

Audible Outputs of Reading Machines for the Blind*

Franklin S. Cooper, Jane H. Gaitenby, and Ignatius G. Mattingly[†]
Haskins Laboratories, New Haven

The generation of spoken English from printed text to serve as the output of a reading machine for the blind is the objective of the research at Haskins Laboratories being carried out under a Veterans Administration contract. A series of listening tests using Compiled Speech texts have been made in recent months, and tests have begun of texts recorded in Synthetic Speech. The purpose of these tests was to find out what blind listeners like (and what they will tolerate) in these forms of machine speech.

Compiled Speech and Synthetic Speech are dissimilar approaches to the reading machine output problem. Compiled Speech, an interim approach, was developed because it was an obvious and accessible exploratory strategy. Sentences in Compiled Speech are built up word by word from a prerecorded spoken vocabulary of 7,200 items. The word assembly is done by computer from an input of punched tape that corresponds to a printed text. (See Fig. 1.) Synthetic Speech, on the other hand, is more promising as a long-term approach. It consists of audible sentences which have been generated electronically from (at present) a typed phonetic input. This temporary input method simulates word retrieval by computer from a large dictionary, stored in the computer memory, of the spelled and phonetic forms of words.

Listeners and Recordings

The listeners were veterans attending the Eastern Blind Rehabilitation Center, Veterans Administration Hospital, West Haven, Connecticut. All were male volunteers, acquired as subjects for the tests through the warm cooperation of Mr. George Gillispie, Chief of the Center, and his assistants, of whom Mr. William Kingsley deserves special thanks. There were a total of eleven subjects, most of them in their early twenties. There were eight hour-long tests (each presented to a minimum of two listeners and to a maximum of four).

Thirty-six sample tapes (twenty-seven different texts or versions of texts) were presented in the course of the tests. Listening time per sample ranged from less than a minute to about ten minutes. Four tapes of Synthetic Speech (three different texts) were used in the beginning phase of this part of the study. An attempt was made to control various characteristics of the readings:

1. Rate of the Speech

The range tested extended from 70 to about 225 wpm. Ten rates within this range were preselected for testing.

*Summary prepared for the Bulletin of Prosthetics Research BPR 10-15, Spring 1971.

[†]Also, University of Connecticut, Storrs.

INTERIM WORD READING MACHINE FOR THE BLIND

PRINT TO AUDIBLE VERBAL OUTPUT USING COMPILED SPEECH

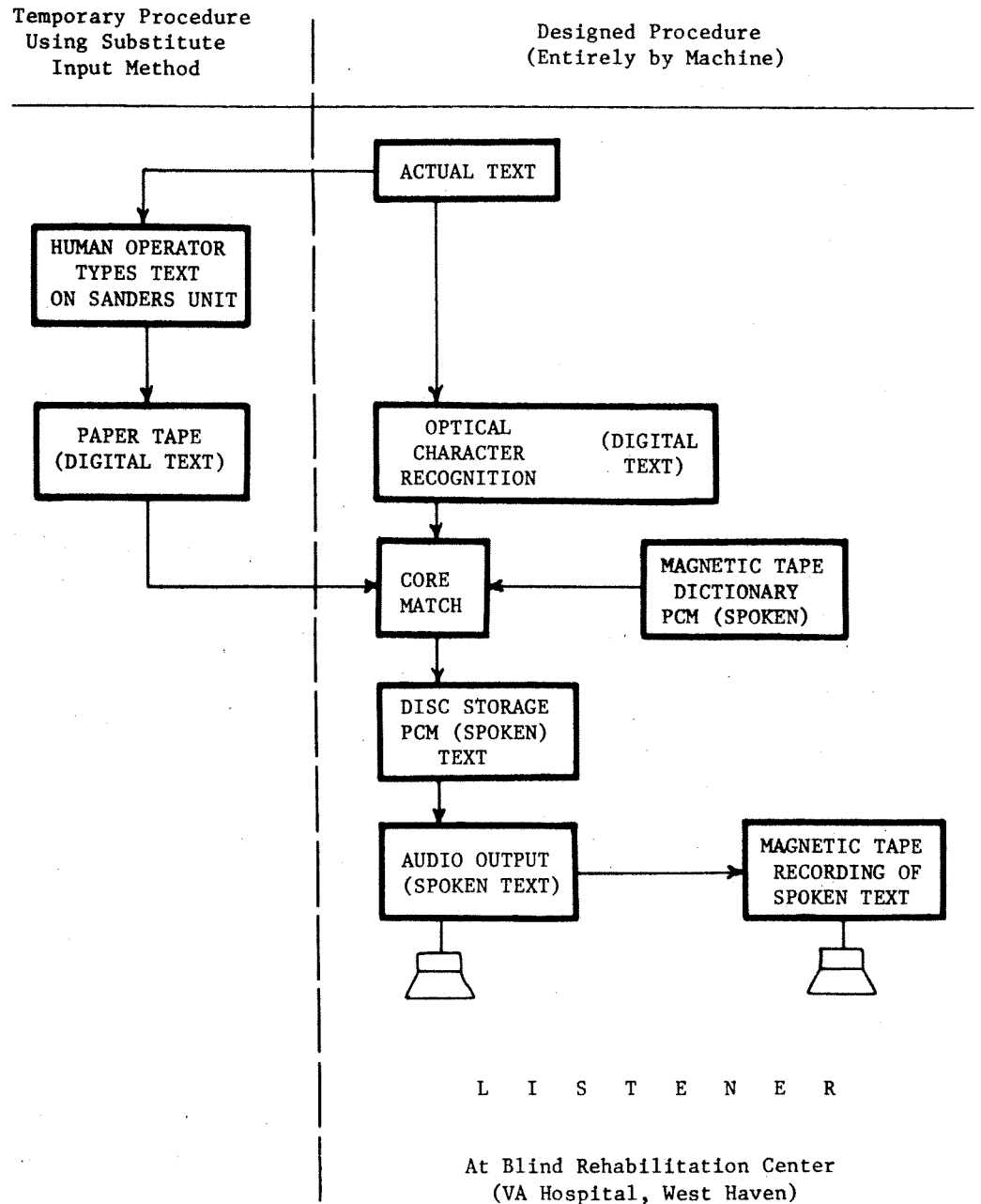


FIGURE 1

2. Type of Rate Manipulation (in Compiled Speech Only)

a. Time-compressed tape

Compression is accomplished by dropping regular segments of the speech at brief, regular intervals. Fundamental frequency remains as it was in the original. The speeded recordings were prepared by the Center for Rate Controlled Recordings, University of Louisville, Louisville, Kentucky. The original rate in Compiled Speech was from 100 to about 115 wpm. The amounts of time compression tested were 60, 65, 70, and 75 percent. Discard segments of 10, 15, and 20 msec were tested at each compression.

b. Capstan-speeded tape

Here, the output rate is changed by using a different capstan. The fundamental frequency rises proportionately to the increase in rate.

c. No rate manipulation

(It should be noted that overall average rates of less than 100 wpm were a function of the amount of spelling that occurred in given texts.)

3. Text (Author and Topic)

Among these were Dickens, Oliver Twist; Steinbeck, Travels with Charley; Pierce, Waves and Messages; a New York Times sports feature entitled "Hit and Write."

4. Amount of Spelling

This variable applies only to Compiled Speech, in which all words of the text that are not contained in the prerecorded vocabulary are spelled, letter by letter.

5. Form of Machine Speech Production

Compiled or Synthetic

The Test Situation

The test sessions took place at the Rehabilitation Center, in whatever room was available. The atmosphere was relaxed and somewhat informal. The investigator began each session with a brief introduction to reading machine research and stressed to the listeners that there were no right or wrong answers to the tests and 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 the eventual improvement of a reading machine output. All the subjects took the task very seriously.

It was thought by the investigator that the listeners' gross reactions--

their free and unprompted comments after each sample--would provide both the broadest and the simplest type of qualitative judgment. The test results given below are a summary of the listeners' spontaneous opinions, as recorded at the time of the tests.¹

Results with Compiled Speech

At its original recorded rate (100 to 115+ wpm), Compiled Speech has errors of intonation and word-phrase stress and timing. Listener opinions indicate that these deficiencies are more evident in time-compressed tapes when the word rate is either above or below the middle of the range that was tested, i.e., preferred rates were on the range of 150-175 wpm--about normal speaking rates. In capstan-speeded speech, rhythmic and inflectional oddities are most intrusive above 120-135 wpm. With no rate manipulation, Compiled Speech is considered too slow for comfortable listening.

Any spelling in a text, at any of the tested rates, is rejected by the listeners. (One subject said that he was unable to follow spelling even when it was done by a human reader, reading the Morning Telegraph aloud to him.)

In the time-compressed tapes, there was a slight preference for the 20 msec discard interval over the 15 msec discard interval, at all rates. The 10 msec interval was unsatisfactory because "warble effects" were produced in the voice quality, though this may be an artifact of the fundamental frequency of the speaker's voice.

The preferred word rates varied with the subject matter of the text and with the author's style. Also, some topics required more spelling than others. A simple chapter from An Introduction to Sculpture by William Zorach, in which there was very little spelling, was of great interest to the blind listeners, although it was played at the relatively slow rate of 110 wpm. A humorous Saroyan story about a vain young man asking for a screen test (in which there were no spelled words) held their attention through many samples of the same text (played at various rates in time-compressed, capstan-speeded, and Synthetic Speech). Yet two other stories by Saroyan (also without spelling) were deemed so dull, regardless of rate, that the subjects could scarcely bear to listen to them. (One story was largely a dialogue; the other, a monologue.)

The length of the speech sample is another factor in the appraisals. Half a minute is too short an exposure for reasonable evaluation if the topic of the text is unknown and if the tape is begun at a random location in the text. A minimum of one minute seems adequate, no matter what the topic or where the tape is begun, at least if the tape rate is within a reasonable range.

¹It must be pointed out that these were preliminary tests. Although the major aim was to have the machine speech evaluated, further purposes were to zero in on aspects of the tapes that should be controlled in later tests and on aspects of the speech itself that should be improved. (The interactions of the variables is an aspect of note, as the results show.)

The overall evaluation of Compiled Speech that emerged from these tests is that it is acceptable at some rates in either time-compressed or capstan-speeded form, but it is not enjoyable. Spelling is its worst feature. Temporal irregularities are also distressing. Listeners doubted that Compiled Speech (with or without spelling) would be tolerated for extended periods of time.

Preliminary Tests with Synthetic Speech

Four samples of speech synthesized by rule have been tested thus far, and the provisional evaluations represent the views of only six listeners. Nevertheless, responses were generally so positive that it seems useful to include a summary of their reactions.

Synthetic Speech was quite easily understood with exposures as brief as half a minute and at rates ranging from 136 to approximately 225 wpm. It was not necessary to time compress or to capstan speed Synthetic Speech tapes because texts at a variety of rates are easily made within the synthesis-by-rule program itself. (Rates higher or lower than those tested are also readily produced by the program.)

After hearing Synthetic Speech, the listeners' comments dealt mostly with the subject matter of the texts, indicating that the prosody (intonation, etc.) was acceptable, or at least not distracting. This was seldom true of Compiled Speech, which evoked comments dealing chiefly with rate irregularities. The one aspect of Synthetic Speech that was faulted was what the listeners called its "accent"--which some of them actively enjoyed. (One man from inland Maine called it "Bostonian"; another, from Brooklyn, thought it "foreign.")

One listener called Synthetic Speech "Great...really cool!" Such unrestrained approval probably reflects assets that are intrinsic to Synthetic Speech but not to Compiled Speech: no words are spelled; rates are equivalent to, or faster than, normal speech; there are normal transitions between words; intonation is naturalistic and flowing; pauses are reasonable; and rhythm is realistic.

It must be noted, however, that two important liberties were taken in the production of these Synthetic Speech tapes. First, the assumption was made that its stored reference vocabulary contained all of the text words encountered. (In the final machine, as now planned, new or rare words will be generated by rules for syllabification and syllable synthesis.) Second, only one of the three texts depended on rules for parsing, i.e., for stress and juncture assignments; the other two tapes had stress signals supplied by a human operator.

Conclusion

Comparison of the appraisals that have thus far been made of the two forms of machine output indicate that Compiled Speech is effectively rejected in favor of Synthetic Speech. Natural voice quality (commented on approvingly in Compiled Speech when played at its original rate) is one obvious lack of Synthetic Speech. Much of the work now under way is aimed at improvements in the naturalness of Synthetic Speech and at the production of additional and longer tests for a more adequate evaluation of its usefulness as a reading machine output.