Reprinted from LANGUAGE
Vol. 33, No. 3, July–September, 1957
Printed in U.S.A.

LINGUISTIC SEGMENTS, ACOUSTIC SEGMENTS, AND SYNTHETIC SPEECH

LEIGH LISKER

University of Pennsylvania and Haskins Laboratories

Linguists, like many other people who feel called upon to talk about language, are in the habit of saying that speech activity is continuously variable in nature.[1] Having said this, they proceed to describe particular samples of speech as sequences of events which they isolate on the basis of either articulatory definitions or acoustic definitions, or both.[2] In carrying out this operation they are not seriously inconvenienced by the continuous aspect of speech: the fact that boundaries between any two of the elementary events composing an utterance cannot be fixed with exactness does not constitute a problem for them. They may occasionally use metaphoric expressions such as 'overlapping phones' and 'slurring', by way of acknowledging the continuously varying character of speech; but the function of such expressions is to justify the keeping of a discrete representation in the face of any demonstration that the physical segmentation of speech is hopeless. (That it is in fact not hopeless at all is beside the point here.) Speech may indeed be continuously varying when looked at in the laboratory, but the basic fact about speech is that human beings can hear it as a sequence of auditory fractions. Linguists may divide an utterance into a larger number of fractions than other listeners do, but even linguists have never been forced to concede the inadequacy of a discrete representation on the ground that it did not account explicitly for the physical continuous character of speech.

The problem of the physically continuous versus the perceptually discrete nature of speech can be put as two distinct questions. (1) Can speech signals be generated by finite sequences of discrete instructions? (2) Can finite sequences of discrete instructions be recovered from speech signals? Both of these questions are answered in the affirmative within the set of postulates of descriptive linguistics, and hence of course are strictly not allowable within that field of investigation. What the procedures of descriptive linguistics can do is enable us to say how many different instructions (phones) are required to generate all the utterances of a language, and how many of these instructions are needed for any one utterance.

When the linguist says that an utterance consists of the sequence [ac] he asserts that, given the instruction [a] followed by the instruction [c], no instructions are required governing the transition from [a] to [c].[3] If, in his judgment,

[1] See Bloomfield, *Language* 76; Harris, *Methods in descriptive linguistics* 25; Hockett, *Manual of phonology* 6.

[2] I exclude from consideration here the proposal that they be isolated by a process of magnetic-tape cutting and splicing which would by-pass the need for phonetic description of either kind.

[3] Certain instructions which the phonetic purist might insist upon—e.g. those regarding tempo of articulation, general posture of the articulating organs when 'in neutral', and general state of muscle tension—are irrelevant to linguistic description.

there is more than a single way of getting from [a] to [c], he simply says that the utterance consists of [abc], implying now that the transitions from [a] to [b] and from [b] to [c] need no specification. If the elements [a, b, c] are auditory fractions, the transitions from one to the next are phonetically determined; if [a, b, c] are segmental phones of a specific language, the transitions may be either phonetically determined or both phonetically and phonologically determined, depending on how the auditory fractions are apportioned among the linguist's segmental phones. The segmental phones, consisting each of one or more auditory fractions, are the segments on which linguistic description is built.

Acoustic analysis. Linguists are committed to the study of those features of speech which have a certain kind of social significance and which can be represented by a finite number of discrete elements. The elements must be interrelated in a way that we call 'structured'. These two general requirements, admittedly vague and perhaps not truly independent of one another, permit the exclusion of some features of the speech continuum from linguistic consideration (e.g. fundamental pitch of the vocal cord tone; voice timbres associated with anger, excitement); but it is sometimes not easy to determine whether a given feature meets the requirements for inclusion in a description. Where the decision depends on the linguist's phonetic judgment, a feature is excluded from consideration only with some feeling of bad conscience. Moreover, the linguist may suspect that linguistically relevant features have been missed. These feelings make understandable the enthusiasm which greeted the advent of the sound spectrograph, for this machine promised easy solutions to problems of phonetic description where ordinary (i.e. impressionistic, intuitive) phonetics provided no certain answers.

When the sound spectrograph was still very young, linguistic papers occasionally paid footnote compliments to its as yet unexplored potentialities;[4] but in practice, acoustic description has not been generally adopted as a technique in phonology. Even the relatively modest proposal that the vowels of a language be specified by their formant frequencies has not been taken up. Delicate matters in the phonetics of specific languages were often not materially clarified by reference to spectrograms; and no one could be sure whether this was because the instrument was in some ways less sensitive than the phonetician, or because the essential information was masked by irrelevant detail, or whether, in fact, the absence of conclusive evidence meant that the problem was not an acoustic one at all. In theory, any acoustic feature judged to be above the threshold of audibility and associated to a statistically significant extent with some phone might be considered an acoustic characteristic or measure of the phone. But we do not have precise enough information concerning the relevance of the various psychoacoustic thresholds to physically complex sounds that have the peculiar significance of speech. In the event that there appear to be many acoustic measures of a phone, there arises the problem of determining the relative importance of each as a cue to identification. This problem cannot be answered by spectrographic studies alone.

[4] For example Pike, Phonemic status of English diphthongs, *Lg*. 23.157 fn. 15; Bloch. A set of postulates for phonemic analysis, *Lg*. 24.4.

From the acoustic examination of speech the question of its segmentability is not easily answered. Spectrograms show a speech signal as changing continuously with time, but its changes are by no means so smooth that boundaries cannot be drawn and hence acoustic segments isolated. However, the sometimes abrupt change-points observable in spectrograms do not invariably mark segments bearing a simple relationship to the linguist's segments. Since workers in the acoustic analysis of speech are primarily interested in determining the acoustic correlates of the latter segments, acoustic boundaries must bear a statable relation to units of perception, quite aside from the acoustic character of the boundaries. In drawing boundaries a choice must be made between a segmentation which groups together all acoustic features contributing to the identification of a phone but permits overlapping acoustic segments, and one in which discreteness is preserved but a many-many relation between acoustic and linguistic segments must be endured. In other words, on the one hand a single linguistic segment may be identified on the basis of cues contained in more than one acoustic segment, and on the other hand a single acoustic segment may provide information for the identification of more than one linguistic segment. If by 'segmentability' we mean that boundaries can be drawn 'perpendicular' to the time dimension, so that the acoustic segments thus formed have either a one-one or a one-many relation to linguistic elements, then speech is not segmentable. If however we admit operations of the kind that the linguist applies in going from auditory fractions to phonemes, then speech may still be segmentable to the extent that a one-many relation exists between linguistic segments and acoustic elements specified by a selected subset of the features contributing to speech perception. By these operations each acoustic segment might be said to supply cues to a single linguistic segment, while any features it contains which have cue value for some other linguistic segment could be considered 'automatic' in the neighborhood of the acoustic segment or segments having a recognized relation to this other linguistic segment. There may well be the objection that we have no right to decide which of several acoustic features that appear regularly associated with a linguistic segment are to be 'zeroed out' until we have evidence that a feature is not necessary for the identification of the segment. The objection is a valid one, but the condition it imposes does not require that a segmentation account for every feature that contributes to the identification of a linguistic element.

Synthetic speech. Since the spectrograph, like the camera, 'doesn't lie' and belongs to no language community, and since no human speaker can vary at will a chosen acoustic feature (except perhaps fundamental frequency and overall intensity) independently of all the others, it seems necessary, in order to isolate and assess the acoustic cues to the identity of a phone, to make use of both human listeners and some kind of speech synthesizer whereby these features can be separately manipulated. In a sense the problem is then one of teaching the speech-synthesizing machine to become a speaking if not a listening member of a language community; as in the case of other language learners, the measure of the machine's success is simply the measure of its social acceptance. The analogy may be considered a bad one on the ground that the synthesizing machine is not subject to the constraints that limit the production of the human apparatus, and that consequently there is the danger that features which the vocal tract cannot produce may appear to have cue value in synthetic speech. This objection, I believe, has only a superficial plausibility; one might, with equal or perhaps greater plausibility, imagine a situation where absence of certain kinds of non-distinctive noise found in speech may cause the devaluation of an otherwise perfectly good acoustic cue. To entertain seriously the proposition that acoustic features having no relation to the articulations of human speech may be assigned articulatory values by a jury of listeners means, I think, to admit the possibility not merely of a new sort of psychoacoustic problem but of a species of miracle.[5]

The human learner trying to match the phonetic product of a speaker may be guided by what he can observe of the articulatory movements of the producing mechanism, such as jaw and lip movements. That is, where circumstances permit, he observes what he can of the movements directly, and infers as much more as he can from the acoustic signal. What is inferred about articulation from the acoustic signal results ultimately from a kind of hit-or-miss experimentation ('babbling') that reveals movements which can be relied upon to produce an acceptable acoustic output. The machine learner—and from now on I refer specifically to the Pattern Playback[6] developed at the Haskins Laboratories—is structurally incapable of observing and matching the movements of a human informant, for anatomically there is no relationship between its moving parts and those of the latter. But it can, with the help of human observers, match its acoustic output directly with the acoustic output of its human model as this is reported by the sound spectrograph. Spectrograms provide information which enables the machine to pass more rapidly through the 'babbling stage', so far as the features most prominent in them prove phonetically important. But because the spectrograph does not separate irrelevant sound from signal, there is no guarantee that a feature is phonetically important merely because it is easily seen in spectrograms, any more than that an easily observed lip movement necessarily has phonemic significance. It appears at present to be true that although some of the acoustic features which enable the machine to say what it 'intends to' in a socially acceptable manner are clearly visible in spectrograms, some very effective acoustic cues are sometimes not discoverable even in the most carefully made spectrographic analyses. Therefore the machine, in learning to produce the various speech sounds, is guided by, but not limited to, what can be learned from them.

A group of investigators have been busy for several years teaching the Pattern Playback to talk American English. They began where all work in acoustic phonetics has traditionally begun—with the isolated vowels of the language. After the machine had been taught to produce these steady-state sounds,[7] they

[5] Peterson, An oral communication model, *Lg.* 31.427, is of a different opinion, and warns that such an eventuality is possible to the point of posing 'considerable danger'.

[6] The Pattern Playback is described by Liberman, Delattre, and Cooper, The role of selected stimulus-variables in the perception of the unvoiced stop consonants, *Amer. jour. of psychology* 65.497–516.

[7] Delattre, Liberman, and Cooper, Voyelles synthétiques à deux formantes et voyelles cardinales, *Le maître phonétique* III.29.30–6 (1951).

next turned their attention to sequences of consonant plus vowel. This sequence type was chosen rather than some other (as consonant-vowel-consonant) in order to minimize the problem of possible transition effects between neighboring sounds and thus to secure the simplest condition for trying to match the elements of the linguist's description at the acoustic level.

The relation between linguistic description and speech synthesis in respect to segmentation may be summarized as follows. Linguists, dealing as they do in discrete symbols, and observing that human beings can be trained to convert strings of these symbols into speech signals and conversely, are led to suppose that one or both of these statements may be true: (1) that somewhere in the communication act, though not in the speech signal itself, there is a flow of discrete elements corresponding to the linguist's symbols; (2) that the continuously varying acoustic output (or perhaps the movements of the vocal organs themselves) can be recorded on magnetic tape, the tape cut up into pieces, and the pieces grouped on the basis of commutation and discrimination tests into the classes which the linguist's symbols represent. Now since the communication act is accessible to observation only in its articulatory and acoustic phases, the first supposition expresses little more than a pious hope. The second statement, on the other hand, being a proposal that the definition of the linguist's elements be based on tape-cutting and splicing, is an example of 'armchair operationalism' that may appeal to the linguistic logician, but leaves anyone cold who has made attempts in this direction. From the acoustic side, spectrograms present patterns that are full of discontinuities, to be sure, but are so far from providing a basis for the linguist's operations that they can hardly be read unless one knows what they are supposed to contain. In the synthesis of speech a major advantage is that acoustic signals with 'built-in' segmentation points are generated. Once these signals may be considered adequate approximations to human speech, the relation between linguistic and acoustic elements can be stated, for we have acoustic segments containing a manageable number of features that may be readily isolated and subjected to phonetic evaluation.

From the data so far derived via synthesis, a few cautious generalizations are possible. (1) A finite number of acoustic segments, each defined with respect to a small number of dimensions, can be arranged in time sequence in such a way that they are perceived as speech. (2) The number of acoustic segments needed for a given speech signal is not smaller, and generally is greater, than the number of phones posited by the linguist. (3) For some phones there exists a many-one relation between acoustic segments and phones, so that a given phone may be said to comprise a class of 'allosones' whose distributions can be described by reference to features in their acoustic environments. (4) The classes of phones established at the motor phonetic level (which agree closely with the classes based on distribution) find a high degree of corroboration at the acoustic level, so that the members of any one class are differentiable by variations within a relatively small number of acoustic dimensions.