Some Rules for Synthesis of General American English

Ignatius G. Mattingly†

Haskins Laboratories, New York City

A system for synthesis of speech by rule is now in operation at the Laboratories. The system consists of a parallel resonance synthesizer parametrically controlled from a laboratory computer, and a computer program. Given a set of properly formatted rules for a language or dialect, and a string of phonemes representing an utterance, the program calculates the required parameter values and transmits them to the synthesizer. The experimenter can type the phonemic string on the computer typewriter and hear the synthesized utterance immediately. Rules have been written for Southern British English and for General American English (GA). Both sets of rules build on earlier studies in synthesis by rule, notably those of Ingemann (reported by Liberman et al., 1959), Holmes et al. (1964) and Mattingly (1966a), and exploit the many investigations of the acoustic cues for speech perception reported during the past 20 years. (See Mattingly, 1968a for bibliography). In previous status reports and other publications some account has been given of the general approach (Mattingly, 1966b, 1967), the computer program (Mattingly, 1968b) and the rules

†Also of University of Connecticut
for Southern British (Haggard and Mattingly, 1968); and an extended account of the system, including a description of the rules for General American, has appeared as a Supplement to the Status Reports (Mattingly, 1968a). We do not propose to repeat that description in detail here, but merely to give some illustrations of how the synthesis-by-rule program facilitates the statement of a few important rules of General American phonology.

It will be recalled that three types of rules are permitted by the synthesis-by-rule program: (1) segmental rules, consisting of one (or more) tables of coded values for each phoneme. Each table specifies the duration of the phoneme and, for all synthesizer parameters except $\text{FØ}$, the steady-state values for the synthetic phonemes and the character of the transitions to and from adjacent phonemes; (2) allophone rules, which make it possible to change the coded values in the segmental rule tables if the phoneme occurs in one or more of a limited set of contexts; (3) $\text{FØ}$ rules, consisting of a number of variables which determine the pitch contour during the synthesized utterance. The segmental rules and the allophone rules are used for synthesis of segmental phonemes; the $\text{FØ}$ rules and the allophone rules for synthesis of prosodic and junctural features.

The workings of the segmental rules can be illustrated by the tables for the GA stops /p, t, k, b, d, g/, represented for purposes of computer input as P, T, K, B, D, G (this notation allows us also to distinguish between "real" phonemes and their synthetic counterparts). The cues which identify these sounds are of three types: (1) those which distinguish stops from other manner classes - the closure and the subsequent release, accompanied by a burst; (2) those which distinguish the voiceless from the voiced stops - the absence of closure voic-
ing, aspiration of the first part of the following phoneme, F1 "cutback"; (3) those which distinguish the three different positions of articulation from one another - the F2 and F3 transitions and the burst spectra. It is not inconceivable that some ad hoc scheme could successfully synthesize the stops without being directly concerned with these cues, but any theoretically interesting set of rules ought to take explicit account of them.

Most of the manner and voicing rules for the stops are provided by the variables in the segmental tables, which determine the excitation, i.e., which of the four possible values of the buzz-hiss parameter B-H2 (silence, buzz, hiss, mixed buzz and hiss) are to be used at each moment. These variables are the "initial B-H2 value", the "final B-H2 value", BH, CV and DV. BH is the proportion of the phoneme duration using the initial B-H2 value when CV = 0. CV is the proportion of BH using closure voicing (during which a stock set of appropriate formant frequency and amplitude parameter values is used instead of calculated values). DV is the duration of devoicing (substitution of hiss for normal B-H2 values) at the beginning of the following phoneme. BH and CV are assigned values between 0 and 1. Thus the excitation of a phoneme could consist of a period of closure voicing, a period during which the excitation is determined by the initial B-H2 value; and a period during which it is determined by the final B-H2 value. Furthermore, the excitation of the first part of the following phoneme can be changed to hiss.

For the stops P, T, K, the initial B-H2 value is silence, the final value is hiss, BH = 14/16, CV = 0. DV = 5 (25 msec) for P and T and 8 (40 msec) for K. Thus there is a silent closure through most of the duration of the phoneme, followed by a short period of hiss which serves as the stop burst, fol-
lowed by aspiration of F2 and F3 of the first 25 (or 40) msec of the following phoneme. Hiss excitation does not excite the first formant of the synthesizer, so that F1 is "cut back" for the duration of the aspiration. Since the first part of a voiceless stop is silence, the steady-state formant amplitude and formant frequency values in the table characterize only the spectrum of the burst. For B, D, G the initial B-H2 value is mixed buzz-hiss, the final value is not used, \( \text{BH} = 16/16 \), \( \text{CV} = 16/16 \) for B and 14/16 for D, G, and \( \text{DV} = 0 \). Thus there is closure voicing for most of the duration of the phoneme, followed (except in the case of B) by a burst with mixed excitation, its spectrum characterized by the steady-state values in the table.

The remaining important manner cue for stops is the transition of F1 towards zero in the preceding phoneme and the rising F1 transition in the following phoneme. The program follows the method of Holmes et al. (1964) in treating a formant transition as a movement between the steady-state formant frequencies of a pair of successive phonemes. The table for one of these phonemes (the choice depends on the manner classes of the two) determines the rate and direction of the transition, and this phoneme is considered the dominant phoneme \( p_{\text{X}} \), while the other is the adjacent phoneme \( p_{\text{H}} \). A formant \( j \) moves according to a simple non-linear rule between the steady-state \( \text{S}_{\text{X}j}(p_{\text{X}}) \) and a value \( \text{BV}_{\text{X}j} \) assigned to the boundary of \( p_{\text{X}} \) and \( p_{\text{H}} \); and between \( \text{BV}_{\text{H}j} \) and \( \text{S}_{\text{H}j}(p_{\text{H}}) \). Stored in the table for \( p_{\text{X}} \) are two quantities \( \text{C}_{\text{X}j}(p_{\text{X}}) \) and \( \text{W}_{\text{X}j}(p_{\text{X}}) \), and \( \text{BV}_{\text{X}j} = \text{C}_{\text{X}j}(p_{\text{X}}) + \text{W}_{\text{X}j}(p_{\text{X}}) \cdot \text{S}_{\text{X}j}(p_{\text{X}}) \).

\( \text{DH}_{\text{X}j}(p_{\text{X}}) \), the duration (in 5 msec time units) of the part of the transition within the adjacent phoneme, and \( \text{DG}_{\text{X}j}(p_{\text{X}}) \), the duration of the part of the transition within the dominant phoneme, are also stored in the table for \( p_{\text{X}} \).
By assigning proper values to these variables the required movement for F1 can be specified. A stop is always dominant, i.e., always $p_z$, and the preceding phonemes are each always $p_n$. The boundaries of the stop correspond to the instant of closure and the instant just after the burst. $c_{F1} = 0$ (100 Hz) for voiced stops and 4 (208 Hz) for voiceless stops. $w_{F1} = 0$, $d_{H_{F1}} = 9$ (45 msec), $d_{G_{F1}} = 0$. In other words, F1 moves down from its steady-state value in the preceding phoneme to 100 Hz (or 208 Hz) at the leading boundary of the stop, and up from the same frequency at the trailing boundary of the stop to the steady-state of the following phoneme. Both transitions take 45 msec and there is no transition within the boundaries of the stop.

The same procedure is used to calculate the F2 transitions which cue place of articulation. For T, D, $c_{F2}$ and $w_{F2}$ are chosen so that $bv_{F2}$ is half the distance between the /t/ locus - 1800 Hz - and $s_{F2}(p_n)$. $d_{H_{F2}} = 9$ (45 msec) and $d_{G_{F2}} = 0$. Thus the F2 transition approaches but does not reach the locus. For P, B, a $bv_{F2}$ lower by about 300 Hz than $s_{F2}(p_n)$ is used. For K, G, the problem is somewhat more complicated. Before front vowels a locus of 3000 Hz works well, but before other vowels no clear locus is defined. An allophone rule is used which provides that before front vowels the values for $c_{F2}$ and $w_{F2}$ are replaced by values appropriate to the 3000 Hz locus. Otherwise the standard $c_{F2}$ and $w_{F2}$ table values, which call for a $bv_{F2}$ about 300 Hz higher than $s_{F2}(p_n)$, are used. These values have not yielded entirely satisfactory results: K WU sounds like /pu/ instead of /ku/. F3 transitions, which we shall not describe in detail, are calculated in much the same way as F2 transitions.
The interplay of allophone rules and $\emptyset$ rules in the specification of prosodic features is best illustrated by the rules for stress and prominence. By stress we mean lexical stress and the stress which subordinates one member of a compound or phrase to another. By prominence we mean the change (usually a rise) in $\emptyset$ which occurs on one or more stressed syllables of a sentence under certain syntactic or semantic conditions. Thus prominence implies stress, but not conversely. The synthesis-by-rule program allows for prominence and three levels of stress: major, minor and low stress (actually, there are doubtless an indefinite number of stress levels, but three levels seem sufficient for almost any sentence). Major stress with prominence, major stress without prominence, and minor stress are marked by appropriate symbols in the input string; minor-stressed syllables are unmarked.

The chief cue for stress, as we have defined it, is vowel length. Thus an allophone rule for major stress replaces $D$ (the standard, or minor stress duration value of a phoneme) by an alternate table value which is usually greater than the standard value, while the allophone rule for low stress replaces $D$ by a second alternate table value which is usually smaller than the standard value. These two rules are not the only allophone rules affecting vowel length: there are rules which increase $D$ for a vowel (as for certain other manner classes) by a specified factor before a voiced consonant, before juncture, and before a pause. Some of these rules add a constant to the previously determined value of $D$; others multiply the previously determined $D$ by some value. It is thus interesting and important to order these rules. This proves to be a delicate business, and is not always successful; a major-stressed vowel in a prepausal syllable before a voiced consonant sometimes turns out to be just too long.
Other allophone rules increase the burst intensity (i.e., $S_{OA}$, the steady-state value of the overall amplitude parameter) and the length of aspiration ($DV$) for voiceless stops before major-stressed syllables, while before low-stressed syllables the burst of voiceless stops is eliminated ($PH = 1$) and $D$ for voiced stops is reduced. After a vowel, and before a low-stressed vowel, a flapped $T$ is produced by allophone rules which reduce $D$, eliminate the aspiration ($DV = 0$) and change the excitation to closure voicing ($PH = 1$, $CV = 1$).

Prominence, like other $\Phi$ effects, is controlled by the intonation table. In the synthesis-by-rule program, calculations depending on the intonation table are carried out in "pitch units," equal to half a semitone; a range of about one and one-half octaves is available. The program measures the slope of the pitch in pitch units/5 msec time unit.

The GA intonation table provides a very gradual pitch slope ($-1/32$) during the "head" of the breath-group, followed by one of three intonation contours. Prominence represents a marked departure from this gradual slope, and is described by five values in the intonation table: $FV_2$, the slope of a prominent syllable after a period of voicelessness; $FV_3$, the pitch change during a prominent syllable if there has been no preceding voiceless period; $FV_4$, the pitch change before a prominently stressed syllable after a voiceless period; $FV_6$, the pitch change after a prominent syllable and before a voiceless period; $FV_8$, the slope of a non-prominent syllable following a prominent syllable.

As the definitions of the variables suggest, prominent syllables in which the voiced portion is preceded by one or more voiceless phonemes are treated differently from syllables before which there has been continuity of voicing since the
previous syllable nucleus. In the first case, the pitch skips up 2 semitones (\(FV_4 = 4\)) above its value before the cessation of voicing, then levels off for the duration of the syllable (\(FV_2 = 0\)). In the second case there is no skip, but a pitch rise of 2 semitones is distributed throughout the syllable (\(FV_3 = 4\)). In either case, when the voiced portion of the syllable is followed by one or more voiced phonemes, the pitch slopes down 2 semitones (\(FV_6 = -4\)). The slope of the next syllable is steeper than that of an ordinary non-prominent syllable (\(FV_8 = -6/32\)).

Thus, since a prominent syllable is by definition one having major stress, it is marked in the GA rules by a pitch rise, by lengthening of the vowel, and by the increased aspiration and louder burst of any preceding voiceless stop.

These two illustrations will give some notion of how the machinery of the synthesis by rule program is applied to the writing of rules for specific areas of GA phonology. Neither these rules nor those which we have not described here are offered as definitive or completely correct; it is felt, however, that reasonable progress has been made in writing rules within the framework of the program and that a great deal of further progress is possible.
References


