A System for Converting English Text into Speech

WILLIAM A. AINSWORTH

Abstract—The feasibility of converting English text into speech using an inexpensive computer and a small amount of stored data has been investigated. The text is segmented into breath groups, the orthography is converted into a phonemic representation, lexical stress is assigned to appropriate syllables, then the resulting string of symbols is converted by synthesis-by-rule into the parameter values for controlling an analogue speech synthesizer. The algorithms for performing these conversions are described in detail and evaluated independently, and the intelligibility of the resulting synthetic speech is assessed by listening tests.

I. Introduction

There are two main methods by which English spelling can be translated into phonemic spelling: the dictionary method and the algorithm method. In the dictionary method the words of the language together with their phonemic spelling are stored in some form, whereas in the algorithm method each letter, or group of letters, is translated into phonemic symbols according to a set of stored rules.

There are many difficulties with the algorithm method. Present day spelling is the result of historical sound changes that have occurred to words derived from a number of languages, with the result that many of the rules of pronunciation have a large number of exceptions. Despite these inherent difficulties, however, the algorithm method does have the attraction of economy. If a computer with a very large backing store is necessary for speech synthesis from text, then the applications of this work will only be possible in a few circumstances where cost is not of paramount importance. On the other hand, if rules that are sufficiently good to produce comprehensible speech can be programmed on a small computer, the benefits of speech synthesis from text can be made available for a great many more applications.

In order to determine the level of intelligibility that can be attained by speech produced entirely by rule, a system is being developed that converts text punched on paper tape into a string of symbols, from which the parameters for controlling a hardware speech synthesizer are calculated. The stages in the process are as follows: 1) segmentation into breath groups; 2) translation into phonemic symbols; 3) lexical stress assigment; and 4) calculation of speech parameter values.

II. Segmentation

The segments of speech are different from those of written language. Speech with pauses between all the words sounds unnatural, and speech with no pauses at all is extremely difficult to understand. As a preliminary to speech synthesis, therefore, it is necessary to segment the speech into the equivalent of breath groups.

This has been accomplished in an approximate manner as follows. Characters are read into a buffer until either a punctuation mark is encountered, or the buffer is filled. If a punctuation mark is found, the string of characters in the buffer is taken as the breath group. If the buffer becomes filled, the sentence is segmented as follows. The words are examined one by one from the end of the buffer to see if the string contains a conjunction. If one is found, the contents of the buffer, up to this point, are taken as the next breath group and the remainder is shifted to the start of the buffer. If a conjunction is not found, a search is made for an auxiliary verb, a preposition, and then an article, in that order. If none of these is found, the buffer is rounded to the next word boundary, and the entire contents of the buffer are taken as the next breath group.

These rules attempt to place breath group boundaries at the following syntactic locations (whichever is encountered first): 1) at a punctuation mark; 2) preceding a conjunction; 3) between a noun phrase and a verb phrase; 4) before a prepositional phrase; 5) before a noun phrase; and 6) after a fixed number of characters have appeared in the input.

The behavior of this scheme depends to a certain extent on the size of the buffer. It was found that by making the buffer about 50 characters long, breath groups of about the right duration were generated. This scheme did not always put the boundaries between segments at the same place that a normal speaker would insert a pause, but it put them in the same place as the writer in approximately 80 percent of the cases.

III. Phonemic Translation

The rules by which the letters were translated into phonemic symbols are shown in Table I. They are intended to produce a received pronunciation dialect of British English. The rules were developed by listing each letter and its most common phonemic manifestations. The conditions under which these manifestations occurred were noted, together with any common exceptions. In cases where the translation of a letter was ambiguous, and where no context rule could be devised that would resolve this ambiguity,

Manuscript received April 30, 1972. This work was supported by the Science Research Council.

The author is with the Department of Communication, University of Keele, Keele, England.

TABLE I

Letter	Phoneme	Letter	Phoneme	Letter	Phoneme
-(a)-	/ə/	VCd(e)d-	/ə/	(g)e	/dʒ/
-(are)	/ɑ/	VCt(e)d-	/ə/	(gh)	/g/
(a)E	/ɛi/	VC(e)d-	/ /	(g)	/g/
(ar)	/ɑ/	(e)r-	/ə/	w(h)	/ /
(a)sk	/ɑ/	wh(ere)	/ɛə/	(ho)v	/hæ/
(a)st	/ɑ/	h(ere)	/iə/	(h)	/h/
(a)th	/ɑ/	w(ere)	/3/	-(i)-	/ɑi/
(a)ft	/ɑ/	(ere)	/ir/	(i)ty	/I/
(ai)	/ɛi/	(ee)	/i/	(i)E	/ɑi/
(ay)	/ɛi/	(ear)	/ir/	(ir)	/3/
(aw)	/ɔ/	(ea)	/i/	(igh)	/ɑi/
(au)	/ɔ/	(e)ver	/ɛ/	t(io)n	/ʌ/
(al)l	/ɔ/	(eye)	/ɑi/	(i)nd	/ɑi/
(a)ble	/ɛi/	(e)E	/i/	(i)ld	/ɑi/
(a)ngSUF	/ɛi/	c(ei)	/i/	-C(ie)	/ɑi/
(a)	/æ/	(ei)	/ɑi/	VC(ie)	/i/
(b)	/b/	(e)r	/3/	(i)	/I/
(ch)	/tʃ/	(eo)	/i/	(j)	/dʒ/
(ck)	/k/	(ew)	/ju/	-(k)n	/ /
(c)y	/s/	(e)u	/ /	(k)	/k/
(c)e	/s/	(e)	/ɛ/	(le)-	/əl/
(c)i	/s/	(f)-	/v/	(l)	/l/
(c)	/k/	(f)	/f/	(m)	/m/
(d)	/d/	(g)e-	/dʒ/	(n)g	/ŋ/
VC(e)-	/ /	(g)es-	/dʒ/	(n)	/n/
th(e)-	/ə/	(g)SUF	/g/	(or)	/ɔ/
-C(e)-	/i/	(g)i	/dʒ/	(o)E	/əu/
-C(e)d-	/ɛ/	(g)et	/g/	(oa)	/əu/
(o)ld	/əu/	c(ow)	/ɑu/	(their)	/ðɛə/
(oy)	/ɔi/	h(ow)	/ɑu/	(th)r	/θ/
(o)ing	/əu/	n(ow)	/ɑu/	(th)	/ð/
(oi)	/ɔi/	v(ow)	/ɑu/	(t)ion	/ʃ/
y(ou)	/u/	r(ow)	/ɑu/	(t)	/t/
(ou)s	/ʌ/	(ow)	/əu/	(u)pon	/ʌ/
(ough)t	/ɔ/	g(o)-	/əu/	(u)V	/u/
b(ough)	/ɑu/	n(o)-	/əu/	(u)C-	/ʌ/
t(ough)	/ʌf/	s(o)-	/əu/	r(u)	/u/
c(ough)	/of/	(o)-	/u/	l(u)	/u/
-r(ough)	/ʌf/	(o)	/o/	(u)	/ju/
r(ough)	/u/	(ph)	/f/	(v)	/v/
(ough)	/əu/	(psy)	/sɑi/	(w)r	/ /
(oul)d	/ʊ/	(p)	/p/	(wh)o	/h/
(ou)	/ɑu/	(q)	/kw/	(wha)t	/wo/
(oor)	/ɔ/	(r)-	/ /	(wa)	/wo/
(oo)k	/ʊ/	(rho)	/rəu/	(wo)r	/w3/
f(oo)d	/u/	(r)	/r/	(w)	/w/
(oo)d	/ʊ/	(sh)	/ʃ/	(x)	/ks/
f(oo)t	/ʊ/	(ss)	/s/	-(y)	/j/
s(oo)t	/ʊ/	(sch)	/sk/	VC(y)	/I/
w(oo)	/ʊ/	Xv(s)	/z/	-C(y)	/ɑi/
(oo)	/u/	V(s)-	/z/	(y)E	/ɑi/
sh(oe)	/u/	(s)	/s/	(y)	/I/
(oe)	/əu/	(there)	/ðɛə/	(z)	/z/

the most common or the most neutral phoneme was generally substituted in all cases. Many of the more common English words were treated as exceptions by including them as individual rules.

These rules were programmed on a small digital computer (PDP-8). It will be seen from Table I that the vowels were a good deal more difficult to translate than the consonants, with the letter O being the most difficult of all.

TABLE II
Analysis of Errors in Phonemic Translation

Source	Error (percent)	Stressed vowels (percent)	Unstressed vowels (percent)	Consonants (percent)
Textbook	8	4.5	2.2	1.3
Novel	11	6.8	3.0	1.2
Newspaper	11	6.9	3.1	1.0

Entries indicate the percentage of words that contained one or more phonemic errors.

These rules were tested by measuring their performance on three 1000 word passages of text. These were deliberately chosen to be on different subjects and written by different authors. One was a passage from a textbook on phonetics, one from a modern novel, and one from a newspaper article on a political theme. The phonemic translations produced by the program had 8, 11, and 11 percent respectively, of their words containing errors.

An analysis was made of these errors. Table II shows a breakdown into stressed vowels, unstressed vowels, and consonants. Consonant errors tended to be the substitution of a voiced consonant for its unvoiced equivalent (e.g., /ð/ for /θ/) or vice versa. Such errors are fairly unimportant perceptually, and though they detract from the quality of the speech produced they rarely change the meaning of a word. Errors in unstressed vowels are also relatively unimportant, especially if the substituted vowel is a near neighbor in the vowel chart of the correct one. Unstressed vowels are normally reduced, so the "distance" between neighboring vowels becomes even less.

The residual errors are the stressed vowels. These are caused mainly by a combination of letters in similar contexts having many phonemic manifestations. For example, EA becomes /i/ in "meat", /ɑ/ in "heart", /3/ in 'year', /ɛi/ in "great", /iə/ in "fear", and /iæ/ in "reality". There are also problems with compound words, and with words such as "bow" and "wind". However, the error analysis shows that cases of these kinds occur only in about 6 percent of the words in a typical text.

IV. *Stress Assignment*

The correct pronunciation of English depends a good deal on placing the stress on the right syllables. In the algorithm method this must be done according to rules. In the present system the following rules were employed. A list of words that are not usually stressed (articles, prepositions, conjunctions, etc.) were stored in the computer. The syllables of all words in this list were left unstressed. If a word was not included in this list, its first syllable was stressed, unless its first syllable was included in a list of prefixes. In that case the second syllable was stressed.

No attempt has yet been made to incorporate sentence stress and any more degrees of stress.

It was found that this scheme placed the stress on the right syllable over 90 percent of the time, but this includes monosyllabic words. The stress was placed on the right syllable in 83 percent of the bisyllabic words, 69 percent of the trisyllabic words, and 56 percent of the longer words in the passages of text mentioned previously.

Two additional rules were programmed at this stage: where a phoneme followed an identical one, the second was deleted; and where one word ended in a vowel and the next began with another vowel, an appropriate glide was inserted between the two.

V. Parameter Calculation

The output of the phonemic translation program was used to generate the parameter values for a terminal analog speech synthesizer by means of a synthesis-by-rule program. The program was similar to that described by Holmes *et al.* [1]. A table of target parameter values and durations for each phoneme is stored in the computer. Continuous parameter values are derived by linear interpolation between the target values. Certain acoustically complex phonemes are composed of more than one element. For example, stop consonants require elements for the silent interval, the burst, and the formant transitions.

Rules for intonation have been discussed by Vanderslice [2], and programmed by Mattingly [3] and Flanagan *et al.* [4]. In the present system a primitive form of intonation was introduced by making the fundamental frequency parameter rise to a peak during each stressed syllable, and fall to a trough midway between successive stressed syllables. An attempt to introduce a natural rhythm into the speech was made by elongating some of the stressed syllables so that the interval between stressed syllables was approximately constant. This is one way of approximating the isochronous foot theory [5]. The rhythm of the speech generated in this way, however, was not noticeably more natural than that based on fixed duration.

The parameter values produced by this program were converted to analog voltages, and used to control a speech synthesizer of the type described by Holmes *et al.* [1].

VI. Performance of the System

The three texts referred to previously were segmented, translated into phonemic symbols, then converted into sounds, and recorded. A hiss was inserted at the end of each breath group segment.

Three subjects were asked to listen to these sounds. They were asked to play the tape recorder until they heard a hiss, then stop the recorder and write down the phrase they had just heard before going on to listen to the next phrase.

The results were scored by counting the number of words correctly written down. The scores obtained were very varied, ranging from 50-90 percent of the words correct. There was no evidence of learning in the scores of the individual listeners, but their overall scores did correlate with their previous exposure to synthesized speech. This suggests that the scores might be improved by training.

In a subsidiary test, the highest scoring listener and the writer listened to speech synthesized completely automatically from unedited typesetting tapes obtained from a newspaper office. Subsequent comparison with printouts of these tapes showed that, despite several typing errors in the original tapes, both listeners identified 80-90 percent of the words correctly.

VII. Discussion

The performance of the letter-to-sound rules appears to be remarkably good. A typical seven word sentence will, on average, contain less than one phonetic error. This level of performance is probably achieved because most of the longer words in English are pronounced according to rule, whereas the common words are pronounced irregularly. The present set of rules ensures that the most common irregular words are treated as special cases, and the correct phonemic translation is generated.

The stress placement could probably be improved by employing the rules of Chomsky and Halle [6]. The present simple rules, however, were fairly successful for the bisyllabic words. This was probably because many common bisyllabic words consist of a monosyllabic word with either a prefix or suffix. The rules ensure that the stress is not placed on the affix.

The listening tests showed that the synthesized speech is not at present good enough for naive listeners to understand. At 50 percent intelligibility the listeners hear a few sentences correctly. Then, perhaps because of a cluster of errors, they lose track for a few sentences. At 90 percent the message is understood almost perfectly, but the written output is slightly different from the input.

At this preliminary stage in the development of the system, the results appear to be extremely encouraging.

References

[1] J. N. Holmes, I. G. Mattingly, and J. N. Shearme, "Speech synthesis by rule," *Lang. Speech*, vol. 7, pp. 127-143, 1964.

[2] R. Vanderslice, "Synthetic elocution," *Working Papers in Phonetics*, vol. 8, Univ. California, Los Angeles, pp. 1-131, 1968.

[3] I. G. Mattingly, "Synthesis by rule of prosodic features," *Lang. Speech*, vol. 9, pp. 1-13, 1966.

[4] J. L. Flanagan, C. H. Coker, L. R. Rabiner, R. W. Shafer, and N. Umeda, "Synthetic voices for computers," *IEEE Spectrum*, vol. 7, pp. 22-45, Oct. 1970.

[5] D. Abercrombie, *Studies in Phonetics and Linguistics.* London: Oxford University Press, 1965.

[6] N. Chomsky and M. Halle, *The Sound Pattern of English.* New York: Harper and Row, 1968.