular word spellings. The benefits are a modest reduction in storage and cost of operation. The storage requirement depends to some degree on how sophisticated the coding of the control bits can be; however, the reduction in storage requirements is not more than a factor of 2 or 3 over System 4, and the operating costs for computer time would be roughly $1–2 per hour of output speech.

Labour and Materials

Obviously, it will require people as well as machines to operate a reader service centre. If the machines are slow, perhaps operating in real time, then one person could probably take care of all the tasks: accepting requests, arranging for books to be brought from library stacks, operating the OCR, speech generator, and tape-recording equipment, mailing tapes to users, and keeping records. At $80–100 per week, salary costs would be $2–3 per hour of output speech. A more realistic approach would use faster machines, operating at twenty to thirty times real time, and so requiring substantially more labour. A system operating at this rate could generate an output recording with four hours of speech on it every ten minutes. Full use of such facilities would require at least two or three people, but the labour costs, distributed across the larger volume, could be less than 50 cents per hour. The principal item for materials would be magnetic tape; this should be no more than $1 per hour to cover both a library retention copy and a tape for the user. Thus, minimum labour and material cost for a service centre attached to a library would range from about $4 to perhaps as little as $1 per hour of output speech, assuming full-time operation.

Possibilities for Time Sharing

The estimates made thus far have assumed full-time rental and use of all equipment, though costs for such items as tape recorders and formant-type synthesizers have been ignored. But it is not realistic, at least initially, to assume operation at the capacity of even the medium size machines considered here; hence, time-sharing or rental time as required on a large computer is probably desirable. In this case, the least expensive system seems likely to be one using rental time or a computer with very large core memory and a fast magnetic tape unit to record the control signals it generates. This part of the operation might well run at 300 times real time. The tapes could then be read into a speech synthesizer at much lower rates, the choice depending on the technical problems of generating speech waveforms at several times real time rates, with probable rates in the range of 8–32 × RT.

The service centre installation would presumably include the digital tape playback, the speech synthesizer, and one or two highspeed, quarter-inch recorders for the final audio signal. Whether or not it will contain the optical character recognizer is an important but difficult question. The comparatively high rental costs (estimated above at $5000–7000 per month) would suggest time-sharing, but physical access may preclude such an arrangement. Thus, the constant costs of a service centre will depend heavily on the specific arrangements that can be made, especially with respect to shared time use of an optical character recognizer, or dependence on Teletypesetter tapes.

REFERENCES

1. Chistovich, L. A. Continuous recognition of speech by man, *Machine Transl. and Appl. Linguistics*, Moscow, 7, 3–44 (1962).
2. Cooper, F. S., Liberman, A. M., Lisker, K., and Gaitenby, J. H. Speech synthesis by rules, *Proc. of Sp. Comm. Seminar*, Royal Institute of Technology, Stockholm, 1962 (Stockholm, 1963).
3. Cooper, F. S. Speech from stored data, *IEEE Convention Record* 7, 137–149 (1963).
4. Cooper, F. S. Toward a high performance reading machine for the blind. *Human Factors in Modern Technology* (New York, 1963).

5. Dewey, G. *Relative Frequency of English Speech Sounds* (Cambridge, Mass. and London, 1923).
6. Feidelman, L. A. A survey of the character recognition field. *Datamation* **12**, No. 2, 45–52 (1962).
7. Gerstman, L. J. and Kelly, J. H. An artificial talker driven from a phoneme input, *J. Acoust. Soc. Amer.* **33**, 835 (A) (1961).
8. Harris, C. M. A study of the building blocks of speech, *J. Acoust. Soc. Amer.* **25**, 962–969 (1953).
9. Holmes, J. N., Mattingly, I. G., and Shearme, J. N. Speech synthesis by rule, *Lang and Speech* **3**, 127–143 (July–Sept. 1964).
10. Kozhevnikov, V. A. and Chistovich, L. A. Rech', *Artikulyatsiya, i Vospriyatiye*, Moscow-Leningrad, 1965. Translated as *Speech: Articulation and Perception*. Joint Publications Research Service, U.S. Dept. of Commerce, Washington, D.C. 223. (JPRS: 30, 543 June 1965).
11. Klemmer, E. T. and Frick, F. C. Assimilation of information from dot and matrix patterns, *J. exp. Psychol.* **45**, 15–19 (1963).
12. Liberman, A. M., Delattre, P., and Cooper, F. S. The role of selected stimulus variables in the perception of unvoiced stop consonants, *Am. J. Psychol.* **65**, 497–516 (1952).
13. Liberman, A. M., Delattre, P., Cooper, F. S., and Gerstman, L. J. The role of consonant-vowel transitions in the perception of the stop and nasal consonants, *Psychological Monographs* **68**, No. 8, 1–13 (1955).
14. Liberman, A. M. Some results of research on speech perception, *J. Acoust. Soc. Amer.* **29**, 117–123 (1957).
15. Liberman, A. M., Ingemann, F., Lisker, L., Delattre, P., and Cooper, F. S. Minimal rules for synthesizing speech, *J. Acoust. Soc. Amer.* **31**, 1490–1499 (1959).
16. Liberman, A. M., Cooper, F. S. Harris, K. S., MacNeilage, P. F., and Studdert-Kennedy, M. Some observations on a model for speech perception, *Proc. AFCRL Symposium on Models for the Perception of Speech and Visual Form* (Boston, 1964) (in press).
17. Lindblom, B. Spectrographic study of vowel reduction, *J. Acoust. Soc. Amer.* **35**, 1773–1781 (1963).
18. Mauch, H. A. and Smith, G. C. Recognition machines: developments at Mauch Laboratories, *Proc. of the Sixth Technical Session on Reading Machines for the Blind*, Veterans Administration (Washington, D.C. 1966).
19. Metfessel, M. Current status of rules for developing an alphabet to synthesize spelled speech, *Proc. of the Sixth Technical Session on Reading Machines for the Blind*, Veterans Administration (Washington, D.C. 1966).
20. Miller, G. A. and Nicely, P. E. An analysis of perceptual confusions among some English consonants, *J. Acoust. Soc. Amer.* **27**, 338–352 (1955).
21. Miller, G. A. The magical number seven, plus-or-minus two, or, some limits on our capacity of processing information, *Psychol. Rev.* **63**, 81–96 (1956).
22. Nye, P. An investigation of audio outputs for a reading machine. National Physical Laboratory, Teddington, England: Auto 8, February 1965.
23. Peterson, G., Wang, W. S-Y, and Sivertsen, E. Segmentation techniques in speech synthesis, *J. Acoust. Soc. Amer.* **30**, 739–742 (1958).
24. Pollack, I. The information of elementary auditory displays. *J. Acoust. Soc. Amer.* **24**, 745–749 (1952).
25. Pollack, I. and Ficks, L. Information of elementary multidimensional auditory displays. *J. Acoust. Soc. Amer.* **26**, 155–158 (1954).
26. Stevens, K. N. and House, A. S. Perturbations of vowel articulations by consonantal context: an acoustical study. *J. Speech and Hearing Res.* **6**, 111–128 (1963).
27. Wickelgren, W. A. Distinctive features and errors in short-term memory for English consonants. *J. Acoust. Soc. Amer.* **39**, 388–398 (1966).