Distribution: limited

UNESCO/SS/3.244.1/h/40
Paris, 29 August 1966
Original English

UNITED NATIONS EDUCATIONAL,
SCIENTIFIC AND CULTURAL ORGANIZATION

INTERNATIONAL STUDY ON THE MAIN TRENDS OF

RESEARCH IN THE SCIENCES OF MAN

AUXILIARY CONTRIBUTION

On the Modern Study of Speech Sounds

by

Morris Halle

(Document to be used intthe preparation of the
study by the special consultants)

ON THE MODERN STUDY OF SPEECH SOUNDS

by

Morris Halle

Department of Modern Languages and Linguistics
Massachusetts Institute of Technology

It is a paradox that in spite of a tradition that can be traced back to classical antiquity or even earlier, phonetics -- i.e., the study of the physical actualization of language and of its perception -- is, in many ways, still in its infancy. The underdeveloped state of the field manifests itself in the fact that the most elementary theoretical issues continue to be debated to this day. Thus, there are quite a few phoneticians for whom it is still an open question whether speech sounds exist, and if so in what sense. Since this is clearly a very basic issue this survey of the present state of our knowledge in this field shall begin with an inquiry into this problem.

About a century ago, when A. M. Bell invented the important phonetic alphabet known as Visible Speech, of which the most important modern phonetic alphabets are lineal descendents, the existence of speech sounds was taken for granted. Utterances were conceived of as being made up quite literally of discrete speech sounds, each with its own specific vocal tract configuration and corresponding acoustical properties, and the process of speaking was believed to consist of a sequence of movements from one fixed configuration to the next. Since it is obvious that the vocal tract configurations are controlled by a very limited set of muscles, it was naturally assumed that all possible configurations of the vocal tract could be specified by specifying a limited number of variables; i.e., the variables controlling

the muscles which determine the vocal tract configuration. In the Visible Speech alphabet, this conception was reflected in the graphic shape of the letters. These letters were constructed of a small number of elementary graphic components, each of which represented a specific variable controlling vocal tract configurations.[1] Although invented primarily for the purpose of teaching speech to English deaf-mutes and the standard pronunciation of the Queen's English to speakers of lesser dialects, the Visible Speech alphabet allowed for the symbolization of speech sounds not found in English. In fact, it was Bell's intention to incorporate in his alphabet all variables known to determine vocal tract configurations in all human languages, for he conceived of Visible Speech as a universal phonetic alphabet.

It should be noted that Bell viewed this alphabet as a device not primarily for recording particular phonetic events, but rather for symbolizing classes of utterances which native speakers would regard as repetitions of one another. We see this clearly in the tests to which Bell subjected his alphabet. These consisted of Bell's making a Visible Speech transcription of utterances in a foreign language or strange dialect. The transcription would then be given to Bell's son, who had not heard the original utterances. The latter would then repeat the original utterance with great naturalness. It is essential to observe that this test did not

---

1. The same idea - though less consistently and less elaborately executed - underlies the Korean alphabet, which dates from the fifteenth century.

entail mimicking the original utterance, only repeating it. We see, thus, that Bell was not trying to record directly observable motor behavior in the vocal tract, but rather to represent it symbolically with the help of symbols that stood for instructions that could be used by a normal speaker to replicate rather than to mimic the behavior in question.

Bell's successors, however, saw phonetics in a somewhat different light. Since at that time, it could reasonably be assumed that there was a very simple, direct relationship between instructions and observable behavior, there was a tendency for phoneticians to think of their transcriptions of utterances as naturalistic records of particular phonetic events. An immediate consequence of this was that attention was focussed on what could be observed directly and the question of how a normal speaker controls his vocal tract so as to produce the behavior observed was pushed into the background. Phoneticians, thus, were deprived of an essential criterion of relevance, and in the succeeding years phonetics was literally inundated by detailed observations of which a very high proportion were, according to Professor Jakobson's witty saying, as little relevant for linguistics as numismatic facts are for the study of finance.

The assumption that the primary objective of a phonetic transcription is to record observable facts concerning specific speech events immediately raised very fundamental questions about the appropriateness of the decisions that are inherent in

an alphabetic notation. For one, the phonetic alphabet presents speech as a sequence of discrete configurations and omits to record what transpires when we move from one configuration to the next. Why should not one do the opposite and omit the stationary parts and represent only the dynamic parts? Moreover, when radiocinegraphic records of speech became available, it was seen that there were few intervals during an utterance when the vocal tract was stationary. Instead, everything appeared to be in motion almost constantly. In view of this it was only natural that phoneticians should begin to question the conception, basic to any alphabetic representation, that utterances are composed of sequences of discrete sounds.

Secondly, the different degrees with which a particular feature can be detected in a given sound are limited only by the acuteness of the observer's senses. A beginner might be able to detect only two levels, e.g., of stress in vowels, where the trained expert can easily observe half a dozen or more. Since clearly individuals differ in the acuteness of hearing, and even a given individual is not always performing at top efficiency, phonetic transcriptions are inevitably subjective, and some phoneticians make it a point to append to every transcription a detailed record of the circumstances under which the transcription was prepared. This face has naturally raised doubts about the general validity of such inevitably subjective observations.

Thirdly, the features that have been chosen to be recorded are not self-evident. Why should one omit to record the speed with which an utterance is made, the voice quality of the speaker, or the so-called "basis of articulation"? Why should one not record the very obvious movements of the epiglottis, of the lower jaw?

In the last hundred years, numerous proposals have been advanced as to how these difficulties might be resolved. None of these proposals, however, has gained general acceptance. More recently it is becoming increasingly clear that the questions raised above are really pseudo-questions which are due to a misconception of the objective of phonetic study. Since phonetics studies speech, it must be concerned above all with understanding the phonetic behavior of normal speakers, with the way in which a normal speaker produces and perceives utterances. Since the behavior of the normal speaker can apparently be understood best in terms of discrete speech sounds and of a particular set of phonetic properties which can assume only a very limited number of values, it is these entities that must play a basic role in the study of phonetics. The fact that these entities may not always be directly observable in particular physical signals but must be inferred from directly observable data -- often by rather complex chains of logical reasoning -- in no way affects their reality. They are every bit as real as the theoretical entities of other fields of inquiry -- as electrons, gravity, valence or conditioned reflexes.

The proposition that the behavior of the normal speaker must be the main concern of phonetic inquiry is now fairly generally accepted, thanks mainly to the work of the different structuralist schools that have arisen in various countries since the late 1920's. A basic tenet of these schools was that this behavior could be understood properly by concentrating on those properties of utterances that are capable by themselves of rendering distinct two otherwise indistinguishable utterances. They declared those properties of the signal that did not function in this way in a particular language as having only marginal interest for an understanding of the normal speaker's linguistic behavior. This appears to have been a mistake for it can be shown that the line thus drawn excludes from linguistic consideration properties of the signal that are clearly part of a normal speaker's linguistic competence, in the sense that they have to be learned by anyone who would have complete command of a given language. The proposal criterion fails to distinguish, e.g., the following two vastly different types of situations:

a) As the articulatory organs move from one configuration to another they naturally will pass through a series of intermediate stages and these intermediate stages will be determined in general by the two terminal configurations; thus, we have a different transition from consonant to vowel in the syllable [tu] than in the syllable [pu]. Since this is part of the normal behavior of the vocal tract the nature of the transition is never

a cue that distinguished by itself one utterance from another. In view of this fact, information about the transition is excluded from linguistic consideration by the proposed criterion.

b) In English, vowels are shortened and laxed in position before voiceless consonants, but not elsewhere; thus the vowel in pat is shorter and laxer than the vowel in pad. Since all English utterances are subject to this rule, the laxing and shortening of vowels cannot be used by itself to signal that two utterances are distinct. Information about shortening and laxing of vowels in position before voiceless consonants must, therefore, also be excluded from linguistic consideration if the proposed criterion is to be taken seriously.

These examples, however, illustrate quite different situations. In the first example, the vowel transitions are the result of physiological limitations under which speech is produced; they reflect the way in which the vocal mechanism operates and say nothing about the language. In the second example, on the other hand, the laxing and shortening of vowels are not a consequence of the normal physiological limitations on speech production, but are rather language specific facts about how utterances in English are to be produced. The differences in vowel transitions are found in all languages and under all conditions, whereas shortening and laxing of vowels before voiceless consonants are not found in all languages, and even in English they are found only when the voiceless consonant belongs to the same word as the vowel (cf. e.g.,

t e italicized vowels in "lay speaker" and "lace pillow"). The p oper distribution of vowel tenseness and length is, therefore, part of the skill that a fluent speaker of English must acquire, it is part of what we mean by his "command of English", whereas the proper distribution of vowel transitions is not similarly learned, it is an inherent property of a person's speech mechanism. In sum, the features discussed in example (a) reflect not linguistic facts, but rather physiological facts. The features discussed in example (b), on the other hand, mirror linguistic facts. Since the proposal to take into consideration in a linguistic description only those features that are capable of differentiating utterances that are distinct, lumps these two cases together, the proposal cannot be accepted. If phonetic investigations are to characterize a fluent speaker's linguistic competence, they cannot be restricted in the manner proposed, but must take into account all properties that in principle could have been actualized by the speaker in a different fashion from the way they happen to be actualized in the language under discussion. These properties together characterize the phonetic capabilities of man as a producer of speech. Since these properties or phonetic features are part of the physiological endowment of the human species, they are linguistic universals.

Limitations of space make it impossible to describe each of the universal phonetic features. For present purposes it will be quite sufficient to state that the total number of features in the set is not large (perhaps thirty), but is larger than it was

thought to be some years ago. The features include such well known phonetic properties as voicing, nasalization, labialization, fronting, retroflection, suction (click), aspiration, and tongue height, etc., as well as various prosodic properties such as stress, falling intonation, high pitch, etc. Instead of describing these features in detail, we shall attempt here to characterize certain general properties of phonetic features, especially properties that are not altogether self-evident.

Phonetic features are the dimensions in terms of which speech sounds are processed by the human organism; these dimensions are not necessarily invoked when the organism processes acoustic stimuli other than speech. That the perception of speech involves mechanisms not utilized in the perception of other acoustical signals is one of the most interesting conclusions to be drawn from recent perceptual studies. In a series of interesting experiments dealing with the discrimination of synthetic speech sounds, it has been found by the workers at Haskins Laboratories (cf. Fry et.al. (1962)) that certain classes of consonantal stimuli are perceived categorially; i.e., stimuli are perceived as being instances of particular consonants, and stimuli within a given region cannot be distinguished from one another, regardless of their physical differences, whereas small differences are perceived when these happen to fall into regions belonging to different categories. Vowel stimuli, on the other hand, do not exhibit categorial perception, especially when presented in isolation. Discrimination of vowel quality seems to be equally

ute within a phoneme or across phoneme boundary. These observations have led some students to suggest that there are two distinct modes of perception, one for consonantal sounds and the other for vowels. This suggestion seems, however, to lack plausibility as it implies

a constant switching on and off of fairly different modes of signal processing in the course of listening to a given utterance. It was noticed that the stimuli that were used in the vowel experiment differed from those used in the consonant in one very significant respect. Whereas the former always sounded like actual syllables of speech, the stimuli in the vowel experiments required a conscious effort to be perceived as speech sounds; without such an effort they sounded rather like complex tones, especially since they were presented to the subjects in isolation. This observation suggested immediately that the tests be repeated with vowel stimuli embedded in speech-like contexts. When this was done by Stevens (1966), it was found that the responses of subjects were categorial, and thus formally resembled those obtained on the experiment with consonants. Given these facts it would appear that the differences in the mode of perception noticed by the Haskins group were not correlated with differences among classes of speech sounds, but rather indicated different modes of perception in the case of speech than in the case of other types of acoustical stimulus. It is important to stress that the stimuli and the task of the subjects were identical in both experiments. What was different in the Stevens experiment was that the stimuli were embedded in speech-like contexts, whereas they were presented in isolation in the Haskins experiment. Hence in the Stevens experiment subjects were forced to perceive the stimuli as speech. Once they were forced into this

mode of perception they naturally made only categorial judgments, for a speech sound belongs to a restricted set of categories and can only be identified in such a manner.

The existence of a specific speech mode in the perception of acoustical stimuli explains why so little has been learned about the perception of speech from experiments with nonspeech stimuli. The essential difference between the speech mode and the nonspeech mode is that in the former the subject can bring to bear in the analysis a great deal of information that is not available to him in the analysis of nonspeech stimuli. In particular, when processing speech stimuli, the subject can incorporate his innate "knowledge" not only of such facts as that the stimuli are discrete, that they are produced subject to the physiological constraints of the human vocal tract as a producer of speech, etc., but he can also invoke his "knowledge" of the phonological rules of the language.

As an interesting example of radically different results in the perception of speech and nonspeech stimuli, consider the following dealing with the perception of stress. It was known already to Bell's contemporary, Ellis, that English utterances manifest a great many distinguishable levels of stress. Thus, the syllables of the compound noun lighthouse door have the stress pattern 1-3-2, where 1 represents primary stress, 2 secondary stress, etc. The noun phrase light house door, on the other hand, i.e., a house door that is not heavy, or dark,

has the stress pattern 2-1-3.

Although at present there is still some question about the precise physical property that signals the phonetic feature of stress, there is no doubt that it is independent of those properties that signal what D. Jones has called the "tamber" of the vowel; i.e., those properties that allow us to distinguish an [i] from an [a], and an [o] from a [u]. It is possible by a special electronic technique to process utterances so as to eliminate from the signal all information about vowel "tamber" while leaving all other information virtually unchanged. Utterances so processed have pitch, length and stress variations, but are, of course, incomprehensible because there is no information about the segmental phonemes that constitute them. When utterances so processed were presented to highly trained phoneticians who were asked to transcribe their stress patterns, it was found that only the location of main stress could be determined with any degree of reliability; the indications of secondary, tertiary and lower degrees of stress were completely unrelated to what was on the signal. There was, of course, no problem about locating these stresses correctly when the original utterances were played back to the subjects. (See P. Liberman (1965).)

This result is not surprising, in view of the notorious difficulty that naive subjects experience in locating non-main stresses in English. However, it is still necessary to explain why there was such a great discrepancy in performance in the two cases.

The most obvious difference between the two experimental situations is that in the first situation the subjects did not know the utterances and hence were unable to utilize in their analysis their knowledge of the stress relations that normally hold in English utterances, whereas they were able to do so in the second situation. The results would appear to suggest that physically, subjects could distinguish no more than two levels of stress in the signal. If they were able to respond reliably to additional levels of stress when the signal was speech this must be due to the peculiar manner in which stress distinctions are processed in speech. We must, therefore, turn briefly to an examination of these.

It has been shown by Chomsky and Halle (1967) that the stress contour of an English utterance can be derived completely given the sequence of segments that constitute the utterance and its immediate constituent structure. The rules which are required for this purpose have three properties that are of importance in the present discussion. 1) Like all phonological rules, the stress rules apply in a definite order which takes account of the constituent structure. A given rule applies first to the smallest constituent, then to the next largest constituent, etc., until the process stops at the boundaries of the so-called "phonemic phrase". 2) When primary stress is assigned to some syllable in a string, the stresses on all other syllables in the string -- if any -- are lowered by one level; i.e., primary stress

is lowered to secondary, secondary to tertiary, etc. 3) The stress rules of English assign only primary stress or alternatively, determine the location of primary stress in the string. The first two properties of the stress rules are universal, they are principles that hold true of such rules in all languages; the third property on the other hand, is specific to the English language. It is readily seen that given the fact that rules apply to constituents in order and the principle of stress lowering, complicated stress contours can be produced by repeated application of a simple rule such as the following that governs the stress relations in English compound nouns:

(1) in compound nouns primary stress is assigned to the rightmost syllable bearing primary stress [redacted] that occurs in a noun preceding the last noun of the compound.

Since monosyllabic nouns in English are assigned primary stress by another rule, that applies first, a compound noun such as light house will contain two nouns with primary stress at the point where rule (1) applies. Rule (1) will then assign primary stress to the noun light and the general principle of stress lowering will lower the stress on house to secondary. A compound noun light house door has a constituent structure ((light house) door). Consequently, rule (1) will apply first to the innermost constituent light house yielding the stress contour $\overset{1}{\text{light}}\ \overset{2}{\text{house}}$, where the integer above the nouns represent the relative stress levels. In accordance with the universal ordering principle rule (1) will apply next to the string $\overset{1}{\text{light}}\ \overset{2}{\text{house}}\ \overset{1}{\text{door}}$, where

door has primary stress by virtue of being a monosyllabic noun. Rule (1) will assign primary stress to the noun light and simultaneously lower the stresses on the other nouns, producing thus the stress contour $\overset{1}{\text{light}}\ \overset{3}{\text{house}}\ \overset{2}{\text{door}}$. Since compound nouns can have more complicated constituent structures than the examples just reviewed, much more complicated stress contours can be generated by repeated application of rule (1).

We return now to the results of Lieberman's experiment on the perception of stress. Since the experimental subjects were speakers of English it must be assumed that they had at their disposal knowledge of the stress assignment rules of English and they were able to bring this knowledge to bear when responding to English utterances.[2] When the utterances were processed, however, so that the subjects could no longer understand them, the knowledge of English stress rules was of little use to the subject, and they were forced to fall back on the stress information that was directly available in the signal. This, however, proved inadequate except in order to determine the location of main stress. The situation thus is rather similar to the one that would prevail if a person were asked to determine whether a given letter was a script "e" or "l" without being provided with enough context to make a unique decision feasible.

---

2. It is clear that the type of knowledge under discussion here is tacit rather than explicit and conscious. Tacit knowledge underlies many, if not all manifestations of skill. Thus, the inability to state Archimedes principle of buoyancy has not prevented people from learning to swim or from building seaworthy boats.

The case just discussed is significant in that it brings out rather clearly certain details of what we have called the speech-mode of perception. In the last example we saw perception influenced by two distinct components, one belonging to the universal properties of language (i.e., the principle of stress lowering) and the other, to the specific properties of English. It is to be expected that if the same stimuli were presented to subjects unfamiliar with English their performance would differ if given utterances in the clear, but should be the same if given the processed utterances. The differences in performance would have to be ascribed to the differences in linguistic background. Precisely how differences in linguistic background would affect performance can at present not be predicted. This would, therefore, seem to be an area where additional data are likely to materially advance our understanding.[3]

The preceding has concentrated almost exclusively on the perception of speech. Phonetics, however, is equally concerned with the production of speech, the physical actualization of utterances, and a great many significant advances have been made

---

3. Interesting experiments showing the effects of prior linguistic knowledge on the perception of signals have been performed by J. Fodor and T.G. Bever, who have conclusively shown that speakers tended to perceive a click superimposed on an utterance as being located between major constituents. Thus in the phrase "she fed her dog biscuits" the click was located before the word "dog" when the story in which the stimulus sentence was embedded required the interpretation that "dog biscuits" was a compound noun; whereas it was located after the word "dog" when the story required a major constituent break between "her dog" and "biscuits". See Fodor and Bever (1965).

in this area. Perhaps the most notable advance is that made in the understanding of the relation between vocal tract behavior and acoustical signal. As a result of the work of Fant (1960) and others, we are able now to calculate the acoustical signal from a description of the gross motor behavior in the vocal tract. Progress has been much slower in understanding the manner in which the vocal tract is controlled in speaking. This slow advance has been due in large part to the fact that phoneticians have tended to regard phonetics as a discipline whose main task is to gather data rather than develop theories about speaking. A more enlightened attitude toward the role of theories in phonetics has become evident in the last few years and the first results of this attitude are beginning to appear.

It is a well-known fact that parts of the articulatory behavior associated with a given speech sound are frequently to be found in the articulation of adjacent sounds. This effect, for which the term "coarticulation" has been coined, has been one of the factors that have tended to cast doubt upon the existence of discrete speech sounds. Recent research conducted by K.N. Stevens and his co-workers at M.I.T. has cast new light on the function of coarticulation in speech. Stevens and his associates have been interested especially in the coarticulation effects on the feature of voicing. As is well known, voicing can be produced in the vocal tract only if the following two conditions are met: the vocal cords must be approximated, rather than held apart, and air must

flow past the vocal cords with a certain minimal velocity. Consider from this point of view the production of a sequence such as [ata], where a voiceless obstruent occurs between two vowels, which are voiced. Since the vocal tract is closed during the articulation of the stop, the pressure in the mouth builds up very rapidly (in 20 msec or even faster) to a point where the air flow from the lungs stops and vocal cord vibrations cease.[4] In other words, in the production of voiceless stops the onset and cessation of vocal cord vibration will coincide with the closure or release of the occlusion and will not require any special gesture such as widening of the glottis.

This effect is made use of by speakers to control the onset of voicing very precisely without precise timing of the requisite articulatory gestures. Consider, for instance, a sequence of voiceless obstruents followed by a (voiced) vowel. If at the beginning of the sequence the vocal cords are far apart, they must be approximated by the time the vowel is articulated. The timing of this approximation movement, however, is not especially crucial. It suffices that it take place at any time before the onset of the vowel, since during the articulation of the obstruents, the presence of the obstruction in the vocal cavity will

---

4. Stops produced with voicing during the stop occlusion require a laxing of the supra-glottal musculature which allows the vocal cavity to expand during the period of closure. This expansion of the cavity lowers the pressure sufficiently for air to continue to flow past the vocal cords.

suppress the air flow and hence prevent vocal cord vibration from arising prematurely. Thus, we have here a clear case of coarticulation, for the approximation of the vocal cords which is a characteristic of voiced sounds, takes place here while the vocal tract is still articulating voiceless sounds.

It appears that many instances of coarticulation are of this type: they are mechanisms whereby certain phonetic events are made to coincide precisely without at the same time requiring precise coordination of the articulatory gestures that produce them. Such behavior, however, is understandable only if the process of speaking is viewed as the production of a continuous physical signal from an abstract representation to which the speaker has access in its entirety so as to enable him to preset his articulators whenever necessary and appropriate. Since such an abstract representation of speech as sequences of discrete entities appears to play also a fundamental role in the perception of speech, it is hardly surprising that speech sounds have been the focal entities in the study of phonetics.

REFERENCES

Chomsky, N. and M. Halle (1967). The Sound Pattern of English, New York. In press.

Fant, G. (1960). Acoustic Theory of Speech Production, The Hague, 1960.

Fodor, J. and T. G. Bever (1965). "The Psychological Reality of Linguistic Segments", Journal of Verbal Learning and Verbal Behavior 4, 414-420.

Fry, D. et.al. (1962). "The Identification and Discrimination of Synthetic Vowels", Language and Speech 5, 171-189.

Lieberman, P. (1965). "On the Acoustic Basis of the Perception of Intonation by Linguists", Word 21, No. 1, 40-54.

Stevens, K. N. (1966). "On the Relations Between Speech Movements and Speech Perception", paper presented to Symposium 23: Models of Speech Perception, XVIII International Congress of Psychology, Leningrad, pp. 68-74.