A Simple Program for Synthesizing British English

MARK P. HAGGARD
IGNATIUS G. MATTINGLY

Abstract—A procedure for synthesizing British English has been developed which incorporates certain economies at little apparent cost in intelligibility, and introduces some rules for allophones which improve quality. Tests are reported with intelligibility around 90 percent, and improvements suggested for the major sources of errors. The assumptions in the procedure for synthesizing speech by rule both summarize and suggest research on fundamental aspects of speech, and thus constitute a model for some aspects of speech production and speech perception.

THIS PAPER describes some progress made in synthesis by rule, using the system described in a previous paper[4] on a computer-controlled formant synthesizer. At this time, the synthesis of British English (RP) is slightly more developed than that of General American (GA), so the rules for the former are described here.

The computational framework is the same as that of Holmes, Mattingly, and Shearme.[2] A very general computation program accepts as input a set of rules and a phonemic specification of the required utterance; it outputs time-sampled values of the parameters that control a parallel formant synthesizer. The detailed rules for generating the speech are, thus, ones of many possible different dialect inputs to the computation program, although that program does have some assumptions and limitations built in. The set of rules for a dialect has three parts: a set of pitch-change characteristics, implemented by the program on receipt of certain prosodic markers and certain segmental features in the phonemic script input; a segment inventory; and a set of allophone rules. We are here concerned with the latter two parts, rather than the pitch functions, which are held constant in the present investigations.

The Segment Inventory

From a linguistic rather than an engineering viewpoint, the object of synthesis by rule is a system with an input symbol inventory equivalent to a phoneme inventory. Given a small set of allophone rules, it is relatively simple to synthesize speech from a phonemic script input. This may, however, require nonphonemic members of the segment inventory. For example, prevocalic and postvocalic /1/ in English are somewhat different acoustically. It would be possible to take the prevocalic case as basic and make the necessary conversions for the steady-state and transition characteristics of postvocalic /1/, thus reducing the size of the segment inventory; but in this particular case (and especially using a nonarticulatory synthesizer), that particular solution would be more wasteful than simply having two /1/ allophones in the segment inventory and using one rule to decide which is appropriate. In the present set of rules, this "stored allophone" solution is also used for back-vowel allophones of the three velar consonants, and lip-rounded or lip-closed allophones of the two loud voiceless fricatives /s, ʃ/. Thus there are six phonemes with major allophones stored in the segment inventory.

Each segment in the tables has seven subgroups of entries corresponding to 1) phonetic classification and duration, 2) excitation, 3) steady-state or "ideal" values of the eight spectrum parameters (frequencies and amplitudes of three formants, overall amplitude, fricative amplitude), and 4)–7) four subgroups of characteristics controlling the transitions in the eight parameters.

Manuscript received October 10, 1967. This work was supported by grants and contracts from the Information Systems Branch of the Office of Naval Research, the Prosthetic and Sensory Aids Service of the Veterans Administration, the National Institute of Child Health and Human Development, and the General Research Support Branch, Division of Research Facilities and Resources, of the National Institutes of Health. This paper was presented at the 1967 Conference on Speech Communication and Processing, Cambridge, Mass.

M. P. Haggard was with Haskins Laboratories, New York, N. Y. He is now with the Psychological Laboratory, Cambridge University, Cambridge, England.

I. G. Mattingly is with Haskins Laboratories, New York, N. Y. 10017 and the Dept. of Linguistics, University of Connecticut, Storrs, Conn.

These characteristics allow specification of continuous transitions as in /w, y/, or discontinuous ones as in nasals. All the table values in the present set of rules have been arrived at as a result of new evaluations in a number of different contexts. Naturally, the agreement with the table data of Holmes *et al.*,[2] and with the consonant formant frequency loci of the earlier Haskins studies is very close.

Given the large number of table entries to specify per segment, the tables must be highly redundant. Most characteristics specified, except the eight steady-state values, could be produced by some simple and general rules operating on an articulatory feature basis; it is appropriate to think of the symmetries in the segment table as the result of just such a more basic set of rules, although the computation program does not actually work that way. There are other redundancies in the tables achieved at the expense of little, if any, reduction in naturalness of the speech. For example, there are no transitions except within vowels and semivowels; there are no formant amplitude transitions except to cue the voicing distinction in fricatives. Thus, a procedure used for producing rule speech, but not used for research purposes, could be simpler than the very general framework of tables and computation used here.

The segment inventory has also been simplified by reduction in size, compared to a phoneme inventory. RP is customarily transcribed with 9 or 10 diphthongs. Being two-element phonemes, these would normally require up to 20 extra sets of table entries. In the present set of rules, there is only one extra element (X), a sort of schwa classified as a semivowel for computing transitions. The economy is achieved as follows. Both first and second elements of diphthongs are related to elements already in the phoneme inventory in their own right (although simply using those elements will not give a good result). The allophone rules are used to create appropriate first and second elements, and the diphthongs are transcribed as two phones: /ay/, /aw/, /ɔy/, /ɔw/, /ɛy/, /iə/, /uə/, /ɔə/, /ɛə/. Allophone rules substitute new transition-controlling characteristics for the second elements, making the transition more gradual; the first element is lengthened, and its formant frequencies modified by the following two ad hoc rules:

$$F_1' = F_1 - 50 \text{ Hz}; \qquad F_2' = F_2 + 225 \text{ Hz}.$$

This procedure does not do justice to the differences, for example, between the first elements of /ɑɪ/ and /ɑʊ/, but it provides intelligible results. Diphthongization of vowels /i/ and /u/ is also provided by including /y/ and /w/ as their second elements, which increases naturalness.

The intelligibility of the output of the set of rules was evaluated in the following way. Monosyllabic words were synthesized for two 55-word rhyme tests in citation form. Each rhyme test comprised examples of each phoneme, except /ʒ/, with consonants other than semivowels in both final and initial positions. The rhyme test response sheet contains five response alternatives for each trial —one correct and four varying minimally from the correct response. In the case of the vowels, these distractors were obtained by proximity in terms of formant frequency; in the case of the consonants, in almost all cases by one feature changes: place, nasality, voicing, and manner (degree of closing of tract). Thus, the response alternatives for *bet* were *pet*, *met*, *debt*, *wet*. Since four of the most likely errors for a given stimulus phoneme were present among the response alternatives, a rhyme test score is a much more stringent test and is not directly comparable with the results from a randomly selected vocabulary of five words. However, the use of the small vocabulary reduces some of the error in large vocabulary testing, i.e., word-frequency effects, inter-individual variability, etc. The ability to focus on particular feature confusions in a type of confusion matrix make the rhyme test a suitable measure for synthesis evaluation. The experimental results will be analyzed in some detail to illustrate the way in which the rules are modified.

Intelligibility Tests on the Rules

Table I shows the results of giving this test to ten speakers of British English. The subjects were drawn from a New York diplomatic office, and were, presumably, of above average intelligence, but none had any previous experience of either psychological testing or synthetic speech. The matrix in Table I is a special type of confusion matrix, in that not all the response alternatives were available on a given presentation, but the pattern of confusions obtained in a complete matrix would not differ greatly, because of the composition of the test. There seemed to be no relation of initial-final difference to phoneme class, so initial and final data are pooled. Finals tended to have fewer errors, presumably because of prepausal lengthening, and possibly the slower transition provided by an allophone rule. There are two types of errors one can look for in such a matrix: those about which it is relatively easy to do something, and those about which it is not. If, for example, /v/ were equally confusable with /m/ and /đ/, there would be little one could do apart from redesigning the synthesizer, as there is not, at present, a nasal channel; raising the transition loci of the second and third formant will lead to a /đ/ response, lowering them to a stronger /m/ response. The occasional unsystematic voicing errors shown by the subjects are of this type also. As all the known voicing cues (voicing lag, burst/friction intensity, duration, length of previous vowel, first formant intensity cutback) are already handled by the rules, it must just be assumed that naive subjects are always going to make a few such errors, as in real speech. It

TABLE I

Confusion Matrix of Response Frequencies Totaled for 10 Subjects over (A) Consonants (B) Vowels

(A)

RECEIVED

SENT	b	d	g	p	t	k	m	n	ŋ	f	θ	s	ʃ	v	ð	z	tʃ	dʒ	w	r	l	y	h
b	33	2	1				3							1									
d		39			1																		
g			38			2																	
p				38	1	1																	
t				1	38	1																	
k			1		2	37																	
m	1						37	2															
n							3	35													2		
ŋ						1		3	16														
f										35	1		1	2	1								
θ										2	30	8											
s											1	31	1			7							
ʃ						1						1	37				1						
v							4	3		6				23	4								
ð	1	2					2	5		2	1			1	26								
z		2						4				2		1		30					1		
tʃ					3												33	2					
dʒ		4									1						8	27					
w																			17	3			
r																				20			
l							2														18		
y																						20	
h						1																	19

(B)

RECEIVED

SENT	i	ɜ	ɑ	ɔ	u	ɪ	ɛ	ae	ʌ	ɒ	ʊ	ɑɪ	oɪ	ɑʊ	oʊ	eɪ
i	16					4										
ɜ		20														
ɑ			19							1						
ɔ		1		19												
u					20											
ɪ						20										
ɛ							19	1								
ae								20								
ʌ								2	17	1						
ɒ				9						11						
ʊ											20					
ɑɪ												20				
oɪ												1	18		1	
ɑʊ														20		
oʊ															20	
eɪ																20

should be remarked that the rules are intended to provide speech at a conversational rate.

The other more interesting class of errors is that which carries implications for the improvement of the rules. Occasional errors in the vowel matrices show that some improvements in durational or diphthongization factors are in order, as the score here ought to be 100 percent (unless subjects are being confused by the orthography). The affricates are frequently confused, on the one hand, with palatal fricatives and, on the other, with the alveolar and velar stops. Here, improvement is likely only by abandoning the limitation that no phone can have two parts. Considering the inexperience of the subjects, the picture in the stops and nasals appears satisfactory, with the possible exception of /n/. There is, however, a lot of confusion in the fricatives; this can be divided into two sources, manner and place confusion. The place confusion of interdentals with labiodentals and alveolars is common even with real speech, and can probably be improved only slightly by elaborating the fricative facilities on the synthesizer; the practical advantage in this may be rather small. More interesting are the manner confusions on voiced fricatives /v, ð, z/ and affricate /dʒ/, which arise from a specific piece of the synthesis strategy. Voicing and friction are continuous through the steady-state portion, there being no pulsed friction. The presence of nonvocalic voicing leads to confusion with nasals, and to offset this, the intensity of friction is rather too high, which, in turn, leads to a deterioration of the voiced–voiceless distinction, even though that is aided by the presence of the other chief cues for fricative voicing. There are three evident solutions; one is to make these phonemes two-part elements, the second element being devoiced friction which would be inhibited in initial position. A second solution is to introduce pulsed friction, but this is restricted by the sampling rate of the speech parameters, unless extensive synthesizer modifications are made, and harmonic components are still necessary; thus, nasal confusions would not necessarily be abolished for the weaker friction phonemes /v/ and /ð/. The third method, in fact the one adopted, is to suspend excitation on the first part of the closed duration of the consonant; thus, brevity and a silent interval abolish nasal confusions, intensity of friction and formants abolish stop confusions, and long voicing lead abolishes voicing confusions. This solution makes the voiced fricatives like affricates in some respects, further forcing the adoption of a two-part element strategy for affricates.

Performance in the semivowels is high, except for some /r/ responses for an intended /w/. These all occurred in only one of the two test contexts, preceding /eI/; it appears the ambiguity is to be removed, not by postulating some allophonic interaction, but by changing the table entries for /w/.

In the vowel confusion matrix, there are two off-diagonal entries large enough to consider. The /I/ responses for /i/ all occurred in one of the two test contexts, preceding an alveolar. The /i/ is diphthongized (narrow transcription might be Iiy), but it appears there is not enough of the highly constricted portion to be registered as significant before an alveolar, with its already high second and third formants. This has been modified.

For the short vowel /ɒ/ which does not occur in General American, there are many long-vowel /ɔ/ responses; all of these occur in only one of the two test contexts, following initial /r/. Data of Lehiste[3] suggest no specific interactions of vowels in this area with initial /r/, so the solution is probably to be sought in modifications in /ɒ/ duration and /r/ transition characteristics.

Thus, explanations and remedies can be suggested for the largest off-diagonal entries in both vowel and consonant matrices, although there will always be a few mistakes in fricative place and voicing of all consonants.

If the sources of error discussed above were to be removed, but still allowing fricative voicing errors, the consonant intelligibility would be 92.5 percent and the vowel intelligibility 97.5 percent. We are proceeding to test these estimates. The estimated intelligibility values probably approach the limit attainable with this combination of system and listeners (although improvement in the naturalness of the synthetic speech is still possible). From the point of view of applications employing connected speech, the residual consonant voicing errors and fricative place confusions should not be troublesome.

Allophone Rules

The rhyme tests are not really sensitive to deficiencies or merits of the allophone rules, although these rules were employed in generating the stimuli to improve quality. We have neglected allophonic effects that might be manifest in the spectrum rather than in durations, and there is evidence that this is appropriate.[1] In the present program, there are three broad categories of allophone rules: 1) partly arbitrary ones, such as those designed to produce diphthongs (though even here simple general rules would have psychological and linguistic implications); 2) nondurational position modifiers, chiefly associated with positional variants of stops, leading, for example, to increased burst intensity and voicing lag before a stressed vowel; and 3) durational modifiers.

The last category is interesting because there are so many influences to be taken into account. An attempt has been made to derive a coherent set of rules for segment durations that will handle all the relevant conditioning influences. Sets of rules have not yet been tested formally, because there are still deficiencies, although they are apparently not so gross as to affect intelligibility. But the allophone rules do make some contribution to the quality of the speech.

The rules in the durational subset

1) lengthen elements in stressed syllables by factors depending on the element class;

2) lengthen prepausal vowel, postvocalic consonants, and final phone by different factors;

3) abbreviate closure in stop clusters;

4) abbreviate semivowels in initial clusters;

5) moderately abbreviate other cluster members;

6) lengthen vowels before voiced consonants.

The last of these rules, although it must be taken into account, is not so important in RP as in GA. Usually, it has to be applied after rules 1) and 2), because its effect is additive in nature, while theirs is approximately multiplicative. Perceptually, combining rules 3), 4), and 5) presents little problem, as sensitivity for duration differences in these cases is not great; but combination of rules 1), 2), and 6) is difficult because all three conditions appear to interact with the distinction between phonemically long and short vowels (which has more force in RP than GA). This creates some problems in the logical format of the actual rules.

Ideally, there would be as many, and only as many, flow paths through the allophone rules as there are major phoneme classes; e.g., vowels, semivowels, consonants. Ideally, the effects of application of a rule would be independent of the application of any other rule in the set when the set is applied in proper order. Here, phonometry is merging with symbolic logic; but that is what an adequate theoretical treatment of allophones would require, and no such treatment has yet been made.

Formal tests are needed to validate the set of allophone rules with respect to the conditioning environments tested, the exact modifications produced, and the logical structure of the set of rules. There are also some assumptions fundamental to the rules that require testing, but to which an intelligibility measure may not be sensitive. We have already been forced to abandon the limitation of no two-part phonemes and of continuous voicing in voiced fricatives. Other assumptions that require testing include the invariance of transition rate with stress and pause, and the assumed insensitivity of consonants to place context. The level of performance of the system appears to be high enough to test these subtler points.

The interaction of allophone rules may seem far from the "simple program" of this paper's title, but it is estimated that less than 25 allophone rules in all are needed to eliminate the perceivable deficiencies of phonemically produced speech. Hence, for either a practical output device or a linguistic model, synthesis-by-rule meets simplicity criteria better than the desperate solution of syllable inventories. (The same may not be true of input devices.) Within the class of synthesis-by-rule schemes, the present one achieves some simplicity by reducing the segment inventory, eliminating nonvocalic transitions, and postulating a few cumulatively applied allophone rules. This much simplicity does not appear to detract from intelligibility.

Acknowledgment

The rhyme tests used were two (slightly modified) from a set of five composed by Patricia Dooley.

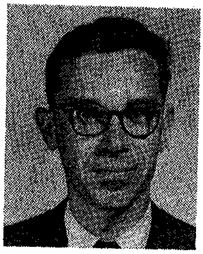
References

[1] M. P. Haggard, "Perceptual study of English /1/ allophones," *J. Acoust. Soc. Am.*, vol. 42, p. 1581, June 1967.

[2] J. N. Holmes, I. G. Mattingly, and J. N. Shearme, "Speech synthesis by rule," *Language and Speech*, vol. 7, p. 127, July–September 1964.

[3] I. Lehiste, "Acoustical characteristics of selected English consonants," *Internat'l J. Am. Linguistics*, pt. 4, vol. 30, July 1964.

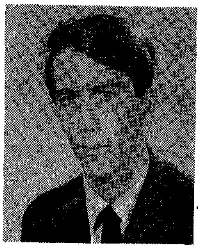
[4] I. G. Mattingly, "Experimental methods for speech synthesis by rule," presented at the Conference on Speech Communication and Processing, November 1967. See also *IEEE Trans. Audio and Electroacoustics*, vol. AU-16, June 1968 (to be published).



Ignatius G. Mattingly was born in Detroit, Mich., on November 22, 1927. He received the B.A. degree from Yale University, New Haven, Conn., in 1947, and the M.A. degree in linguistics from Harvard University, Cambridge, Mass., in 1959.

He has taught at Groton School (1947–1948) and Yale University (1950–1951). From 1951 to 1966, he was employed by the Department of Defense. Since 1966, he has been associated with Haskins Laboratories, New York, N. Y., and has taught at the University of Connecticut, Storrs. As a Guest Researcher at the Joint Speech Research Unit, Eastcote, England (1963–1964), he became interested in speech synthesis by rule, and has published a number of papers on this subject.

Mr. Mattingly is a member of the Technical Committee on Speech of the Acoustical Society of America, the Linguistic Society of America, the Linguistic Circle of New York, and Phi Beta Kappa.



Mark P. Haggard was born in London, England, on December 26, 1942. He received the M.A. degree in psychology at Edinburgh University, Edinburgh, Scotland, in 1963, where he also did some work on speech waveform correlates of neurological disorders in the Department of Phonetics. He received the Ph.D. degree from Cambridge University, Cambridge, England.

In 1966–1967 he worked at the Haskins Laboratories, New York, N. Y., on the synthesis of speech. He is currently engaged in lecturing at Cambridge University, where he is also Fellow of Corpus Christi College. His chief research interests are psychoacoustics, speech production, and speech perception, on which he has published a number of short papers.

Dr. Haggard is a member of the Acoustical Society of America.

Reprinted from IEEE TRANSACTIONS
ON *AUDIO AND ELECTROACOUSTICS*
Volume AU-16, Number 1, March, 1968
pp. 95-99