PROBLEMS IN MACHINE CONVERSION OF PRINT TO "SPEECH" [a]

Jane H. Gaitenby, A.B.

Haskins Laboratories, Inc.
305 East 43rd Street
New York, N.Y. 10017

Applied research that is linguistic in character is required by a variety of enterprises and institutions these days. The research to be reported here has been made under contract to the Prosthetic and Sensory Aids Service of the Veterans Administration. The Veterans Administration has supported a number of research projects for the purpose of developing reading machines for the blind; and among them have been studies in electronic instrumentation, in psychology, in linguistics, and in cross-disciplinary fields. Each of these investigations has been concerned with the problem of converting the printed word to a tactile or auditory output. Part of the general research problem is optical in nature (because print must be recognized electronically), part deals with printed-symbol-to-sensory-symbol correspondence, and another part deals with human perception of units of sounds. A specific problem area is the analysis of the structure of spoken English, in view of the fact that the printed word is speech at one symbolic remove. Since blind people in general are obliged to approach all handwritten or printed or illustrated material either through braille or through the intercession of a human reader, a reading machine of some kind is an obvious need.

Leaving the consideration of cost aside, we can assume that the best reading machine will be one that converts printed text to an output that is as much like real speech as possible. In short, the machine should talk. This ideal machine should produce completely natural sentences, and to do this it should have the ability to vary intonations and pauses appropriately for specific texts. Ideally, it should be replete with variable voice quality, such as one shade and timing of voice for business letters and another one for romantic novels—or letters! But such a perfect machine would require a gigantic storage capacity, starting with tens of thousands of words. It would also require a remarkable program to manipulate its memory, to account for all manner of related nuances: grammatical, semantic, and intonational—in order to duplicate the associational memory and variable voice

[a] Given before the XIth Annual National Conference on Linguistics sponsored by the Linguistic Circle of New York, 12 March 1966, and printed in this issue with the permission of the author.

of the human being. Bear this ideal machine in mind—and the fact that it *is* an *ideal* machine.

Speech has been synthesized at Haskins Laboratories and elsewhere by rules applying to very small speech units, of phonemic or syllabic size, generally. A logical developmental step was to experiment with sentence production from word units that had been pre-recorded by a human speaker. A device called the interim word-reading machine was built for the Veterans Administration at Haskins to test the feasibility of generating sentences from single spoken words. Although a successor to that machine is already under way, many of the problems encountered in the course of outfitting the first interim device with a vocabulary, and other equipment, remain; so I will confine my remarks to the machine that has already served much of its purpose, but is soon to become ancestral.

The well-named *interim* word-reading machine at Haskins deals only with the word storage, retrieval, and output side of the reading machine problem, and omits the optical scanning operation (which is not our contractual obligation). My concern has been to get words recorded for the machine's storage that can be played out one after another, to (presumably) sound like sentences. The print-scanning function has been by-passed in the machine by simulation. The contents of a printed text are typed on a Flexowriter. The punched-tape resulting is simply a letter-to-digit conversion, and this is the form of information put into the machine. This is the input.

Now about the vocabulary storage in the machine. Each separate spoken word that has been previously recorded is stored on magnetic tape, and along with each word on the tape is its digital spelling, or code. When the code for a given word is sensed on the *punched* tape by the machine, that word is searched out on the stored spoken vocabulary *magnetic* tape. The word is matched and played back if it is stored, and is recorded on another tape at the same time, where it is added to the rest of the words in the order commanded by the original text. (If the word is not stored, it must be spelled out, letter by letter, unfortunately.) When the contents of the entire text have been accumulated, this new tape, of spoken words, is played back to the listener. This is the output of the word reading machine, and the entire process is, in fact, a conversion of print to sound. I will let you judge to what degree that sound is speech-like, before back-tracking over certain linguistic problems that the tape will illustrate. The sample you will hear now is about a minute long. It will be played for you at the rate in which the individual words were originally spoken and recorded. The long words which are spelled out at the end will probably surprise you. (By this I mean that you may have trouble understanding them.)

[TAPE: PART I, "This is speech produced word by word . . ."]

That was an output of 81 words per minute, which is slow compared to normal speech—although it is not a great deal slower than my rate. This sample and the ones to follow later were compiled by a manual method, while the machine itself was still under construction; but these are the actual words from the 7200-unit vocabulary (of words, letters, numerals, punctuation, and a few suffixes) available to the machine's own storage. The original words were spoken by John T. Wadsworth, in a long series of short recording sessions made in the course of one year, for the most part.

After the words were recorded, the original tape was copied and edited. Then each word was separately mounted on a card like this one. (There are samples to be passed around.) These test sentences you have heard were generated by playing the word cards in sequence through a machine (Language Master) that plays tapes from cards instead of reels, and then each word was recorded in direct sequence on another machine. The sentences *as such* are therefore synthetic, since they were never spoken as sentences by the original speaker. The output you have heard, and those to come, demonstrate by negative evidence, that real speech is produced not word by word, but in continuous groups of words that are compatible in respect to tempo, loudness, and melody.

Every problem in applied research has constraints. For this machine, one—and only one—intonational version of each vocabulary word could be stored. This is a severe restriction, because in normal speech a given word may occur in many very, *very* different prosodic forms. Also, just one pronunciation was allowed per printed form, but homographs are a minor problem, compared to the fact that each word had to be stored in a single, frozen, stress and intonation form. It is clear that the one version recorded should be a highly probable spoken form for a highly probable type of occurrence in printed form.

We chose the vocabulary to be stored from the Dewey and the Thorndike and Lorge lists of the most frequently printed English words. There were no published data on probable spoken forms of the vocabulary words, although there were some helpful reports in the acoustic literature dealing with measurable correlates of stress and intonation by such investigators as Bolinger, Fry, and Denes.

Dr. F. S. Cooper of Haskins gave us a take-off push in the right general direction and some fine instruments to work with. We made exploratory acoustic and perceptual tests of real and synthetic speech, and approached the machine's speech problem with the hypothesis that: very frequent phrases and polysyllabic words were structurally similar in a way that might be useful for us. That is, the syllables of a polysyllabic word have a persistent stress relationship, even in varied intonational environments, and the syllables of the most frequent phrases also tend

to be regular in their stress relationships. We gradually learned something about the prosodic components of stress in words or phrases—intensity, frequency, and duration—and decided to "program" our human speaker, if possible, to make the word recordings using prescribed stresses and intonations. Now the question was, what stress to prescribe for what. We turned to the study of probability of word occurrence in printed and spoken form.

The most frequent type of phrase in English texts is prepositional. Articles, prepositions, and a connective are by far the most common English words. Most highly frequent words are monosyllabic. The stress of the most frequent words is usually very low. Most phrases begin with a preposition, as mentioned before, and most phrases and sentences end with a noun. Nouns carry much information and are generally prominent in the speech chain. And so on with the other grammatical classes . . .

We also examined so-called "intonation" and were soon convinced that it is acoustically reflected in durational shifts as well as in frequency and intensity changes. Otherwise, we knew very little except that syllables lengthen immediately before a pause, that pitch and intensity peaks usually start high and tend to decline toward the end of an utterance, and that pauses for punctuation are variable in length and of course have structural significance. (Some of these aspects of intonation almost certainly have a physiological base.)

Facts or observations of this sort, along with counts of form class sequence taken from texts in daily papers, books, and periodicals, indicated that the spoken lexicon should be recorded on the basis of a word's grammatical function. Grammatical function is correlated to some extent with word stress category, and a trained speaker could, with effort, produce words at a prescribed relative pitch, length, and loudness. An underlying assumption was that the stress prescriptions themselves would be valid, and that they would be consistently reproducible by a human speaker on demand, and over a large span of time.

Hand-out One (Fig. 1) is a diagram of probable grammatical sequences in printed texts. The size of each circle indicates the projected relative prominence of a word as a given part of speech, used in writing the stress prescriptions. Although nearly all the form classes in English can be—and often are—preceded or followed by almost any of the form classes, the arrowed lines shown between any *two* classes stand for statistical likelihood of sequence.

Using the probability rationale for the manner in which word classes were to be spoken, required that each of the 7,200 words be classified as a member of a particular grammatical class, before it could be suitably recorded. This was a stumbling block.

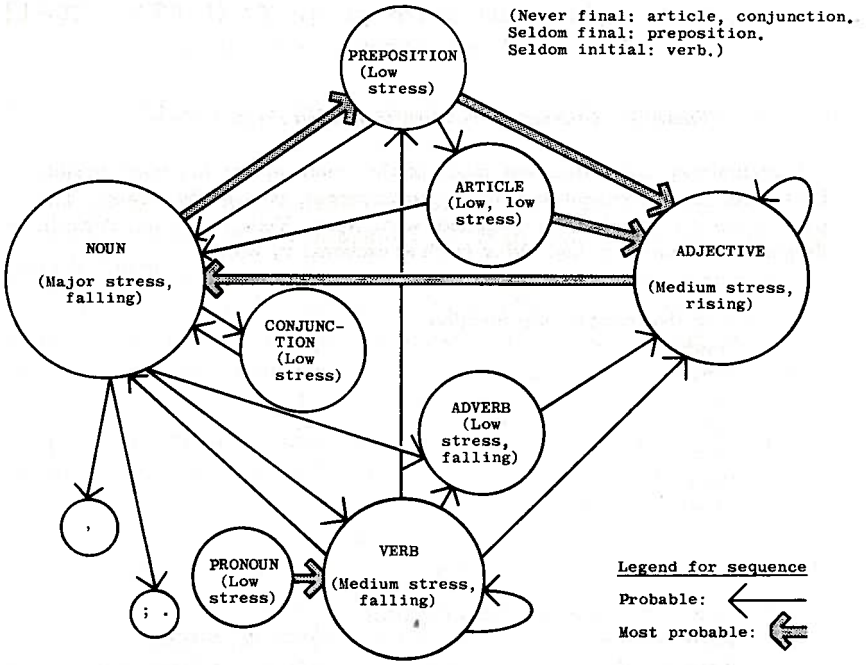

FIGURE 1. Hypothesized most probable grammatical and intonational sequences in English texts.

Hand-out Two (Table 1) suggests the problem by showing a breakdown of the thousand most frequent words by *potential* membership in a grammatical class. About half of the words can function in more than one role, and that role can be determined only by context. We therefore classified the multifunctional words by their most *probable* function, through educated guesswork and by intuition as native speakers. Impossible-to-classify words were put into the most neutral stress group, along with main verbs, whose stress in a sentence seems to be unpredictable. The prescriptions were written, and the recordings were made.

Again you may judge the output for yourselves, now bearing in mind that the sentences should be listened to not only to test the intelligibility, but also to note the words as stress types.

The recording speaker did follow the directions explicitly about 99 percent of the time, and the texts used have actually been a random test of the stored 7,200 word vocabulary. The questions to think about when listening to the tapes are: To what degree is normal intonation approximated in these tapes, and in what kinds of cases is it least normal, and finally, what *is* normal intonation?

TABLE 1.—*Grammatical Functions of the Most Frequent Words*

A preliminary breakdown was made of the commonplace functions possible for the 1027 most frequent words (i.e., *written* forms), as given by Dewey. The parts of speech used here as categories were Noun, Verb, Adjective, Adverb, Preposition, Connective, and Other (such as exclamatory word):

A) Words that have a *single* function,		
Nouns	196	
Verbs	140	
Adj.	81	
Adv.	48	
Prep.	17	
Conn.	8	
Other	1	
	491	
B) Words that have *two* potential functions,		
Noun or Verb	215	e.g., "being"
Noun or Adj.	120	"three"
Verb or Adj.	50	"open"
Adj. or Adv.	26	"better"
Adv. or Prep.	18	"to"
Adv. or Conn.	8	"since"
Conn. or Other	6	"though"
Adv. or Other	4	"really"
	447	
C) Words that have *three* potential functions,		
Noun, Verb, Adj.	46	e.g., "present"
Noun, Adj., Adv.	7	"first"
Noun, Adj., Conn.	3	"either"
Verb. Adj., Prep.	3	"near"
	59	
D) Words that have *four* potential functions,		
Noun, Verb, Adj., Adv.	6	"last, left, set, front, further, back"

[TAPE: PART II "Every city has its old guard . . ." Current book sample, from Zinsser's *The City Dwellers.*)

PART III "The North Wind and the Sun . . ." (well-known phonetic exercise)]

The paragraphs below have been added for readers of the Bulletin of Prosthetics Research, to compensate for their inability to hear the tapes speak for themselves.

The best feature of the interim reading machine output is the approximation of naturalistic intonation, particularly in prepositional phrases. Since phrases of this type occur far more often than any other syntactic structure in texts, the overall acceptability is high.

The most unnatural aspect of the synthetic sentences is the presently unavoidable number of spelled words, occasioned by the restriction on the size of the stored vocabulary. When, in the midst of the verbal output, the listener suddenly hears a spelled word, he is more or less bewildered and his comprehension of the sentence suffers. Spelling is not typical of either conversation or of reading and actually represents a total shift from the medium of speech to the medium of writing. Spelling is therefore destructive to the intelligibility of the whole text despite the fact that it is necessary in a mere 5 percent of the words of a normal text. In addition, the unrecorded words which must be spelled in the output are the most infrequent (i.e., least expected, least guessable) words. And they are—on the average—exceptionally long or peculiarly spelled, "difficult" words. Furthermore, each letter of a spelled word is a whole syllable, and thus the output word rate is severely slowed down by spelled words. Another handicap in spelled words is that the letters, as spoken, lack natural intonation and spacing, which a good spontaneous reader would provide if he were obliged to spell. Each of the 26 letters had to be recorded in only one intonational form, just as each of the 7,200 words was recorded in a single spoken version. Because there are even fewer restrictions on the place of letter occurrence in a word than are on the place of word occurrence in a phrase or sentence, advance prescriptions for specific letter intonation were not attempted. The letters were merely recorded in groups by rhyming syllables, e.g., "A, J, K, . . .," at a sustained single level, in the hope that in sentences they would rapidly be perceived as letters, distinct from the words spoken as units in their immediate environment. A very brief pure tone signal also precedes and follows each spelled word in the program for the output—to signal the abrupt shift from whole words to letters on the auditory track—and this seems helpful.

Studies have been and are under way to formulate a program for conversion of printed words to audible, recognizable syllables—which are in-

tended to circumvent the spelling problem. It is well-known, however, that there are few one-to-one spelling-to-sound conversions in most of the letters and syllables of English (especially where stress is concerned), so this experimental method presents problems of its own. It remains a challenge.

The second generation word reading machine now being developed, sired by the interim machine and a computer facility, will store words as control signals rather than as waveforms as at present. In that machine intonation can be manipulated and programed by synthesis, and certain of the present output irregularities in stress, timing, and inflection, will be overcome. The word rate can be speeded up as the phrasing is improved, and this is a requisite for blind listeners. At the same time our understanding of real speech phenomena will increase.