SPECIAL ARTICLE

"Special Article" in the Journal of Rehabilitation R&D identifies material which addresses some matter of urgent or broad scientific concern to many of our readers. We believe that the example presented below is of importance not only to those involved in the development of sensory aids, but to all of those who are involved in the kind of problem-solving, "targeted" research that is so characteristic of the field served by this publication.

As in this case, the Journal's "Special Articles" will rarely if ever be found to follow the format of a standard scientific paper, but will invariably have been reviewed by appropriate members of the Editorial Board and by ad hoc reviewers selected for their experience and stature in the field and the relevant disciplines.

Evolution of Reading Machines for the Blind: Haskins Laboratories' Research as a Case History

FRANKLIN S. COOPER, Ph. D.
JANE H. GAITENBY, B.A. [a]
PATRICK W. NYE, Ph. D.

Haskins Laboratories
270 Crown Street
New Haven, CT 065116699

Reading machines for the blind are now an accomplished fact. They are not as good or as widely available as eventually they must be, but they are demonstrably useful. Not many years ago the construction of such machines was only a goal.

The main part of this account deals with work that was done by the Haskins Laboratories under research contracts funded by the Veterans Administration (VA). This research, which spanned two decades, played a significant role in achieving a better understanding and solution of the reading machine problem. However, the period of VA support is only the middle chapter of a longer story which begins at least 50 years earlier.

INTRODUCTION

The quest for a machine that can open the world of ordinary books to blind readers dates back to the 19th-century discovery that the electrical resistivity of selenium is influenced by light. Many technical applications followed that discovery, including at the turn of the century an apparatus for reading specially-prepared "photophonic books." But only now nearly 80 years later, do we have the first devices that may reasonably be called reading machines for the blind. They achieve that goal in the sense—and to the extent—that a blind user can himself read a variety of printed materials without unreasonable expenditures of time and effort; moreover, there is a reasonable expectation that reading machines will become affordable by individual users.

There have been many proposed solutions to the reading machine problem. Most have been abandoned, though some existing devices dating back to earlier efforts may continue to be used because they meet special needs and are comparatively affordable and transportable. Their major shortcomings are that reading is very slow and much training is required to learn the machine's "language." Nonetheless, it is usual to denote as reading machines all those devices that convert printed text into some kind of auditory or tactile signal, regardless of level of performance or requirements for special training. These devices deserve their name because they give the blind user independent access to personal papers and the like, even though they can offer only limited access to the larger world of books.

It is often useful, because of the difference in level of performance, to set apart the new generation of devices by calling them "high-performance" reading machines. Are they indeed high-performance devices and is the reading machine problem now solved? Or are the new devices only another plateau? The history of the field suggests a cautious answer despite major gains in speed and ease of reading. Indeed, the story of technologies of all kinds has the repeating theme of new approaches that lead to rapid attainment of a new plateau of performance, followed by steady but less dramatic gains attained by conventional refinement. It may be useful to characterize uneven progress of this kind as the normal technological cycle of revolution and evolution.

The potential for a revolutionary gain in reading speed, and for access to ordinary books, has been realized by two innovations: the use of optical character recognition (OCR) for input and synthetic speech for output. However, there has not yet been enough experience on routine tasks to establish the true usefulness of such machines to blind readers: Can the remaining faults be remedied by routine refinement or do the limitations lie deeper? Knowing the nature of the prob-

[a] Present address: Route 66, Huntington, MA 01050

lems and the reasons for past successes and failures provides a background against which the present achievements may be viewed in perspective.

We shall describe, as a case history, the work of one research group concerned with a single central aspect of the problem. There are advantages in the case history approach: experiments can be described in reasonable detail and in terms of the ideas that prompted them; also, in the present instance, there is freedom to concentrate on high-performance reading machines and to avoid the obligation a reviewer otherwise would feel to devote comparable attention to other types of machines (3, 9, 27, 28, 29, 42, 46, 47) that fill other kinds of needs.

This approach allows the evolution of the underlying ideas to be discussed from a coherent point of view, and so offers a potential gain in clarity, though at some risk of bias. However, tracing the evolution of ideas has a difficulty that is different from that of tracing the development of devices: An old idea, when replaced by a newer one, does not become merely a seldom used artifact, as devices do, but rather reverts to nonexistence—it becomes almost literally unthinkable. Similarly for reading machines. Now that they can talk, it hardly seems possible that this ability was not always the ultimate goal—that there was a time not so long ago when the very possibility of speech as an output signal was novel, or that somewhat earlier even the need for such a signal had not been realized.

It is against such a background of evolving concepts that today's reading machines should be seen, in order to appreciate their merits and assess their limitations.

HISTORICAL REVIEW

Research on reading machines at Haskins Laboratories began in the mid-1940s. A literature survey at that time turned up many attempts to use photoelectric devices as aids to the blind, and one complete development (the Optophone) that had been carried from inception and production to full evaluation. Earlier attempts to devise reading machines, such as the photophonic books (60) of V. de Turine, required specially prepared texts in which the letters were represented by small transparent squares. When the page was scanned by an opaque mask with apertures for the letters, a selenium cell and associated circuits responded to the transmitted light and produced an audible signal for each letter. The primary disadvantage of the system was the need for specially prepared materials, a limitation that also flawed other reading systems proposed during the following three decades.

THE OPTOPHONE

The Optophone had a profound effect on the development of reading machines. In its earliest form, this device was merely an aid to the blind in locating the light from doors and windows, and was called the Exploring Optophone. Invented in 1912 by Fournier d'Albe, it was soon modified to give information about the patterns of letters on the printed page. An early version of the Optophone was demonstrated to the British Association in 1913. In a public demonstration in 1917, a reading speed of three words per minute was attained. The original instrument had mechanical crudities that made it difficult to use and generated a continuous sound, even across the blank spaces between letters or words. Shortly after World War I, the firm of Barr and Stroud made many improvements in the device and converted it into a "black-reading" Optophone which generated sounds only from black areas of each letter (2). Manufacture of the device was undertaken, and in 1923 Miss Mary Jameson, an early and very apt student, gave a public demonstration in which brief passages were read at 60 words per minute.

The Optophone as engineered by Barr and Stroud embodies the best technical practice of its period. It is a precision instrument of about the same size and complexity as a portable sewing machine. The book to be read is placed face downward over a curved glass plate and a mechanical scanning mechanism. A line of type is scanned with five vertically arrayed points of light, as indicated in Figure 1. The beams of light are chopped by a rotating disc with perforations so spaced as to generate the musical notes G, C', D', E', G'. Individual notes or chords are heard only when the corresponding beams encounter black areas of a letter. Thus, the h in Figure 1 is shown generating the single note E', which was preceded by a four-note chord and is to be followed by the three-note chord C'D'E' and then by a silence preceding the next letter, i. Some of the chord sequences for individual letters are quite distinctive but others are much alike, as, for example, a, e, o, and c. It was not claimed by the makers of the Optophone that

individual letters could always be readily recognized, but that "when the alphabet has been learned, the motif for each letter is recognized as a whole, and later in the reader's practice the more extended motifs for syllables and even words will become familiar to his ear."

There was substantial enthusiasm for the Optophone, particularly in England, as a result of Miss Jameson's performance, although her exceptional gifts enabled her to achieve reading rates far above those of other students. The principal difficulties appeared to involve ambiguities in the identification of the letters, especially when they occurred in rapid sequences. Even long training did not overcome this problem and did not, to any substantial degree, realize the expectation that recognition of larger patterns for syllables and words might replace letter-by-letter reading. Confusions were especially likely if the lines of type were not accurately aligned with the scanning mechanism, and correct alignment was not easily achieved in spite of ingenious mechanical arrangements. Interest in the device had substantially subsided by the end of the twenties, though Miss Jameson continued to use her personal Optophone for many years.

The Optophone was an achievement in the evolution of reading machines, and we should consider its lessons: If a reading machine for the blind is to be useful, it must use the same printed materials that sighted people read; and what is wanted is a machine that can be operated—and owned if possible—by the individual blind reader. The central problem was thought to be the technical one of generating distinctive sounds from the printed page. This was solved fairly adequately despite some ambiguities as to letter identities. Yet that solution was not useful to blind readers. The underlying reasons for this failure were not fully understood until long afterward.

BRAILLE, TALKING BOOKS AND VISAGRAPH

Meantime, practical aids to reading developed rapidly along other lines as well. In this country, the decade of the thirties saw the use of both Braille and the Talking Book become widespread (26). Technology and Federal funding were decisive factors in both cases. For Braille, an appropriation to provide books for the blind brought an end to the long and sometimes bitter disputes about what kind of embossed type or raised-dot code should be accepted as a standard. This was 100 years after Louis Braille had invented the system that bears his name. His basic system had won out over embossed type because it was easier to read, and over other dot systems because his could be produced by comparatively simple machines or even by a blind individual using a simple perforated guide.

The Talking Book lagged behind Edison's invention of the phonograph by half a century, and did not follow automatically even from the resurgence of that device in the twenties. The phonograph and its records in their commercial form were poorly adapted to the reading needs of the blind. In fact, it took a combination of events to make Talking Books a reality (34). In 1932, a grant from the Carnegie Corporation enabled the American Foundation for the Blind to develop suitable recording methods, reproducing machines, and mailing containers. Joint action by the Foundation and the Congress launched a library service for distributing Talking Book records and machines, many of the latter built under a W.P.A project. The service has been continued by the Library of Congress and fills an important need, especially of the older blind for whom Braille would be difficult to learn and not rewarding for pleasure reading.

The thirties saw another notable development carried through to a working device but abandoned because it failed to meet the real needs of blind users. The Naumburg Visagraph (45) used a cylindrical scanner-embosser to convert the black and white patterns of the printed page into enlarged raised replicas on a sheet of aluminum foil. In a series of tests, blind readers found the letters too difficult to comprehend with any ease. For this fundamental reason the Visagraph failed to become a viable reading aid, even though it had two significant advantages: books could be embossed on demand and it was as easy to reproduce diagrams, formulas, and the like as to copy letter-text.

By the nineteen-forties, Braille books and Talking Book recordings offered some partial access to the wealth of libraries. But the limitations were severe. Braille required

FIGURE 1
Tone generating method of the black-reading Optophone.

much learning and only the exceptionally skillful reader could match childhood rates of visual reading. Embossed books and recordings were both cumbersome and obtainable only from libraries. Worst of all, the selection of titles was severely limited because the total number of books in any category remained very modest. Ironically, the Optophone and the Visagraph—the two devices that might have provided unrestricted access to books—were already museum pieces.

HASKINS LABORATORIES' RESEARCH, PHASE ONE: WORK FOR THE COMMITTEE ON SENSORY DEVICES

The end of World War II brought changes of many kinds, including a new approach to aids for the blind. University research groups, organized and funded by the Office of Scientific Research and Development (OSRD), had been strikingly successful in applying science to the development of weapons and in expanding the technological base. With many blinded veterans returning from the war, Dr. Vannevar Bush sought to use his organization's prowess on their behalf. Guidance for the effort was put into the hands of a Committee on Sensory Devices (CSD) made up of physiologists, a psychologist, and physicist, under the chairmanship of Dr. George W. Corner. Meeting first in January 1944, the CSD chose to concentrate on guidance devices and reading machines, the two main needs of the blind to which the new technology might apply. It was evident quite soon that matching technologies to needs would be a novel undertaking in which the CSD would need facilities for working out preliminary developments. The Haskins Laboratories, a small nonprofit research institution, was placed under contract as a central laboratory to serve the CSD in exploratory research and in recommending industrial contractors for more extensive development tasks. Dr. Paul A. Zahl served the Laboratories as principal investigator and shared the direction of the research with Drs. Caryl P. Haskins, Franklin S. Cooper, and Alvin M. Liberman.

The charge to Haskins Laboratories was quite general and provided for a close working relationship with the CSD. The Laboratories' efforts were about equally divided between guidance devices and reading machines. Most of the guidance device developments were done by industrial contractors; evaluation of the devices with blind subjects was carried out by Haskins Laboratories. Research on reading machines was done almost entirely by Haskins Laboratories except for a parallel arrangement between the Committee and Dr. Vladimir Zworykin of the Radio Corporation of America (RCA) Laboratories. The CSD also undertook two additional developments: the improvement of optical magnifiers for persons of limited visual acuity, and improvements of the Visagraph, primarily for the production of enlarged embossed images of diagrams, prints, etc.

The entire program, from initial planning to final reporting, lasted less than 4 years—due primarily to shifts in government organization and patterns of funding, starting with the dismantling of the OSRD. However, there was a deeper reason as well, namely, a growing pessimism about early breakthroughs. Although, in each of its four lines of research, one or more devices had been brought to a first stage of practical trial, none of them had achieved striking success in meeting the needs of the blind.

A candid assessment of the CSD's accomplishments and a thoughtful analysis of the lessons learned from its work appear in a report written by its chairman (15). Commenting upon the CSD's emphasis on the early development of devices, Dr. Corner notes the sense of urgency (due partly to wartime conditions) that had to be seized before it waned, and also a prevalent belief in the potential usefulness of actual devices, however crude, in obtaining realistic responses from blind subjects. He adds, "Whatever may have been the wisdom of its course, the Committee therefore promoted more engineering and less psychology than it would have done if its activities had been paced at the peacetime rate and if the problems were in the field of pure science. One thing has surely been gained in this way of handling the program; it is the realization by physicists, engineers and mechanical inventors that when a machine is to act upon a man there are always going to be biological and psychological limitations that outweigh all the mechanical difficulties."

READING MACHINE RESEARCH AT HASKINS LABORATORIES

The program of research (9) for the Committee on Sensory Devices began early in 1944. The Laboratories' previous work had been on problems in the field of radiation biophysics and on the motion-sickness component of traumatic shock; also, in electro-optics as applied to densitometry and color photography. It was clear that the new work on aids for the blind would be concerned primarily with man-machine interactions. Indeed, the CSD had stressed the importance of approaching the problem from the point of view of the needs and psychological capabilities of potential users—in short, basic research rather than a gadget development program.

Analzying the Problem of the Optophone

It was necessary as a first step to recruit psychologists, to share in the work and then to attempt a careful analysis of the problem itself. A good starting point was to review the history of the Optophone. Why, in spite of careful engineering and intensive training of its users, had it failed to be useful? Did its faults lie mainly in the mechanism, or in the audible signals it generated, or possibly in the users' insufficient training? Both experimental work and pencil-and-paper analyses were undertaken. One of the original Optophones, borrowed from the museum collection of the American Foundation for

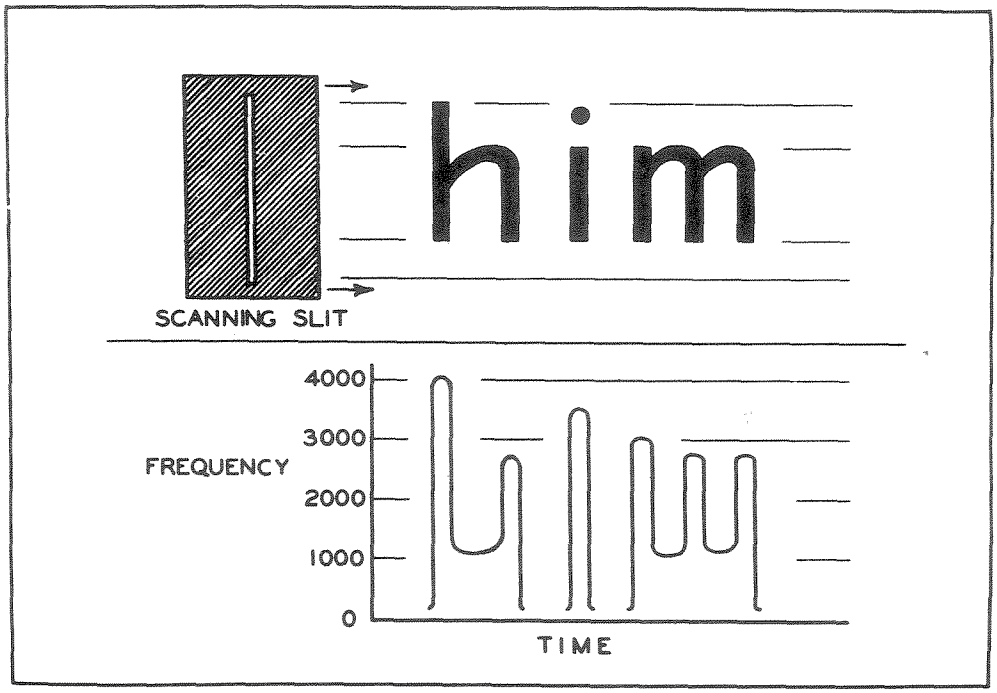


FIGURE 2
Tone generating method of the FM-SLIT reading machine (above), and frequency-time plot of its output (below).

the Blind, was put back into operating condition. Careful listening to its sounds confirmed old reports that, though the signals were reasonably distinctive, confusions often occurred among certain groups of letters. Perhaps the most striking impression was that one had been listening to a very substantial amount of text when, in fact, only a few words had been scanned. In a way this is not surprising because each letter generates three or four distinctively different chords when scanned slowly, as it must be if it is to be distinguished from other groups of chords that are only slightly different.

The sounds from the original Optophone were compared with recordings of a simulated Optophone made by Dr. Zworykin's group at RCA. For engineering convenience, the RCA device did not use a series of separate beams but rather a single spot of light that oscillated rapidly up and down across the letters as it moved slowly from left to right. The vertical sweep was synchronized with a frequency-modulated oscillator, so that tones of higher or lower frequencies were generated from the upper or lower parts of letters, just as in the Optophone. Thus, the signals from both instruments contained almost identical information about the black and white patterns of the letters—and yet the audible effect was very different. The RCA simulation had a harsh buzz (at the vertical sweep rate) that dominated the signal and gave the impression that identifying the letters would be even more difficult than with the musical tones from the Optophone.

A third comparison was made with a device—simulated in the early tests—that looked at the letters through a narrow vertical slit and used the total amount of black thus seen to control the frequency of a tone. This tone could vary between 100 Hz and 4000 Hz or drop to silence between letters and words. Figure 2 shows the scanning method and resulting signal for this FM-SLIT device. The output seemed to have about the same complexity as that of the Optophone and to share the characteristic that some letters had distinctive sounds whereas other groups of letters were ambiguous.

Inherent Limits on Speed of Reading. But was confusability the principal problem? If so it might be possible, with sufficient ingenuity, to generate distinctive sounds even from letters that were visually similar. Another possibility, though, was that a different kind of limitation would prove to be decisive. Pencil-and-paper analyses suggested that the rates at which letter sounds could be followed by a listener would be seriously limited, regardless of how distinctive the individual sounds might be.

It is well known that clicks or other brief sounds are heard as separate events when the repetition rate is low. As the rate increases, the character of the sound changes first to a buzz (at about 20 sounds per second) and then to a tone of rising pitch. Even if the brief sounds are not identical, they can hardly retain their individual character without merging into a buzz as the rate increases. With the Optophone, there are on average about three different chords per letter, which means that about five or six letters per second (one average English word per second) would be an absolute upper limit on letter-by-letter reading. One can readily be convinced that the 60-word-per-minute rate is unacceptably

slow, simply by reading aloud at one word per second. The actual performance of any such device would be far below that rate even after much training, as may be inferred from long experience with International Morse Code. That code provides an almost perfect parallel, since each letter is represented, on the average, by about three dots or dashes per letter. This leads again to an estimate of about 60 words per minute as an upper limit, which is consistent with existing world records for code reception. As for the effects of long training, even expert operators of commercial radio stations send and receive at only 30 to 40 words per minute.

Thus, both theory and broad experience with International Morse Code suggest that even the best of letter-reading devices will be limited to 20 words per minute or so for the average reader—hardly a tenth of the rate at which sighted people read.

Early Experimentation

That was a discouraging prognosis, but even so there were reasons why it seemed desirable to explore letter-reading devices with some thoroughness. One was that any reading, even at limited rates, was better than none at all, and especially if a device could be simple and cheap enough to give the blind person independence in reading personal correspondence, sorting papers, and the like; besides, there was no obvious alternative to devices that operated on a letter-by-letter basis. A second reason was the hope, not entirely disproved by Morse Code, that the signals for letters would somehow coalesce into word-size units, just as the developers of the Optophone had hoped that its signals might be heard as words after sufficient practice. The ways in which sounds can combine to give auditory patterns had been little investigated and so it seemed premature to conclude that no combination of sounds could possibly be found that would meet this requirement.

Constructing and Simulating Various Devices. The experimental approach was accordingly aimed at trying out as many kinds of reading machine signals as one could reasonably devise. For practical reasons, the machines had to be simulated; also a reasonably simple standard listening test had to be devised.

This was done by developing a screening test that contained eight common four-letter words, and a device by which the signals corresponding to these words could be produced without building a working model of each machine. The simulation technique made use of a general-purpose scanning device, with specialized signal generating circuits for each new kind of reading machine. Disk and sound-on-film recordings were made to serve as test materials for psychological evaluation. The scanning device was a 16-mm movie projector, modified to move the film slowly and continuously past the film gate. The letter text, photographed onto the film along its length, could then be projected so as to move slowly across a scanning aperture behind which were eight lenses, photocells, and audio-generating circuits. It was then quite simple to "try out" any kind of Optophone that had eight or fewer scanning beams. Other kinds of reading devices could be simulated by combining the photocell signals in various ways.

The signals characteristic of a number of different letter-reading machines were simulated by these means. Initial tests of the size and orientation of the scanning aperture seemed to show that a rather narrow slit worked best, although some machines were tried in which the slit was divided into sectors. For a single slit (with all eight photocells connected together), the audible signals were modulated in a variety of ways. For example, amplitude-modulated signals of a fixed frequency proved to be very monotonous and not distinctive. Frequency modulation of different wave shapes (sawtooth, square, and sine waves) showed that sine waves gave the least disagreeable sounds. For frequency-modulated tones, the best results were with a frequency swing from 100 to 4,000 Hz, with larger steps at the high-frequency end of the scale. A system of this kind, referred to as the FM-SLIT system, was tried extensively in later tests and was the basis of a portable machine built by the RCA Laboratories.

Attempts were made to "enrich" the signal, for example, by allowing the upper half and lower half of a letter to modulate separate signals, or by generating hisses and clicks from the risers and descenders of such letters as b and p. Some of these modifications seemed to add to the distinctiveness of the signals, but they always increased the perceived complexity.

Assessing Performance. Comparative tests were run on the more promising simulations. A limited set of words (eight of the four-letter words which occur most frequently in English) were recorded in a rote learning format, and the rate at which they could be learned when presented in various random orders was determined. Some kind of comparison signal was needed; it seemed obvious that speech could be taken as the upper bound on expected performance but that actual spoken words would be altogether too easy. So a synthetic language (which came to be known as Wuhzi) was devised. It was based on a transliteration of written English which preserved the phonetic patterns of words and so made the new language pronounceable. The results of these comparative tests are shown in Figure 3 for eight simulated machines and for Wuhzi. Clearly Wuhzi was best; it was learned rapidly and gave near-perfect scores within the first 15 to 20 trials. The Optophone and FM-SLIT machine (which were given further extensive tests) performed less well. All the other machines were distinctly inferior to these two, though in some cases this was contrary to one's intuitive impressions about the signals. Also, for the RCA machine, performance would probably have been more nearly comparable with the Optophone and FM-SLIT machines if the available test recordings had been from the device in its final form. The screening tests also allowed comparisons at different reading speeds (50, 100, and 150 words per minute) as shown in Figure 4. Difficulty in learning increased rapidly with reading rate, but the quantitative data are probably not reliable because extraneous factors may well have been serving as cues, since the number of words was so limited.

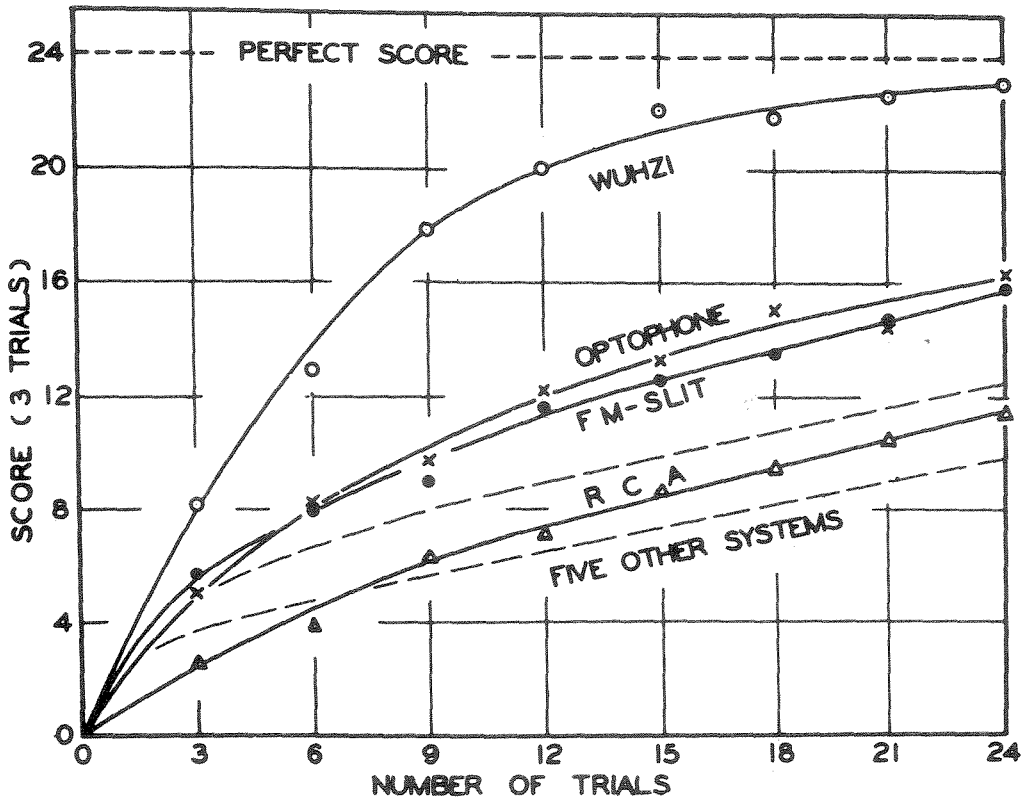


FIGURE 3
Performance on comparative test of various (simulated) reading machines.

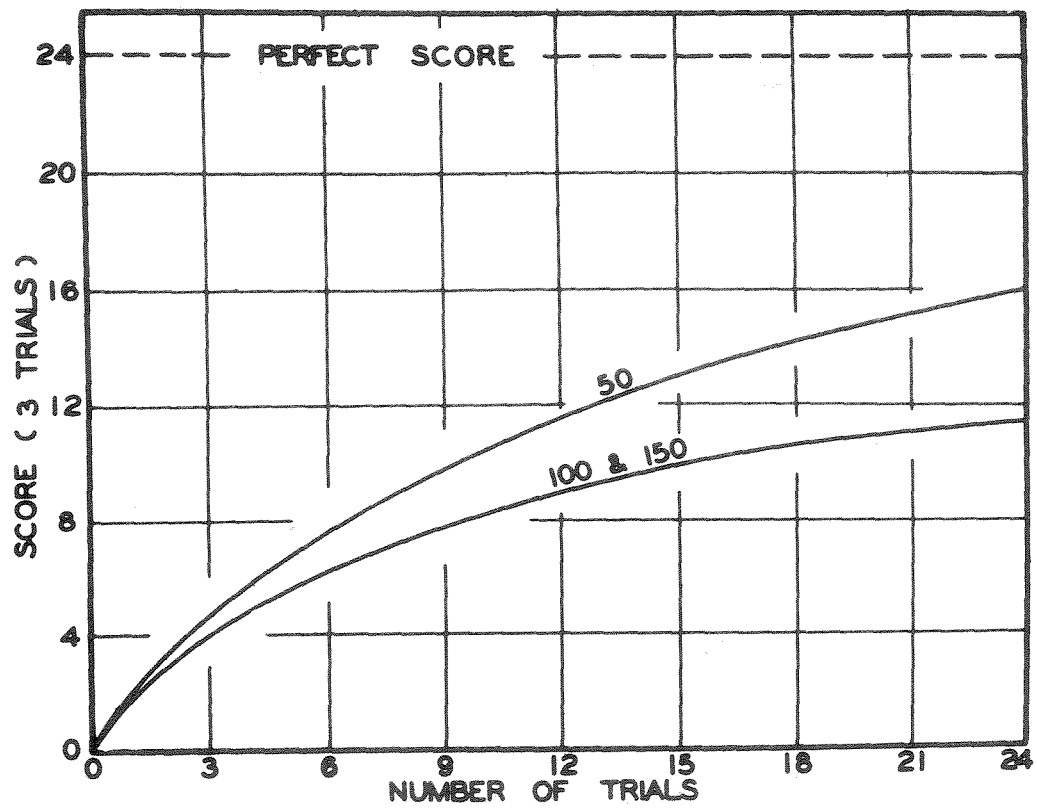


FIGURE 4
Performance versus presentation rate (50, 100, and 150 words per minute) for FM-SLIT reading system.

The screening tests were supplemented by semiproficiency tests for several of the machines and by extended training on a working model of the FM-SLIT machine. The semiproficiency tests used recordings of simple sentences made up from a vocabulary of about 50 common words. The objective was to allow each subject to attain an intermediate degree of proficiency over a period of 6 to 10 hours practice. The extended training tests of the FM-SLIT system were aimed at finding out how proficient a subject could become after long practice with an actual device.

The subject was seated before a table and used a hand-held device, with or without mechanical guides, to scan enlarged film images of letters and of sentences from 4th and 5th grade school books. Learning was slow and the average reading rate attained at the end of a 90-hour training period was 4.2 words per minute, with no significant gain in reading speed during the second half of the period. Analyses of the mistakes indicated that difficulty in the unambiguous identification of individual letters was a factor in limiting the reading speed; that is to say, subjects used much time in repeatedly rescanning some of the letters and words.

A single value for reading rates can be misleading unless test conditions are fully specified; moreover, since conditions are rarely the same for tests done in different laboratories, comparative reading rates are not very informative. Even within a given test format, there can be wide variability in reading rate due to fortuitous combinations of initial letters and context. Thus, in the proficiency test described above, an examination of the reading rates for successive single lines of connected text showed that occasional lines are read at speeds well above the average, though still slow by ordinary standards. The most probable rate, however, was in the range of 2–4 words per minute. Moreover, there was little gain in reading speed betwen the first half of the training period and the latter half. This is one basis for the conclusion that a plateau in reading speed had been reached.

The parallel work by RCA Laboratories gave results that were only a little more encouraging. Tests with the Type A machine (based on the Optophone) used three blind subjects, although only one was carried to saturation in reading speed (at about 190 hours). The attained level for this one subject was approximately 20 words per minute. Letter recognition with random presentation tended to level off at about 80 percent after 40 to 50 hours of practice. This same device, and one similar to the Haskins Laboratories FM-SLIT system, were tested independently at the Naval Medical Research Institute using test materials that were recorded on discs in a learning format and presented at a steady rate of about 12 words per minute, so reading rate was not a variable. Of five subjects, the best attained a score of 60 percent accuracy at the end of 10 days; average scores for the other four subjects were much lower.

The general conclusion from all these tests does not differ from historical results with the Optophone and experience with International Morse Code: The speeds attainable with devices of this general class are too low to be generally useful for continuous reading, though they might be acceptable for certain restricted tasks.

Reanalyzing the Problem

While these efforts were underway to improve performance from simple letter-by-letter reading machines, an increasing part of the Laboratories' attention was given to further analysis of the problem and to more sophisticated approaches. An attempt was made to examine and classify the various ways in which a reading machine might operate. Both the principle on which the mechanism might work, and the nature of the sounds that might be produced, were considered. The resulting classifications are different enough so that it is useful to consider both in parallel.

As to sounds, it has been mentioned above that enriching each letter's output with enough features to be distinctive to the ear is almost sure to prolong each word; at higher rates it will cause words to mix into an indistinguishable buzz. And yet spoken words evade this limitation. How can this be? The answer might be that there are typically only three or four distinctive sounds (phones) per word (rather than per letter) and that these sounds merge smoothly into one another to give a unitary impression of the entire word. A desirable goal, then, would be a word-reading device, one that would generate a "speechlike" output. Just what is meant by "speechlike" in this context is a topic to which we shall return after a look at how mechanisms might be classified.

The assumption implicit in all of the mechanisms described thus far is that the optical shape of the printed letter will be translated directly into an acoustic shape for recognition by the ear. Might it not be possible to use the shapes of printed words in much the same way, to build a word-reading machine? Some kind of optical or electrical integration across the letter elements of the word would be needed, but the integrated information could generate sounds for the word that vary less rapidly than the letter rate. On the above bases, we classified all machines that operate on the shapes of letters or words as direct translation machines and divided the group into non-integrating (letter-by-letter) and integrating (word-reading) machines.

Since letters have identities as well as shapes, there was the possibility in principle—though not then in practice—that letter identities might be recognized, in which case there would be much greater freedom in assigning sounds to them than when the letter shape per se must be translated into sound. Such machines were classified as recognition-type machines. The letter identities could, by direct keying, generate sounds which might be the letter name or the sound usually given to it in "sounding out" words. Another possibility would be to accumulate the letters for an entire word and use programmed keying to generate a distinctive unitary sound for the entire word. Technologically, all of this seemed very far in the future, but we gave much thought to the kind of sounds that might be generated and how useful they might be. In fact, the development of the synthetic language Wuhzi was intended, in part, to demonstrate

that if words of an arbitrary kind could be pronounced, then they could be learned as a new language—in one sense, a dialect of English, inasmuch as meanings and syntax are preserved, though sound similarities are not. Moreover, programmed keying with sounds that bore some resemblance to usual letter sounds might indeed make this dialect recognizable as English, even though many words would have bizarre pronunciations because of spelling-to-sound disparities.

Experimental Approaches

Several kinds of experimental work were undertaken to explore these more exotic types of reading machines: (i) two direct-translation, integrating types of word machines were built at the Laboratories and preliminary tests of them were made; (ii) RCA Laboratories was encouraged to undertake development of a recognition-type spelling machine; (iii) simulation studies were started to find out whether letter sounds might serve as a replacement for letter names (spelling); (iv) and a program of basic studies was begun to find out just what acoustic characteristics would make a sound truly "speechlike."

Word-type Machines. Neither of the two integrating-type direct-translation devices showed much promise. One, dubbed the Vowel Generator, produced a signal by mechanically chopping the image of several successive letters along the line of type, with major emphasis on the letter just coming into view. The signals were vowel-like in character and changed smoothly and continuously across the complete word, but they were completely lacking in consonant character and seemed rather indistinct. In a second machine, we attempted to correct this difficulty by generating signals of a consonant-vowel-consonant character for each four letters of the word (or less, at the end of the word). The change of sound character was to be determined by a cyclic switching operation, triggered by successive letters and interword spaces. The signals, as simulated, indicated that such a machine would have the fatal defect that the mechanical rhythm would dominate all other aspects of the signal, and so no further work was done on this device.

The RCA Recognition Machine, built as a bench model, utilized a scanning operation similar to that in the RCA version of the Optophone. However, the photoelectric signal served as input to a function matrix where it was matched against scanning patterns for the different letters of the alphabet. A match between input and matrix identified the letter, and this actuated one of a set of very brief tape recordings to sound out that letter's name. This experimental model was completed at the very end of the CSD program, so test results were meager. Recognition of letters was reasonably successful and successive letters in a line of type could be scanned and identified at a maximum rate of 48 to 60 words per minute, set by the magnetic tape announcing system. There were some difficulties with ambiguities between letters, and in maintaining alignment between type and scanning head. Also, when the letter sounds were recorded at speeds of 50 words a minute or so, the letters sounded as though they had been clipped, and since all letter sounds were equally long, the rhythm pattern was very pronounced. Overall, the development demonstrated feasibility for a letter-recognition approach and confirmed the expectation that reading rates could be improved somewhat over direct translation methods, though probably not beyond 50 to 60 words per minute.

It seemed reasonable to expect that a substitution of letter sounds for spelling (in which the names of the letters are themselves complete syllables) would have advantages as the acoustic output for a recognition reading machine. The sounds, of course, would have to merge smoothly into each other and yet be distinct enough to identify the letters. Could a blending of this kind be achieved?

Phonetic Summation. We undertook to answer that question by recording the letter sounds and reassembling them in new combinations for new sentences. The simplest, but most effective, of the experimental methods was to splice together short pieces of sound-on-film recordings to form the new sentences. For technical reasons, this had to be done by cutting the sound segments from one piece of film, assembling them end to end in a long narrow printing box, and then making a contact print for playback on a 16-mm film phonograph. (Today, with magnetic tape, or computers, the technical problems would be far simpler.) The primary difficulty, though, was not a technical one. It was one of isolating that part of a sound recording (made from spoken words or sentences) that represented the individual letter sounds. Another problem was that the sound segments all had to be of the same duration if they were to be used by a mechanism such as the RCA Recognition Machine, whereas the actual sounds of speech differ widely in duration.

The experimental result was quite clear: sentences generated in this way were unintelligible. The letter sounds were difficult to identify unambiguously, they did not blend, and the rhythmic pattern (due to equal durations) was a dominant feature. The possibility that the poor result was due to faulty splicing was excluded by cutting apart a recorded sentence, and then resplicing it. The reconstituted sentence was entirely intelligible and hardly distinguishable from the original recording. The failure of our one attempt at "phonetic summation" did not, of course, prove that speech sounds could not be combined into a speechlike stream, but it did suggest that this might prove difficult to achieve.

The core of the difficulty was that very little was known about the nature of speech sounds—about the acoustic parameters that cause a sound to be "speechlike." Certainly not enough was known to serve as a guide in devising an output for a reading machine, even one sophisticated and costly enough to provide letter identifications as a basis for generating the sounds. A program of research was undertaken in the final year of our work for the Committee to study speech sounds from this point of view. That work will be discussed in a following section since it was central to the next phase of the Laboratories' research program.

LESSONS FROM THE CSD PROGRAM

When Haskins Laboratories' program of research on

reading machines under CSD sponsorship began in mid-1944, it was oriented toward basic research on human factors in reading by ear. Just 3 years later, all of the research other than report writing came to an end, primarily because there was little prospect of achieving a practical working device or technological breakthrough within the next year or so.

In what sense, if at all, do the 3 years (1944–1947) of research represent a plateau in the evolution of reading machines? It is true that none of the devices—either the models built at the Laboratories or the fully engineered ones built by RCA—have survived except in museum collections, but it may be reasonable to claim that a deeper understanding of the problems was attained and a clear direction set for future research. As compared with the development of the Optophone 20 years earlier, the underlying problem was seen in a different way. For the Optophone, the problem had been seen as the technical requirement that print be converted into sound; in the CSD program, the objective was to match sounds from reasonably simple devices to the needs and capabilities of blind listeners. By the end of the CSD program it was clear that some kinds of sounds were inherently unsuitable, and that the reasons for this went beyond those that had been considered limiting for the Optophone. Moreover, it had become evident that the only kinds of sounds for which high performance could be expected would be sounds that were speechlike. Just how such sounds could be generated, and the complexity of the mechanisms needed to make them, were not well understood; but the direction in which a solution might be sought had been indicated.

The following paragraph from our report to the CSD (8) in mid-1947 makes clear the extent, and the limitations, of the understanding we then had about the overall problem: "One of the principal conclusions to be drawn from the work done thus far is that a successful reading machine must present its information in word-like units, not letter-by-letter. The development of machines which will do this requires prior knowledge of the physical characteristics of sound patterns which give a unitary impression. Spoken languages are made up of such units and, accordingly, a device which can yield speechlike sounds would appear to have a good chance of success. Moreover, recognition-type machines are inherently capable of generating a dialect which should resemble spoken English to a degree. It is clear that the ultimate success of the entire reading machine program (i.e., the development of either a recognition or an integrating type of translation machine) depends on basic information about the physical characteristics of speech-like sounds."

From what we now know about reading machines that paragraph appears both prophetic and quaint. No one now quarrels with the idea that a high-performance reading machine needs to be based on knowledge about speech, or that its output cannot be presented on a letter-by-letter basis. But nowhere in the paragraph does it appear that spoken English itself was envisaged as a reasonable objective for reading machine development. The most that could be foreseen, given the limitations imposed by the knowledge and technology of the time, was that it might be possible for a machine to recognize letter identities, and if it did, to convert the letters into phonetic equivalents that would "sound out" the words in an English-like dialect, though only if a way —not then evident—could be found to make the sounds merge together in a speechlike manner. Even such a machine would have pushed the knowledge and technology of the time to their limits.

PHASE TWO: RESEARCH ON SPEECH SYNTHESIS

For nearly 10 years, the research at Haskins Laboratories turned away from a direct concern with reading machines to more basic studies of speech and speech-like sounds. However, these studies eventually led back to the reading machine problem, and participation in the VA research program. Consequently, some account of the intervening events is appropriate here.

WHY IS SPEECH SO FAST AND EASY?

The principal thing that changed over the intervening decade was the nature of the problem. Increasingly, during the latter part of the CSD program, it was asked: Why did speech sounds serve so well as an acoustic signalling system? Speech was far better and faster than the best arbitrary sound codes that could be devised. Moreover, the limitations observed were able to be rationalized. Why did they not apply to speech? Could long experience and the use of word-size units make that much difference? Or did the sounds of speech match the ear's perceptual capabilities in some special and especially efficient way? So long as reading machines were the focal problem, the efficiency of speech was simply a well-known fact that could serve as a yardstick for other signals and proof that easy, speedy reception was possible.

The termination of the CSD program, followed by modest but long-term support from the Carnegie Corporation, left Haskins Laboratories free to concentrate on speech itself—on how something so complex acoustically could be perceived so easily and so fast. The physical complexity of speech had just become fully evident in the sound spectrograms published in 1946–47 by the Bell Telephone Laboratories (BTL) (52,53). But complexity was not all. One might have expected to see distinctive patterns corresponding to what were, to the ear, highly distinctive sounds. There were patterns in the spectrogram, to be sure, but they lacked obvious correspondences: They were different for the same word when spoken in different contexts or by different speakers; moreover, there was not a sequence of separable patterns corresponding to the sequence of obviously disjunct sounds. The real puzzle—given such seemingly muddy signals—was how speech could be perceived at all!

EXPERIMENTS ON SPEECH

The experimental approach taken was to use spectro-

grams as if they were recordings, intended to be played back to a listener, but with one difference: Changes could be made in the patterns before they were turned back into sound. By listening to the effects of such changes, it could be found what parts of a pattern were important in identifying the sounds of speech. The great advantage of spectrograms for such an analysis-synthesis strategy was that the information was laid out in conceptually manageable patterns. The disadvantages were that complex instrumentation was needed and had to be built—first a spectrograph to yield patterns to be worked on, and then a playback device for listening to the patterns, before and after modification.

Sound Spectrograph. The construction of a spectrograph and of a Pattern Playback was started in the final year of the CSD program as a way to discover just what acoustic characteristics of speech would make it "speechlike" and therefore likely to be useful in a reading machine. The principal reason for building a spectrograph was that the BTL model was simply not available, and not likely to be so for several years. Another reason was that it had a very limited dynamic range, adequate for visual inspection but not for playback with even moderate fidelity. It was supposed, from what was known about the effects of amplitude distortion, that a dynamic range of 30–40 db would be desirable; also, a spectrographic transparency was needed for use in the playback device. All of this meant a complete redesign of nearly every component of the BTL spectrograph. By the end of the CSD reading machine program, spectrograms on film had been made that were more or less comparable with the BTL spectrograms.

During the next few years, the spectrograph was reworked several times (10). The initial use of acetate discs for recording the sample to be analyzed (with 1.8 seconds of speech recorded on a single re-entrant groove) gave way to 12-second recordings on magnetic tape. This allowed three average sentences per spectrogram on film 7 inches wide by 7.2 feet long. The combination of a variable-intensity cathode-ray tube as light source, and a Photoformer[b] to linearize tube and film characteristics, allowed recording as spectrograms the (preemphasized) spectral intensities linearly as optical densities over a 36-db range. It was later thought to be a poor reward for the effort involved that this turned out to be far more precision and range than was required and, even more ironic, that the direct use of film spectrograms for playback was not the best way to experiment on speech.

Pattern Playback. The development of a playback device for spectrographic patterns also went through several stages. In that case, though, the care and refinements that went into the final instrument paid solid dividends and, in fact, the Pattern Playback is still used occasionally.

[b] Photoformer: A device that employs a cathode-ray oscillograph and negative feedback from a phototube to generate an output voltage that is an arbitrary function of the input voltage. The function is given by the shape of a partial mask on the cathode-ray tube.

The initial design, of which a "quickie" variant was built in the final days of the CSD program, used both the spectrogram on film and a set of sound tracks on film to modulate a beam of light. The spectrogram allowed light to pass where there had been energy in the speech spectrum at a particular moment; then, this light was again modulated at audio frequencies corresponding to the spectrogram. A photocell collected and mixed the various components to give a composite audio output. The sine-wave modulations were recorded onto a rectangular sheet of film as a sequence of sine-wave soundtracks, stacked vertically in order of increasing frequency. This was wrapped around a transparent cylinder that also carried the spectrographic transparency. Thus, rotation of the cylinder past an illuminated slit served both to scan the spectrogram and to generate the sine-wave modulations of the light that was then transmitted to a phototube.

There was nothing wrong with this arrangement in principle, but it had very serious practical flaws. Not nearly enough light came through the two films to give usable audio signals; in fact, the signal-to-noise ratio was so bad that almost nothing could be heard except noise.

In a second version, a number of changes and improvements were made (6,11). To improve the signal-to-noise ratio, a powerful mercury arc was used as a light source and a multiplier phototube was used as the pickup device. The two optical modulations were separated by a lens system. Audio frequencies comprising all the harmonics of 120 Hz up to 6000 Hz, were generated by a large tone wheel driven at 1800 rpm. Speech-rate modulations were provided by a spectrogram made into a belt and scanned at its own time scale of 7.2 inches per second. A number of detailed refinements were introduced, such as linearization of the tone wheel modulator by predistorting the sine-waves used to record it; also, elimination of the buzz from residual modulated light by a cancellation circuit. A further feature that proved to be very important was that the spectrogram (used as a transmission modulator) could be replaced by a reflection modulator. This was a clear acetate belt on which patterns could be copied in white paint from the spectrogram; likewise, freehand patterns of any kind could be converted into sound, just as if they were spectrograms.

INITIAL EXPERIMENTS WITH SPECTROGRAPH AND PLAYBACK

The spectrograph was in operation well before the Playback was completed, and a number of spectrograms had been made of a list of sentences (the so called Harvard sentences), that were designed for testing the intelligibility of speech in noise. The first question to be asked, once the Playback was ready to operate, was the very elementary one: Would it talk at all, and if so, how intelligibly? Theoretically, there was every reason to suppose that if one resupplied the approximate frequencies indicated on a spectrogram, then the resulting sound would be very much like speech and ought, therefore, to be intelligible. To be sure, the reinserted frequencies did

not match exactly those from the real speech, but rather were a substitute set drawn from the first 50 harmonics of a fundamental frequency of 120 Hz. The pitch of the synthetic speech would, therefore, be strictly monotone regardless of how the sentence had been spoken, but the spectral variations ought to be about right. In fact, the Playback did talk very well when it was given transmission versions of the Harvard sentences. The speech quality was poor—rather noisy and a little rough—but there seemed little question about intelligibility. Formal tests with naive listeners (11) gave scores of about 95 percent. Some preliminary experiments with overlays that blocked out parts of the spectrographic patterns were not very instructive, partly because the speech quality was then so poor and partly because the effects on intelligibility were difficult to estimate.

Some of the difficulties seemed inherent in transmission spectrograms so the alternate mode was used—one in which the Playback could work by reflection from patterns painted in white on a clear acetate belt. It was found unnecessary to copy the spectrographic patterns in detail; all that was really necessary was to preserve the features which were visually most prominent and then, largely on a trial-and-error basis, to make further changes that improved intelligibility. Paintings of the same 20 sentences prepared in this way gave intelligibility scores of about 85 percent. This was not quite as good as for the original transmission spectrograms, but the voice quality was better—even quite acceptable—and one could tell almost immediately by ear whether a particular change in the painted pattern gave a gain or loss in intelligibility.

SEARCH FOR THE ACOUSTIC CUES

It was at this point, in the early nineteen-fifties, that serious research on the nature of speech and its perception could begin. Our colleagues, Pierre Delattre and Alvin Liberman, carried through a series of studies that provided a solid experimental basis for the new field of acoustic phonetics (12,16,39).

What they set out to do was to find the acoustic cues —those parts of the spectrographic pattern that were principally responsible for a listener's judgment that he had heard one particular speech sound rather than another. They did this by working with syllables rather than sentences and by using sets of syllables that represented phonetic classes of sounds, e.g., the voiceless stops, or nasals, or fricatives. Then they varied the patterns, one aspect at a time, and asked naive listeners to identify the resulting sounds. In this way, after several years and many thousands of patterns, they were able to find the two or three principal acoustic cues for each of the consonants and vowels of English.

Only a beginning had been made on this task by the summer of 1956 when the research was reported at a conference on reading machines that was organized by the VA, and (somewhat later that year) when discussions began on the research that Haskins Laboratories might do for the VA. Before turning to an account of those events, it may be useful to relate the Laboratories' work to the research on speech that was underway elsewhere, and then to give a few examples of our research findings about the nature of speech (40).

There was, in the late forties and early fifties, an upsurge of interest in experimental work on speech. Much of it had been sparked by Homer Dudley's Vocoder (18,20) and Voder (19), the wartime development of the sound spectrograph, and Martin Joos' insightful little book on "Acoustic Phonetics" (35). These developments and the Laboratories' own demonstration of speech synthesized from simplified spectrograms, led in late 1949 to the first of a series of four speech conferences at MIT. Indeed, in 1955 and 1956, there were speech conferences at San Diego and Christchurch, England, as well as at MIT. By about this time, several groups had developed speech synthesizers of various kinds, some of which could generate quite natural-sounding speech [c]. One of the highlights of the meeting at MIT in the summer of 1956 was an on-stage conversation between Walter Lawrence's Parametric Artificial Talker (38) and Gunnar Fant's Orator Verbis Electris (22). Each repeated its demonstration sentence with an amusing array of pitch modulations.

The work at Haskins Laboratories on the acoustic cues with the Pattern Playback was making rapid progress by the summer of 1956. It was by then well-known, from visual studies of spectrograms, that the consonants and vowels so clearly heard in speech were not at all evident to the eye; in particular, the temporal stretches that were heard as vowels did not usually show the steady-state "characteristic tones" attributed to them in the twenties and thirties. Also, the consonant stretches seemed to evade simple characterization; they were often heard just where the spectrographic patterns were weak or changing rapidly and also in different ways in different contexts. But if one painted a copy of only the most prominent features of the real spectrographic pattern—essentially, a cartoon version—the Pattern Playback would "speak" it almost as clearly as if all the rest of the pattern were present.

---

[c] These synthesizers used resonant circuits to generate the formants and so could mimic the pitch changes characteristic of human speech, thereby adding an important dimension of naturalness. As the early versions of PAT (38), DAVO (55), and OVE (23,24,25) evolved in the late nineteen-fifties and early sixties, some read their control parameters from plastic tapes, much as spectrograms were read with the Pattern Playback and our own pitch-controllable Vocoder Playback (Voback) (7). As it turned out, improvements in naturalness made little contribution to the search for the cues.

So the first important finding was that intelligibility was carried by an underlying simple pattern, which meant that the speech signal could be drastically simplified with little or no loss. But this only sharpened the question about where the consonants and vowels were, or rather, how to characterize them. Were the rapid up-and-down excursions of the formants[d] merely connecting links between the "real" consonants and vowels? Or, did these transitions (as they had come to be called) themselves carry important information?

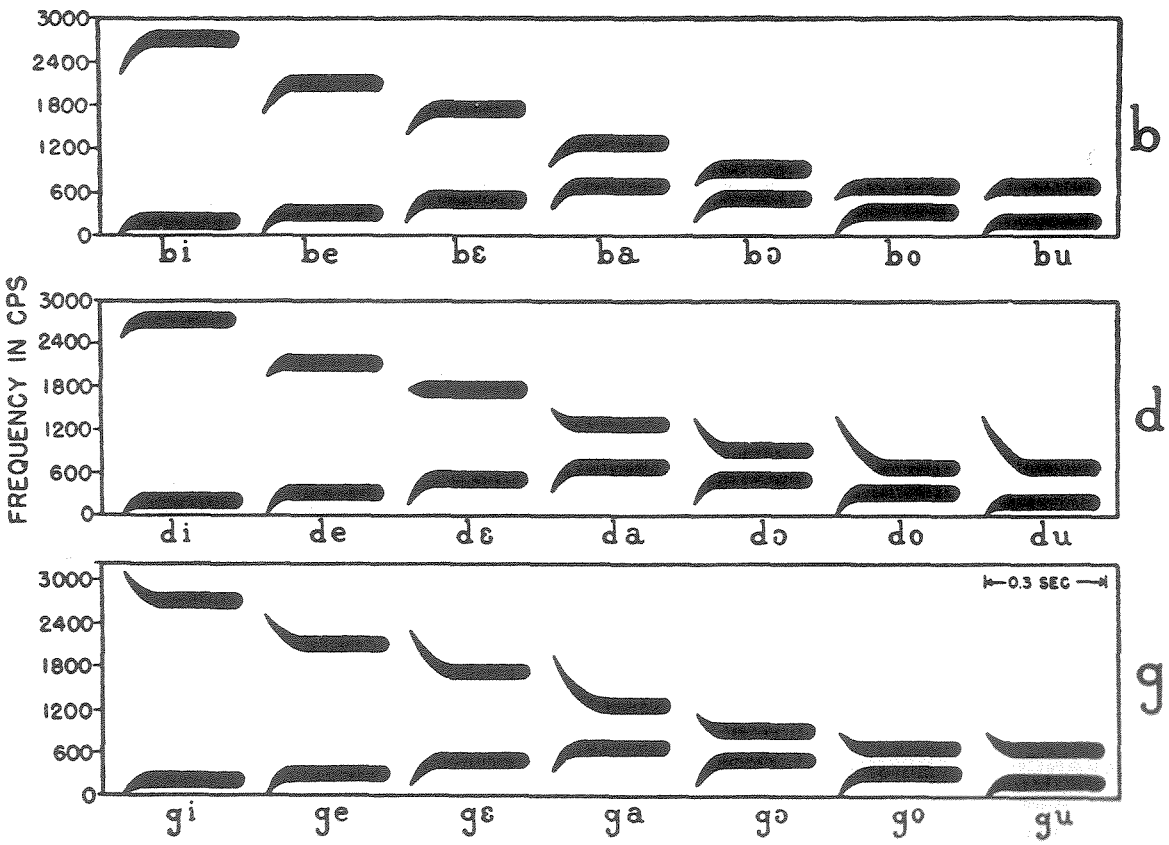
Some of the earliest experiments at Haskins Laboratories were with syllables such as ba, da, and ga that showed these transitions to a marked degree. The Laboratories had already confirmed that the vowels could be represented (to the ear) by two or three steady-state formants and that the vowels differed one from another only in their formant frequencies. So all kinds of formant transitions were painted onto the beginnings of the first and second formant pattern for, say, the vowel a. When the sounds from these patterns were played (in randomized order) to naive listeners, they had no trouble in labeling them as ba, or da, or ga. Their responses indicated two things: not only which transitions corresponded to each of the three consonants, but also that the transitions did indeed carry much information.

Experiments of the same kind with other vowels gave comparable results (Fig. 5), except that each vowel had its own preferred set of transitions for b, d, and g. However, comparisons across vowels revealed a rather simple principle from which the various transition patterns could be derived (Fig. 6): The second formant for each of the three consonants seemed to arise from its own "locus" frequency and then—except for an initial brief interval of silence—to move briskly to the vowel's second-formant frequency, whatever that might be; and, for all three consonants, the first formant started from a very low frequency (16).

In comparable experiments, it was found that the systematic changes, mainly at the start of the first formant, would produce the voiceless stops, p t k, or the nasal stops, m n ŋ; also, that the same changes could be

[d] A formant is a frequency region in which there is a relatively high concentration of acoustic energy. Formants are usually referred to by number, counting from low to high frequencies.

FIGURE 5
Synthetic spectrograms showing second formant transitions that produce the voiced stops b, d, and g with various vowels.



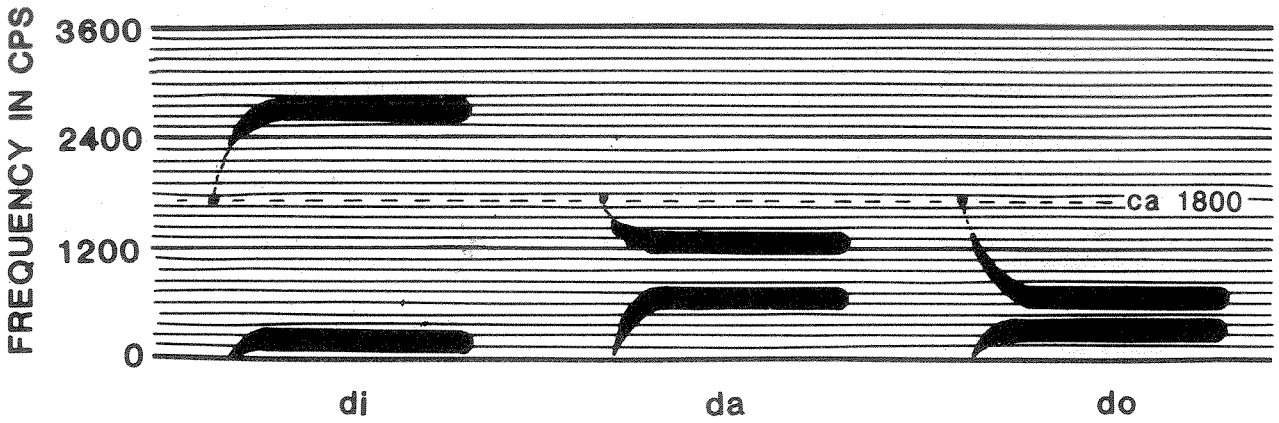
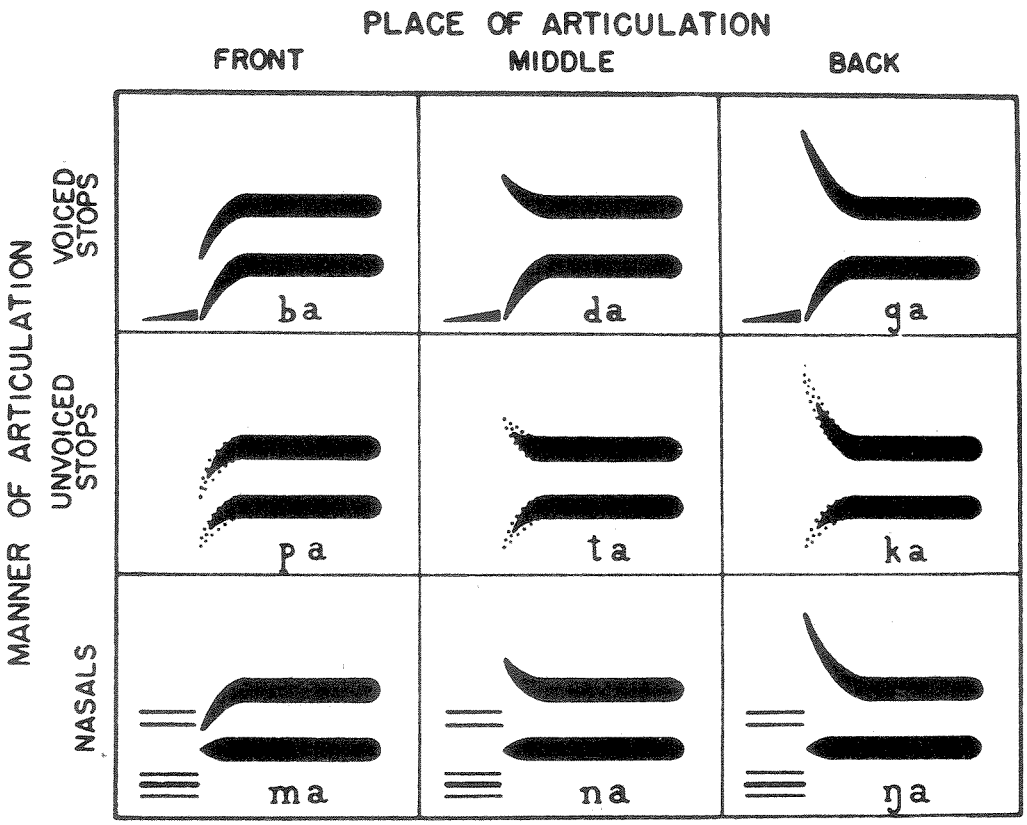


FIGURE 6
Spectrogrphic patterns for d with three vowels, showing extrapolations of the second formants to a common locus frequency (for d) at 1800 Hz.

FIGURE 7
Spectrographic patterns that illustrate the transition cues for the stop and nasal consonants in initial position with the vowel. The dotted portions in the second row indicate the presence of noise (aspiration) in place of harmonics.



applied to a full range of vowels. Thus, results to this point could be summarized (for a given vowel) in a 3x3 array of the acoustic cues (Fig. 7), with the x-axis and y-axis corresponding to the conventional phonetic dimensions of manner and place of articulation (40).

RELEVANCE TO SPEECH SYNTHESIS

The close correspondences between acoustic cues and articulatory dimensions had important implications for an understanding of speech perception, and this set the direction of much future research (13). However, a different aspect of the results proved to be more directly relevant to the reading machine problem, namely, that the acoustic cues were essentially independent of each other and that they combined freely to give the full set of stop and nasal consonants. Notice what Figure 7 tells us: We can start with only three different manner cues and three different place cues and combine them to get nine different consonants; further, if we use these same triads of place and manner cues with the formant frequencies for the seven vowels of Figure 5, we can get 63 different syllables.

If this same combinatorial principle applies to the acoustic cues for the remaining consonants of American

English—as further research showed that it did—then one would need to know only a limited set of cue-recipes to undertake the synthesis of words and sentences never before seen as a spectrogram. It is, in fact, possible to do so, though the doing is not quite as simple as the above discussion would imply. Pierre Delattre became quite adept at this form of "synthesis-by-art"; one of his early creations is shown in Figure 8. Clearly, he had in his head an implicit set of rules to guide his painting. If those rules could be made explicit, then anyone skilled with a paint brush could do speech synthesis by rule.

PHASE THREE: RESEARCH FOR THE VETERANS ADMINISTRATION

BEGINNINGS OF THE VA PROGRAM

There had been earlier conferences on sensory aids for the blind, but it was at the Fourth Technical Session, in August of 1956, that an active research program began to take shape, and it was only a few months later that the research program at the Haskins Laboratories—the focus of the present account—got under way.

The first of these conferences had been held in 1954, and others followed at nearly yearly intervals. They reawakened interest in reading machines for the blind, although most participants still saw the problem in terms of how to generate from the printed page a set of letter-by-letter sounds, comparable in a general way to

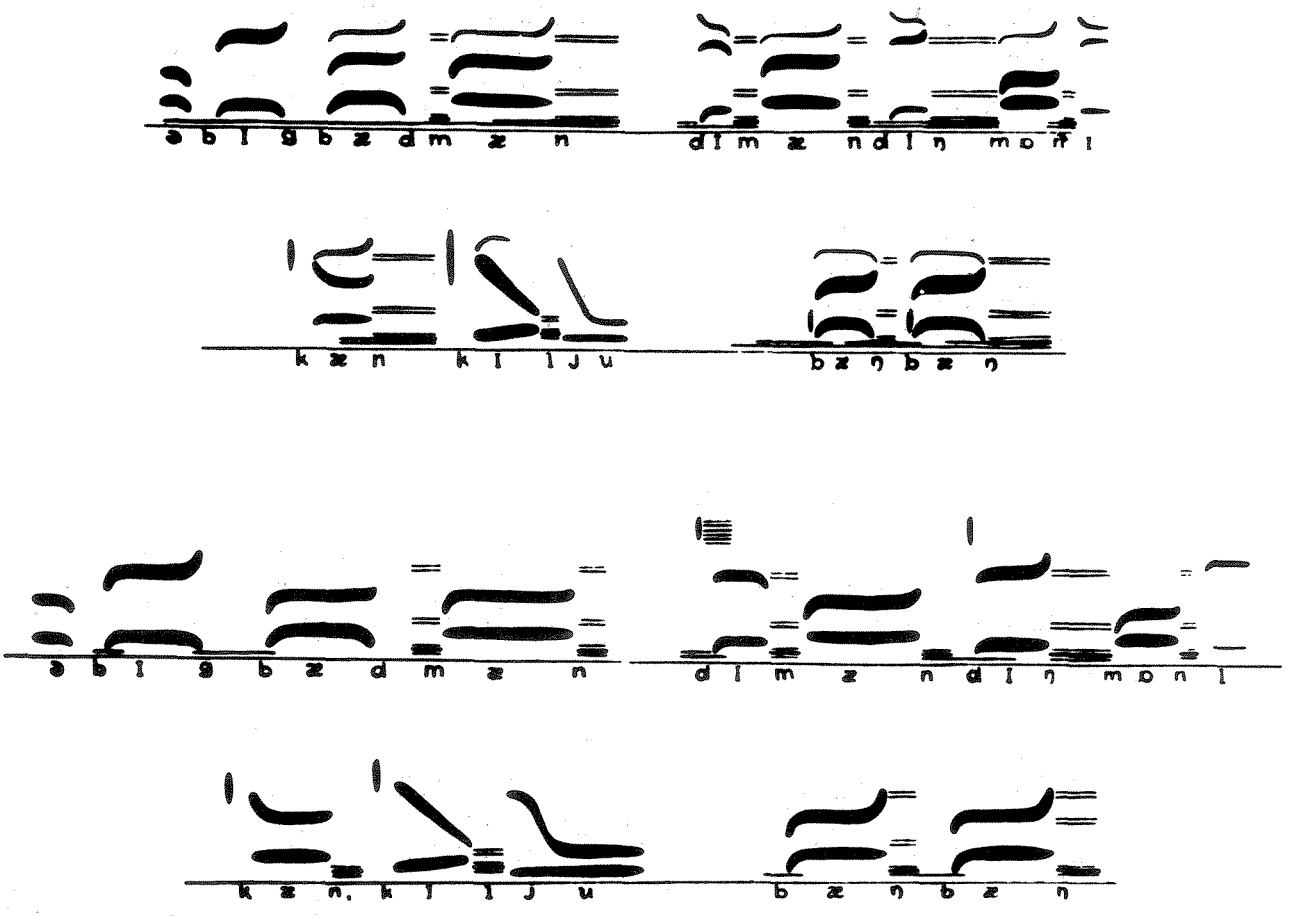


FIGURE 8
Two versions of a sentence employing principally stop and resonant consonants. The lower version is a first draft which was painted directly from the typewritten text ("A big bad man demanding money can kill you. Bang, bang."), in accordance with the "rules" derived from experiments on acoustic cues. Revisions by ear resulted in the upper version. Both were highly intelligible when converted into sound by the Pattern Playback.

Morse Code. By that view, the technical problems were not trivial, nor was the task faced by the blind person in learning an arbitrary acoustic code—but those problems had somehow to be lived with and overcome, since no other kind of reading machine seemed feasible.

A second view of the problem was that the principal conclusion from the CSD research—that arbitrary letter-by-letter signals simply would not do—might have to be accepted in spite of the technical complications that this conclusion implied. The worst complication was thought to be that the machine would have to recognize the printed letters in order to generate acceptable signals from them. Optical character recognition was then in its infancy, so this view of the reading machine problem seemed to erect a second high barrier; or, to put the matter affirmatively, there were now the two problems of devising a simple optical character recognizer, and then of teaching it to speak aloud the letters it had recognized.

The third view, put forward by the Haskins Laboratories, was that even these two technical problems—OCR and a letter-by-letter output—were not the main hurdle; rather, that the central problem was one of matching the acoustic signal to the listener's perceptual capabilities, and further, that this required the acoustic signal to be, at the very least, "speechlike". This view changed, over the course of the conferences, to the far more demanding requirement that the audible output must be speech itself.

The need for a speechlike output was presented at the Fourth Technical Session in a paper on "Synthetic Speech and the Reading Machine Problem." The paper also surveyed the various types of reading machines then thought to be possible, even though some seemed visionary. It now serves to show how much—and how little—was really understood at that time about reading machines and especially the output problem.

The three views of the reading machine problem were the basis for the three phases of the VA program of research, which appreciated the limitations of the acoustic code and spelling approaches, but saw also that the difficulties in generating speech from print would take years to solve. A practical program, it was believed, must have earlier and more certain payoffs even if the resulting devices might have limited capabilities. The principal contractor under the first, short-range phase of the program was the Battelle Memorial Institute, which was charged with developing and testing a device to generate arbitrary acoustic signals from print. Battelle was to build on the earlier work with Optophones and the RCA A-2 Reading Machine (1).

There were two middle-range projects: a major one, assigned to Mauch Laboratories, was to devise a machine that could recognize printed characters reasonably well and generate a spelled output (56); a smaller contract was given to Professor Milton Metfessel, University of Southern California, to press ahead with his work on a "spelling-bee" output that would gain reading speed by using very brief sound segments for the letters (44).

The long-range phase of the program, assigned to Haskins Laboratories, also had two parts: one was to build a machine with which to test the usefulness of Compiled Speech, i.e., a "spoken" output made by splicing together standardized voice recordings of words to form sentences; the second was an open-ended study of speech and speechlike signals to find out what kind of artificial speech would work best in a reading machine and how to generate it, assuming that one had available the output of an optical character recognizer.

RESEARCH ON AUDIBLE OUTPUTS

Objectives—Although the two tasks undertaken by Haskins Laboratories were formally distinct, they had a common purpose: to arrive at the best choice of audible output signals for a high-performance reading machine for the blind. There was not, at the beginning of the program or at any later point, the intent to design and build the device itself. This restriction on program objectives was due in part to the realization that an optical character recognizer would be an essential part of a high-performance reading machine, and the belief that commercial needs would make OCR devices available by the time the output problem had been solved; furthermore, engineering development was neither a strength nor an interest of the Laboratories.

It was not at all clear what kind of audible output offered the most promise, provided only that it was speechlike: synthesis seemed to have the potential for natural, flowing speech, though only if a great deal more could be learned about how to synthesize from a phonetic transcription. Even then, the often peculiar relations between the letters and sounds of words might mean that synthetic speech would always have bizarre pronunciations. An obvious competitor was compiled speech; it could avoid these strange pronunciations by using a human speaker to supply the correct sounds, whatever the spelling. But speech compiled from word recordings would have its own language problems: A single, fixed pronunciation would have to serve, even for words which a human would speak differently and with different inflections when they occurred in different contexts. Then, too, there are so many words! No recorded dictionary of a practical size could contain them all. Spelling would be a possible way to deal with the exceptions, but would be disruptive if it occurred very often.

On balance, though, compiled speech from a Word Reading Machine seemed the surer solution and the one on which practical efforts could be started without delay. Conventional tape-splicing techniques would permit initial studies of some of the language problems, though such methods would obviously be too slow and laborious for the production of the paragraph-length texts that would be needed to assess comprehensibility and acceptability by blind listeners.

This is why one phase of the research program consisted of a contract for the construction of an Interim Word Reading Machine (IWRM) to serve primarily as a research tool in studies of the language problems inherent in compiled speech. The device was to operate semiautomatically in a laboratory environment. It was not intended to be a production device for the prepara-

tion of recordings in volume; therefore, a simple design with medium speed components would suffice and design compromises could be made, so long as they did not adversely affect the quality of the voice recordings.

Pending completion of the IWRM, the research part of the program was to concern itself with audible outputs of all reasonable kinds. This included the language problems of compiled speech; it included also work on the rules by which a machine could synthesize "spoken" English when the letters on the printed page were identified for it, as they would be in a library-type reading machine. The studies were to start with the information then available about speech synthesized from hand-painted patterns; then to adapt these results to the synthesis of speech by a machine and (later) introduce automatic "corrections" for English spelling. Another goal was to devise a "speechlike" output suitable for a personal-type (direct translation) reading machine. This would require a careful study of letter shapes to find elements which could be easily identified by a simple machine and a study of the best way to assign speech sounds to those elements. Much of the experimental work—at least initially—would be done with the existing Pattern Playback equipment. Eventually, the plan was to build a laboratory device to produce reasonable quantities of synthetic speech and speechlike sounds in order to test their usefulness as reading machine outputs.

The formal objectives of the two parts of the Haskins Laboratories program are summarized in the above paragraphs. They remained the general guidelines throughout the two decades that followed, though there were several shifts in emphasis as new information and new techniques emerged. Progress was uneven, shifting from one aspect of the program to another, and some early hopes fell by the wayside, victims to competing solutions. For all these reasons, a brief chronological survey may serve as a useful introduction to more detailed accounts of the several lines of study.

Chronological Overview

Exploration of Alternatives—The first phase of the program, beginning in 1957, was concerned mainly with the competing claims of various audible outputs. It was not until 1970–71 that a clear choice could be made between the two principal contenders.

The work on compiled speech consisted of a series of small studies of such things as monotone versus inflected speech, rate of speaking, manipulation of stress, and the like. These studies continued at a steady pace, answering many of the questions, but always hindered by the need to rely on slow and laborious manual methods. These methods gave way, by the end of the sixties, to a computer facility that made possible the easy "reading aloud" of page-long texts.

The design and construction of the IWRM progressed rapidly to a point where the device was fully designed and more than half completed. At that point, the funds ran out, and although the work was carried to completion it had to be done at low priority, and at a pace that eventually made the device obsolete. Its functions were, in fact, taken over at the end of the sixties by the computer-based system mentioned in the preceding paragraph.

Progress on speech synthesis by rule, like that on the IWRM, progressed rapidly at first but then slowed, though for different reasons. The initial surge came from the work of Frances Ingemann and the fact that she had several years of research results on the acoustic cues as a basis for her work. She was able, within the first year, to organize all of this material into a set of rules for synthesis that had not previously existed, and that were fully explicit instructions on how to paint (for the Pattern Playback) the control patterns that would "speak" any desired sentence. The next advances came more slowly, but as with compiled speech, the program gained momentum again by the end of the sixties when computer facilities and an additional body of knowledge about speech had become available.

The mid-sixties were a period of uncertainty as to just where the program should go. Progress was slow on both compiled speech and speech synthesized by rule, needing (as was later learned) the technical assistance that only computer methods could provide. During those years, considerable effort was put into a new way of generating speech that seemed to evade some of the difficulties of both compiled and rule-synthesized speech. However, by 1971, interest in this new variant —called Re-Formed Speech—had succumbed to the good progress that was then being made in the synthesis of speech by rule.

Thus, by about 1970, the field of possibilities had been canvassed, with only two candidate methods surviving: compiled speech and speech synthesized by rule.

Automating Speech Production; Evaluation Studies. The objectives of the work then shifted from studies of compiled and synthetic speech to ways of obtaining fairly long passages of each type. These were needed in order to test for intelligibility and acceptability and, indeed, to make a choice between the two kinds of speech output. There had been sufficient success with both types of output to inspire thought about the kind of Library Service Center that might be set up within a very few years to provide recorded books on demand for blind veterans. This was envisaged as a central facility that the VA would itself set up and manage, with technical advice and assistance from the Laboratories.

The evaluation studies made it clear that the contest between compiled and rule-synthesized speech had been won conclusively by synthetic speech. Hence, efforts were shifted almost entirely to automating speech synthesis by rule, even though that was a substantially harder and longer job than generating compiled speech by machine methods.

User Evaluations of Synthetic Speech and Plans for a Reading Service Center. By 1973 it was possible to report that "from a purely technological viewpoint, the automated production of speech from printed text is wholly feasible. Indeed a prototype system exists at Haskins Laboratories". This did not mean, however, that all the problems were solved. It was felt that synthetic speech was reasonably intelligible and acceptable; for

example, short stories played to a naive audience would be understood and appreciated even though some words and names might be missed. Thus the output was reasonably acceptable despite its machine accent. But just how intelligible the speech was, or what sounds and words were giving the most trouble, or whether in a more general sense the synthetic speech would satisfy a serious reader after the novelty had worn off—these questions could not be answered.

The first step in answering these questions was to make quantitative, controlled studies of word and sentence intelligibility, and later, of the comprehension of paragraph-length passages. The second step was to start preparing for in-depth user tests aimed at testing both the utility of reading machines in real life situations and the improvement of synthetic speech in response to user comments. Since it seemed by then unlikely that the VA would organize facilities for these user trials, plans were made jointly with the University of Connecticut to set up a Reader Service Center for blind students. It was planned to provide the students with synthetic speech recordings of assigned readings from their textbooks; also, work at the Laboratories pressed ahead on mechanizing the synthesis-by-rule procedures so that substantial quantities of recorded synthetic speech would be available.

Final Phase. By 1975, it was concluded (somewhat reluctantly) that these cooperative plans for a Reading Service Center to serve blind students and to evaluate and improve reading machine performance would have to be abandoned for lack of funding, even though the technical and human facilities were in hand. The research was turned, instead, to improving the quality of the speech synthesized by rule and, in particular, to developing a new and a better speech synthesis algorithm. The quantitative evaluations of Phase Three had shown that the intelligibility of the synthetic speech was good enough for easy comprehension of simple, straightforward materials, but that listening to it put a heavy load on the comprehension of more complex (textbook) materials. Hence, further work on the rules for synthesis would have been required in any case.

By the end of 1978, it was becoming evident that some kind of reading machine—as distinct from a library-based reading center—would soon be feasible, but with further compromises in a speech quality which was already only marginally adequate.

The foregoing overview has sketched the chronology of the reading machine research that Haskins Laboratories did for the VA. There were several simultaneous strands that can now be recounted separately and in somewhat more detail.

Compiled Speech

The sections that follow deal with some of the main areas of research on compiled speech and its language problems, as they were investigated by essentially hand methods. The account turns then to the development of a machine for doing the compilation automatically. An account of the final competitive tests between compiled and synthetic speech will be deferred until the evolution of speech synthesized by rule has been described.

Preliminary Experiments. Linguistic research on compiled speech began with an applied program of purposely modest size. The task was to record a small spoken-English vocabulary from which small test sentences could be built. There was only one significant constraint to be observed in recording the vocabulary: Only one spoken version of each spelled word could be stored for use, although that single version could be employed more than once in a sentence.

An important consideration, in the effort to compose a usable store of single tokens of spoken words, was the fact that a naturally spoken sentence is a multiword unit. All naturally spoken sentences are delivered with intonation—a variable and varying prosodic feature that extends across word boundaries, and even across phrase boundaries. This fact would complicate the attempt to generate whole, "life-like" sentences from "frozen" words which would have to appear in the same acoustic shape in every context (i.e., with unchanging pitch and pitch contour, duration, intensity, and phonic color). Nevertheless, the precise nature of the complications had to be ascertained.

The initial test began with recordings of a magazine article that had been read by a male talker and recorded on magnetic tape. The talker, who spoke with reasonably normal American speech, read the selection in four ways: in normal intonation and in a monotone, producing each of these at a normal rate and at a slow rate. Next, the individual words of the recordings were "edited apart" by listening to the tapes and marking word boundaries. Once isolated, the words (on tape snippets) were mounted separately on "Language Master" cards (which permit the separate and successive playing of small bits of speech), and were re-recorded in various grammatical arrangements to test the compatibility of the vocabulary when heard in new sentence structures.

Informal listening tests of the manually-compiled sentences by members of the Laboratories' staff produced the following observations:

Prosodics (The melody, timing, and loudness of speech)

1. A word's acoustic shape normally changes according to its verbal and intonational context.
2. A word in prepausal position must be acoustically longer than it is in other positions.
3. Polysyllables are never normally spoken in a monotone.
4. Listeners feel that pitch is the primary cue to stress and intonation.

Grammar

1. Articles and prepositions are usually less prominent (perceptually and acoustically) than other parts of speech.
2. When the vocabulary is recorded by the talker, certain highly frequent words must be spoken many times, in a variety of ways, so that the most probable (or most neutral?) form of each word can be selected for the basic vocabulary supply. A case in point is the

most frequent English word, the, which has four main possible pronunciations; another example is which, which can play more than one grammatical role.

Punctuation

A short interval of silence (e.g., 750 msec) in the output can substitute for a printed comma and a longer silence (1750 msec) can suggest a period. These durations work well for the somewhat slow rate that the particular talker used, but they might have to be changed for speech at other rates.

The talker's manner of speaking

1. If the vocabulary is spoken in a monotone, the words are fairly compatible when transplanted into sentences, but they are dreary and slow. Listeners find monotone delivery of text too terribly dull to endure for more than a very few minutes.
2. An intentionally undramatic (but not monotone) reading produces quite good words for recombination into new sentences.

Some of the observations noted above were made on the basis of negative evidence. In attempting to make sentences from single prerecorded words it was easy to discover important features of normal speech by their sometimes jarring absence in the trial sentences. For example, a word put into prepausal position (at the end of a sentence) was often heard as much too short, although it was heard as sufficiently long when located elsewhere in a sentence.

Not all the results of the preliminary linguistic study surprised the investigators, although some did. An attempt to address some of the problems pointed out by the observations—especially with respect to prosodics—was made in designing the IWRM and later, even more successfully, in the computer-implemented speech-synthesis-by-rule system devised by Mattingly. Other problems, such as the multigrammatical roles of English words, which are encountered in generating speech from print, still remain to be solved. It seems unlikely that a solution to this problem can be found until computer programs for parsing a text and analyzing its meaning become more sophisticated than they are today.

The Search for Prosodic Descriptors. To complement these early experiments with compiled speech, a study of the acoustic properties of stress and intonation in real speech was undertaken. A pilot test, employing the same talker, was run to establish procedures for later data acquisition. Speech analysis was performed using spectrograms, waveform traces, and fundamental frequency contours recorded on 35-mm film.

Provocative problems were encountered in trying to measure syllable duration, intensity, and even fundamental frequency. (How could perceptually important dynamic events be measured and described acoustically? Who could say where syllables began and ended, when they visibly flow together in the acoustic record of speech?) An element of arbitrariness was inescapable in deciding what was the significant aspect to measure. In the end, the peaks of the syllable intensity and frequency contours were selected as the principal descriptive features of these parameters, whereas for syllable duration, acoustic amplitudes augmented by listening served as descriptors of the syllable boundaries.

Using these descriptors, the prosodic aspects of three long sentences spoken by each of four adult talkers (including a female with a low-register voice) were analyzed acoustically. The measured items, consisting of some 400 syllables, were made by tedious manual methods, there being no other way available at that time.

One observation that emerged from the prosodic study led to the hypothesis that polysyllabic words and highly frequent phrases share a common prosodic property, that is, a persistent stress relationship among their component syllables. A further observation indicated that the direction of combined prosodic feature movement (up or down, from one syllable to the next) was the acoustic key to word accent (lexical stress). These ideas were tested in an experiment in stress perception that was run concurrently, using as stimuli brief syllables of synthetic speech whose frequency, duration, and intensity components were controlled and manipulated. In formal listening sessions, 10 staff members selected the more prominent (stressed) syllable in each of 64 syllable pairs. The results showed clear evidence that the prosodic features are additive in stress perception, as the descriptive study had suggested. The experiment did not reveal how stress and intonation could be separately defined, however, yet it could be said that fundamental frequency and intensity peaks do tend to diminish across a long utterance, and that syllable duration rises before a pause.

Preparation of a Larger Lexicon. Regrettably, those characteristics, no matter what their generality or importance for naturalness, could not be used by the IWRM in generating compiled speech, since it required that a single recorded version of a word (with its set pattern of pitch, loudness, and length) must be used on every occasion. The best goal attainable appeared, therefore, to be one of making word recordings that would be neutral (i.e., most adaptable to all sorts of contexts), and yet fairly natural (consistent in tempo, smooth in articulation and not monotonous). It seemed reasonable to hope that an impression of normal sentence stress would be supplied by the listener, much as it is by the reader of a printed text, largely on the basis of syntax and word order.

If, however, word order is contradicted by abnormal stress relationships among the (rearranged) recorded words, ambiguities or confusions in comprehension result. Hence, in order that the words might be recorded and stored in the lexicon in their most congenial forms, the effects of abnormal stress call for an examination of the words in respect to their overall frequency in written English, as well as in respect to their most frequent grammatical and phonological environments and semantic functions.

A statistical study of English words was, therefore, begun with a scrutiny of the Dewey (17) and Thorndike

and Lorge (59) lists of syllable and word frequencies. A list of about 7000 of the most frequently used words was drawn up for the IWRM vocabulary. The grammatical usages possible for each word were listed. The results of this study were both enlightening and, in a way, discouraging: The diverse grammatical functions, especially for the most frequent English words, make obvious the difficulties to be overcome in the conversion of print to speech by machine. Thereupon, a grammatical investigation of a number of randomly selected texts (portions of novels, newspapers, magazines, and personal letters) was made with the intention of learning which part-of-speech sequences (syntactic structures) most often occurred.

It was found that the prepositional phrases occur with overwhelming frequency in texts of all sorts. The first words of prepositional phrases are words of absolutely greatest frequency—a preposition (e.g., of, in, with, by and to) is most often followed by an article (of which only three exist in English: the, a, and an)—words usually spoken with a very low stress. A prepositional phrase ends in a noun (as do sentences, in most cases). Nouns receive relatively high stress; also, nouns terminating prepositional phrases (or sentences) are either potentially or actually prepausal, and so usually exhibit a falling pitch contour and declining loudness.

Based on such observations, "prescriptions" were evolved for the manner in which the vocabulary for compiled speech should be spoken.[e] Fundamentally, the rules relied on (i) the probability that a given part of speech would occur in a certain grammatical context, and (ii) the probability that a given part of speech plays a patterned role in intonation. By referring to acoustic and perceptual analyses of real speech, along with reference to the experimental sentences in compiled speech, it became possible to describe objective intonational data in terms that a talker could use in subjectively monitoring his own speech when producing the huge lists of words required for the compiled speech lexicon.

After a number of try-outs for the role of talker, a male graduate student in linguistics was chosen to perform the difficult task. Working part-time weekdays for about 13 months, he recorded the nearly 7200 lexical items (in one-hour sessions), following the very exacting instructions for speaking the words. (These had been grouped in a long series of scripts by the initial sound of words, by number of syllables, and by part of speech.) Nouns were delivered at normal pitch, with falling intonation, at normal speed and loudness; verbs at a slightly lower pitch level, faster and less loud than nouns; (most) adjectives at the pitch of verbs, but with rising intonation, etc.

[e] This way of generating the words for compiled speech probably accounts for the reasonably good results we obtained with sentences and paragraph-length texts, even at nearly normal speech rates (see infra). A less optimistic view of compiled speech was taken by Stowe and Hampton (57) on the basis of intelligibility tests of words spoken in isolation at slow and fast rates but without special attention to the manner ("prescription") of their production.

The talker—a diligent, talented, and tireless speaker—managed to comply with these prescriptions. When his job was completed, thousands of word recordings had been collected that were deemed compatible in pitch, loudness, and length. A small team of assistants kept pace with the daily recordings. One person edited each hour-long tape to isolate the words; another one or two people manufactured Language Master cards that carried the individual words as separate spoken items; finally, the editor punched a small hole fore and aft of the spoken word on each card. (The holes, plus a photoswitch, were used to control another recorder that was specially modified for start-stop operation.) In all, about 1.3 miles of adhesive-backed magnetic tape was edited, cut apart, and mounted on the (homemade) Language Master cards. Thus, the lexicon was gradually assembled.

We now backtrack slightly to the period just preceding the above recording operation, to mention two matters of importance to the structure of the vocabulary—missing words and helpful suffixes.

Missing Words posed a problem, no matter how large the recorded lexicon, since some words that had not been included in the storage would inevitably occur. In the originally proposed lexicon (6000 words) it had been estimated that some 5 percent of the words in an ongoing text would be missing. The practical solution for that problem was to add the spoken letters of the alphabet to the lexicon, so that spelling aloud would replace the missing vocabulary item. Although each of the 26 letters of the alphabet was spoken rapidly (and very carefully) prior to storage, each one was unavoidably one whole syllable long (and w was even longer). This meant that the overall word rate of a sentence declined considerably when even one word had to be spelled. Moreover, words requiring spell-outs were longer, on the average, than the (high-frequency) words that constituted the recorded vocabulary—resulting in greatly reduced word rates in any sentence that needed several spelled words. Still another negative feature of the spelling procedure was the fact that the missing words were the least predictable ones in the sentences, and therefore caused comprehension problems for the listener. Worst of all, listeners found it irksome and hard to shift quickly from the medium of speech to the medium of spelling.

Helpful Suffixes, on the other hand, provided a way to increase the effective size of the lexicon very substantially, simply by adding a few extremely frequent (spoken) suffixes:

[s] as in hats or writes	[ɪŋ] as in heading or writing
[z] as in heads or rides	[t] as in looked
[ɪz] as in roses or rises	[ɪd] as in wanted

Thus, for example, a word stored only as a singular noun could easily be generated in its plural form (e.g., hat + s), or, a regular verb in the lexicon could be inflected (e.g., look + s; look + t) by adding the appropriate sound to the base of the word. (Rules were written for analyzing the word into base and suffix.) In turn, this

study led to the writing of preliminary rules for converting spelling to sound. These rules worked for most of the vocabulary, with the exeception of only those words having highly irregular pronunciations. (The general letter-to-sound rules were modified later and written as rules for the automatic pronunciation of surnames.)

During this study, an intriguing fact came to light: It was found that some very frequent suffixes (such as -ation) have fixed stress, and tend to "predict" the stress shape of the preceding syllables in the words to which they are attached. This observation was tucked away for future reference (when automatic lexical stress prediction might be wanted) along with a list of the "stress-stable" suffixes, and of prefixes that might also be used for stress prediction when suffixes were either non-predictors or altogether absent from a word.

Early Preparation of Compiled Speech Texts. With a spoken vocabulary mounted on some 7000 cards, we were now in a position to generate very many different sentences and long connected texts. It must be remembered, though, that the generation of compiled sentences in this early part of the project relied on manual retrieval of the Language Master cards, and manual transfer of the single word recordings from the Language Master machine to the stop-start re-recording device. Although it took hours to compile a thousand words of text, a large variety of literature was duly sampled. Selections from, for example, Bertrand Russell's writings, recent novels, obscure Russian novels, the news and sports page of the NY Times, random sections of Time Magazine, and personal letters were converted to compiled speech—and subsequently appraised by a variety of listeners, ranging from the Laboratories' staff to visiting scholars (some of whom were blind).

The listeners' consensus was that compiled speech was generally intelligible. The voice was pleasant, but the delivery was often a bit dull, partly because the word rate was on the slow side (about 120 words per minute, if no spelling occurred in the selection; much slower when words had to be spelled). And spelled words interfered drastically with comprehension. There was also, of course, a certain choppiness in the delivery —unavoidable when "canned" words were abutted to build sentences. This confirmed our belief that a really satisfactory reading machine for the blind would have to deliver speech that was truly continuous. Also, naturalistic intonation is a sine qua non of continuous speech, whereas compiled speech was, at best, a mild caricature of normal delivery.

Word Duration and Speech Rhythm. Nevertheless, despite obvious shortcomings, compiled speech continued to be studied and it proved to be instructive in a number of ways. One very obvious problem concerned word duration and speech rhythm; clearly, they were interrelated, and both were deficient in the compiled speech. The problem was a challenging one because duration is affected by numerous factors which, if better understood, could lead to the writing of better rhythmic rules for speech at a variety of rates—and also because speech rate is a prime concern of blind people who must do their reading by listening.

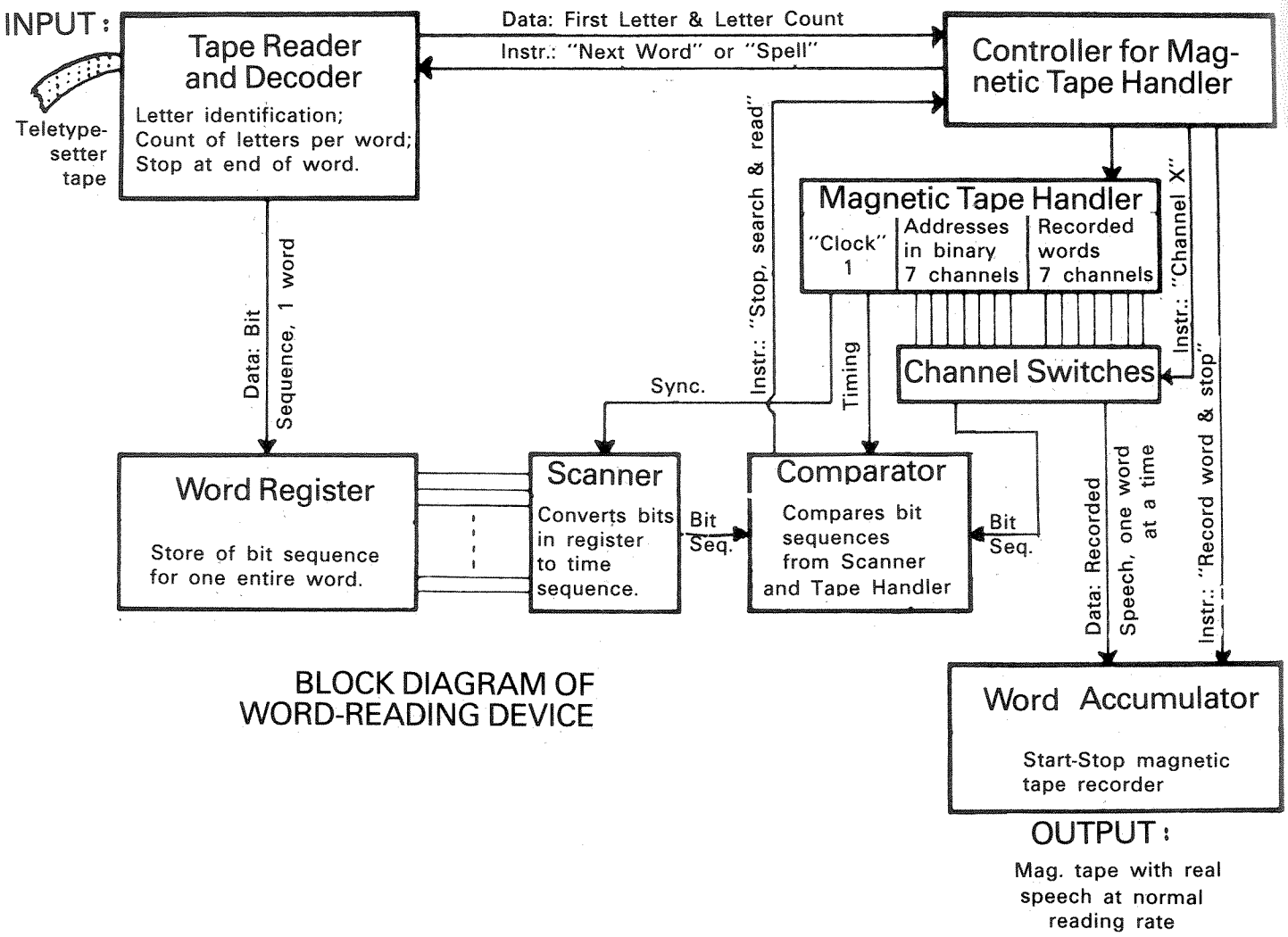
A series of studies on segment, syllable, word, and phrase duration in continuous speech was undertaken at about this point and led to a paper entitled "The Elastic Word" (30). Aside from illustrating the durational flexibility of various linguistic units, that paper also demonstrated that native speakers of English closely share durational patterns in their speech, a fact that underscored the need for very carefully specified rules for duration in synthetic speech (and other modes of output). This requirement may be seen in retrospect to have foreshadowed the early obsolescence of speech compiled from the durationally inflexible vocabulary of the IWRM.

An Interim Word Reading Machine

It was clear from the beginning of the program that some kind of machine would be needed to produce long recordings of compiled speech, i.e., to perform automatically the equivalent of many thousands of tape splicings. The overall design was fairly simple and straightforward: the device used Teletypesetter (TTS) tapes as input and accumulated voice recordings, word by word, as its output; it had to have a sizeable dictionary so that only a few words would need to be spelled; also, it had to operate automatically, reliably, and with a minimum of supervision. Actually, quite a number of design considerations were involved in blending these requirements into a single machine.

The Interim Word Reading Machine was an interim device only in the sense that it bypassed such major engineering problems as character recognition and real-time access to a large memory. Teletypesetter tapes (available to us from Time Magazine), provided a large amount of input material that would otherwise have had to come from character-recognition equipment. The need for fast access to a large memory was also evaded because the stored recordings were not read out immediately (as they would have been in a real-time device), but were transferred to a start/stop recorder that could wait as long as necessary for the next word to be found. The quality of the output speech was not affected by these compromises; the only penalty was speed, since the IWRM required hours to generate a speech recording that lasted only minutes.

Operation of the Interim Word-Reading Machine. The operation of the device is illustrated in Figure 9. A word from the TTS input tape is read into the Decoding unit, where each character is interpreted and either rejected (as relevant only to typecasting) or accepted and stored as a digital code. A search of the Dictionary tape can now proceed. The identity of the first letter of the stored word is used by the System Control unit to select just one of the 14 available pairs of tracks on the Dictionary tape. One track of this pair contains the digital addresses of words that begin with the same letter as the target word; the other track contains voice recordings of the corresponding words. The search proceeds at high speed, with the digital addresses from the dictionary tape being compared, bit by bit, with the target address stored in the shift register of the Scan-



BLOCK DIAGRAM OF WORD-READING DEVICE

FIGURE 9
Operation of the Interim Word Reading Machine, showing functions performed by the component units and (separate) paths for data and instructions.

ner-Comparator unit. Synchronization is checked (or reestablished) by clock and framing pulses from a clock track.

An exact match between the two addresses means that the desired word has been found. Accordingly, the transport of the Dictionary tape is shifted abruptly from fast forward to slow reverse in order to transcribe the voice recording onto the quarter-inch tape of the Word Accumulator, with due attention to the duration appropriate for the audio version of the word. It now remains only to return the Dictionary tape to its home position at the middle of the tape and to initiate the reading of the next text word from the TTS tape. In practice, the entire cycle required, on the average, about 10 seconds to yield about one-third of a second of speech: i.e., the IWRM operated at about one-thirtieth of real time.

The Need to Spell; Specialized Vocabularies. What happens if an exact match is not found? Since the words on each track of the Dictionary tape are ordered by word length, the search for a word that is, say, five letters long need proceed no further than the first word that has six letters. Moreover, search time is further reduced because the words that are used most frequently, usually the shorter words, are examined first. Failure to find the target word means, in the simplest case, that each letter must be sounded out. An alternative is possible, one that would certainly be needed in a full-scale word-reading machine: Failure to match the address in the main Dictionary tape would initiate a second search in a track pair reserved for specialized vocabularies. It was planned that the IWRM would test the usefulness of this procedure.

The vocabulary of 6,000 words—later increased to 7,200—was chosen as a design compromise among several factors: complexity of the tape-handling equipment, cost of recording the Dictionary tape, and adequacy of the vocabulary as indicated by the frequency with which missing words would have to be

spelled. Some idea of the trading relation between vocabulary size and frequency of spelling can be had from these rather rough estimates: 50 percent spelling rate for a vocabulary of 100 words; 25 percent for 1,000; 10 percent for 3,000; 5 percent for 6,000; 1 percent for 15 to 20,000 words. (The number of different words in Webster's Collegiate Dictionary is about 60,000; more than 600,000 are claimed for Webster's New International Dictionary.) Thus, we expected that the IWRM would have to spell about one word in each twenty words of running text.

Instrumentation. The design of the IWRM was fairly conventional. A Friden paper tape reader was used to transmit the TTS characters directly to a relay decoding tree and transistorized shift register for temporary storage. The tape-transport mechanisms for both the Dictionary and the Word Accumulator were fast start-stop units that moved their tapes from bin to bin. The inch-wide Dictionary tape was searched for digital addresses at 60 inches per second, and its audio recording was read out and copied at 3.75 inches per second.

The Scanner-Comparator unit proved to be by far the most complex and expensive part of the entire reading machine. The circuit complexity was due in part to the dual requirement that the Scanner-Comparator serve in recording the tape initially, as well as later in finding and playing back the dictionary entries. The construction of circuitry of this kind would today be considered fairly trivial; indeed, the entire operation would probably be relegated to a microprocessor. But when the IWRM was built, the commercial modules typical of second generation computers were not yet available, and we had to build our own printed circuit cards (including even etching the cards). Likewise, both of the tape-transport mechanisms had to be built in the Laboratories' own shop.

Since the functions to be performed by the System Control unit depended on the detailed structure and function of all the other units, its design was deferred until those other units were built. In fact, the design was eventually executed in software for a small computer.

By mid-1958, the design constraints for the above components had been determined, and by mid-1959, all of the design and about two-thirds of the construction had been carried to completion. However, the fixed-price contract funds were exhausted by this time and, although the Laboratories eventually carried the development to completion at their own expense, progress was slow after mid-1959.

Demonstration of an Operating System. A functionally complete and operating system was demonstrated to the VA in December, 1965. The IWRM searched for the words of a sentence in a small trial dictionary, found the words, and assembled the recordings into a connected sentence on a word accumulator. The speech quality was acceptable. However, the IWRM was not then in deliverable form as a completed device, nor did the Dictionary tape contain the full 7200-word vocabulary (then on Language Master cards).

A decision to terminate the project at this point was made on the basis of a number of considerations: the most cogent were that the system was already technically obsolete and that the substantial amount of additional work needed to put it in final form and to record the dictionary tape would be largely wasted, since the same result could be obtained by computer simulation of the system (as, indeed, it was).

Compiled Speech by Computer Simulation of a Word Reading Machine. By 1969, the IWRM had been simulated on a medium sized computer. Some hardware peripherals had to be designed and built for this work, in particular a pulse code modulation (PCM) system for converting the analog speech wave into digital form; however, most of the effort went into programming the various operations. The system described below was largely created by one of the authors' colleagues, Dr. George Sholes.

With the 7200-word dictionary recorded on conventional digital magnetic tape, the process of generating a passage of compiled speech from a punched paper tape input is as follows: the punched paper tape (corresponding to about one typewritten page of text) is read into the computer and each word is assigned a number corresponding to its serial position in the text. Next, the digital magnetic tape is searched from beginning to end to find "matches" between words stored on it and words of the input text. Each record on the dictionary tape consists of a brief heading that contains the spelling of the word, followed by a much longer section that contains the digital version of the spoken word. The heading is compared with every word in the input text while the audio part of the record is being stored in core memory. If no match is found, then the dictionary tape continues to run and the next audio record is written over the last one; when a match is found, the audio part of the record is rewritten onto a disk file, in a sector numbered to correspond with the serial number of the word from the text. (Since this same word might appear several times in the text, the search is carried to the end of the text and the audio part is written into corresponding sectors for all other instances of the word.) Then the search of the dictionary is resumed.

In this way, the disk file comes to contain the audio counterpart of each text word in text order, except for those words of the text which were not matched by the dictionary tape. Such words are given a distinctive code and their spelling is entered into the disk file so the word can be spelled at the proper time (from letter recordings also contained in the disk file.) The final operation is to read the disk file serially and regenerate (and record) the speech using the PCM output system.

Paragraph-length texts were produced, using the digital word dictionary and punched paper tape input for the text. Speech quality was exactly comparable with that obtained by manual methods, except that it was free from the clicks between words that had sometimes marred the earlier recordings. In short, the IWRM then existed in computer-simulated form, and operated successfully.

Summary and Conclusions. The original engineering concepts for the hardware IWRM appear to have been sound and were in fact realized, although at a much

later date than had been planned and under circumstances that made it seem wise to terminate construction of the device at the stage of a demonstrated working system.

In retrospect, several factors contributed to this final outcome; perhaps the principal one was a failure to appreciate fully the complexity of the device. This led to negotiated funding under a fixed-price contract that was about half as much as was actually needed. The consequent lack of funds slowed the work. External events also played a hand. The period from 1957 to 1962 was one of extremely rapid technological advances, away from vacuum tube circuits to solid state electronics and to the development of cheap modular circuits for handling digital information. Thus, in June 1958 when the Scanner-Comparator unit was being designed, one could not have bought suitable printed circuit cards except at prohibitive prices; yet by the time the unit was built and working on the bench, modules were so plentiful and so inexpensive that it seemed foolish ever to have fabricated them at the Laboratories. Finally, computer methods were becoming so inexpensive and were so superior in flexibility that one would not then have considered building a hardware device. Indeed, the objectives of the contract were soon met completely by computer simulation, as the foregoing section relates.

The Evolution of Speech Synthesized by Rule

We knew, when the Laboratories program of research for the VA began in 1957, how to get reasonably intelligible speech from the Pattern Playback even when we did not have a real spectrogram to copy. We called this "synthesis-by-art" because it depended on long familiarity with painting the patterns that had been used in the search for the acoustic cues. Would it be possible to write down recipes, or rules, that would enable someone who lacked that experience to paint equally good patterns? What would be the underlying structure of such rules? And was enough known about the cues, in a reasonably quantitative way, to make the rule writing possible? These were the problems that faced Dr. Frances Ingemann when she joined the program late in 1956 to apply her linguistic skills to this task.

The central problem was one of units—how big should they be? Clearly, words were too big and there were too many of them. Words served well for compiled speech, but only because a human speaker knew how to generate large numbers of them. But for synthesis, one would need to have long and complicated rules for each word, hence thousands of such sets of rules for a usable dictionary.

Syllables would seem a better choice, or even half-syllables (formed by cutting at the middle of the vowel). Most of the work on cues had, in fact, been done with either CV or VC syllables; moreover, no more than a few hundred half-syllables would be needed for a rather good approximation to normal English.

The phoneme was another possible choice and, though much work had been done with syllables in searching for the acoustic cues, we had interpreted our findings as cues for the phonemes (with the tacit understanding that these phonemes were not to be found as separate and independent parts of the speech signal). Phonemes had the advantage that there were only about 40 of them for English, so the number of rules would be manageable. However, the cue description of a given phoneme was different for each different neighboring phoneme with which it might be paired, and this would require either very complicated rules for the individual phonemes or a second set of rules to deal with interrelationships. While this was not as simple a situation as one might desire—and there are other complications not yet mentioned—it seemed the most promising approach available and it made direct use of the research findings about cues. Certainly, that research had shown how futile it was to treat speech as if the underlying units could be shuffled around as moveable type is in printing.

Dr. Ingemann did find, though, that a phoneme-based rule system could be very considerably simplified by taking account of the subphonemic dimensions (features) according to which phonemes organize themselves into groups such as the stop consonants (according to manner of production) or the bilabial consonants (according to the place of production). Perhaps the best way to see the structure of the rules is to consider an example. Figure 10 shows the kinds of rules needed to synthesize the word "labs"—synthesize in the sense of creating a pattern for the Playback according to precise and explicit instructions. The two dimensional structure of the rules is clear from the upper half of the figure; thus, for each of the four phenomes there is a set of conditions (reading down the columns) that need to be realized simultaneously. Likewise for each of the four rows, the interrelations among neighbors are specified (implicitly) in terms of the formant loci.[f] The labels on the rows—manner, place, voice, and position—are familiar subdimensions from articulatory phonetics, and it is the decomposition of the rules that buys simplicity for the system. Thus, the specific phoneme specified by a column is the only common member of the various groups of phones for which manner, place, and voicing rules have been given. The actual rules for, say, manner of production are written for whole classes of phones and so there are only as many such rules as there are classes—not individual phonemes. The same is true for place, voicing, and position rules. Even though several rules must be used, the total number of rules can be substantially less than the number of phonemes.

At the end of 1957, Frances Ingemann had, in fact, written a recipe book for speech synthesis by rule (SSBR) which incorporated all that we then knew about the acoustic cues. It was sufficiently explicit for the synthesis rules to be used by anyone, and the resulting speech was, for the most part, fully intelligible, though woodenly machinelike. She presented a demonstration recording to a meeting of the Acoustical Society of

[f] Thus, in proceeding from consonant to vowel, the locus specifies that formants should begin at frequencies characteristic of that consonant, and then proceed within a specified time to the formant frequencies characteristic of the vowel. This defines the "transition" between the two phonemes.

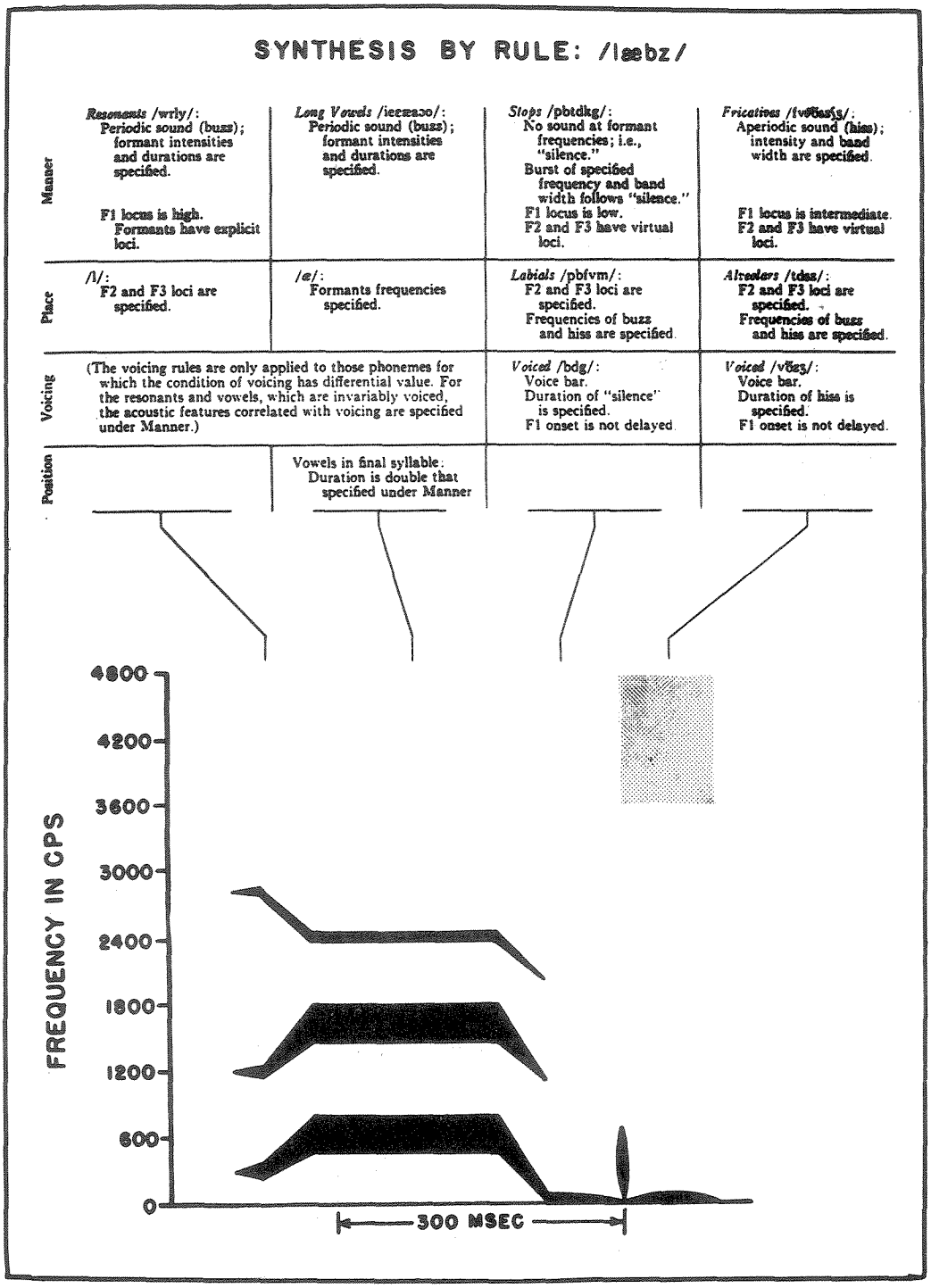


FIGURE 10
Table illustrating the rules for synthesizing the word "Labs", and the pattern derived therefrom for use with the Pattern Playback.

America (33) and later co-authored the definitive paper in the field, "Minimal Rules for Synthesizing Speech", that was read (by invitation) before the Acoustical Society of America by Alvin Liberman (41).

The Search for Naturalness. The initial success in formulating rules that would generate intelligible sentences from strings of phonemes had come quickly and easily, in part because it exploited almost a decade of background research. But revising the rules to make that speech sound reasonably natural, and a little more intelligible, was a slow and discouraging task, and it was nearly another decade before synthetic speech by rule showed much promise as an output for reading machines.

The difficulties were of several kinds: lack of knowledge, lack even of a clear definition of the problem, and lack of instruments that were adequate to the task. In the earlier work on cues for the sounds of speech, there were clear criteria for knowing when a significant variable was being manipulated and when an answer had been found. This was not true for naturalness and there was no real understanding of the relationship between acoustic variables and speech quality. It was not even clear how much of the blame for poor quality was inherent in the hardware synthesizers then being used and how much was due to the signals that controlled them. We knew, of course, that the Pattern Playback had a number of limitations that might well affect naturalness; the most obvious was the total lack of pitch modulation. There were other synthesizers of the formant-generator type that could manipulate voice pitch and they made very nice vowels, but they did not generate natural sounding speech. It was, therefore, a real milestone when John Holmes succeeded, after several months work, in synthesizing a single sentence that was literally indistinguishable from the voice recording with which he had started—thus proving that poor speech quality should not be blamed on the hardware.

Although limitations of knowledge and equipment were genuine difficulties, some of the faults of the speech synthesized by the original set of rules were so obvious that there was little doubt about what should be done to correct them. For one thing, the timing was all wrong, since all syllables were about the same length and gave the impression that the speech had been set to a metronome. This led us, and others, to study the relative durations in natural speech in order to write rules that would give our synthetic speech a more natural rhythm. Even with the Pattern Playback, it was quite possible by modulating vowel durations to make the stresses fall on the right words. It was less easy, but still possible, to write rules that would do this on the basis of the phonetic structure of the sentence (and the punctuation of the written text).

A related problem was how to synthesize unstressed syllables in such a way that they would be unobtrusive and yet not lose all character. This led to further work on the shifts in vowel formant frequencies that are a part of destressing.

Intonation was of course another aspect of the synthetic speech which could enhance—or destroy—its naturalness. A good deal was known, in a descriptive way, about changes in voice pitch during ongoing speech, but it was difficult to sort out the changes that were being used to signal stress from those that were linked to the syntax. Without this information, it was difficult to do much about rules for intonation, though it was quite clear that wrong intonation was a serious defect. Our work in this area depended initially on a synthesizer that used painted patterns—in much the same way as the Pattern Playback did—to control the output half of a vocoder, and thereby gain control over the pitch as well as the spectrum of the synthetic speech.

The Computer: A New Tool for Synthesis By Rule. Some progress was being made in our laboratory and elsewhere in dealing with these problems of naturalness but the pace was slow, in part because experimentation with hardware synthesizers was cumbersome. The situation began to change as computers became available. The rules for synthesis, once programmed, could then be used to generate many trial texts and so quickly show where difficulties might lie. The Bell Telephone Laboratories took the lead in this development and, in 1961, Kelly and Gerstman described and demonstrated "An Artificial Talker Driven from a Phonetic Input" at a meeting of the Acoustical Society of America (36). This was a tour de force combination of computing skills that were then being developed at BTL with the knowledge about acoustic cues that Gerstman had gained from his participation in the research at Haskins Laboratories. In 1964 Holmes, Mattingly, and Shearme at the Joint Speech Research Unit (JSRU) in England developed a mechanized system for generating synthetic speech by rule (31). By 1966, Haskins Laboratories had acquired its own computing facility and had built a computer-controlled formant synthesizer. Also in 1966, Ignatius Mattingly joined the Haskins staff and undertook (as a thesis project) to program this equipment to generate spoken American English by rule. He was in a position to draw on his earlier work at JSRU as well as the work at Haskins Laboratories, and by 1968 he had completed his thesis project (43).

It is interesting to note how dramatic was the change that computer facilities made possible. The following quotation is from a conference report that one of the authors of this paper gave on "High-Performance Reading Machines for the Blind" at St. Dunstan's, London, in June 1966 (58). In commenting on the merits and limitations of synthetic speech (which then seemed less promising than some form of compiled speech), the paper concludes, "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, 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."

By 1968 pessimism about the prospects for using synthetic speech in a reading machine had changed to optimism, largely on the basis of Mattingly's successful undertaking. True, there was much yet to be done and

it was still not clear whether a reading service for the blind, if it were to be established within the next few years, should use compiled speech or the new synthetic speech. Definitive tests with potential users had still to be made. But it was clear that synthetic speech must be given serious consideration.

Mattingly's SSBR program used, as input, sequences of phonetically-spelled words interspersed with stress and juncture symbols. Three levels of stress were recognized (high stress, mid stress, and no stress); they were reproduced in the output speech as increases in syllable pitch, loudness and duration. The juncture symbols that marked phrase boundaries indicated the pitch contours that the computer should use in synthesizing that phrase. A group of acoustic-phonetic rules, expressed in tabular form and capable of alteration by an experienced user, were responsible for carrying out a conversion of the input string into a set of 15 synthesis control parameters. The rules specified the trajectories that the control parameters should take to produce consonants and vowels and, in addition, the overlapping effects produced by coarticulation in fluent speech. These control parameters dynamically manipulated the formant-type speech synthesizer. In addition, Mattingly developed an executive program that made it relatively easy to revise and/or supplement the rules.

Abandonment of Speechlike Output and Re-Formed Speech

We were not alone in clinging to the hope that it would be possible to bypass the very considerable technical problems in making a high-performance reading machine based on letter recognition and the use of spoken English. How much simpler it would be if only the device could find, in the shapes of letters, enough information to generate acceptable sounds! We were convinced that these sounds had to be "speechlike" in the sense that they could be pronounced easily by a speaker of English, though the result might well be a jabberwocky language. Our early experiments with one such language, WUHZI, had convinced us that it could be learned fairly easily.

The hidden difficulty, and the one that eventually led us to abandon the whole idea, seems simple in retrospect. If one considers that very many commonly used words differ from each other by only a single letter, then it is clear that the shapes of these words will not differ very much either. Hence, one would need quite detailed information about shape features—almost as many bits of information as would be required for complete recognition of the letters. To be sure, some bits could be saved by using a limited inventory of phonemes in synthesizing the artificial language and one might take advantage of regularities in the way words are constructed; even so, a rough calculation suggests that one could expect no more than a 20–25 percent reduction in the information that would have to be extracted from the word shapes.

Re-Formed speech, essentially a hybrid between compiled and synthetic speech, was a child of the mid-sixties—when the compiled speech seemed feasible but not very good and synthetic speech promised to be fairly good but seemed not very feasible. The main difficulty with compiled speech was that the voice-recorded words were not flexible, as they needed to be to fit gracefully into sentences. The trouble with synthetic speech was that too much remained to be learned about how to build a speech signal from the ground up. However, we did know, from work on bandwidth compression devices, how to analyze spoken speech into the formant tracks that correspond roughly to paintings for the Pattern Playback. So why not store these formant tracks (from spoken words) instead of storing waveforms? We could then compile these control parameters for the words into sentences and generate ongoing speech with a formant-type synthesizer. All of the component steps were known to work, at least reasonably well, and there were advantages: most importantly, the stress and intonation of the individual words could be manipulated to make them fit the requirements of the sentence; also, the control signals could be stored much more compactly than the waveforms (by a ratio of about 1 to 20), and this would permit digital storage and ready adaptability to computer control of the entire process.

Actually, we did quite a little work on this kind of speech, and generated just enough of it to demonstrate that the process would work and that the speech would be fairly good. But the breakthrough on synthetic speech came at about this time, so work on the compromise method was dropped. In retrospect, this was almost certainly the correct decision, though there are limited applications for which synthesis from stored control signals has real utility (54).

Comparison of Compiled Speech and Speech Synthesized by Rule

By the time Haskins Laboratories had completed its move from New York to New Haven (mid-1970), the output options for a reading maching had been reduced to compiled speech and speech synthesized by rule. We knew how to generate both, but it was not clear which would be the better choice. Comparative trials of compiled speech and speech synthesized by rule were run, using tape recordings of various texts. The twofold purpose of the proposed tests was to learn more about blind persons' expectations concerning reading rates, subject materials, voice quality of the machine speech, overall tolerance of the two types of audible output—and whatever else might be important to them. Conveniently, and very cooperatively, Mr. George Gillispie, Mr. William Kingsley, and their associates at the VA Eastern Blind Rehabilitation Center in West Haven, Connecticut, agreed to seek out volunteers among the blind veterans at their facility to serve as listeners in these field trials.

For reasons of simplicity, the tests were run at the VA Center. A total of 11 subjects participated—all male and most of them in their twenties. There were eight hour-long tests of 27 different texts, each presented to a minimum of two listeners and some to as many as four. The conditions were somewhat informal; the tests took place in any available room with any available volunteers (although the subjects were usually scheduled to avoid conflict with the Center's own programs). The investigator began each session with a brief introduction

to the reading machine research and stressed to the listeners that there were no right or wrong answers; that, in fact, no answers as such were needed—only candid comments on anything about the tapes that they cared to mention. It was made clear that the purpose of the tests was to improve the reading machine output. All the subjects took the task very seriously.

Several variables were manipulated in presenting the tapes:

1. Form of machine speech (compiled or synthetic);
2. Speech rate (ten rates within a 70 to 225 word per minute range were used.);
3. Rate manipulation (by simple speed up or by Time-compressed Speech. The Compiled Speech texts were processed by the Center for Rate Controlled Recordings, University of Louisville, Louisville, Kentucky, where they were time-compressed by 60, 65, 70, and 75 percent.);
4. Text (author and topic, i.e., Dickens, Oliver Twist; Steinbeck, Travels with Charley; Pierce, Waves and Messages; sports articles from newspapers; several Saroyan stories.); and
5. Amount of spelling (applicable to compiled speech only).

At the end of each session, the reactions of the blind listeners were collected and summarized. For Compiled Speech, the preferred rates varied with the topic and the author's style. Also, certain topics involved more spelled words than others. (Spelling was deplored by all listeners.) When the speech was time compressed, the preferred rates were in the 159–175 words-per-minute range (i.e., normal speaking rates). However, monologues and dialogues were not enjoyed in this form of speech. The length of speech sample had an effect on the acceptability of the output; for example, half a minute was inadequate for an evaluation (if the topic of the text was unknown and if the tape was begun at a random location in the text), but a minimum of one minute seemed to be sufficient to make an appraisal if the rate was within a reasonable range. The overall evaluation of Compiled Speech was that it was acceptable at some rates in either time-compressed or capstan-speeded form—but was not enjoyable. Spelling was its worst feature. The temporal irregularities were annoying. Listeners doubted that such speech could be tolerated (with or without spelling) over extended periods.

Synthetic Speech (in which no spellings appeared) was quite easily understood with exposures as brief as half a minute and at rates ranging from about 135 to 225 words per minute—that is, from slow to fast speaking rates. Listeners' comments dealt chiefly with the subject matter of the texts, indicating that intelligibility and prosody were acceptable, or at least not distracting. The one aspect that was faulted was what the listeners called its "accent."

Comparisons of these early appraisals of Compiled Speech vs. Synthetic Speech indicated, therefore, that Compiled Speech was effectively rejected and Synthetic Speech was quite enthusiastically accepted.

The Evaluation of Speech Synthesized by Rule

Prospects for Reading Machine Applications. This phase of the Laboratories' reading machine research began in 1970 when the results of comparative tests of compiled speech and speech synthesized by rule from a phonetic input showed the latter to be clearly superior. As has been noted earlier, the main objective of the Laboratories' research program was limited to the development of an acoustic output that would be suitable for use in a reading machine for the blind. The results obtained with SSBR in listening tests had made it clear that this goal was very close at hand. Moreover, in conjunction with research aimed at improving the overall performance of the synthesis method, it was apparent that some effort should now be made to obtain equipment and to prepare software to produce phonetic texts for speech synthesis by rule (SSBR) input directly from the printed page. Not only would such equipment and software be needed in any complete reading machine, but user acceptance tests would almost certainly require quantities of "spoken texts" that could only be generated by a fully automated system.

Thirteen years earlier, at the outset of the VA program, although optical character recognition had been in its infancy it seemed safe to assume that commercial needs for OCR equipment would soon multiply and ensure the rapid development of low-cost multifont optical readers. However, by 1970 it had become apparent that the OCR developments, still essential to the success of reading machines, had not proceeded at the pace expected. While in part this delay may have been due to an underestimation of the difficulty of developing an economically viable multifont print recognizer, it was also in large part due to the unanticipated direction that the commercial demand for character recognition equipment had taken. Over the preceding decade, the need for very fast and accurate numeral-recognition systems designed to read magnetic or optical characters—usually printed but sometimes handwritten—had continued to grow at a rapid pace spurred by demand from the banking and credit card industries. In the broader commercial sector, the development of automated stock and inventory control systems tended to call for the automatic recognition of a larger set of printed characters including alphabetics. However, a pervasive difficulty of all these applications is that accuracy must be maintained for the enlarged character set in environments that typically produce poor print quality and crumpled documents. As a practical compromise, special typefaces were designed specifically to make it possible for OCR machines to function with typewritten materials composed and handled in offices and warehouses. Machines designed to recognize these special typefaces cost in the region of $50,000 and were unable to function satisfactorily on the wide variety of fonts found in newspapers and books. On the fringes of the OCR industry in the early 70's there were, however, a few multifont readers that had been designed and built for military intelligence and other specialized applications in the publishing and information retrieval fields. These more-versatile machines all shared the trait of being

about an order of magnitude higher in cost (probably because the development costs were high, electronic components cost more than they do today, and small market demand did not allow these costs to be spread over a large number of units).

Therefore OCR equipment with the versatility needed for application in a reading machine did exist but was not really available. Meanwhile, yet another problem lay in the path between the printed page and the generation of a speech output—finding a suitable algorithm for converting the printed alphabet into phonetic symbols. Here the problem had either a simple solution that imposed practical limits on the size of the vocabulary, or a more complicated and, at the time, unproven solution which promised fewer restrictions on vocabulary size. The former solution was represented by the straightforward dictionary look-up procedure which, for an unrestrained selection of text, would require that the phonetic equivalents of some 500,000 words be stored. The latter solution was represented by a procedure that derives the phonetic form of any English word by analytical means. Work on such an algorithm was underway at MIT by a group headed by Jonathan Allen. This effort led eventually to a complete (computer-based) text-to-speech system called MITalk (4,5). The Allen method involved the decomposition of words into affixes, prefixes, and root forms, then finding their phonetic equivalents and assembling the phonetic spelling. Less storage space seemed likely to be required, despite the need to store the root forms and an exception list. Estimates were that the roughly 20,000 items that had to be stored could be used to generate an English vocabulary many times that size.

Considering the state of development of both OCR equipment and orthographic-to-phonetic conversion capabilities, there appeared in 1970 to be clear grounds for optimism about the practical nature of the task of building a reading machine. But it was also clear that the building of a reading machine would be expensive (at least initially) and that it would be bulky—particularly in view of the fact that the MIT work was at that time unfinished and that letter-to-phoneme recoding by direct dictionary look-up appeared to be the better choice for a prototype machine. Thus, our assessment of the situation during this period led us to the conclusion that the first reading machine would probably be installed either in a VA hospital, on a college campus, or in a large library associated with a dense population center where the level of demand would justify the costs of the equipment and its operation.

Initial Studies of SSBR Performance. With the issue of whether a reading machine could be built no longer in much dispute, the question of whether it would meet the human factors requirements began to dominate. Speech synthesized by rule had been shown (in short passages) to be sufficiently like natural speech to be understood even by groups of naive listeners. Moreover, it was known that comprehension improved with a little listening practice. But exactly how intelligible was the synthetic speech when compared with natural speech? Would listeners tolerate the imperfections of synthetic speech when they were obliged to listen to long passages and recall the content? These were questions that clearly needed to be asked in order to evaluate whether the construction of a pilot reading machine center based on an urban college campus or library could be economically justified. The group at Haskins Laboratories, therefore, turned its attention to a study of the man-machine interface.

An exploratory study of the speech-acceptability issue was carried out with the help of blind students at the University of Connecticut. Ten recorded passages totaling 2.5 hours of listening time were drawn from text books in psychology and psychiatry as well as works of ancient and modern literature. The style of these texts ranged from simple prose to more elaborate syntactic constructions demanding the use of memory for embedded clauses, and requiring analytical thought to extract the content. After listening to SSBR recordings of these passages, the blind students offered their comments, which contained broad agreement on five points:

1. The simple prose was intelligible but the subject matter of the more complicated material was difficult to understand;
2. The stress and intonation aspects of the speech were impressive and helpful;
3. The "nasal" quality of the synthetic speech was unpleasant;
4. The rate of presentation was too slow,[g]
5. Long and often unfamiliar polysyllabic words were recognized with ease, while monosyllables embedded in sequences of other short words were among the items that were most often missed.

Thus, our preliminary probe into listener acceptance pointed to two main areas of concern: (i) The poorer intelligibility of monosyllabic words compared with multisyllabics, and (ii) the interaction of speech intelligibility with the complexity of the subject matter being read. More information was needed about these topics. However, new techniques of inquiry had to be found because the methods of the preliminary study contained two serious weaknesses. The first was that the data were wholly subjective. Thus, while the listeners' comments clearly indicated that synthetic speech was more difficult to understand than natural speech, they did not indicate how much more difficult it was, or provide a quantitative measure of the listeners' performance. Such figures of merit for synthetic speech compared with natural speech would also be needed in gauging the progress made with future improved versions of synthe-

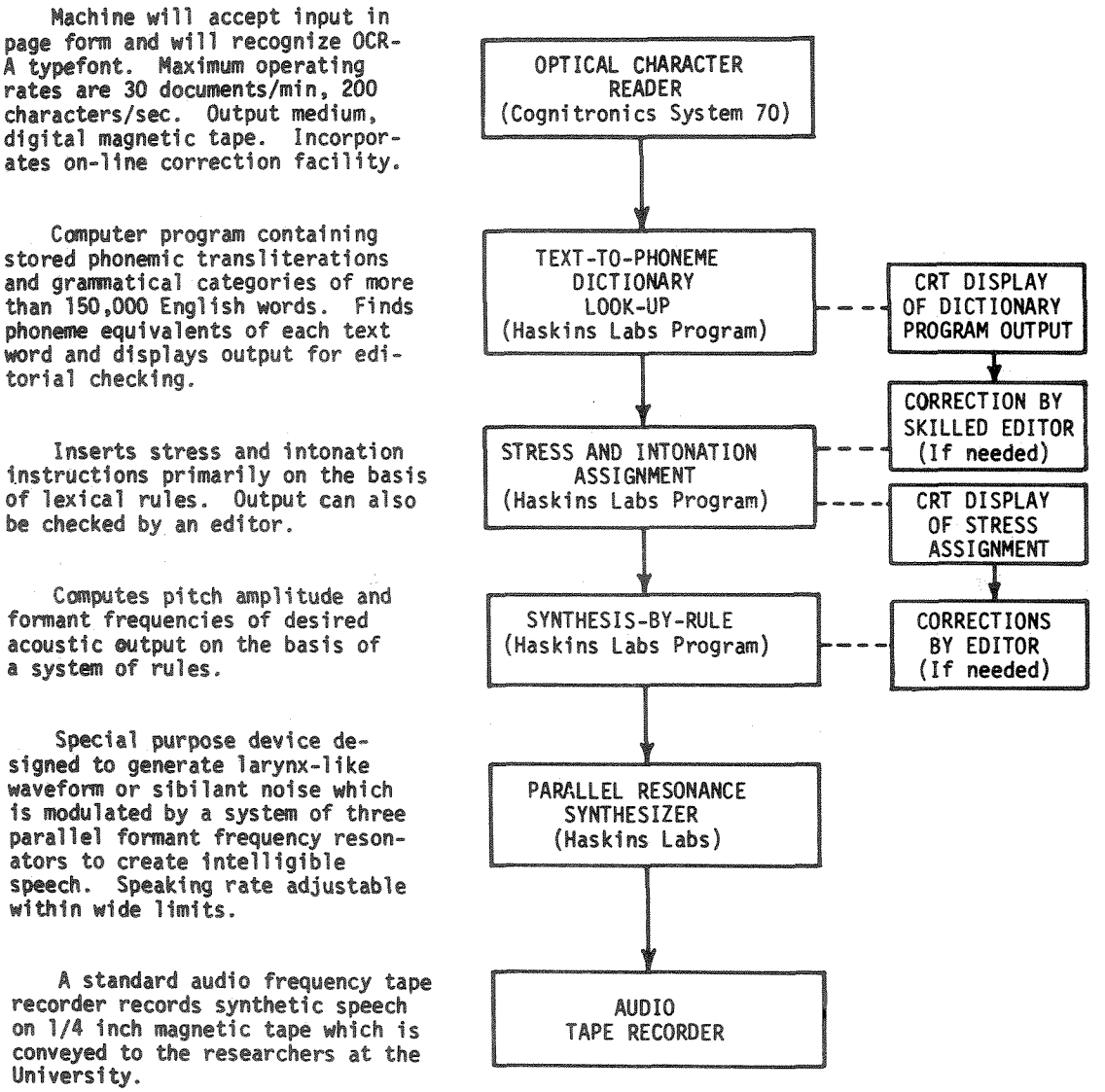
[g] The speaking rates varied from 101 to 156 words/minute. The latter is within the norm for human speech but the long silences (2-8 sec) between some sentences in these early recordings made the overall rate seem slow. These unnecessary silences were eliminated in later recordings.

sized speech. The second weakness lay in the volume of reading matter employed in the study. Owing to the fact that the test materials had to be typed in phonetic script by hand, the procedure was sufficiently slow that the volume of reading matter that could be supplied was too small to permit an investigation of practice or fatigue effects.

Development of a Prototype Reading Machine. To overcome the shortcomings of these preliminary studies, we sought to assemble the components of a laboratory prototype reading machine that would produce substantial amounts of synthetic speech more or less automatically. Figure 11 provides a diagram of the Laboratories' text-to-speech prototype processor. An OCR system (purchased with money granted to the Laboratories by The Seeing Eye, Inc.) served as the primary input stage of the text processor. This OCR system, manufactured by the Cognitronics Corporation, read upper-and-lower-case typescript in an OCR-A typefont that could be generated on a regular IBM "golfball" typewriter. Thus, although special input text was needed, it could be prepared by ordinary typists. Moreover, these typists could do their work at locations remote from the Laboratories and at rates that were much faster than those achieved by even the most skillful phonetic typists. In addition, the use of typewritten texts saved computer time because, unlike the preparation of phonetic texts, the typing could be done independently of the computer.

The typed page was then "read" by the OCR device, giving a sequence of machine-readable alphabetic char-

FIGURE 11
Operation of the Prototype Reading Machine. The system was employed to generate substantial amounts of speech synthesized by rule for use by students and in evaluation studies.



acters. These were converted, at stage two of the process, into a sequence of phonetic symbols and stress marks by direct dictionary lookup, using a phonetic dictionary that was made available to us by the Speech Communications Research Laboratory, Santa Barbara, Calif., and installed in the Laboratories' computer by our colleague, Dr. George Sholes. This dictionary contained the phonetic equivalents of more than 150,000 English words with their syllabic stresses marked according to the three-level system employed by the SSBR program. A CRT monitor provided the operator with an opportunity to examine the output of the lookup procedure. Any words not found in the dictionary were displayed on the screen spelled in orthographic form so that the operator could intervene and supply the missing phonetic equivalents. Following dictionary lookup, a third stage was responsible for modifying the lexically-defined stress and inserting intonation marks on the basis of a system of rules applied to the punctuation of the original typescript. At the fourth stage of the process, the phonetic string became the input to the SSBR computer program and was converted into fifteen parallel streams of digital control signals for the specially built terminal analogue speech synthesizer that was mentioned earlier. The speech was recorded on magnetic tape for use in evaluation trials.

In short, this prototype reading machine would do, in the laboratory environment, everything that a "real" reading machine would do in a library environment, with one exception, namely, that it required typewritten material as its input. This was a limitation that could readily have been eliminated, though at rental costs for a multifont OCR machine which could not be justified for the experimental use we envisaged, which was to supply recorded tapes for the blind student subjects at the University of Connecticut.

Modified Rhyme Test. In parallel with work on a prototype reading machine, work was started on measuring the intelligibility of synthetic monosyllables (49). The first experiment employed a version of Fairbanks' Rhyme Test (21) which is known as the Modified Rhyme Test, or MRT (32).

The MRT involved the use of 300 monosyllabic words, grouped together into batches of six which rhymed with one another. The test was prepared in both synthetic and natural speech versions and was administered in closed form. Listeners were provided with typewritten lists of the rhyming words, and their task was to listen to one presented word (selected at random) from each six-word group shown and to identify it by circling in pencil the item which most closely resembled the word that they had heard. Thirty sighted students from the University of Connecticut were engaged as listeners. The overall intelligibility scores were found to be 92.5 percent for synthetic speech and 97.3 percent for natural speech—the difference indicating the margin for improvement. The latter figure agreed quite well with data obtained by other workers on natural speech. Word initial /v/ and final /r/ in particular—as well as the labial, labiodental, and dental fricatives in any position—were the least intelligible phones.

The MRT was a useful test in that it provided for the first time a measure of comparative performance for synthetic speech with respect to an ideal level—natural speech. However, the test itself proved to have a number of deficiencies that revealed themselves only after the results had been analyzed. For example, the extrapolation of many of the results to normal English speech is difficult because of intrinsic limitations of the MRT test itself: The individual consonants do not appear an equal number of times, nor in all vowel environments, nor in an appropriate balance of initial and final positions. Thus, the infrequent occurrence of some phones, combined with the fact that the response data exhibited a marked learning effect, might have contributed to low intelligibility scores for those phones. Moreover, the fact that the words were presented in isolation made the speech unnatural. Therefore, we sought to devise a new test in which the phonetic constituents would appear in varied environments with relative frequencies that were more similar to those found in English.

Test Results with Nonsense Sentences were obtained from a testing procedure designed to meet the above objectives. The test obliged the listeners to recall words placed in sentences that were syntactically normal but meaningless. It was dubbed the Syntactically Normal Sentence Test (SNST).

The test employed 126 nouns, 63 adjectives, and 63 past-tense verbs—all monosyllables selected from the first 2000 most frequently used words in English (50). Words from each of these categories were randomly selected to create 200 meaningless sentences of the form "The (adjective) (noun) (verb) the (noun)." These sentences were recorded in both naturally-spoken and synthesized speech as groups of 50 sentences, with a 10-second interval between the sentences. During this interval, the 32 sighted listeners were required to write down the sentence they had just heard, using ordinary English spelling. Because the test was open in form and the sentences lacked semantic context cues, the task of transcribing them proved to be considerably more difficult than responding to the MRT, even though the naturally spoken sentences were properly articulated and the synthetic sentences had coarticulation built into them.

The response errors were analyzed into two main classes:

1. Phoneme errors, which could be substitutions of vowels or consonants for other phonemes, (e.g., "fat" for "sat," "sat" for "sad," etc.); or insertions of one or two phonemes in an otherwise correctly reported word (e.g., "paved" for "paid"); or deletions, which are the omission of vowels or consonants in otherwise correctly reported words.

(2.) Word errors, which could be words left unreported (i.e., omitted words) and transpositions, or words which were correctly identified but in the wrong position within the nonsense sentence. Word location within the sentence was also examined as a possible factor in the number of errors made.

We will pass over the detailed results of our analysis

and remark on two general but important observations. First, the results from the SNST demonstrated that the task of recalling sentences in which the words are coarticulated (but lack any semantic content) provides quite a sensitive test of synthetic speech performance. Second, the lowest number of recall errors on synthetic speech was made on adjectives (occupying the initial test word position) whereas the highest number of errors was made on nouns (those in the second-word position). This error pattern contrasted markedly with that found in natural speech where the verb (third test word position) proved to be the most misreported word. The reasons for this observation were not discovered by further analysis of the data, but the errors indicated the existence of a trading effect between memory load (known to vary with serial position) and the extra attention (or cognitive effort) needed to identify synthetic speech sounds.

In summary, a comparison of the results of the MRT and the SNST showed that the margin of difference between listening performance for synthetic speech and natural speech increased significantly for the more demanding SNST. The average error rate for natural speech in the SNST was about 5 percent compared with 3 percent in the MRT, while for synthetic speech it was 22 percent on the sentence test compared with 8 percent on the isolated MRT words. It must be noted, however, that the figures for the SNST include errors of all kinds ranging from words totally omitted to minor phonetic errors that may well have been corrected had the words appeared in meaningful contexts. Also, the reporting requirements were different: the MRT used a closed response set while the SNST demanded open responses. All of these considerations would lead one to expect higher error rates for the sentence test than for the isolated word test. The important point is that, as the task gets harder, the errors increase at a faster rate for synthetic speech than for natural speech. Thus, the results demonstrated both the sensitivity of the testing procedure and the need to focus further attention on the improvement of methods of synthesis.

Studies of the Comprehensibility of Synthetic Speech. While analytical studies typified by the MRT and the SNST provided useful information about synthetic speech on a microscopic level, there was an evident need to examine it on the macroscopic scale, i.e., to ask about overall performance on longer passages. It was already apparent that synthesized speech was sufficiently intelligible to enable information to be conveyed from the printed page to an untrained listener at speeds in excess of those offered by any existing reading aids for the blind. An exploration of user acceptability and performance issues was fully warranted.

The first plan for a field evaluation of synthetic speech using the Laboratories' prototype reading system was outlined in a paper by Nye, Hankins, Rand, Mattingly, and Cooper (48) published in 1973 and given in detail in a proposal submitted to the U.S. Office of Education, Bureau of Education for the Handicapped (BEH). The plan called for a combined effort by the University of Connecticut and Haskins Laboratories to provide a pilot reading machine service to a group of about 20 blind students at the University of Connecticut. Texts required by the blind students in their regular courses were to be prepared in typewritten form and converted into synthetic speech using the Haskins Prototype Reader; the texts were also to be converted into page-embossed Braille at the University of Connecticut Computation Center using the MIT DOTSYS III Braille Translation program. The goals of the plan were to obtain data on the usefulness of such a service to blind students, and on the usefulness of Braille versus synthetic-speech materials. This was to have been achieved by determining how much actual use was being made of the services and the relative proportion of the demand for synthetic speech and Braille. The proposal was approved, but because at that time there were budget uncertainties for many Federal agencies, the promised funding was repeatedly delayed until the BEH then administratively eliminated the project. The opportunity to carry out the plan was lost.

Meanwhile, with support from the VA, studies of listening comprehension with synthetic speech from the Prototype Reader were continuing on a more modest scale (51). The testing technique employed, as a measure of comprehensibility, the time taken to answer questions on the contents of synthesized and naturally spoken texts.

Two equally difficult passages of text were selected from a standardized reading test, each approximately 12 minutes in duration. One text was recorded, either in synthetic speech from a then-current version of the SSBR algorithm or in speech from an older synthesis program, while the other text was recorded in natural speech. The synthesized speech was generated either from a hand-edited phonetic script or from a phonetic text derived automatically (i.e., without editorial intervention) from orthographic input. After a single listening to one of the texts, a multiple-choice 14-item questionnaire was administered to each listener, and the time taken to provide as many answers as the listener could recall was noted. The listeners were then allowed to replay all or parts of the text as many times as was necessary to allow them to fully complete the questionnaire. This additional time was also noted.

The results showed that there were no significant differences between synthetic and natural speech as to the aggregate times taken to answer questions after hearing the passages for the first time. However, the listeners did take a significant 1.75 minutes longer to answer the remaining questions relating to synthetic speech passages during the second listening opportunity. The results obtained with different synthesis algorithms indicated that listeners performed somewhat better with the newer SSBR algorithm than with its predecessor, and that their performances with the hand-edited text produced only a slight improvement over that produced entirely by machine.

In conjunction with that comprehension study, a paired-comparison preference test was run in which each listener selected his preferred form of synthetic speech from all possible contrasting pairs. The test re-

sults showed that the various speech outputs ranked in the same order on the preference scale as they had in the comprehension study. This suggested that there is a strong relationship between listener preference and listener performance and, therefore, the greater the extent to which the speech can be made to sound natural the greater is the gain to be expected in listener performance.

The same comprehension test was used on a later occasion (14) to contrast performance on easy versus difficult texts. Two new texts of greater difficulty were chosen in addition to the two original (easy) texts. The two additional passages covered technical subject matter from the fields of anthropology and geology. The two were also of roughly equal difficulty. Recordings were made of each text "spoken" either in synthetic speech or by a human speaker at the same rate of delivery. The text durations ranged from 12–14 min. Timing observations obtained while the listeners answered the questionnaire showed that on average they required 7.5 min for human speech and 11.7 min for synthetic speech. As expected, the answering times for both natural speech and synthetic speech increased with text difficulty, and, more significantly, the differences in time for natural and synthetic speech increased with text difficulty. Thus, the results confirmed the impressions of some of the early listeners to synthetic speech, namely, that the difficulty of understanding the content of a passage of text does increase more rapidly with the complexity of that content when synthetic speech replaces natural speech.

A Pilot Reading Machine Service to Blind Veterans

The Laboratories' contact with staff at the VA Eastern Blind Rehabilitation Center at West Haven, Connecticut, was reestablished for another study of listener reactions to computer-generated speech. On this occasion, at the suggestion of veterans in residence at the Center, the daily columns of Ann Landers were converted into synthetic speech, recorded and sent to West Haven for listening and responses.

The original texts were obtained from the local newspaper publisher in the form of Teletypesetter tapes and read into the Laboratories' computer with a specially modified reader. However, variations in tape conformation introduced by the different machines that punched them caused numerous errors and subsequent delays while corrections were made. As a result, only 1.5 hours or so of synthetic speech were generated during the project—less than had been anticipated. Nevertheless, the project was valuable for two reasons. First, it provided an opportunity to evaluate duration as a supplemental cue for stress. Second, the informal style of Ann Landers' column involved a number of syntactic structures that the stress assignment algorithm could not ade- quately handle. Thus, in some cases the sentences were ambiguous unless the main stress was applied to just the right word, so corrections had to be made by hand. In other cases, typographical devices such as boldface printing were used instead of punctuation. This also required intervention since the dictionary lookup program made no distinction between typefaces and had to depend entirely on formal punctuation to assign stress and intonation. Performance was therefore liable to be erratic when the Prototype Reading Machine was operating in automatic mode.

William De l'Aune, Ph.D., and the research staff of the Blindness Center conducted the listening sessions in an informal atmosphere. However, despite the best efforts of the VA staff, the test procedures did not gain the wholehearted cooperation of those patients who were in residence at the time. The patients seemed reticent, possibly because they were uncertain as to whether their own intellectual abilities, rather than the performances of the speech passages, were really what was being examined. Consequently, they showed a distinct preference for making general comments about the quality of the speech rather than answering questions that would indicate how much they had understood. The results were, for these reasons, somewhat disappointing.

Improvements in Speech Synthesis by Rule

The initial development of a new SSBR program was perhaps the most important work performed in the final years of the Laboratories' VA-supported research. This program made a significant departure from principles embodied in the earlier program by abandoning the use of a hardware synthesizer for final speech output and by placing greater emphasis on the syllable as the unit of production.

Although the practical advantages of real-time synthesis were highly valued during much of the earlier work, the difficulty of modifying the hardware (whose speed of response made real-time synthesis possible) demonstrated its inflexibility for research purposes—particularly when the drive to improve speech quality made the need for synthesizer adjustments more acute. Therefore, in later work, algorithms similar to those employed by Klatt (37) were employed in a software synthesizer programmed in FORTRAN on the Laboratories' PDP-11/45 and VAX computers to simulate the sound generators and resonators of the original hardware. The chief advantage of a software synthesizer is that the components can be easily rearranged so that any desired synthesizer structure can be assembled. This flexibility allows the experimenter, within minutes, to make design modifications that would take many hours, were they to be attempted in hardware. There is a penalty, however, in generating the speech: a software synthesizer introduces an unavoidable delay of several seconds while the program computes the speech waveform.

The present SSBR program (also written in FORTRAN) is called SYLSYN (for Syllable Synthesis). Organized in terms of phonetic syllables, the program provides a more direct representation of coarticulatory effects in their spectral and temporal aspects than was possible with the earlier SSBR programs, which were based on phonetic segments. The input to the program is a transcription of syllable features. The rules are stored in a disc file which is accessed by a special subroutine during synthesis. These rules relate the feature transcription to a specification, as a function of time, of each of

the various influences that shape the syllable. In conjunction with target values specified in the rules, these influence functions are used to determine the parameter values of the software synthesizer which, in turn, produces the digital waveform that is converted into an audio signal. So, by editing the rules file, the user can modify not only the rules for synthesis but also the characteristics of the synthesizer itself.

SUBSEQUENT DEVELOPMENTS in the EVOLUTION of READING MACHINES

The research project on Audio Outputs of Reading Machines for the Blind at Haskins Laboratories formally came to an end in September, 1978, while work on completing the new SYLSYN program and other related research was still underway. The end to the project was the consequence of a policy decision made by the VA to withdraw its support of further research in this area. The VA had funded a wide variety of short- and long-range reading machine research projects in different institutions over a period of more than 20 years. Having begun to fund research on the development of a speech output at a time when the building of a talking machine was a highly speculative venture, the VA had been consistent in its concern for the endeavor by promoting conferences and the publication of results. By 1978, however, those who had followed recent developments could hardly have regarded the VA's withdrawal of research support with surprise and, at the Laboratories, the news was not entirely unexpected.

Starting in the early 1970's, several technical developments and legislative enactments of importance to the blind and other handicapped persons combined to create a climate of opportunity for entrepreneurs interested in providing devices and services for the disabled. A very few years after the development of Mattingly's successful SSBR program using an input of phonetic symbols and intonation marks, a synthesizer requiring similiar input, implemented in compact hardware form, was offered commercially by the Federal Screw Works with the name of VOTRAX. At about the same time, Telesensory Systems, Inc., with Federal assistance, made its first successful entry into the marketplace with a reading aid for the blind that used a tactile output. The supply of such products and the effort to develop them received an additional impetus from the Rehabilitation Act of 1973 (which was to be further enlarged by major additions enacted into law in 1978). Then, in 1974, Kurzweil Computer Products, Inc., began a vigorous effort to marshal a combination of Federal and private support for the development of a personal reading machine based on an optical recognizer (recently built by that company), the VOTRAX synthesizer, and existing knowledge about speech synthesis by rule. Finally, the technical trends of the 70's towards sharply lower costs for integrated electronic circuitry of steadily escalating complexity, culminating by 1976 in the ready availability of microprocessors, fueled an atmosphere of rising technical expectations among the handicapped as well as the desire of engineers to meet those expectations.

Thus, it was easy to foresee the likelihood of a swing away from research and toward an effort to apply the available technology and existing knowledge that research efforts over the years had accumulated. Whether this knowledge will prove sufficient to permit current reading machines to find a significant number of useful applications is still unknown.

What can be stated with assurance, however, is that the problem of machine-to-man communication as encountered by the blind reader is still far from being completely solved. Despite the great advances that have been made since the invention of sound producing reading machines at the beginning of this century, the intelligibility and comprehensibility of the speech now being generated is still in need of further improvement. All speech synthesized by rule from text, whether produced in well equipped laboratories or produced by commercially available reading machines, is unmistakably unnatural. Its articulation is imprecise and its intonation and syllabic tempo are faulty. Subject matter is more difficult to understand when spoken synthetically than it would be if spoken naturally. With much still to be done, the research of the Laboratories into synthetic speech is continuing—currently with support provided by the National Science Foundation. With this support we hope to continue to make contributions that will benefit the blind reader.

THE READING MACHINE PROBLEM IN RETROSPECT

The case history we have recounted spans nearly four decades and draws upon the experience of almost as many earlier decades. The account deals not only with events over this span of years but also with changing ideas about the nature of the reading machine problem. When Haskins Laboratories first encountered that problem in the mid-1940s, the brilliant technical achievements of World War II seemed to offer the early prospect of a personal and portable reading machine. But many facets of the problem, and of its solution, were not at all foreseen. It is only now, 40 years later, that this expectation is nearing fulfillment.

One thing not foreseen was the inability of listeners to cope with the arbitrary letter-by-letter sounds that could be produced by simple mechanisms. An aspect of the solution that was unforeseen until the mid-fifties was that machines might someday be able to talk, as well as read, like people; nor was it foreseen until scarcely a decade ago that there would be any possibility of such sophisticated performances by mechanisms of very modest size and cost.

Why has it taken so long for all this to happen? For one thing, we often see—and come to expect—that technology leaps ahead of its scientific base, and so seems to make sudden great strides. But it can leap only so far, and therefore progresses, on the average, only as fast as does the underlying science. Moreover, that science, as it concerns reading machines, has had only meager support over most of its course. In the present case, although a generous share of the research budget of the VA's Prosthetic and Sensory Aids Service was provided, the level of funding was often the limiting factor in pressing ahead with the research; indeed, a project of such complexity could hardly have been carried forward at all had it not been able to draw on the equipment and technical skills provided by parallel research on speech that the Laboratories were doing for other Federal agencies (Department of Defense, National Institutes of Health, and National Science Foundation).

But the pace of technology itself also set limits on the evolution of reading machines. Most of the time, it was a matter of asking the current technology to deal with tasks that were at the limits of what was possible without excessive cost; often, this pointed the work toward what was then possible rather than what was truly desirable, and so led to effort along lines that had to be abandoned only a few years later. This was true, for example, of all the construction work done on an Interim Word Reading Machine; it was true also of the work on speech synthesis by rule, which languished for seeming lack of promise until computers became available as Laboratory devices. Likewise, the very same explosive developments in microelectronics that have made possible today's compact text-to-speech reading machines also made suddenly obsolete the carefully planned efforts to set up a Reader Service Center for blind users.

Perhaps we should ask, not why progress has been slow, but how it happened at all. The problems to be solved were indeed difficult and time consuming. Few industrial research projects could have survived so long a maturing; the time scale to which they are geared is usually measured in years, not decades. Even Government support for research can cope with such long-term projects only when there are individuals in Government who have both the vision and the persistence to defend the undertaking.

Basic research is plainly essential to the development of devices such as reading machines for the blind. Only basic research could have led to speech synthesis by rule and to the demonstration that SSBR was the right choice as output signal for a high-performance reading machine. But is basic research sufficient to solve the entire problem? Probably not, and for a variety of reasons. For one thing, the kind of people and the kind of organizations that deal naturally and well with basic

research do not usually have the temperament or skills to handle the entrepreneurial job of bringing a device to market. The Government, for its part, lacks effective mechanisms for bridging the gap between the research it supports and the finished devices that embody that research; that is to say, between research and procurement—both of which the Government does do—there is much development and testing that is done only by private industry, when it is done at all. Fortunately for the users of reading machines now and in the future, there has been this kind of entrepreneurial effort ■

Acknowledgement: The authors would like to thank Eugene F. Murphy, Ph.D., recently retired from the Veterans Administration, for his vision and his confidence that better understanding would lead to a solution of the reading machine problem. We also appreciate the VA's support of this research, which spanned a period of two decades.

REFERENCES

1. Abma JS: The Battelle Aural Reading Device for the Blind. **In** Human Factors in Technology (chapter 19, pp. 315–325) E. Bennett, J. Degan & J. Spiegel (Eds) New York; McGraw Hill, 1963.
2. d'Albe EE: The optophone: an instrument for reading by ear. Nature, Lond., 105:295–296, 1920.
3. Allen J: Electronic aids for the severely visually handicapped. CRC Crit Rev Bioengng 1:139–167, 1971.
4. Allen J: Synthesis of speech from unrestricted text. Proc IEEE 64:433–442, 1976.
5. Allen J:Linguistic-based algorithms offer practical text-to-speech systems. Speech Technol 1:12–16, 1981.
6. Borst JM: The use of spectrograms for speech analysis and synthesis. J Audio Engng Soc 4:14–23, 1956.
7. Borst JM & Cooper FS: Speech research devices based on a channel vocoder. J Acoust Soc Amer 29:777(A), 1957.
8. Cooper FS & Zahl PA: Research on Guidance Devices and Reading Machines for the Blind: A Final Report of Work Done at the Haskins Laboratories between February 15, 1944 and December 31, 1947 under the auspices of the Committee on Sensory Devices, The National Academy of Sciences (Appendix W, p.8). New York, Haskins Laboratories, 1947.
9. Cooper FS: Research on reading machines for the blind. **In** Blindness: Modern Approaches to the Unseen Environment (chapter 32, pp.512–543) P. Zahl (Ed). New Jersey, Princeton University Press, 1950 (Reprinted 1963 & 1973. New York, Hafner Press).
10. Cooper FS: Spectrum analysis. J. Acoust Soc Amer 22: 761–762, 1950.
11. Cooper FS, Liberman AM & Borst JM: The interconversion of audible and visible patterns as a basis for research in the perception of speech. Proc Nat Acad Sci Wash 37:318–325, 1951.
12. Cooper FS, Liberman AM, Borst JM & Gertsman LJ: Some experiments on the perception of synthetic speech. J Acoust Soc Amer 24:597–606, 1952.
13. Cooper FS, Liberman AM, Harris KS & Grubb PM: Some input-output relations observed in experiments on the perception of speech. Proc 2nd Intern Cong on Cybernetics, Namur, Belgium, 930–941, 1958.
14. Cooper FS, Liberman AM, Gaitenby JH, Mattingly IG, Nye PW & Sholes GW: Research on audible outputs of reading machines for the blind. Bull Prosth Res BPR 10-23, 331–335, Spring 1975.
15. Corner GW: The committee on sensory devices. **In** Blindness: Modern Approaches to the Unseen Environment (chapter 28, pp.431–442) P. Zahl (Ed). New Jersey, Princeton University Press, 1950 (Reprinted 1963 & 1973. New York, Hafner Press).
16. Delattre PC. Liberman AM & Cooper FS: Acoustic loci and transitional cues for consonants. J Acoust Soc Amer 27:769–773, 1955.
17. Dewey G: Relative Frequency of English Speech Sounds. Cambridge, Mass, Harvard University Press, 1950.
18. Dudley H: The vocoder. Bell Labs Record 17:122–126, 1939.
19. Dudley H, Riesz RR & Watkins SA: A synthetic speaker. J Franklin Institute 227:739–764, 1939.
20. Dudley H: The carrier nature of speech. Bell Sys Tech J 19:495–515, 1940.
21. Fairbanks G: Test of phonemic differentiation: the Rhyme Test. J Acoust Soc Amer 30:596–600, 1958.
22. Fant G: Speech communication research. IVA (Sweden) 24:331–337, 1953.
23. Fant CGM: Modern Instruments and methods for acoustic studies of speech. Proc Eighth Intern Congr of Linguistics (Oslo): 282–358, 1958. (This report deals also with a num-

ber of other synthesizers, including POVO and DAVO (MIT) and Voback (Haskins Labs)).
24. Fant G & Martony J: Speech synthesis. Speech Transmission Laboratory, Stockholm, QPSR 2:16–18 & 18–24, 1962.
25. Fant G, Mártony J, Rengman U & Risberg A: OVE II synthesis strategy. Proc. Speech Comm Sem Stockholm II: Paper F5, 1963.
26. Farrell G: Avenues of communication. **In** Blindness: Modern Approaches to the Unseen Environment (chapter 21, pp. 313–345) P. Zahl (Ed). New Jersey, Princeton University Press, 1950 (Reprinted 1963 & 1973. New York, Hafner Press).
27. Fender DH: Reading machines for blind people. J Vis Impair and Blindness 77:75–85, 1983.
28. Freiberger H & Murphy EF: Reading machines for the blind. IRE Trans Prof Group Hum Fac Electron HFE-2:8–19, 1961.
29. Freiberger H & Murphy EF: Reading devices for the blind: an overview. **In** Human Factors in Technology (chapter 18, pp.299–314) E. Bennett, J. Degan & J. Spiegel (Eds). New York, McGraw Hill, 1963.
30. Gaitenby JH: The elastic word. Haskins Laboratories Status Report on Speech Research SR-2:3.1–3.12, 1965.
31. Holmes JN, Mattingly IG & Shearme JN: Speech synthesis by rule. Language and Speech 7:127–143, 1964.
32. House AS, Williams CE, Hecker MHL & Kryter KD: Articulation testing methods: consonantal differentiation with a closed-response set. J Acoust Soc Amer 37:158–166, 1965.
33. Ingemann F: Speech synthesis by rule. J Acoust Soc Amer 29:1255, 1957.
34. Irwin RB: The Talking Book. **In** Blindness: Modern Approaches to the Unseen Environment (chapter 22, pp.346–352) P. Zahl (Ed). New Jersey, Princeton University Press, 1950 (Reprinted 1963 & 1973, New York, Hafner Press).
35. Joos M: Acoustic Phonetics. Lang Monogr 23, Language, 24:2. Suppl., 1948.
36. Kelly JL & Gerstman LJ: An artificial talker driven from phonetic input. J Acoust Soc Amer 33:835, 1961.
37. Klatt DH: Software for a cascade/parallel formant synthesizer. J Acoust Soc Amer 67:971–995, 1980.
38. Lawrence W: The Synthesis of Speech from Signals which have a Low Information Rate. **In** Communication Theory, (chapter 34, pp.460–471) W. Jackson (Ed). London, Butterworths, 1953.
39. Liberman AM, Delattre PC, Cooper FS & Gerstman LJ: The role of consonant-vowel transitions in the perception of stop and nasal consonants. Psychol Monographs 68, 1954.
40. Liberman AM: Some results of research on speech perception. J Acoust Soc Amer 29:117–123, 1957.
41. Liberman AM, Ingemann F, Lisker L, Delattre PC & Cooper FS: Minimal rules for synthesizing speech. J Acoust Soc Amer 31:1490–1499, 1959.
42. Mann RW: Technology and Human Rehabilitation: Prostheses for Sensory Rehabilitation and/or Sensory Substitution. **In** Advances in Biomedical Engineering vol. 4 (pp.209–353) R. Kenedi (Ed). New York, Academic Press, 1974.
43. Mattingly IG: Synthesis by Rule of General American English. Ph. D dissertation, Yale University. (Issued as a supplement to Haskins Laboratories Status Report on Speech Research.) 1968.
44. Metfessel MF: Experimental studies of human factors in perception and learning of spelled speech. Proc Int Congr on Technol and Blindness (pp.305–308) L. Clark (Ed). New York, American Foundation for the Blind, 1963.
45. Naumburg RE: A bookprint reader for the blind. Sci Amer 145:113, 1931.
46. Nye PW: Reading aids for blind people—a survey of progress with the technological and human problems. Med Electron Biol Engng 2:247–264, 1964.
47. Nye PW & Bliss JC: Sensory aids for the blind: a challenging problem with lessons for the future. Proc IEEE 58:1878–1898, 1970.
48. Nye, PW, Hankins JD, Rand T, Mattingly IG & Cooper FS: A plan for the field evaluation of an automated reading system for the blind. IEEE Trans Audio Electroacoust AU-21:265–268, 1973.
49. Nye PW & Gaitenby JH: Consonant intelligibility in synthetic speech and in a natural speech control (Modified Rhyme Test results). Haskins Laboratories Status Report on Speech Research SR-33:77–91, 1973.
50. Nye PW & Gaitenby JH: The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. Haskins Laboratories Status Report on Speech Research SR-37/38:169–190, 1974.
51. Nye PW, Ingemann F & Donald L: Synthetic speech comprehension: a comparison of listener performances with and preferences among different speech forms. Haskins Laboratories Status Report on Speech Research SR-41:117–126, 1975.
52. Potter RK: Introduction to technical discussions of sound portrayal. J Acoust Soc Amer 18:1–3, 1946. (See also the five related articles that follow this introduction.)
53. Potter RK, Kopp GA & Green HC: Visible Speech. New York: van Nostrand, 1947.
54. Rabiner LR, Schafer RW & Flanagan JL: Computer synthesis of speech by concatenation of formant-coded words. Bell Sys Techn J 50:1541–1558, 1971.
55. Rosen G: Dynamic analog speech synthesizer. J Acoust Soc Amer 30:201–209, 1958.
56. Smith GC & Mauch HA: The development of a reading machine for the blind: summary report. Bull Prosth Res BPR 10-6, 98–124, Fall 1966.
57. Stowe AN & Hampton DB: Speech synthesis with prerecorded syllables and words. J Acoust Soc Amer 33:810–811, 1961.
58. Studdert-Kennedy M & Cooper FS: High-Performance Reading Machines for the Blind: Psychological Problems, Technological Problems and Status. **In** Sensory Devices for the Blind (pp.317–342), R. Dufton (Ed), London, St Dunstans, 1966.
59. Thorndyke EL & Lorge I: A Teacher's Word Book of 30,000 Words. New York; Teachers College Press, 1968.
60. Turine V de: Photophonic books for the blind. L'Eclairage Electrique 31:16–19, 1902.