*THE COGNITIVE REPRESENTATION OF SPEECH*
*T. Myers, J. Laver, J. Anderson (editors)*


358

# Phonetic Representation and Speech Synthesis by Rule

IGNATIUS G. MATTINGLY

*University of Connecticut and Haskins Laboratories*

While a computer model of performance in speech production is certainly conceivable, the title of this seminar, "Speech Synthesis Programs as Models of Speech Production," seems incorrect as a characterization of existing systems for synthesis of speech by rule. Insofar as these systems have had more than the purely pragmatic goal of converting written utterances into spoken utterances, the objective has been a more modest one, but one that is nonetheless directly related to the theme of this Symposium, "The Cognitive Representation of Speech." The objective has been to elucidate the nature of the phonetic representation shared by speaker and hearer (and recorded, after a fashion, in a conventional phonetic transcription) and the relationship of this representation to physical events. A simulation of the mental activity underlying actual speech production would be a far more ambitious project, though one well worth attempting.

Synthesis-by-rule has also, of course, considered other interesting questions: the relationship of the conventional orthography to the phonemic representation, and the relationship of the latter to phonetic representation. But its central concern, surely, has been to relate the phonetic representation either to the articulatory or to the acoustic parameters of continuous speech. The practitioner of synthesis-by-rule cannot avoid this concern, for his rules must somehow specify the physical correlates of the elements of the phonetic representation, and his program must necessarily include some kind of algorithm for getting from one phonetic element to the next.

The simplest conceivable synthesis-by-rule system would be one in which each phonetic element is associated with an articulatory or acoustic segment of specified duation (or of a sequence of such segments). Since the inventory would be small, prerecorded segments of natural speech rather than parametrically-specified synthetic segments could be used. This system would have only one "rule": concatenate segments corresponding to successive phonetic elements. Such schemes have indeed been proposed (Harris, 1953), but it was appreciated quite early by practitioners of speech synthesis by rule that the acoustic record (and *a fortiori* the articulatory record) contained intervals that could not reasonably be viewed as steady-state segments correlated with single phonetic elements, and that these intervals (which were regarded as "transitions" connecting "true" steady-state segments) were perceptually extremely important (Liberman, Ingemann, Lisker, Delattre & Cooper, 1959).

Thus, the synthesis-by-rule systems developed in the 60's (see Mattingly, 1974, for a review), though differing in detail, had in common the requirement that transitions, as well as steady states, had to be described in the synthesis rules. In the synthesis-by-rule system described by Holmes, Mattingly & Shearme (1964), for example, each "phoneme table" specifies, for each parameter, not only a steady-

state value but also a contribution to a "boundary value," and the moment-by-moment values during a transition are calculated by linear interpolation from the steady-state value for the current phonetic element to the boundary value, and from the boundary value to the steady-state value for the following element. The interpretation of the phonetic representation implicit in these systems was thus consistent with the concept of segmentation proposed by Pike (1943): speech consists of "level segments," during which one or more articulatory "crests or troughs of stricture" are maintained, connected by "glide" segments, during which one or more strictures are being applied or released. Successive elements of the phonetic representation referred to successive level segments, from which the glides could be predicted.

This approach to synthesis-by-rule, though demonstrably capable of producing speech that was by conventional measures intelligible (Nye and Gaitenby, 1973; Pisoni, 1979), ran into three related sorts of difficulty. First, it can be observed in natural speech that the acoustic "steady-state" segment associated with a particular phonetic element, and likewise the transitions from the preceding and to the following segments, vary according to the immediate phonetic context in ways that are clearly of perceptual importance. For practical purposes, this variation could be dealt with by writing allophone rules to modify the phonetic specification according to context, or by allowing the transition from the preceding steady-state segment and the transition to the following steady-state segment to overlap, eliminating the current steady-state segment (Holmes, et al., 1964). Either approach was an admission of a deficiency in the Pikean view of phonetic representation (from an articulatory standpoint, the deficiencies of the Pikean view are even more glaring since the simultaneous maintenance of several constrictions over a period of time — an articulatory "steady-state" — is quite unusual).

Variations that depended on non-adjacent phonetic context were less easily dealt with. Since the algorithms used in these systems considered only two consecutive phonetic elements at a time, this kind of contextual variation could be handled only by frankly *ad hoc* procedures. Though it was generally assumed that such variation was on the whole of little perceptual importance, its very existence posed further problems for the Pikean view.

The final, and theoretically most damning, difficulty with the Pikean approach was that the durations of steady-state and transitional segments are subject to extensive contextual variation and had to be assigned completely *ad hoc.* Even these *adhoc* assignments were notoriously unsuccessful in producing realistic speech-timing patterns. But what is at issue is not just whether the durations for phonetic elements could be adequately specified within a segment-by-segment framework (or even by a more elaborate framework involving higher-order prosodic units), but whether the elements of a phonetic representation can be said to have durations at all.

The fact that some synthesis-by-rule systems have respectable intelligibility scores should not lead anyone to suppose that these problems are of no practical importance. It is clear that the speech produced by these systems, though intelligible places a much greater load on short-term memory than does natural speech. Nye and Gaitenby (1974) investigated the ability of listeners to recall immediately after presentation semantically anomalous but syntactically acceptable sentences (e.g., "The wrong shot led the farm") under two conditions. In one condition, listeners heard naturally-spoken sentences; in the other, sentences synthesized with the Haskins Laboratories synthesis-by-rule system. Pisoni (1979) later used the same test in an evaluation of Klatt's (1976) synthesis-by-rule system. The percentages of words correctly reported were for natural speech,95%; for the Haskins system (Mattingly, 1968; Kuhn, 1973), 78%; and for the Klatt system, 78.7%.

Though both synthesis-by-rule systems have conventional intelligibility scores close to natural speech, it would seem that the unnaturalness of synthetic speech, in particular, perhaps, the failure to represent coarticulatory variation adequately and the unnaturalness of the timing of acoustic events, seriously interferes with short-term memory coding.

The objections to segmental models have often been pointed out, and it is rather surprising that practitioners of synthesis-by-rule have paid so little attention. Menzerath & Lacerda (1933), Joos (1948), Fant (1962), Liberman, Cooper, Shankweiler & Studdert-Kennedy (1967) and many others have made it clear that a many-to-many relationship between phonetic elements and acoustic segments, however defined, is the norm rather than the exception in speech, and that — in Fant's words — "several adjacent sounds [i.e., acoustic segments] may carry information on one and the same phoneme, and there is overlapping insofar as one and the same sound segment carries information on several adjacent phonemes" (1962:9). Nor is it the case that the many-to-many relationship is to be attributed solely to the merging into a single acoustic stream of effects due to several articulators, for multiple phonetic influences simultaneously affect the movements of an individual articulator (MacNeilage & Scholes, 1964; MacNeilage & DeClerk, 1969).

In other words, information in speech is transmitted in parallel, and it is this fact that makes possible higher information rates for speech than would be possible in a truly segmental process. But parallel transmission appears to present no difficulty for the speech-perception mechanism. On the contrary this mechanism seems to be specialized for decomposing an encoded period of speech into the component phonetic influences, and is rather less at home with isolated consonants or isolated vowels (Liberman, et al., 1967).

Both acoustic and articulatory models to account for the many-to-many relationship between phonetic elements and acoustic events have been proposed. Joos regarded phonetic elements as "overlapping innervation waves" and showed that the formant trajectory for a $C_1VC_2$ syllable could be decomposed into a vowel "layer", an initial-consonant layer and a final-consonant layer (Joos, 1948:109,125). In the model proposed by Ohman (1967) to account for observed dynamic changes in vocal-tract shape in $V_1CV_2$ syllables, the predicted shape depends on a time function, a vowel-shape function, a consonant-shape function, and a coarticulation function associated with the consonant. Even for values of the time function where the consonant-shape function predominates, the predicted shape function depends also on the vowel function, to the extent determined by the value of the coarticulation function. Moreover, if the vowel-shape function is not constant but varies depending on the time-varying values of phonetic features, that is, if $V_1 \neq V_2$, the predicted shape at any point in the $V_1CV_2$ sequence depends on these feature-value time functions as well as on the consonant-shape and coarticulation functions. An observed vocal-tract shape during a VCV utterance is thus interpreted as the result of superimposition of a consonant shape tolerating a certain degree of coarticulation upon a changing vowel shape.

It is worth noting that the Joos and Ohman models are essentially "prosodic." That is, they treat elements of the so-called segmental sequence in the way that prosodic features are conventionally treated. The overlapping and layering of prosodic features is usually taken for granted: it would be quite unconventional to propose a division of the signal into stress and intonation segments. These models are also quite consistent with earlier attempts by phoneticians and phonologists to treat one or another "segmental" element as if it were a prosodic feature (see, for example, Hockett's discussion of "componential analysis", 1955: 129ff).

However, it is not sufficient to regard the sounds of speech merely as an inventory of "innervation functions" or (as we prefer to call them) "phonetic-influence functions" that may overlap with one another freely and to an indefinite extent. Phonetic elements are perceived as ordered, and if this ordering is not to be attributed to the existence of successive segments, some other explanation is required. Moreover, there are obvious restrictions on the co-occurrence of overlapping patterns, and corresponding restrictions on the perceived ordering of phonetic elements.

The basis for these restrictions becomes obvious if we consider what combinations of phonetic influences can in fact be effectively transmitted in parallel. If, in the utterance [pla], the onset, constriction and release for [l] were to occur entirely during the period of closure for [p], the [l] would have no acoustic correlates. But if the [l] release is delayed until after the [p] release is well advanced, information about [l] (as well as about [p] and [a]) is available both before and after the [l] release. There has to be some means of guaranteeing that this second pattern will in fact be the one that is used. Again, in stop sequences of the form $V_1 S_1 \ldots S_n V_2$, information about stops $S_2 \ldots S_{n-1}$ will be present if the release of $S_j$ is delayed relative to that of $S_{j-1}$. But the period of constriction for $S_j$, because the constriction is maximally close, will convey only manner information, and the burst will convey place information about $S_j$ itself, but little, or no information about any other phonetic element. Hence there will be no effective parallel transmission except for the periods when the $S_1$ constriction is being applied during the constriction for $V_1$ and the $S_n$ constriction is being released during the constriction for $V_2$. Thus length of stop sequences has to be severely limited, as is the case in all languages.

The general articulatory prerequisite for parallel transmission would appear to be that the constrictions for one or more closer articulations must be in the process of being released or applied in the presence of constrictions for one or more less close articulations. In terms of this formulation, the conventional ranking of manner classes according to degree of closeness (obstruents, nasals, liquids, glides, vowels) corresponds to a ranking according to the degree to which information can be encoded during the release or application of the constriction, and the inverse of this ordering, to the degree to which information can be encoded during the period of maximal constriction (Holmes et al. (1964) exploited this ranking of the manner classes to a limited extent in their synthesis-by-rule system). If parallel transmission is to be maximized, then the articulations of speech must be scheduled so that periods during which constrictions are released in rank order alternate with periods during which constrictions are applied in inverse rank order. This is of course exactly what is accomplished by the syllabic organization of speech. It would seem, therefore, that the syllable has more than a phonological or prosodic role: it is the means by which phonetic influences are scheduled so as to maximize parallel transmission.

The perception that phonetic elements are ordered now has an obvious explanation. This perception does not arise from the detection of successive segments, or even of the successive releases or applications of constrictions. It is rather the ranking of the manner classes itself that governs the percept. That is, the listener interprets the available acoustic data in terms of a framework of expectations about the structure of the syllable based on the ranking of manner classes.

Interpreting syllable structure in terms of the manner-class ranking is in itself hardly novel: Jespersen (1926) proposed such a ranking, based on "sonority," as the basis for an account of the syllable. But the argument here is that if segments are to be replaced in a phonetic model by phonetic-influence functions, syllable structure is essential for efficient speech communication and not simply a concomitant linguistic structure.

At Haskins Laboratories, we are developing a new synthesis-by-rule system in which acoustic parameters depend on the interaction of overlapping phonetic influences, and the timing of these influences is determined by the structure of phonetic syllables. For various practical reasons, we have chosen to use the acoustic parameters of a terminal-analog synthesizer rather than articulatory ones, but an essentially similar approach could be used with an articulatory synthesizer.

The phonetic elements that are considered to influence the acoustic character of the syllable in our system are the vowels of the current, preceding and following syllables, the initial consonants of the current and following syllables, and the final consonants of the current and preceding syllables (higher-level prosodic elements have not as yet been taken into consideration). With each such phonetic influence is associated a rank that depends upon the manner class of the element, and within manner class, upon temporal order; a set of target parameter values; and a time-function, ranging in value between 0 and 1, that represents the weight of the influence relative to the combined weight of all lower-ranking influences.

A phonetic-influence function is defined from the beginning of the preceding syllable to the end of the current syllable, in the case of syllable-initial articulations, or from the beginning of the current syllable to the end of the following syllable, in the case of syllable-final articulations (in this way, intersyllabic influences are taken care of). An influence function has a growth period, during which it has the form $I_t = ke^{\beta t}$ (cf. Lindblom, 1963), a possible steady-state period of duration $h$ during which $I_t = 1$, and a declining period during which $I_t = ke^{-\gamma(t-h)}$. The rate at which an influence grows (or declines) depends on $\beta$ (or $\gamma$), its effective onset time on $k$. Syllables may vary in duration according to their phonetic structure and the value of $k$ is adjusted accordingly.

Given $I_{i,t}$, the strength of the $i$th-ranking influence at time $t$, and $T_{i,j}$, the target value for the $j$th parameter associated with this influence, the parameter value reflecting the $i$th and lower-ranking influences is

$$V_{i,t,j} = V_{i-1,t,j} + I_{i,t}(T_{i,j} - V_{i-1,t,j})$$

Taking as $T_{0,j}$ the target value for the vowel of the previous syllable, the parameter value reflecting all influences can be calculated iteratively. At any particular instant, the weight of most possible influences will be zero or near zero, and computation is speeded by neglecting these influences.

The variables of this algorithm that are associated with influences of elements of each manner class are defined by an ordered set of rules. These variables include the target parameter values, the increment to syllable duration attributable to the element, the duration of the steady-state period, the times relative to the notional beginning (or end) of a syllable, when the strength of an initial (or final) influence equals .5, 1, and .5 again ($\beta$, $\gamma$ and $k$ are determined from these time-values). The definitions of these variables in the rules are conditional upon particular patterns of feature-values that might be specified in the phonetic description of the syllable. Before the parameter values are computed, the pattern of feature-values in each rule is compared with the actual phonetic description. If the rule applies, the algorithmic variables mentioned are defined according to the rules. Since the rules are ordered, a variable may well be redefined by one or more subsequent rules.

We feel that this scheme reflects more clearly the essential character of the relationship between the phonetic representation and acoustic events than our earlier synthesis-by-rule system, or other systems in which a Pikean segmentation was assumed. We hope that it will make possible the production of at least equally intelligible and more natural and more understandable synthetic speech.

*Acknowledgement* Support from the National Institutes of Health (NICHD) and the University of Connecticut Research Foundation is gratefully acknowledged.