Phonetic Aspects of Time and Timing*

Leigh Lisker+

ABSTRACT

By a definition narrow enough to exclude acoustic and physiological aspects of speech behavior, phonetics is reduced to the descriptive practice of linguists, whose judgments on the physical nature of a speech signal are primarily auditory and sympathetic proprioceptive. These judgments are for the most part embodied in a special alphabet of indeterminate size, each element of which is defined with reference to some particular state of the vocal apparatus. In general, a dimension of time is not included in the set of auditory and articulatory properties by which the different states are specified. Since, in all but a negligible number of cases, speech signals are said to not involve a single state of the vocal apparatus, but rather a sequence of such states, this sequential ordering is explicit recognition of a temporal dimension. But the time-ordered elements are themselves "timeless" unless the linguist determines that varying the duration of one or more of them serves to signal a semantic—that is, a linguistic—distinction. At this point, one of the two segments said to differ significantly in duration will often be judged to have a duration "inherently" determined by its other properties, while the other will be characterized as "long" or even "overlong." Aside from duration as a property ascribed to the segments constituting a speech signal, there are temporal aspects of speech that are less often given an explicit representation in the linguist's transcription; these are at best indirectly indicated by the so-called "junctural" marks and stress markers. One temporal aspect of speech that is regularly ignored is the feature of rate of articulation, for within certain ill-defined limits speech tempo is ad libitum.

Let me begin with a preamble to explain my understanding of "phonetic aspects of time and timing," in the present context. That understanding is to a considerable degree determined by a factor that is itself temporal, or at least temporal at one remove. I have in mind the spacial arrangement of the discussion titles on the program I was given, and my belief that it follows the conventions of written English and thus signals our chairman's wish that I speak

*This paper was presented by invitation of the 100th meeting of the American Speech and Hearing Association, Washington, D.C., 21-24 November 1975.

+Also University of Pennsylvania, Philadelphia.


113
first, with Katherine Harris, Dennis Klatt, and Peter MacNeiIlage to follow in that order. Therefore, I supposed, in preparing for this morning's business, that our discussion of the temporal organization of speech activity would begin with a consideration of certain aspects to be called "phonetic" and then go on to physiological and acoustical data and theories bearing on our topic. This order seems to imply that physiological and acoustical data comprise what are in some sense nonphonetic aspects of the speech process, and while I prefer to think of phonetics as deserving a much broader definition, it is both convenient here, and regrettably close to the general practice of some language scholars in discussing language behavior, to restrict the scope of my contribution so as to exclude in particular the subjects which Katherine Harris and Dennis Klatt will be addressing. That leaves the perceptual aspect for me to talk about--itself a broad enough subject to include a good deal that many of us might want to exclude from phonetics.

Since the scholarly caste that has for longest concerned itself with speech is the one of linguists, even though some who are called linguists would deny the study of speech activity a place in their discipline, I will for now take as reports on the phonetic aspects of speech timing those observations on the temporal properties of speech that linguists include, more or less systematically, in their language descriptions. Such observations, like most others referring to physical properties of a language in its speech guise, reflect judgments by the observing linguist as to the physical attributes characterizing speech events as readings of particular strings of linguistic items. Physical properties are most often defined in articulatory terms, sometimes in acoustic, but the judgments are almost entirely based on auditory input without overt reliance on any observational data obtained under laboratory conditions. That linguists' phonetic judgments are to an extent based on such data seems undeniable, but it is not usual to find them informed by a knowledge of the latest laboratory findings. This is understandable when we remember that many linguists are not primarily interested in precise physical descriptions, but rather in devising spelling systems that meet certain criteria, only one of which is that its letters bear a statable relation to physically describable aspects of the classes of speech signals they are designed to represent.

However, this does not make the linguist's phonetic transcription a fully explicit physical description. First of all, it represents speech by a linear array of discrete letters, so that as description it misrepresents speech in a serious way. Second, the physical properties represented by the transcription are primarily those to which distinctive function is attributed by the linguist; if some others are represented as well, this is, as Bloomfield (1933) put it, "due merely to chance observations...by an observer with a good ear," exercising a skill of "little scientific value." The linguist's representation embodies a partial physical description, but despite a possible implication of Bloomfield's comment, the disclaimer of completeness is no gesture of modesty. The linguist claims to know, from observing speech behavior in the interview situation, just what it is in the physical signal that the native speaker-hearer must produce and attend to in order that the signal be correctly interpreted. The incompleteness of the physical specification is dictated by the linguist's assertion that not all features of the signal and the signal-generating activity are linguistically significant, and that the linguist's technique of observation and analysis suffices to identify those features that are. In one undoubtedly influential view, that of Chomsky and Halle (1968), it is asserted that it is indeed linguistically irrelevant whether the linguist's phonetic statements correspond to physically
attested fact; in their *Sound Pattern of English* a hypothetical speaker-hearer is invoked whose beliefs concerning the phonetic properties of his language are what a phonetic transcription should represent. We may be permitted to note, though only in passing, that the ideal speaker-hearer whose phonetic intuitions are to be represented seems to be very well aware of the acoustic and physiological studies to be found in MIT's Quarterly Progress Reports, and may even bear a suspicious resemblance to one of the authors of *The Sound Pattern of English*. It might well be the case, therefore, that the phonetic notions of this speaker-hearer are not immutable.

At the heart of the linguist's practice of phonetic transcription, and serving as the principal bearer of his phonetic judgments or assessments of the ideal speaker-hearer's intuitions, is an alphabet of unknown but possibly finite size, to each letter of which is assigned a function as the referential of a physically defined set of vocal-tract configurations or its acoustic consequences. In defining the value of each letter of this alphabet, reference is made to a smaller set of parameters by which the state of the vocal tract is to some degree specified. In using this alphabet to spell a speech signal, successive vocal-tract configurations are identified and appropriate letters are arranged in a linear left-to-right order, which corresponds to the temporal order of the observed vocal-tract states. Each state represented has a temporal-order relation to every other state represented by the letter sequence, and the expression of this temporal relation is obligatory. This is trivially so because the only allowed spacial relation between letters is either left or right placement. Despite the fact that the letters stand for incomplete specifications of vocal-tract shape, no two of them may be simultaneously applicable, each being appropriate for a unique and unspecified time interval. Or, if you prefer, the duration is specified as being equal to that of one "segment," the duration of which is not further specified. Presumably each vocal-tract state represented is maintained over the duration of the segment, though it is not clear that this is necessarily the claim in all cases. Only the order of succession of the different states is represented—of necessity—with one segment succeeding another without overlap and without the intervention between any two immediate neighbors of a third requiring representation on linguistic grounds.

In addition to the letters that represent temporal segments, there are others that have, along with grammatical and intonational meanings, some significance as temporal markers. These are the several so-called juncture signs, as well as those indicating levels of stress. The juncture marks, which correspond very roughly to word-space and the punctuation marks of standard orthography, indicate places in the temporal sequence where, together with other phenomena, there may also occur ritardandos and even brief pauses, especially if they coincide with certain grammatical boundaries. But none of these so-called suprasegmental indicators is exclusively or even primarily temporal in reference, and demonstrations of the need to employ them in phonetic transcription generally focus on variations in pitch and loudness. Marks for stress, which for many linguists mean relative loudness, also have secondary temporal meaning; the presence of a mark of high stress usually can be taken to imply a local increase in segment durations, and, at least for English, the intervals between successive high stresses in a speech stretch are said to be of roughly constant duration. Thus, the placement of high stress marks may be said to govern the relative tempo with which the segments are produced within the utterance, in the same way that the vertical lines marking off the measures of musical notation
tell us that all the notes of one measure are to be performed within a time span equal to that occupied by all the notes within any other measure in the same text. Modern musical notation is more explicit on the matter of timing, of course, and stress placement in musical performance is not rigidly tied to the measure, but some phoneticians occasionally make use of the musical measure to represent temporal regularities observed in speech. In the view of many linguists, however, such regularities are not distinctive in language, and hence have no place in a phonetic transcription, however important they might be for the global characterization of the phonetic properties of speech generally, or of one language as against others. Except insofar as juncture and stress markers provide some guidance to tempo, the task of performing a piece of phonetic transcription is very like that of the musician asked to sight-read an unfamiliar piece from a medieval neumatic score, which indicates nothing of the individual notes but their relative pitch and sequencing. The lack of explicit timing information or instructions has its advantages for both kinds of performance, allowing scope for individual variety of expression. For the performers of speech and music the freedom of choice implied by the notations is probably wide enough to permit readings of the same score that are different enough to convey different messages to a listener. For the musician a notation that fails to specify segment duration allows one kind of temporal latitude if the musician is a flutist—the segments can be given durations at will. For the slide-trombonist segment durations are also ad libitum, and there is the additional freedom to determine how rapidly to shift from one pitch to the next in glissando playing. Producing speech is more like playing the trombone than the flute, and phonetic transcription does not prescribe how rapidly the shift from one vocal-tract state to the next is to be accomplished. I have probably pushed the analogy much too far, for it is fair to object that musical notation is a set of instructions for performance more than a description, while phonetic transcription is more a description than a performance. As a set of instructions, the phonetic transcription will have an adequacy that depends, I suspect, less on its degree of specificity than on whether or not the "score" it presents is familiar to the reader. Even if the score as a whole is novel, it must be made up of parts that are familiar if it is to be performed correctly. At the very least, the reader must be a practiced producer of fluent speech in order to implement the score as intended by the transcriber.

As a model of speech, the linguist's graphical representation suffers from inadequacies that are well-known: a speech signal does not consist of a sequence of sounds, each fixed for some unspecified duration and separated from its closest neighbors by intervals of near-zero duration; but it is perhaps unfair to charge the linguist with responsibility for such a model merely because his transcription practice seems to presuppose it. In fact, likely enough, the linguist is well aware that it is wide of the mark, and is only too ready to accept the contrary view of speech as a process, everywhere continuous, which possesses no properties that provide a physical basis for segmentation. The static definitions of vocal-tract states that he provides as interpretations of the transcription represent, then, outputs of a particular kind of sampling of this continuous signal, where the number of sampling points is determined by the number of perceived "change points" in the signal, but is pretty much independent of duration. In short, the transcription is the output of a special kind of "A to D" converter whose sampling rate is not temporally specified, the interval from one sampling point to the next depending roughly on when a perceived change in signal quality comes along. Instead of supposing the speech signal to consist of a succession of states, each maintained for some finite
time interval corresponding to the segment, one can instead say that each segment or letter of the transcription represents a state of the vocal tract which must be achieved or approximated within some time interval, and that this interval, though not necessarily the state which characterizes it, has both a finite duration and a specified place in the temporal sequence of states. The duration over which the state characterizing the segment is maintained may or may not be as great as the total duration of the segment, whatever that might be defined to be; the linguist as auditor will order segments with respect to duration, independently of what the laboratory phonetician may say about the duration over which the specified vocal-tract state is maintained. Because in fluent speech it is not unusual for that duration to be close to zero, it seems clear that we cannot hope to account for the linguist's judgments (and those of the rest of us as well) of segment duration simply by measuring the durations of steady-state intervals that might be discovered here and there in the speech signal. Since a good deal of the literature on speech timing is devoted to reporting durations of phonetic segments—when, in fact, what is being talked about are durations measured between acoustically specified change-points in the speech signals—a close relation between these measurements and the listener's judgments of segment duration must be established before those measurements can be claimed to reflect phonetic aspects of speech activity, at least in the narrow definition of phonetics I am assuming at the moment. In other words, before we can justify referring to durations between physically specified events as equivalent to vowel durations, for example, we must do what the psychophysicists did to establish the nature of the relation between pitch and fundamental frequency or to connect loudness with sound pressure level and frequency. In short, we must confront our old friend the segmentation problem. As we know, this is not so much a question of how to segment a signal, which is everywhere continuous, but rather where to cut the signal, amply possessed of discontinuities, so that the pieces derived can be claimed to correspond reasonably to the listener's segments.

Let us look now at the linguist's representation of speech as a sequence of segments, defined by reference to states either aimed at or manifested by the vocal tract or the homunculus that runs it, with segment durations not specified. This reticence as to the temporal dimension of the segment is tacit admission of the freedom to perform what is linguistically the same speech piece with tempos varying over a considerable range; moreover, no claim is made that the relative durations of segments are constant with changes of tempo (Gaitenby, 1965). But is it in fact true that relative duration is never specified by the linguist's description? Of course not. In the description of some languages, the linguist finds it useful to distinguish members of a particular phonetic class with respect to a temporal dimension; for example, Thai is said to distinguish between short and long vowels (Abramson, 1962); for Estonian both vowels and stop consonants come in three grades of duration (Lehiste, 1970); English vowels are either short and lax or long and tense. Where a difference in length is considered to be distinctive, the linguist may elect to represent the longer of a pair as a sequence of two like segments, thus by implication recognizing that a single segment possesses one unit of duration. Sometimes, however, a special kind of segment is devised, whose only characteristic is that it has the length of one segment, all other properties being given by the specification of an immediate neighbor, usually the one directly preceding it. The long vowel or consonant involves, then, a particular kind of reduplication. Whether the observation that a particular vocal-tract state is maintained sometimes for a longer interval and sometimes for a shorter one to be represented by one spelling device or another, has most often been decided by criteria not primarily phonetic in nature.
(Sometimes there may be some phonetic basis for asserting that the extra-long duration of an articulatory position must be analyzed as a sequence of repeated gestures.)

Uncertainty as to the number of segments over which a single position is maintained is not the only problem encountered in dealing with a temporal dimension at the segmental level; the same uncertainty may arise in connection with the evaluation of what are clearly recognizable sequences of—at least—one level of phonetic description. The notorious example of this is the case of stop-fricative sequences that may be accorded the status of one-segment-long affricates if there seem to be strong phonotactic (that is, distributional) reasons to do so, in which case an especially close temporal relation between its sequential components may also be discovered. There are other such examples: Are the so-called "prenasalized" stops of certain west African languages "really" one or two segments? Are the Russian palatalized consonants sequences of consonant and /y/? Sometimes, but apparently not very often, there is a genuine convergence of phonetic and phonotactic considerations; in Polish two linguistically distinct stop-fricative sequences differ phonetically in ways that allow some justification for calling one of them a single segment and the other a sequence. Thus it would appear that recognizably different vocal-tract states in immediate succession are not invariably allotted to two segments with only one possible temporal relation; that relation may be characterized as one of "close" or "open" transition, or, in the case of vocal-consonant sequences, "close" versus "loose nexus." Now perhaps we should be inclined to look for and find differences in degree of coarticulation to support a particular answer to the question of "one segment or two?"

Apart from cases where the linguist is forced to recognize a temporal feature because it appears to play a linguistically distinctive role quite like the features by which vocal-tract shape is specified, there are occasions where contextually conditioned variations in segment duration are recognized. The greater duration of vowels preceding voiced stops is marked in phonetic transcription, but that added duration (as compared with the durations of the same vowels before voiceless consonants) is not said to constitute another segment, and both the linguist and the phonetician are motivated to discover some basis, phonetic in the broad sense, for considering it to be a consequence of coarticulation. Similarly, the durational difference between the English vowels /r,u/ and /i,u/ is ascribed to the laxness of the first pair as contrasted with the tenseness of the second. Similarly, the brevity of the apical flap of American English is a consequence of the small force of articulation exerted in its production. Some kinds of temporal variation at the level of the segment that have been reported appear to have escaped attention; for example, the greater durations of initial fricatives as compared to final, or the greater durations of final nasals as compared to initial.

Observations of this last kind, which relate relative duration to position within the segment sequence, are in effect, assertions that there must be postulated units larger than the individual segment for which temporal regularities may be stated. The smallest of these is the syllable (only phoneticians who look at physiological and acoustic data worry about the organization of consonant-vowel and vowel-consonant sequences), a unit whose usefulness in phonetic description is acknowledged in the same measure as its resistance to definition is deplored. Linguists tend to solve this problem by believing in the syllable as a phonotactic unit with no phonetic standing, while phoneticians incline to
describe it as the basic element of speech organization. In this view the positional variation to which a phonetic unit conforms is, first of all, that of position within the syllable. In fact, it would seem that the durations of the segments composing a single syllable are changed sufficiently from their hypothesized "inherent" durations for the syllable to be the elementary temporal unit, with "inherent" or baseline durations assigned to segments defined purely with respect to their status within this unit. This, for the linguist, is so far from being a controversial statement that I think I might justly be charged with beating a horse that was stillborn; no linguist's discussion of speech timing has ever proposed a direct relation between utterance duration and the number of segments composing it. But a fairly direct relation between duration and syllable number appears to be intuitively acceptable, whether or not linguists make an explicit statement on the matter. The acceptance of this relation underlies the practice of defining the somewhat elusive quality of speech that we presume is chiefly temporal in nature, namely speech tempo, as corresponding more to a measure of syllables than to segments per unit time.

The same belief—I would suppose—underlies the distinction made between languages like Spanish, which exhibit this feature of "syllable timing," and languages such as English, whose contrary tendency to "stress timing" seems to require explanation as a departure from the expected. In English, we are told, the constant duration intervals into which an utterance can be analyzed are marked by stress (as has already been mentioned). In effect, this says that the durations of utterances are determined by syllable count, but not all syllables count. So far as I know, however, no one has proposed that speech tempo for English be equated with a measure of the duration separating adjacent stressed syllables. What has sometimes been reported (and this makes such a measure less appealing) is that syllables that are stressed at one tempo may be produced with noticeably reduced stress when tempo is increased, suggesting a tendency to keep interstress durations constant over a range of tempos. Of course, with all the importance that has been ascribed to the syllable as a unit of speech organization, both in production and in perception, it is remarkable that the linguist's writing system fails to represent this unit any more directly than do the more widely known alphabetic orthographies (and I suspect it would create at least as large a class of problem readers if put to more general use). Perhaps the fact that linguists have followed the alphabetic rather than the syllabic model in their writing practice comes from the general exclusion of temporal aspects in specifying speech, but it does seem odd, nevertheless, that a fundamental unit is not explicitly represented.

We come, finally, to the aspect of speech referred to as its "rhythmic" quality, which everyone seems to agree is an all-pervasive feature. From time to time linguists appeal to rhythm as a factor that determines stress placement in the case of, for example, lexical items like fourteenth, whose stress contour is variable with context; and linguists have sometimes, for example, characterized languages as "machine-gun-like" in effect. The basis for this conviction that we all share—namely, that speech can be described as rhythmic and that it is profitable to discuss the temporal organization of the process without first deciding whether any exists—doesn't seem so obvious as to be undeserving of a final remark. It is this; if all speech is rhythmic, it is certainly true that some speech is more obviously rhythmic than other, the most well-regulated speech being the perfectly metrical performance of a child's chant or poetry reading. If prose speech differs from these in temporal organization, is the difference one of kind or degree of regularity in timing? The poet's art bends speech to
aesthetic effect; some of the stuff it fashions probably is unchanged from nature (that is, the phoneme stock), but some is creative transformation or even possibly additive. How much of our conviction that rhythm is a characteristic of natural speech represents a metricization, and how much a metrification, of the object of all our attention? I wish I might end on the note of that ringing question, but more soberly suppose that we shall learn, from studies that examine data and not just the sometimes stray observations of the linguist—that speech activity may be described as at least, or at best, "quasi-rhythmic" in nature.

REFERENCES


