Syllable Synthesis*

Ignatius G. Mattingly†

ABSTRACT

A scheme for synthesis by rule based on the phonetic syllable is described. A syllable-feature specification of the utterance to be synthesized determines a pattern of articulatory influences; these influences in turn determine the parameter values of the synthesizer.

For quite a long time, as the first slide (Figure 1) may remind you, my colleagues at Haskins Laboratories have been insisting that speech is a code, and that the encoding unit is the phonetic syllable. As the result of the merging of various coarticulatory influences, the correlates of the phonemes at the acoustic level are, in the vivid phrase of Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967), "overlapped or shingled, one onto another," yielding "irreducible segments of approximately syllabic dimensions." This observation should, indeed, be generalized to include the articulatory level as well (MacNeilage and DeClerk, 1968). In this view of the syllable [which of course goes back at least to Stetson (1951)], my colleagues have been encouraged by the findings of Koshevnikov and Chistovich (1965).

But the appeal of the phonetic syllable as an encoding unit does not rest merely on empirical observations as to the unsegmentability of anything smaller. There is not time to make the theoretical case for the syllable at length, but I would at least point out how nice it would be if it were possible to order freely the units of an ideal phonetic transcription at each prosodic level. Because of phonotactic restrictions, this condition is clearly out of the question if these units are conventional phonetic segments, but seems quite reasonable if the units are phonetic syllables. Though overlap between the physical manifestations of adjacent syllables occurs, the principle of free ordering in the transcription will be preserved as long as such overlap is predictable from the specification of the individual syllables.

From this point of view, the syllable is a cyclic process, passing from onset to peak to offset as the vocal tract moves from a more closed to a more open to a more closed configuration. The process can be realized in many

*This paper was presented at the 92nd Meeting of the Acoustical Society of America, San Diego, California, 15-19 November 1976.

†Also University of Connecticut.

Figure 1: Interaction of consonantal and vocalic influences in [bæg]. After Liberman (1972).
different ways, depending upon the phonetic choices of the speaker. In the
ideal phonetic transcription of an utterance, these choices are the values
assigned to syllable features of the sort suggested by Fujimura (1975,
1976), and phonetic segments have no formal status. Thus, not only the
difference between [pa] and [ba], but also the difference between [pa] and
[pla] depend upon a feature selection. The articulatory and acoustic
consequences of a particular choice are determined by the syllabic process
and may, in principle, extend throughout the physical manifestation of the
syllable.

I hope it is clear that I am speaking of phonetic rather than of
phonological syllables. If there are such things as phonological syllables--
the matter is unsettled--they do not in general correspond one-to-one with
phonetic syllables, any more than phonemes correspond one-to-one with phones.
And, by the same token, if phonological syllables do not exist, the case for
phonetic syllables is unaffected. A representation in terms of syllable
features at the phonetic level is a priori entirely consistent with a
segmental representation at the phonological level, and would not necessarily
entail any fundamental revision of generative phonology. One of the motiva-
tions for generative phonology, in fact, is that phonological units do not
necessarily correspond with phonetic units (Chomsky, 1964).

What I have been saying places a heavy explanatory burden on the concept
of the phonetic syllable, and it will be credible only to the extent that
syllabic processes can be shown to be orderly and more explicit. Thus, the
case for the syllable will be enhanced if phonotactic restrictions, as well
as much of what is now regarded as allophonic variation, can be interpreted
as arising naturally from inherent properties of syllables. Synthesis by
rule is an attractive tool for this undertaking.

Recently, we have begun work at Haskins on a new synthesis-by-rule
program. In this new scheme, the phonetic syllable has a central role. The
input to the synthesis program is a phonetic transcription of an utterance in
syllabic, rather than segmental feature values. At present, the features are
binary, which simplifies the transcription, but we have no strong commit-
tment to binarity at the phonetic level. A series of ordered rules relates the
feature values of the transcription to the variables used in the routine that
calculates parameter values for Rod McGuire's software simulation of OVE III
(Liljencrants, 1968). Since the program has as yet no phonology, it is not
at present a practical vehicle for synthesizing quantities of text. Nor has
consideration yet been given to stress and intonation, though the syllable
plays a crucial role in these matters.

In the routine for calculating parameter values, the character of a
syllable is considered to be determined by numerous influences: the vowels
of the previous, current and following syllables, the final consonants of the
previous and current syllables, and the initial consonants of the current and
following syllables. With each such influence is associated a set of target

---

1 Fujimura (1976) Syllable as the unit of speech synthesis. Unpublished
memorandum, Bell Laboratories.
parameter values and a curve that represents the extent of the influence over time. An influence curve is a modified exponential function of the form \( \kappa e^{-\beta t} \). In this function [similar to the one used by Lindblom (1963) in his well-known model of consonant-vowel coarticulation], the coefficient \( \kappa \) determines the effective time of onset of an influence, and the exponent \( \beta \) determines its rate of growth. On phonetic grounds, these properties of the function are appealing, since one would expect both the shape and the relative timing of the various influence curves to be significant variables. Of course, there are other functions that might possibly have been used instead. The value of the function is restricted to the range 0...1, since it is used as a weight, and at a certain time \( \tau(x) \) after the notional beginning of the syllable cycle, \( \beta \) becomes negative, so that the influence will begin to diminish. The target values, and the values of \( \kappa, \beta, \tau(x) \), and other variables are assigned by the rules.

It might be objected that the notion of an "influence" simply reintroduces the phonetic segment in a new guise, particularly when I refer to the influences of consonants and vowels, and employ the conventional terms for manner classes. But unlike phonetic segments, influences are not linearly ordered; their temporal relationship is more complex than that. And "consonant," "vowel," and the various manner class terms are to be understood not as segment categories, but as labels given to various recognizable aspects of the syllabic cycle by which they are defined.

Because of our particular interest in the temporal patterns of events within the syllable, we have provided various ways to control these patterns in the program. As we have just seen, \( \kappa \) controls the effective onset of an influence; by manipulating this variable, different degrees of consonantal and vocalic coarticulation may be provided. Since the moment when an influence begins to diminish is a variable, articulatory holds for stops and fricatives can be represented. Moreover, each influence can potentially increase the duration of the syllable by a certain amount. If such an increment is called for, the onsets of syllable-final and following-syllable influences are postponed by appropriately reducing their \( \kappa \) values.

The actual parameter values for a particular 5-msec sample of speech are derived by an iterative calculation. The influences are regarded as ordered, from vowels to fricatives to stops. At each iteration, the value computed for a parameter is

\[
V_i = V_{i-1} + I_i(T_i - V_{i-1})
\]

that is, the weighted sum of the target values associated with the influence and the value computed at the preceding iteration, the relative weighting being determined by the value of the influence function at that point in the syllable. (At the first iteration, the target value for the vowel of the previous syllable serves as the seed value \( V_0 \).) Because of the large number of influences, the burden of calculation would be considerable were it not

---

\(^2\)Our investigations of syllable duration are reported in a paper read by Linda Shockey at an earlier session (Shockey and Mattingly, 1976).
Figure 2: Influence functions for current vowel and new vowel for [a] preceded by [ə] and followed by [ɛ], and resulting formant trajectories.
Figure 3: Effect of adding influence of diphthong on syllable shown in Figure 2.
Figure 4: Effect of adding influence of current initial glide and following initial glide on syllable shown in Figure 3.
that at any one time, many potential influences are inactive, that is, have values near zero, and may simply be ignored.

The next three slides (Figures 2, 3, 4) illustrate how the overlapping of various influences is realized. This slide (Figure 1) shows an [a] assumed to have been preceded by [ə] and followed by [ɛ]. The curve with black circles in the upper portions of the slide shows the increasing influence of the [ə] at the expense of the [a] of the preceding syllable. The curve with white circles shows the increasing influence of the [ɛ] of the following syllable at the expense of the [a]. The lower portion of the slide (Figure 1) shows the target formant frequencies for all three vowels and the formant movements resulting from their influence.

In this slide (Figure 3) the influence of a final palatal glide is interposed, in addition to the other influences, to give the diphthong [ar], and the formants change accordingly.

Finally, in Figure 4, the influences of an initial [y] glide in the [ar] syllable and of an initial [w] glide in the following syllable are superimposed upon the other influences.

This way of calculating parameter values will be recognized as a generalization of the method used by Holmes, Mattingly and Shearme (1964) and by the earlier Haskins programs for calculating formant transitions (Mattingly, 1968a, 1968b; Kuhn, 1973), in which the "boundary value" used as a basis for interpolation was the weighted sum of the target frequencies of two adjacent phones. It is also analogous, as Tim Rand has pointed out, to a series of filters, each of which corresponds to an "influence."

The scheme, as described so far, is quite general, and could be implemented in terms of articulatory gestures, or vocal tract shapes, or formant movements, depending upon the choice of parameters. The most interesting and satisfying implementation would be the articulatory one, but because we are anxious to explore temporal questions as soon as possible, we are beginning with an acoustic version.

REFERENCES

Acoustics Speech and Signal Processing ASSP-23, 82-87.
Holmes, J. N., I. G. Mattingly, and J. N. Shearme. (1964) Speech synthesis
Kozhevnikov, V. A. and L. A. Chistovich. (1965) Speech Articulation and
Perception, trans. from Rech' Artikulyatsiya, i vospriyatiye (Moscow-
Kuhn, G. M. (1973) A two-pass procedure for synthesis by rule.
In The Neurosciences Third Study Program, ed. by F. O. Schmitt and


