From: Handbook of Sensory Physiology. Edited by R. Held, H.Leibowitz, and H.-L.Teuber. Vol.8: Perception. New York: Springer-Verlag, 1978.

Chapter 5

232

Phonetic Perception

ALVIN M. LIBERMAN* and MICHAEL STUDDERT-KENNEDY,**
New Haven, Connecticut (USA)

With 6 Figures

Contents

To include a chapter on phonetic perception in a handbook like this is to assume that the process is not wholly accounted for by such principles as we might find in research on the perception of nonspeech sounds. It is appropriate, then, that we here offer support for that assumption. We will

* Also University of Connecticut and Yale University.
** Also Queens College and Graduate Center of the City University of New York.

not examine all relevant considerations, only those that bear most directly on the relation between the information in the acoustic signal and the listener's perceptual response to it; in our view, those are the most pertinent. Nor will we analyze such arguments as there are for the opposite assumption – namely, that auditory mechanisms are sufficient – though we will, as is proper, refer the reader to relevant papers[1].

Phonetic perception is what happens when, on hearing speech, a listener recovers the phonetic message. That message consists of the meaningless segments we perceive as consonants and vowels. These are ordered in strings, organized into larger units, and carried on a prosodic contour. The segments, both consonants and vowels, are called "phones"; among the larger units are syllables; the relevant aspects of prosody are stress and intonation. We must distinguish between the perceived phones and the more abstract phonologic forms that underlie them. Thus, the final segments in "cats" and "dogs" are different phones – voiceless [s] in "cats" and voiced [z] in "dogs" – yet at a more abstract phonologic (or morphophonemic) level they are the same. Our concern will be with the less abstract phones and their relation to the still less abstract sounds. Also, to keep our task within bounds, we will deal only with the segmental aspects of phonetic structure, including the organization of phones into syllables, though perception of prosody presents interesting, perhaps even similar, problems.

Students of language commonly assume a complex, grammatic relation between meaning and its phonetic vehicle, but often disregard the further complications that arise in the conversion to sound. They tend rather to suppose that the phonetic segments (or their constituent features) are represented discretely in the signal, as if by an acoustic alphabet. If that were so, perceiving phones would be like perceiving any other sounds; there would be no special problem of phonetic perception and no reason for this chapter. There is evidence, however, that the sounds of speech are not an alphabet on the phones, but a complex and grammatic code. In the first section of the paper we will place that code in the larger scheme of language and identify its important characteristics.

If it is true that the phones are linked to the sounds by a special code, we should suppose that extracting the phones from the sounds would require a correspondingly special decoder. In the second section we will give reasons for supposing that such a decoder may exist.

There is, of course, an alternative to grappling with the problems created by the peculiar relation between sound and phonetic message: We can try to evade them. Indeed, we might suppose that phonetic perception does not occur, that the segments of the phonetic level are mere fictions, invented by linguists for their convenience, with no functional significance in language or in its psychophysiology. In that view the listener would go directly from sound to some meaningful segment (for example, word), bypassing the phonetic and phonologic structure entirely. To justify our concern with phonetic perception, we will, in the third section, argue that phonetic (and phonologic) structure plays an important role in language and is, in fact, recovered by the listener when he comprehends what is said to him.

[1] For reviews of speech perception research from several points of view, see DARWIN, 1976; PISONI, in press; STEVENS and HOUSE, 1972.

A. Special Nature of the Speech Code: Function, Form, and Key

For anyone who would understand the perception of speech, the salient fact is that the perceived phones are related to the sounds by a peculiar grammatic code, one of several that link sound to meaning. To grasp the nature of that code, it is useful to view it as part of the larger grammar (see, for example, MATTINGLY and LIBERMAN, 1969; LIBERMAN, 1970). For that purpose, we will divide grammar—and language—in two. Making the cut at the phonetic level, we will look first toward meaning and then, in the other direction, toward sound.

I. Function of the Meaningless Phones and of the Grammatic Codes Linking Them to the Meaningful Message

To see what grammar accomplishes, and thus to appreciate the role of the meaningless phones, we should first consider the shortcomings of an agrammatic mode of communication (see LIBERMAN et al., 1972; LIBERMAN, in press). In that mode, there would be a straightforward connection between message and signal. Instead of grammatic rules like those that build longer and more complex structures (syllables and sentences, for example) out of shorter and simpler ones (phones and words), there would be only a list of all possible messages and their corresponding signals. Obviously, such a mode would work well enough if there were reasonable agreement in number between messages and signals. But, just as obviously, there is no such agreement: The number of messages we have to send is vastly greater than the number of holistically different signals we can efficiently produce and perceive, especially if we are committed to signaling with sound. In short, an agrammatic mode of communication would limit the number of possible messages to the small number of distinctively different sounds we can produce and perceive. The consequence would be that most of what we want to express with language would be inexpressible.

We should suppose, then, that one function of grammatic codes is to restructure the information in the messages so as to make it compatible with our sound-signaling ability, and thus to match the potentialities of the message-generating intellect to the limitations of the vocal tract and the ear. But why two grammars, syntax and phonology, and why the two kinds of segments, meaningful and meaningless, they govern? What is the function of this dual structure, characteristic of all languages, and especially of the meaningless, phonologic portion that concerns us in this chapter? Why not, in a simpler world, have only a syntax—rules that organize and reorganize segments (words, for example) that are meaningful? Such a language could, from a logical point of view, evade the limitations imposed by the paucity of different segments, since it

would be possible, even with a small set, to construct an infinitude of messages. A phonology-free language would, of course, have to make do with a small vocabulary, but that is not, in logic, a devastating limitation. For to the extent that we can organize our semantic space by a hierarchy of features, a small vocabulary might nevertheless suffice for many of the things we want to talk about (OGDEN, 1967). But specifying a particular thing would, at best, take a lot of talking and listening, given the properties of vocal tracts and ears, and it would require, in addition, that one's mind work in ways that may be uncongenial to it.

At all events, no language does get along with a very small vocabulary. Vocabularies tend to be large and to grow ever larger (but see KLIMA, 1975, pp. 247–270). To achieve these large vocabularies, given the limited number of signals we can command, languages use a very few meaningless segments—two to three dozen, in most cases—to construct a large number of meaningful ones: hence, phonology. Taken together, then, syntax and phonology serve as a kind of interface, joining an intellect, which initiates, comprehends, and stores messages, to a vocal tract and ear, which produce and receive the sounds by which those messages are conveyed (MATTINGLY, 1972; LIBERMAN, 1974).

II. Function of the (Grammatic) Speech Code Linking Phonetic Message to Sound

Perhaps the need for grammatical recoding has ended with the production of the phonetic message. If so, the final link to speech could be agrammatic—a unit of sound for each segment of the message—and thus of no special interest to either the linguist or psychologist. But the phonetic message is only a stage in the grammatic process that connects meaning to sound. Further and still quite drastic restructuring is necessary. To see why, we need only put the most obvious requirements of phonetic communication against the capabilities of the ear and the vocal tract. Although that has been done in earlier publications (LIBERMAN et al., 1967; LIBERMAN et al., 1972; STUDDERT-KENNEDY, in press), we nevertheless offer a brief review here.

Two requirements of phonetic communication are of special interest: The phones must be communicated at a high rate, and their order must be properly apprehended by the listener. With regard to rate, it is obvious enough that language is more efficient the more rapidly it is communicated. It is only slightly less obvious that language is hard to understand when it is communicated too slowly. Slow communication can create difficulties because the meaning of the longer segments is distributed in complicated ways among the shorter segments they comprise. Hence full comprehension of a sentence, for example, must wait on completion of a structure that is formed by the words. The requirement about order follows from the use of a small number of phonetic segments. If we are to keep the number of segments per word within bounds, we must respect order: a word like "dam" must be distinguished from its mirror image, "mad".

It is plain that if the phonetic segments were transmitted agrammatically—that is, each phone by a discrete segment of sound—the requirements of phonetic communication would not be met. We could neither speak nor listen as fast as we need to—and, indeed, do—nor could the listener keep the segments in their proper order. Speaking rates vary considerably, but they reach 20–25 phones/s, at least for short stretches. Presumably, it would be impossible to speak that rapidly if, as in an agrammatic mode, the gestures were made discretely, one for every phone and each in its turn. And even if the speaker could articulate that fast, the listener could not resolve the sound segments that would result; at 20–25 acoustic segments/s, the units of sound (hence phones) would merge to produce, in perception, a buzz or pitch. Moreover, the listener would have difficulty identifying the order of such discrete sound units, even at rates low enough to permit him to resolve them. Given the results of research on nonspeech sounds (WARREN et al., 1969; 1976b), we should suppose that he could distinguish permutations of segments, but only on the basis of overall differences in the perceived pattern, not by assigning each segment to its own place in a sequence. Surely, then, the grammatic restructuring that makes communication distinctively linguistic cannot end with the production of the phonetic message. At least one more grammatic conversion is necessary if the message is to be transmitted and perceived efficiently.

III. Form of the (Grammatic) Speech Code Linking Phonetic Message to Sound; Fit of Form to Function

In the conversion of abstract phones to concrete sounds there is a restructuring of information, designed as if to match the requirements of phonetic communication to the properties of the vocal tract and the ear. Though much that is important about this conversion remains to be learned, enough is known to enable us to see some of its important characteristics. Thus, we know that the segments are first broken down into something like the well-known articulatory features of place of production, manner of production, and voicing[2] (for an explication, see, for example, LADEFOGED, 1971). As speech is produced, those separate features are assigned to the appropriate and more-or-less independent parts of the articulatory apparatus; the component gestures made by those parts are organized into preplanned coding units longer than a phonetic segment; and the organized complex of gestures, representing features of each of several successive phonetic segments, is produced simultaneously or with considerable overlap. The result is that the coding unit—roughly a syllable in many cases—comprises segments whose component gestures (features) are thoroughly interleaved (COOPER et al., 1952; FANT, 1962; LIBERMAN et al., 1967; COOPER,

[2] In the case of the consonants, place of production refers to where in the mouth—lips, alveolar ridge, or velum, for example—the consonant constriction is made; manner of production refers to an articulatory maneuver—velum closed or open, tract totally closed or only partly closed, for example—that is characteristic of phones with the same place of production; voicing distinguishes classes of phones having the same place and manner according to the state of the vocal cords—open or closed—at the beginning of the gesture.

1972; STEVENS and HOUSE, 1972; STUDDERT-KENNEDY, 1975a). We will call that arrangement by its common name, coarticulation.

Coarticulation enables a speaker to produce phonetic segments at rates considerably higher than the rates at which he must change the states of his articulatory muscles (COOPER, 1972). Thus, he speaks faster than he could if each phonetic segment were represented by a unit gesture, produced in its proper turn as one of a sequence of gestures. But coarticulation has consequences for perception as well, enabling the listener to evade just those limitations of the auditory system we referred to earlier. Consider, again, that if the phonetic message were transmitted agrammatically—that is, one acoustic segment for each phonetic segment—then the temporal resolving power of the ear would make it impossible to perceive speech at the rates that we do, in fact, commonly attain. But, as we have seen, the relation between phonetic message and sound is not agrammatic in that sense. Rather, coarticulation effectively folds information about several successive phonetic segments into a single stretch of sound. Moreover, the overlapped activity of several different articulators, e.g., lips and tongue, will often affect the same parameter of the sound, e.g., second formant[3]. At any chosen instant of time, therefore, each acoustic parameter is (commonly) carrying information about more than one phonetic segment. (For fuller discussion, see LIBERMAN et al., 1967.) That being so, the limit on rate of phonetic perception caused by the temporal resolving power of the ear is no longer set by the number of phonetic segments transmitted per unit time, but by the considerably smaller number of acoustic segments into which those phonetic segments have been encoded. Just how much saving is effected in this manner depends, of course, on the size of the encoding unit; and that will surely vary according to the nature of the contiguous phones in the string, rate of articulation, and other factors that we only dimly understand. But a significant amount of encoding will almost always occur—most obviously within the syllable—and it will, at every rate of articulation, effectively reduce the number of discrete acoustic segments that must be perceived.

Consider, now again, the difficulty the auditory system would have in identifying the order of phonetic segments at even moderate rates of speech if each phonetic segment were represented by an acoustic segment. But inasmuch as the phonetic segments are not so represented, the problem of identifying order of discrete acoustic segments does not arise (DAY, 1970; LIBERMAN et al., 1972; COLE and SCOTT, 1974; DORMAN et al., 1975a). Recall how successive phonetic segments are encoded into the same stretch of sound, and imagine, for example, simple cases like [ba] and [ab]. If these syllables are produced at moderately rapid rates of articulation, it will be true of both acoustic patterns that information about the consonant and the vowel is carried simultaneously from the beginning of the sound to its end. But given that the articulatory gestures have opposite directions in the two cases—from closed (consonant) to open (vowel) for [ba] and from open (vowel) to closed (consonant) in [ab]—the acoustic shapes of the two acoustic syllables will be different. Indeed, they will be

[3] A formant is a peak in the resonance curve of the vocal tract. The center value of this peak, specified in Hz, is called the formant frequency.

mirror images: For [ba] the formants will be rising throughout; for [ab] they will be falling. Thus, information about the order of phonetic segments is present in the acoustic signal, not as discrete events in ordered sequence, but as variations in shape or form (LIBERMAN, 1976).

IV. Key to the Speech Code

Suppose the speech code were entirely arbitrary. In that case, a perceiving device could only match the signal against a dictionary of auditory templates, just as if it were using a code book. Of course, the templates could not correspond to segments the size of phones but would, rather, have to be at least as large as the coding unit that encompasses the acoustic consequences of coarticulation. As we remarked earlier, we do not know exactly how large that unit is or how stable it might be in the face of variations in speaking rate, word and phrasal stress, and other conditions of articulation. We can only suppose that, at the smallest, the unit would have to be of approximately syllabic size, since there is normally so much coarticulation within syllable boundaries.

But the speech code is not arbitrary; there is a key that unlocks it. To see the nature of the key, and how it makes sense of the relation between message and signal, we need only remind ourselves that the peculiarities of the speech code are just those that are introduced by the speaker as he lends himself to the processes by which the message is encoded in the sound. When those processes are understood, their consequences can hardly appear arbitrary. Thus, the key to the code is in the manner of its production. We should remark parenthetically that in this respect speech is like the rest of language and different from most other processes: All the complications of language that the hearer must cope with are only those that, as speaker, he "knows" how to introduce; the complications of nonlinguistic perception, on the other hand, are typically not owing to the hearer (or viewer) but are, rather, external to him. At all events, the processes by which speech is produced make it possible to understand the relation between acoustic signal and phonetic message, however peculiar that relation might be.

Although knowing how speech is produced enables us to see why the complications of the code should be peculiar in the way they are, it does not provide an automatic decoding procedure. Thus, we now understand enough about the speech code to be able to synthesize speech by rule (INGEMANN, 1957; LIBERMAN et al., 1959; KELLY and GERSTMAN, 1961; KELLY and LOCHBAUM, 1962; COOPER, 1962; for a summary, see MATTINGLY, 1974). That is, we can build a mechanism that accepts as input a string of phonetic symbols and then, as output, delivers speech. Using rules for the conversion that can be either acoustic or articulatory, the synthesizer produces speech that is imperfect—reflecting our imperfect command of the code—but rather highly intelligible, nevertheless, and reasonably acceptable. Now if we could simply turn those rules around, we should have a working model for speech perception. Unfortu-

nately, the rules for synthesis, like all grammatic rules, work in only one direction: downhill; they take us from message to signal but not the other way. Perhaps there are rules that go in either direction, but they have not yet been found. Thus, to suggest that a listener might use the rules as a key, is only to imply some kind of connection between perception and production, of which more later; the underlying mechanism is, at present, unknown.

B. Special Processes of Phonetic Perception

Surely the most parsimonious way to account for phonetic perception is to invoke only those mechanisms that are more or less common to mammalian (or primate) auditory systems (see MILLER, in press). Can we suppose, then, that such processes are sufficient, or must we look to specializations of various kinds? If specializations do exist, are they in the form of auditory devices that are tuned to respond to the phonetically relevant parts of the speech signal? Or are they more accurately characterized as integral parts of a system, more linguistic than auditory, that is specialized to deal with the peculiarities of grammatic codes? In this section we will consider whether both such types of specializations might exist—the one to deal with the purely acoustic characteristics of the perceptually important parts of the signal and the other to cope with the grammatic code relating the signal to the phonetic information it conveys.

I. Auditory Specializations for Extracting the Phonetically Relevant Information From the Speech Signal

Many important attributes of the speech signal, including some that carry a heavy load of phonetic information, are not physically salient. For example, although most of the linguistically important information is contained in the lowest three formants, the acoustic energy is not tightly concentrated there but is, rather, smeared diffusely over the entire speech spectrum. Or again, despite the fact that consonants carry a far heavier load of segmental phonetic information than do vowels, they are signaled by far less acoustically prominent portions of the spoken syllable. Thus, formant frequency shifts (transitions) that carry important, even essential, information about consonantal place of articulation often make excursions of hundreds of cycles in some 30–40 ms. Since humans seem to have no difficulty in extracting that information, one is led to wonder

whether there may not be devices in the auditory system specialized for that purpose. These devices would be analogous, perhaps, to the feature detectors found in other species.

We have in mind the example of the cat, in which WHITFIELD and EVANS (1965) found single cells ("miaow" cells) responsive to the rate and direction of frequency change. WHITFIELD (1965) pointed out the possible relevance of this finding to the perception of formant transitions in speech when he suggested that such units might be ". . . a final link in the mechanism . . . by which speech-like and similar signals are processed (p. 247)." If WHITFIELD is correct, we would have, not an auditory specialization for language, but rather a general auditory device (perhaps typical of mammals) that is exploited by humans for linguistic purposes.

In fact, an auditory mechanism specialized for language may be difficult to demonstrate, since we obviously cannot apply to humans the electrophysiologic techniques that have been used on animals. It may, however, be possible to approach the matter indirectly, as, for example, by extending to speech, adaptation procedures originally developed in studies of vision. The first to do this were EIMAS and CORBIT (1973). With synthetic syllables (for example, [ba] vs. [pa]) ranging along an acoustic continuum, these investigators used the techniques of adaptation to produce shifts in the position of the perceptual boundary. The results led them to speculate that their procedures had affected a pair of binary phonetic feature detectors, and that adaptation or fatigue of one detector functionally sensitized its opponent. Subsequent work (see COOPER, 1975, for a review) demonstrated analogous effects for other consonantal feature oppositions. If these effects were truly on phonetic features, they would only provide additional evidence for the "reality" of such entities and offer still another method, though potentially a most useful one, for defining their boundaries.

More relevant to our concerns here, therefore, are adaptation studies like the one by BAILEY (1973), which showed that the effect decreased with a decrease in spectral overlap between adapting and test syllables. This suggests that if feature analyzing systems were indeed being isolated, the features were auditory rather than phonetic (for a relevant discussion, see ADES, in press). The finding by BAILEY assumes considerable importance from our point of view, because there is apparently no other kind of evidence for the existence of feature analyzing systems of an auditory sort. Unfortunately, the matter appears not so simple. Further investigation has shown that the degree of adaptation is contingent on so many other aspects of the synthetic continuum, including intensity (GANONG, 1976) and fundamental frequency (ADES, in press), that one may, in the end, be led to doubt the feature interpretation altogether. Perhaps, then, the achievement of the work on selective adaptation will have been to demonstrate the operation of distinct perceptual channels rather than the existence of feature detectors as such. Nevertheless, the investigators may have found a method for exposing processes that respond to linguistically significant parts of the speech signal, and thus to have made possible the discovery of auditory specializations for language.

II. Linguistic Specialization for Recovering the Phonetic Message

Even if auditory detectors of the kind just discussed do exist, they could do no more than extract from the acoustic signal those features that are phonetically relevant. They might thus solve problems created by the fact that speech is, in certain respects, a poor signal, but it would presumably remain to some other device, more phonetic than auditory, to deal with the different fact that speech is a special code. As we were at pains to point out earlier, the peculiar characteristics of the code arise from the way speech is produced, in particular, from coarticulation. We should suppose, then, that the distinguishing characteristic of the phonetic device would be that it somehow makes use of that circumstance (COOPER et al., 1952; LIBERMAN et al., 1952). For the present, the emphasis should be on the word "somehow"; we do not wish to speculate about the underlying mechanism, if only because we cannot offer relevant data. But if there is a device that behaves, by whatever means, as if it "understood" how speech is produced, then we should expect to find evidence for a link between perception and production. Indeed, it would be just such a linkage that would clearly characterize phonetic as against auditory perception (STUDDERT-KENNEDY, 1976; LIBERMAN and PISONI, in press).

In the sections that follow, we will identify several kinds of support for the assumption that there is a phonetic perceiving device and, correspondingly, a phonetic mode of perception. Some of that support is indirect in that it depends on our inability to account for certain phenomena of speech perception in terms of what we now know of how the ear works and what it commonly does; but some is more direct, being based on putative differences between auditory and phonetic perception and, in some cases, on the apparent links to production that characterize the phonetic mode.

1. Coping With the Segmentation

If there were an acoustic criterion that could directly divide the speech stream into segments corresponding in size to the phones, then we should see no need to invoke other-than-auditory processes. No matter how complex in structure the acoustic segments might be, we should suppose that correspondingly complex auditory processes would be equal to the job. As we have seen, however, one of the characteristics of the speech code is that the phonetic information is distributed in curious ways through the sound. This is the most striking disparity between acoustic signal and phonetic message and, from the standpoint of a perceiving device, the most troublesome. Indeed, the disparity is greater than our characterization of the speech code might have implied, since the sound segments do not map onto the phones either in the way they divide or in the way they group. Thus, rapid switches in sound source during the articulation of successive phones may spread the information about a single message segment through several acoustic segments (FANT, 1962, 1968), as when

stop-consonant closure and release into a following vowel yield a brief silence, an explosive release, a period of aspirated noise, and a more-or-less abrupt voice onset. On the other hand, coarticulation may, as we have previously noted, cause the information about several message segments to be collapsed within a single segment of sound.

The severity of the problem is evidenced by the fact that it has resisted solution for many years, as much by those concerned with speech synthesis as by those working on automatic speech recognition. Both groups have been driven to acknowledge that segments the size of phones are not to be found *as segments* in the acoustic stream; the irreducible acoustic unit is of approximately syllabic dimensions, just as we would expect, given the very earliest result of research with synthetic speech (COOPER et al., 1952; LIBERMAN et al., 1952). In the first attempt to "synthesize" speech by commuting (and concatenating) segments of sound excised from prerecorded utterances, HARRIS (1953) found that the "building blocks" had to be larger than phones. Other investigators (PETERSON et al., 1958) later reported some success in producing speech by concatenating prerecorded segments, but the segments they required were a numerous and varied assortment of syllables and "phoneme dyads." Significant improvements in this method of synthesis have recently been made by FUJIMURA (1975, in press), though again the unit must be larger than the phone. And now, even in synthesis by rule, MATTINGLY (1976) has found it advantageous to preorganize the phonetic segments into syllables and then use those larger units as input to his synthesis program. As for the work on automatic speech recognition, it has long been plain that segmentation into phones by a straightforward acoustic criterion is hardly possible (HYDE, 1972), though segmentation into syllables can be done reasonably well (MERMELSTEIN, 1975).

The foregoing considerations and facts suggest that phones are not directly perceived but must rather be derived from a running analysis of the signal over stretches of at least syllable length. There is ample experimental evidence that this is so.

Consider, for example, the matter of segment duration and its role in the perception of phones. It is known that, in English, the contrast between voiced and voiceless stops in syllable-final position (e.g., [ab] vs. [ap]) can be determined by the duration of the preceding vowel (DENES, 1955; RAPHAEL, 1972). But what happens, then, if that vowel is itself preceded by the consonant-vowel transitions appropriate to, say, [b], as in [bab] vs. [bap]? Does the listener pay attention only to the duration of the preceding vowel? Presumably he cannot do that if, as we have suggested, the transition cues for the consonant simultaneously carry information about the vowel. And, indeed, he does not. According to recent experiments (Raphael et al., 1975), the duration used by the listener to determine voicing in the final segment includes all, or almost all, of the transition cues for the consonant in the initial segment.

Given only that result, we might suppose nevertheless that the listener takes one part of the acoustic signal as consonant and another part as vowel, provided we further suppose that the voicing of a syllable-final stop is determined by the sum of the durations of consonant plus vowel. At least two other experiments

suggest, however, that the listener does not compute consonant and vowel durations on different parts of the syllable. In one of these experiments[4], listeners were asked to adjust the duration of a steady-state vowel to match the duration of the medial vowel in a stop-vowel-stop syllable whose formants had parabolic trajectories. As determined by that simple and direct technique, the perceived duration of the medial vowel was found to include a significant portion of the consonant-vowel transitions.

The other experiment dealt with duration as a cue for the perceived identity of a medial vowel and, simultaneously, with the voicing of a final stop, for example, [bɛt], [bæt], [bɛd], [bæd].[5] The results clearly imply that the listener did not assign one part of the syllable duration to the vowel and another part to the consonant. Rather, it was as if he used the whole duration of the syllable, but used it twice: once to determine the identity of the vowel and again to determine whether the syllable-final stop was voiced or voiceless.

That the information about the phonetic segments is spread through the syllable is indicated also by evidence that the flanking transitions in a CVC syllable are used to judge the identity of the medial vowel. For example, OCHIAI and FUJIMURA (1971) recorded natural, but distinctly articulated words and observed no errors of vowel identification. However, when they presented 50-ms portions gated from the vowel centers, listeners' judgments frequently shifted in directions that could be explained by contextual assimilation. Even more striking are the results of STRANGE et al. (1976). They recorded nine vowels spoken in isolation, and the same nine vowels spoken in various CVC frames. Despite the increased acoustic complexity introduced by a dynamic syllable structure, listeners correctly identified the vowels significantly more often when they were presented in a consonantal frame, even a variable one, than when they were presented in isolation. Thus, for the purpose of identifying the vowels, the perceiving system used those parts of the syllable that also contained information about the consonants. That is yet another reflection of the complex relation in segmentation between signal and message. But it also shows that though the perceptual target is a vowel, for which static formant frequencies are often assumed (PETERSON and BARNEY, 1952), the perceptual system nevertheless prefers the dynamic configuration of a syllable, perhaps because it can then take advantage of the many constraints inherent in the way the vocal apparatus works when it coarticulates.

Considering all that is known about the peculiar disparity in segmentation between perceived message and transmitted signal, we suppose that the appropriately segmented percept lies at some remove from the immediately given auditory pattern, and that it is recovered by processes different from those the auditory system is ordinarily called on to provide. As for the possibility that such special processes make reference to production, we can offer no direct evidence about segmentation as such, only the observation that to find the segments, it must help to understand where they were lost.

---

[4] MERMELSTEIN, LIBERMAN, and FOWLER: personal communication.

[5] MERMELSTEIN: personal communication.

2. Phonetic Interpretation of the Sounds of Speech

We should now look more directly at some phenomena of speech perception that depend, presumably, on the same decoding processes that perform the segmentation but pertain more closely to what those segments, once retrieved, sound like to a listener. Do they sound like other sounds or do they not? And when not, is there evidence of a link to production?

a) Impressions of the Difference Between Auditory and Phonetic Modes

To convey a feeling for what we mean by the suggestion that the sound of speech is different from the sound of nonspeech, it may be useful to describe several phenomena that are part of the experience of people who work with synthetic speech. One of these is reflected in an observation made by investigators who used the Pattern Playback, an early research synthesizer that converted handpainted spectrograms and other designs into sound (COOPER, 1950; COOPER et al., 1951). Having succeeded in constructing highly schematized spectrograms, like the one at the top of Figure 1, that nevertheless produced intelligible speech, the investigators thought to take advantage of the flexibility of the Playback in order to destroy the intelligibility of speech by a novel and, they assumed, uniquely effective procedure: Instead of drowning the speech in noise, which was the usual way, they would "mislead" the ear. To that end they added to the spectrogram "false" formants, always continuous with the "true" formants, that improperly connected and extended the proper components of the acoustic pattern. An example is shown in the middle and at the bottom of Figure 1. In fact, as the reader can see, the eye is misled. But the ear was not. When the altered pattern was converted to sound, the listener heard the original phonetic message against a loud background of variously pitched whistles. It was as if the perceptual machinery had separated the acoustic effects that a vocal tract can produce from those it cannot. At all events, the effect was of two qualitatively different kinds of perception—articulate, monotone speech in the one case, complex and very bad "music" in the other.

Much the same kind of phenomenon, though on a smaller scale, can be produced, not only on a device like the Pattern Playback, but also on the more modern parallel-resonance synthesizers now in common use. An example is seen in the contrast between the initial stop consonants of the syllables [ba] and [ga]. As shown in Figure 2, a sufficient acoustic cue is the direction of the second-formant transitions, rising for [b] and falling for [g]. Now, given our knowledge of psychoacoustics, we should suppose that those cues would sound like rising and falling glissandos or like chirps of different pitch, depending on how rapidly the formants moved on the frequency scale. And, in fact, when we present the formant-transition cues by themselves, as shown in the inset of Figure 2, that is exactly how they do sound (MATTINGLY et al., 1971; SHATTUCK and KLATT, 1976). But what do we say, then, about the fact that those same transitions are heard in the context of speech as the abstract linguistic events we can only describe as [b] and [g]? Of course, the transition cues are isolated in the one case but part of a larger, if otherwise constant, pattern

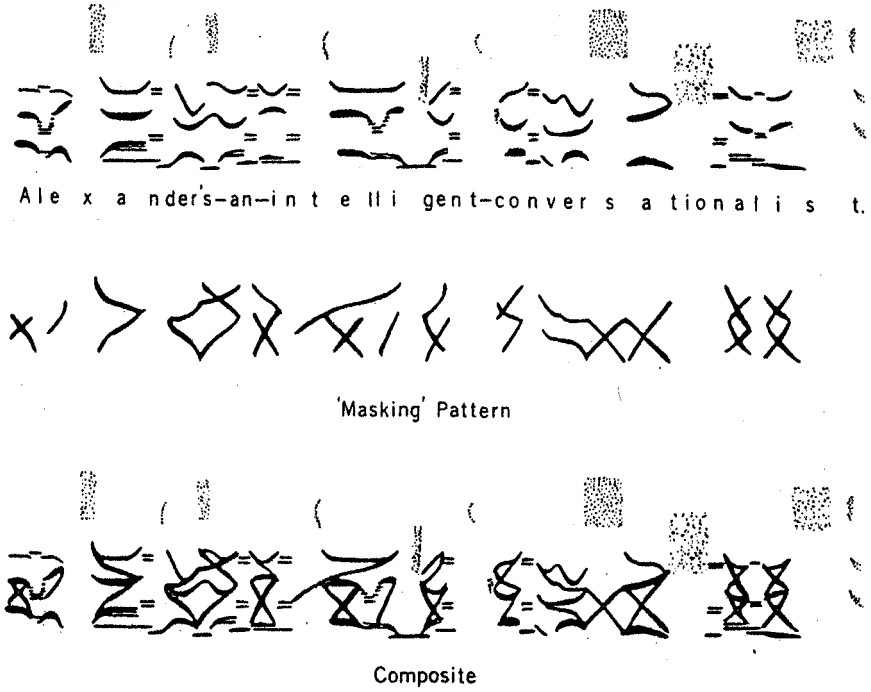

Fig. 1. At the top, a hand-drawn spectrogram appropriate for synthesis of a sentence; in the middle, a pattern intended to "mask" the sentence by "misleading" the ear; and, at the bottom, the composite of sentence and "mask", which produces, again, the perception of the sentence, plus a dissociated set of whistles and noises

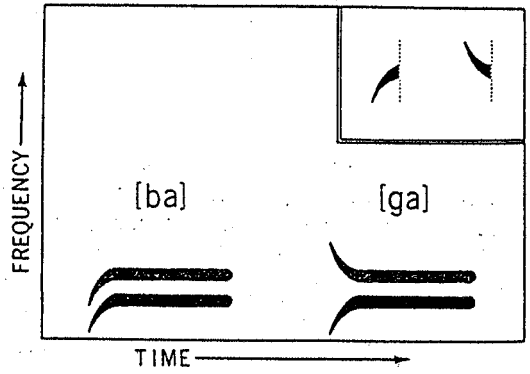

Fig. 2. Spectrographic patterns sufficient for synthesis of [ba] and [ga]. Inset: The second-formant transitions that cue the perceived difference between the syllables, but sound, in isolation, like chirps

in the other, so that we might attribute the difference in perception to some kind of auditory interaction.

But even when the transition cues are in exactly the same acoustic context, it is possible to hear them, simultaneously, as phonetic stops and auditory chirps. That effect was created by RAND (1974) in the following way. Into

one ear he put all of the first formant and the steady-state parts of the second and third formants, while into the other ear he put just the transition cues (of the second and third formants) that distinguish [ba] and [ga], being careful to synchronize them properly with respect to the rest of the pattern. Though there is but one context—and indeed one brain—the formant transitions will, in this situation, often simultaneously produce two very different perceptions: the syllable [ba] (or [ga]) and a rising (or falling) chirp.

Essentially the same kind of effect has been created, though successively now instead of simultaneously, as part of an experiment designed by BAILEY et al. (1977) to permit comparison of speech and nonspeech perception. The stimulus patterns are similar to those commonly used in research with synthetic speech in that they contain transitions appropriate to several stop consonant-vowel combinations, followed by vowel steady-states; they differ from those normally used in that the formants are replaced by pure tones, one for each formant and set to its center of energy. On first being presented with such patterns, listeners hear them as a complex of tones, but after some time they begin to hear them as speech. We will not here presume to report on the results of the experimental comparisons that the study was designed to permit; we only remark the phenomenon, which is that there is a striking difference in subjective impression, depending on whether the listener is perceiving the stimulus patterns as tones or as speech; thus, it offers yet another way to gain a general appreciation of the perceptual differences between speech and nonspeech.

At all events, it is just such qualitative contrasts in perception as we have described here that can convey to a listener a direct impression of what we mean by the distinction between auditory and phonetic modes. We turn now to some relevant experimental observations.

b) Acoustic Cues as a Source of Information About What the Speaker's Vocal Tract Did

Those aspects of the speech signal that, when varied, cause phonetically significant changes in perception are known as "acoustic cues." It is to those cues that we should now look, because we find there the clearest evidence for the link between perception and production that characterizes perception in the phonetic mode. No single piece of evidence is, by itself, wholly convincing; it is only the pattern that tells. For when we view the data in the light of known or imaginable auditory processes, we see a number of unconnected facts that require, apparently, an equal number of ad hoc assumptions. If we apply Occam's razor, however, we find a more-or-less comfortable fit to the single assumption underlying this chapter: that the acoustic cues are processed, not only in the auditory system, but also at some more abstract, phonetic remove; there, an appropriately specialized device uses the articulatory information provided by those cues to shape the listener's perception of what the speaker said.

α) *A Simple Example.* To see how an acoustic cue—silence—might provide information about a phonetically important gesture, we should consider the fol-

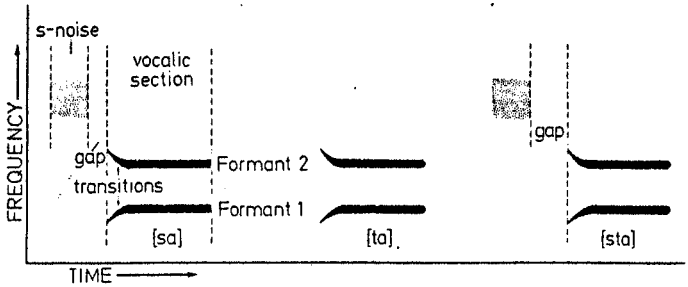

Fig. 3. Schematic spectrograms illustrating the importance of silence for the perception of a stop consonant: [sa] becomes [ta] when the noise is removed, or [sta] when a silent interval of appropriate length is introduced between the noise and the rest of the syllable

lowing facts about fricatives and stop consonants. A speaker cannot produce a stop consonant without closing his vocal tract for a brief period, and he cannot close his vocal tract without producing a period of silence. Hence, silence might be important to the perception of stop consonants, especially if the perceptual processes "know" that stops require closure and that closure results in silence. It is relevant, then, to discover that in the perception of stops silence is, in fact, an important condition.

Suppose, for example, we record the fricative-vowel syllable [sa]. As shown schematically in Figure 3, the acoustic pattern consists of a patch of noise, associated with the fricative, followed by a vocalic section. The vocalic section begins with the formant transitions characteristic of the fricative [s] when coarticulated with the vowel [a]; there follow, then, the steady-state formants characteristic of the (drawn-out) vowel [a]. It should be noted about the formant transitions at the beginning of the vocalic section that they are also appropriate, at least approximately, for the stop consonants [t] and [d], which have the same place of production as [s]. Now if we remove the patch of noise, listeners will commonly hear [ta], not [a]—that is, they will hear a stop consonant where none was before. If we now replace the s-noise in such a way as to create a silence of about 50 ms between it and the vocalic portion, listeners will again hear the stop, this time in [sta]. We should say, parenthetically, that the same kind of effect can be obtained starting with a stop-vowel syllable like [ta]. In that case, putting s-noise immediately in front of the syllable will cause the listener to hear [sa], not [sta]; if the listener to hear [sta], we must create a short period of silence between the s-noise and the vocalic section.

We see in this example that silence has just the sound we should expect it to have, given the assumption that it tells the listener whether or not the speaker closed his vocal tract long enough to have produced a stop consonant. But, surely, there might be other, perhaps more parsimonious, assumptions. We note in this connection that our examples conform to the paradigm for auditory forward masking, so we should take account of the possibility that the transition cues are simply being masked when the noise is too close to

them; or we suppose, more vaguely, that there is some (not previously discovered) auditory interaction between silence and the transition cues which causes us to hear the peculiar sound of a stop consonant.

But there is considerable evidence that such alternative assumptions will not hold. Note, first, that in fricative-vowel syllables like the [sa] of our example, it has been found that the formant transitions contribute significantly to the perception of the fricative (HARRIS, 1958; DARWIN, 1971). We should suppose, therefore, that the transition cues are "getting through"—that is, they are not being masked by the s-noise. It is only their (phonetic) interpretation as fricative (when the silence is relatively short) or stop (when the silence is relatively long) that is affected.

More evidence of the same kind comes from a study of selective adaptation by GANONG (1975). There, the first step was to measure the shift in the (perceived) boundary between [b] and [d] caused by adaptation with the syllable [dɛ]. Then, a patch of s-noise was placed in front of the [dɛ] so that it sounded, as in our example, like [sɛ]. When that syllable ([sɛ]) was used as the adapter, the effect on the [b-d] boundary was found to be just as great as it had been with [dɛ]. From that it follows not only that the transition cues were getting through—that is, that they were not being blocked by the noise when they were perceived as [sɛ] rather than as [dɛ]—but that they were getting through in full strength.

A third kind of evidence comes from a comparison of how the transition cues are perceived when, in an acoustic context otherwise like that of our example, they are in or out of a proper syllable (DORMAN et al., 1975). The syllable consisted of a patch of s-noise followed by a vocalic portion that was either [pɛ] or [kɛ]. With the noise up close, listeners reported hearing [sɛ], not [spɛ] or [skɛ]; [spɛ] and [skɛ] were perceived only when there was an appropriate interval of silence between the noise and the rest of the syllable. In the other (nonsyllable, nonspeech) condition, the transition cues were isolated from the rest of the vocalic section, in which circumstance they sounded like chirps of different pitch and could easily be identified on that basis; then they were placed, as in the speech patterns, after the patch of s-noise. In that condition—that is, when heard as chirps—the transition cues were correctly identified even when there was no silent interval separating them from the noise. Thus, they were not significantly masked by the noise, but, just as important from our point of view, their perception was not changed in any qualitative way—that is, there was no apparent interaction among noise, silence, and transitions.

Much the same kind of result has been obtained with stops in syllable-final position (DORMAN et al., 1975). First, it was established that in the disyllables [bɛb dɛ] and [bɛg dɛ], listeners could correctly perceive the syllable-final stops [b] and [g] only if there was a sufficient period of silence (approximately 60 ms) between the syllables. Then, the second-formant transitions that were the only acoustic difference between the [b] and the [g] were isolated from the rest of the pattern of the first syllable, in which circumstance they were heard as two quite different chirps, and presented, as in the first condition, before

the syllable [dɛ]. Listeners correctly identified the chirps most of the time, even when there was no silence at all between them and [dɛ]; the amount of masking was relatively slight, nothing at all like the total effect that had occurred in the case of the speech sounds, and there appeared, again, to be no interaction-caused change in the phenomenal "quality."

So much, then, for the possibility that silence is a necessary condition for perception of stops because it prevents masking of the transitions or because it collaborates in some auditory interaction with them. We turn now to the fact that in the absence of transitions and other stop-consonant cues, silence can be a more nearly *sufficient* condition for perception of a stop.

Suppose we insert the appropriate amount of silence between the noise of a fricative and a vocalic section so structured that no stop is heard when it is presented by itself. Begin, for example, with the syllable [lit], then put a patch of s-noise in front of it. In that case, the resulting syllable is perceived as [slit] if there is no silence between the noise and the vocalic section, but as [split] if the silence is increased sufficiently (DORMAN et al., 1976; ERICKSON et al., 1977). For a simpler example, consider that an appropriate amount of silence inserted between a patch of s-noise and the vowel [i] will produce [ski]; a similar arrangement with [u] will produce [spu] (SUMMERFIELD and BAILEY, 1977). Notice, too, in these last cases that silence is not only a sufficient cue for stop consonant manner but that the "place" of the perceived stop (whether [k] or [p]) is different, of which more later.

Silence has also been shown to be a sufficient condition for distinguishing fricative from affricate both in syllable-initial and syllable-final positions. Thus, one can record the word "say" and the word "shop" and then convert between "say shop" and "say chop" by varying the interval of silence between the two words (DORMAN et al., 1976). Or one can record "dish" and convert it to "ditch" by introducing an appropriate amount of silence between the vocalic part of the syllable and the fricative noise at the end.[6]

The foregoing considerations all imply that the perception of silence in our simple example is not only auditory but also phonetic. As a phonetic percept, it conforms to a fact about the speaker's production—namely, that a stop consonant cannot be produced without closing the vocal tract. Of course, such conformity could occur only if there were a phonetic perceiving device specialized to make use of the information about articulation, and if there were, correspondingly, a phonetic mode of perception.

β) Equivalence in phonetic perception of different acoustic cues produced by the same articulatory gesture. It is a commonplace of speech and speech perception that different acoustic cues may have equivalent effects in phonetic perception. That is of interest because the cues are often so different acoustically that it is hard to conceive how they might be related from an auditory point of view. The relevant facts fall into several classes; we will here offer samples of each.

Perhaps the simplest class comprises those ubiquitous cases in which there are multiple (and distributed) acoustic consequences of the same articulatory

[6] Raphael and Dorman: personal communication.

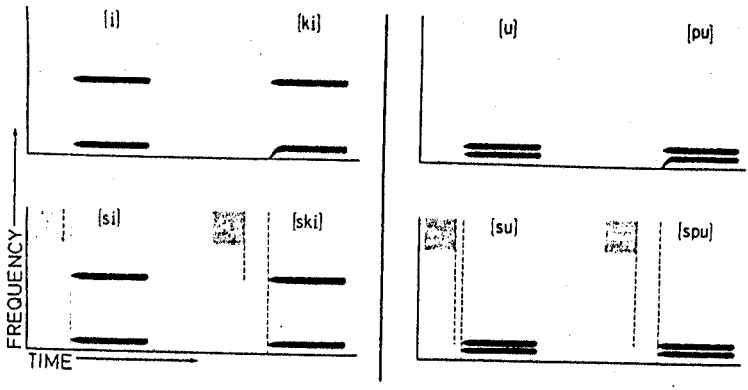

Fig. 4. Spectrographic patterns that illustrate how two very different acoustic cues—a transition of the first formant (top half) and an appropriate interval of silence (bottom half)—are phonetically equivalent in the perception of stop consonants

gesture. Consider again the example of the preceding section that is owed to SUMMERFIELD and BAILEY: An appropriate interval of silence between a patch of s-noise and the vowel [i] (or [u]) causes the listener to hear [k] in [ski] (or [p] in [spu]). We now represent that fact schematically in the top half of Figure 4. In the bottom half we represent the companion fact, discovered in earlier research on the "locus" of the stops, that a rising transition at the beginning of the first formant of [i] (or [u]) will also cause a listener to hear the stop [k] in [ki] or ([p] in [pu]) (DELATTRE, et al., 1955). Now we note the perceptual equivalence of about 60 ms of silence, which is the cue in the top half of the figure, and the rising frequency modulation at the beginning of the first formant, which is the cue in the bottom half, and we ask what that amount of silence and that kind of sound could possibly have in common. Nothing, we should think, when we consider them from an auditory point of view, but in articulation they have an obvious bond. To say [ski] (or [spu]), rather than [si] (or [su]), the speaker must close his vocal tract, which produces the silent interval; and then he must open it, which produces the rise in frequency of the first formant. Thus, the two very different cues are the distributed acoustic results of an essential component of the stop-consonant gesture. Given that they sound alike—either can produce the perception of stop consonant—we should suppose it is because they refer to the same articulation.

For this same example, it remains to take account of the fact that the perceived stops had two different places of production, velar in [ki] (or [ski]) and labial in [pu] (or [spu]). We note, first, that energy at frequency levels corresponding to the second-formant levels of [i] and [u] is appropriate for closure of the vocal tract at the velar and labial places, respectively. That helps us to understand why [i] becomes [ki] (or [ski]) and [u] becomes [pu] (or [spu]) when sufficient cues for the stop manner are added. But notice now a fact that is more relevant to our present purposes, which is that these differences in perception of place of production occur in the same way regardless of how

the manner dimension was signaled. Thus, our two very different acoustic cues—silence and sound—are equivalent, not only in their ability to produce the perception of manner, but also in the way they combine with the other information in the signal to produce the perception of phonetic place.

Given our assumption of a link between production and perception, and given that a linguistically significant gesture almost always has multiple acoustic consequences, we should expect to find many other instances of phonetic-perceptual equivalence among cues that are very different in acoustic-auditory terms. Just how many must depend on how finely we dissect the acoustic signal into separate cues, and how often, in experiment, we play the cues off against each other. Relevant studies have already made an impressive record. It reaches back in time to an extension by LISKER (1957b) of an earlier study (LISKER, 1957a) on the voicing distinction in poststress position (as in "rabid" vs. "rapid"). Having determined in the earlier work that duration of intersyllabic silence is an important voicing cue, LISKER then found that specifiable amounts of that temporal cue could be traded for specifiable settings of spectral cues (extent of appropriate transitions of the first formant at the end of the first syllable and the beginning of the second). Now, in a recent experiment on the distinction between fricative-vowel and fricative-stop-vowel, SUMMERFIELD and BAILEY (1977) have established and precisely measured the equivalence of silence on the one hand, and, on the other, such spectral cues as the frequency at which the first formant starts and the extent of the first-formant transition.

There is also evidence of equivalence in phonetic perception among different kinds of temporal cues. Referring again to LISKER's experiment, we note his finding of an equivalence between duration of intersyllable silence and the duration of the first syllable of the word. In a recent experiment[7] referred to earlier, on the distinction between [dish] and [ditch], there is an equivalence between the duration of silence separating the vocalic position of the syllable from the noise and the duration of the noise portion of the fricative (or affricate). Also new is the discovery of a similar equivalence between duration of silence and duration of noise in the contrast between fricative-vowel and fricative-stop-vowel.[8] In all these cases time is traded for time; but in the one period of time there is silence, in the other sound.

In the spectral domain, too, equivalences among different cues are not hard to find. For example, an early paper (COOPER et al., 1952) presented preliminary evidence for the separate contributions of several acoustic cues to the perception of the [m—l] distinction, among others. Later, it was shown more clearly that in the perception of place of production in stops, second-and third-formant transitions made independent constribution (HARRIS et al., 1958; HOFFMAN, 1958). In the current literature is a particularly elegant study of the voicing distinction by SUMMERFIELD and HAGGARD (1977) that reports an equivalence between the starting point of the first formant and the variable known as "voice-onset-time" and shows explicitly how these acoustically disparate cues are related in articulation. A somewhat similar result with two voicing cues—fre-

---

[7] Raphael and Dorman: personal communication.

[8] Bailey, Summerfield and Dorman: personal communication.

quency of the fundamental frequency and voice-onset-time—has been found recently by MASSARO and COHEN (1976) [cf. HAGGARD et al. (1970)], though an articulatory basis was not made explicit.

Having offered several examples of the equivalences in phonetic perception between different acoustic cues that are the consequences of the same articulation, we should bring this section to a close. But not without first saying that it is hard to know where the list of relevant examples should end. Should we, for example, include the kind of equivalence that is found between spectral cues for syllable-initial consonants and the duration of the syllable[9], or between silence as a cue (for voicing, or place, or gemination) and the tempo of the surrounding speech (PICKETT and DECKER, 1960; PORT, 1976), or between the setting of the second-formant transition as a cue for the stops and the position of the first formant (RAND, 1971)?[10] It is when we try to answer that question, and thus to define the boundaries of the phenomenon we are here considering, that we see most clearly how unsatisfactory from a theoretical point of view is the notion of acoustic cue. We find it useful, even necessary, when we want to refer to those pieces of sound that an experimenter varied and found to be effective. But if the cues are to be fitted into a conceptual frame—as something other than items in a list—we should regard them as information about the behavior of a speaker's vocal tract.

So far, we have considered only those different acoustic cues that are phonetically equivalent because they are the common products of a single articulatory gesture. These are, perhaps, the least complex and most telling of the instances that imply a link between speech perception and speech production. But they are not the only ones. Equally numerous are the cases in which there is phonetic equivalence between acoustic cues that are very different because the phone they signal is produced in different contexts (LIBERMAN et al., 1967; but see STEVENS, 1975). In these cases, too, we suppose that a common articulation is responsible for that which is common in the perception. Of course, such articulations as these can hardly be identical in all particulars, since they are linked to the gestures for the surrounding phones, and these change, of course, with each new context; the commonality can only be seen in terms of shared components, whether end targets or inferred motor commands (for relevant discussion, see MACNEILAGE, 1970). But given such articulatory similarity as there may be, gross differences in acoustic signal can and often do arise with changes in context, primarily as a consequence of coarticulation. It is the more important, then, to give some attention to these context-conditioned variations in the cues because, as we said in an earlier section, coarticulation is the essence of the speech code.

To illustrate how acoustic cues that vary because of phonetic context are nevertheless equivalent in phonetic perception, we choose an example that shows two kinds of contextual effects—one that depends on variations in the identity of the phone following the target phone, and another having to do with variations in the position of the target phone in the syllable. The example is the pair

[9] SUMMERFIELD: personal communication; MILLER and LIBERMAN: personal communication.
[10] Also, BAILEY: personal communication.

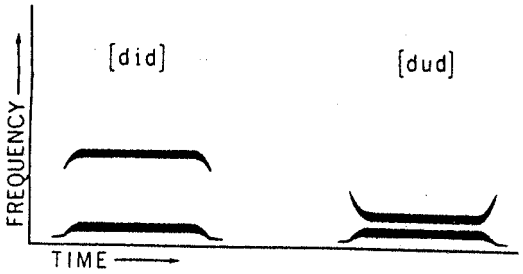

Fig. 5. Spectrograms sufficient to produce the syllables [did] and [dud], illustrating the variation in acoustic cues for the stop consonant [d] that can occur as a function of vowel context ([i] vs. [u]) and position in the syllable (initial vs. final)

of syllables [did] and [dud], shown schematically as two-formant approximations in Figure 5 and taken from the results of early experiments on the stops (DELATTRE et al., 1955). (These patterns are appropriate, and reasonably sufficient, for synthesizing the intended syllables.) Having noticed that the lower (first) formant is the same in the two cases, we fix attention on the higher (second) one. We see there that, as a consequence of coarticulation, a phonetic alteration limited to the middle (vowel) segment of a consonant-vowel-consonant syllable does not change only the middle portion of the sound; rather, it changes the entire second formant. The transition cues for [d] are therefore in very different positions in the spectrum, being relatively high in frequency for [did] and low for [dud]. Moreover, the transition cues for stops in corresponding positions in the syllable are opposite in direction—for [did] they are rising in initial position and falling in final position, but for [dud] they are falling in initial position and rising in final position. Of course, the inference we would draw from these cases is much the same as that we draw from those in which the context was fixed and the disparate acoustic cues were the products of exactly the same gesture: The cues are presumably interpreted by a phonetic device that acts as if it knew how they were produced. But if the device has that ability, then it can conceivably do more than just "hear through" the context-conditioned variation in the cues so as to arrive at the canonical form of the phonetic segment; it might also be able to take advantage of the fact that such variation produces a special kind of redundancy in the signal and provides important information about such aspects of the phonetic structure as sequential order, juncture, linguistic stress, and tempo. If so, then the acoustic variation that is produced by articulation (and coarticulation) in different contexts would not be an obstacle to perception but a considerable help and, correspondingly, a most important characteristic of the speech code.

γ) *Nonequivalence in phonetic perception of an acoustic cue produced by different vocal tracts: ecologic constraints of a phonetic sort.* Given that phonetic perception is somehow shaped by what a vocal tract does, as we have suggested it might be, we should ask: whose vocal tract? Common sense suggests that it can hardly be that of the listener, nor yet that of the speaker; most plausibly,

it must be some abstract conception of vocal tracts in general. We should expect, then, that the phonetic device would behave as if it knew, for example, that two vocal tracts can do what one vocal tract cannot. In that case, acoustic cues might have one effect or another, depending on whether they were produced by one speaker or by two. That such ecologic considerations are important is indicated by experiments.

One experiment dealt with the perception of the syllable-final stop in the example of [ɛb dɛ] vs. [ɛg dɛ] that we described earlier. There, it will remembered, listeners could hear the [b] or [g] only if there was a sufficient interval of silence between the syllables, presumably because the phonetic perceiving device "knew" that the speaker could not have produced both stops without closing his vocal tract for a certain period of time. But two vocal tracts—one saying [ɛb] (or [ɛg]), the other [dɛ]—can produce the disyllable [ɛb dɛ] (or [ɛg dɛ]) with no silence at all between the two syllables. The experiment revealed that listeners behaved accordingly: When a single speaker produced both first and second syllables, a silent interval of some duration was necessary for perception of the syllable-final stops, but when one speaker produced the first syllable and another the second, listeners heard the syllable-final stops even when there was no intersyllabic silence at all (DORMAN et al., 1975b).

The other experiment dealt with the distinction between fricative and affricate (in "shop" vs. "chop") that we also described earlier. In that case, inserting a sufficient amount of silence between "say" and "shop" caused the listener to hear "chop." Our assumption was that this occurred because the silence informed the listener that the speaker had closed his vocal tract, as he must to produce the affricate. But two vocal tracts—one saying "now say" and the other "chop"—can produce "now say chop" with no silence at all between "say" and "chop." Thus, with two speakers, the size of the interval of silence provides no useful phonetic information. The results of the experiment suggested that the listeners' perceptions took account of that fact. Starting with "now say" and "shop", and given a silent interval appropriate for "chop," listeners did indeed hear "now say chop" if there was only one speaker; but if there were two, the listeners heard "now say shop" at all intervals of silence (RAPHAEL et al., 1975).

Those results imply that the vocal tract to which the perception is linked is a very abstract one indeed, as we should have expected. But they also provide additional support, of a rather different kind, for the hypothesis that some such link does indeed exist.

δ) *Addition of equivalent acoustic cues: algebraic sums in the phonetic mode.* The claim that two very different acoustic cues are equivalent in phonetic perception is largely based on the experimental demonstration of a trading relation between them. Thus, it has been determined that some number of milliseconds of a temporal cue is equal to some particular setting of a spectral cue. An implication is that the two cues together will summate algebraically to enhance or reduce the perceived phonetic contrast, depending on just how they are combined. We believe this to be worth remarking, because cues that are algebraically summed would have positive and negative signs only in the phonetic domain, or so it would seem. An example may show why.

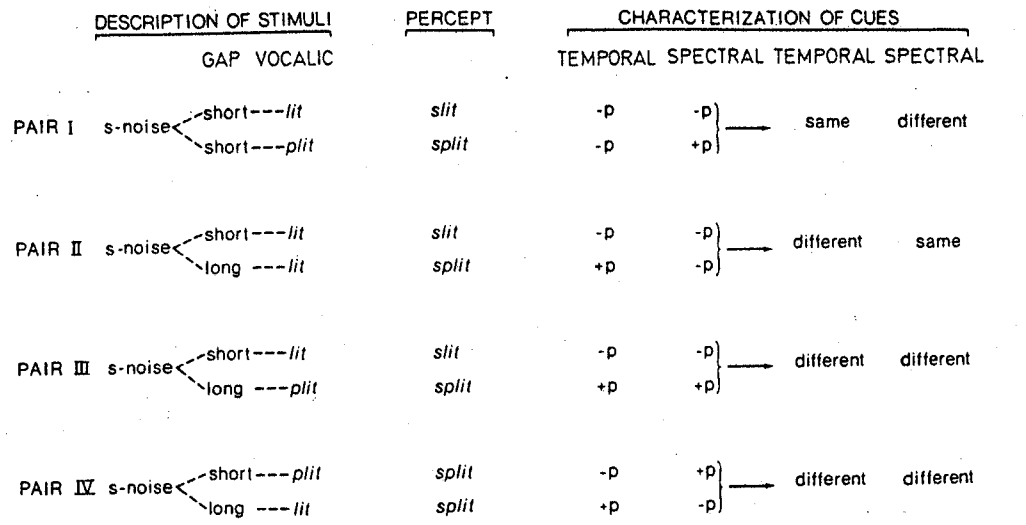

Fig. 6. Diagrams illustrating how spectral and temporal cues separately produce the same phonetic distinction (Pairs I and II) and how, taken together, they either enhance that distinction or reduce it (Pairs III and IV)

Recall the fact, described earlier, that an appropriate period of silence inserted between an s-noise and the syllable [lit] will produce [slit] if the interval is relatively short but [split] if it is sufficiently long. Given that we can, of course, also convert [lit] to [plit] by appropriately changing the spectrum at the beginning of the vocalic syllable—specifically, by altering the formant transitions—it follows that we can use the spectral maneuver to interconvert between [slit] and [split] while holding the temporal cue fixed (ERICKSON et al., 1977). Those facts are diagrammed in Figure 6 as Pairs I and II, where we characterize the cues as "minus p" or "plus p" to indicate the way they bias the perception. In this case, as in the others we described earlier, we see how a phonetic distinction can be produced by either of two cues, one spectral, the other temporal. In pairs III and IV, both the spectral and temporal cues differ between the members of the pair, but in different ways. In the one case (Pair III), the combination enhances the perceived difference, while in the other (pair IV), it permits the minus and plus biases to summate (algebraically) so as to produce two percepts (split) which are the same or very little different (LIBERMAN and PISONI, in press).

To appreciate the significance of the perceptual addition exemplified in Figure 6, we should think of it as a paradigm for comparative studies with nonhuman animals, a potentially enlightening endeavor because phonetic perception and the algebraic summation that goes with it exist presumably only in creatures that speak. Others would perceive the stimuli of Figure 6 in an auditory way. Hence, they should find the pairs that differ by two cues (III and IV) to be more discriminable that those (I and II) that differ only by one, and, further, the pairs with two-cue differences should be almost equally discrimina-

ble. Note, incidentally, how relatively easy it would be to test that hypothesis, not only with animals but with human infants: The measurement—relative difficulty of discrimination—is surely one of the easiest to make, and the order of difficulty to be expected from nonhuman animals is very different from that already obtained with us human beings.

c.) Nonequivalence in phonetic perception of the same or similar acoustic cues. Just as the processes of speech production cause different acoustic cues to be correlated in articulation and (hence) equivalent in perception, so also, if in a somewhat more complex way, do they sometimes cause the same cue to be uncorrelated in articulation and (hence) different in perception. An early instance of this was seen in the first "synthetic" experiment on the stops, where it was found that a burst centered at 1440 Hz was perceived differently in front of different vowels (LIBERMAN et al., 1952). Subsequently, much the same effect was found with real speech (SCHATZ, 1954). More recently, the general effect has been confirmed, though with better methods for controlling the stimuli, but now it is seen that the exact nature of the effect varies somewhat depending on just how much of the "real" burst is used and just where it is placed in time with reference to the vowel.[11]

Another example concerns silence, about which we have already heard so much. Having seen earlier that it is a cue for the perception of phonetic segments, we should note now that it is effective in regard to all three phonetic dimensions: manner, voicing, and place. In connection with manner, we should remember that an appropriate amount of silence, placed between the noise of a fricative and a vocalic piece of sound, will produce the perception of a stop consonant, the perceived "place" of the stop depending on the nature of the vocalic section. We should also remember that, in similar fashion, silence will produce the affricate manner when introduced appropriately between, for example, the word "say" and the word "shop." In regard to voicing, we saw earlier that variations in the duration of intersyllable silence will convert a poststress voiced stop (as in "rabid") into voiceless (as in "rapid"), and vice versa. Now we turn to the dimension of place, and point out, as we had not before, that in a disyllable like "rabid", reductions in the duration of intersyllable silence will cause the listener to hear "ratid"—that is, a stop with a different place of production (PORT, 1976). This perceptual change correlates—not accidentally, according to our hypothesis—with the fact that a speaker closes his vocal tract for a shorter time when he says "ratid" than when he says "rabid." Given the utterance "rabid" and an artificially shortened silence between the syllables, it is as if the listener heard "ratid" because his phonetic perceiver knows that the speaker could not have said "rabid" since he did not close his vocal tract long enough. In sum, then, a single acoustic dimension, duration of silence, produces contrasts on each of three phonetic and perceptual dimensions—manner, voicing, and place. That curious situation arises because the very different kinds of articulations—indeed, the different sets of muscles—that underlie the independence of those dimensions in the phonetic and perceptual domains happen to converge on a single acoustic dimension.

[11] RAPHAEL, DORMAN, and LIBERMAN: personal communication.

Perhaps the reader will have noticed that we did not specify the amounts of silence that are appropriate in the aforementioned cases, and he will quite naturally wonder if they are within the same ranges for the three phonetic and perceptual dimensions. We did not specify because the appropriate durations vary according to how the other relevant cues are set, and much of this remains to be worked out. It is reasonably clear, even now, that the durations of silence for manner and voicing overlap greatly. For place, there probably is some overlap with voicing, depending on just how the other cues are set, but at this moment the relevant data have not been gathered.

Since we have, up to this point, looked only at the segmental aspects of phonetic structure, it may seem inappropriate that we should now broaden our view to glimpse those other aspects that pertain to prosody and syntax. But the temptation to do so is great because there is, at just this juncture, a very natural and interesting connection. The point is that the duration of a syllable conveys information not only about the identity of the phonetic segments it comprises, as we have already seen, but also about the tempo (rate of articulation), degree of linguistic stress, and position in the syntactic frame. We do not know how the separate contributions to duration are sorted out in perception, but, as KLATT (1976) has pointed out, considerations of simple logic suggest that the perceiver can hardly arrive at his decisions in some particular order, one at a time, since each decision would appear to depend on every other one. In any case, it does appear that, in production, these several aspects of the message are encoded into the same aspect of the signal, and then, in perception, properly recovered.

C. Is Phonetic Perception Necessary?

So far we have assumed the perceptual reality of phone-size segments. In this final section we propose to justify that assumption.

I. Preliminary Remarks

No one, of course, doubts that linguistic utterances are perceived as sequences of wordlike, or morphemic, segments. But the processes by which these segments are extracted from the acoustic signal are far from certain. Do we, in perceiving speech, pass directly from the overall acoustic shape of the constituent morphemes to their syntactic and semantic attributes, or do we, rather, first analyze at least some portion of each utterance (with the possible exception of nonpropositional greetings, interjections, and expletives) into its phonologic components,

and only then proceed to syntax and meaning? Certainly, the phonologic attributes of each morpheme are available to consciousness. But what is the form that gives access to the listener's lexicon? Is lexical storage analog and isomorphic with gross auditory shape, or is it digital and isomorphic with phonologic structure?

We should make clear from the outset that the precise form of any possible morphemic sound pattern is not our concern. We have spoken until now of phones, since we take a phonetic representation to be at the first remove from the auditory signal and to be perceptually available, even though not attended to in normal listening. However, for the present discussion, it is a matter of indifference whether the representation is assumed to be a feature matrix, a sequence of phones, or a sequence of more abstract phonologic segments. Our only concern is whether the form is segmented or unsegmented.

II. Evidence Against Segments Smaller Than Syllables

Consider, first, the grounds for believing the perceptual representation to be unsegmented. Foremost is the fact, to which we have repeatedly alluded, that phonetic segments are not discretely arrayed in time as are letters of the alphabet in space, but are, rather, transmitted simultaneously or with considerable shingling. This fact alone has led some students to abandon the phone as a perceptual unit in favor of the context-sensitive allophone (WICKELGREN, 1969; but see HALWES and JENKINS, 1971) or the syllable (MASSARO, 1972; WARREN, 1976a, 1976b).

A second line of argument draws on reaction-time studies, demonstrating that listeners, asked to monitor a word list or sentence, display successively shorter reaction times as the target item increases in duration from phone to syllable to word (SAVIN and BEVER, 1970) and even to sentence (BEVER, 1970), suggesting a perceptual progression from larger unit to smaller rather than the reverse. The solution to this paradox was provided by MCNEIL and LINDIG (1973), who showed that reaction times are, in fact, shortest for the items of which a list is composed or, in other words, for those items to which the experimental situation has drawn the listener's attention (see also FOSS and SWINNEY, 1973). RUBIN et al. (1976), have elaborated these conclusions, arguing that such monitoring experiments do not measure the time taken to process the targets perceptually, but rather the time taken to bring them into consciousness (cf. STUDDERT-KENNEDY, 1974, p. 2366). This does not, of course, preclude the possibility that normal processing entails unconscious access to the lexicon through phonologic analysis, since, in all likelihood, these experiments have no bearing on normal perceptual processes at all. However, it does invite the reflection that the several attributes of a morpheme—its phonologic components, syllabic structure, syntactic and semantic markers—may all be simultaneously available to the listener, once access to his lexicon has been granted by overall acoustic (or, in reading, visual) shape (WARREN, 1976b).

Finally, a broad line of argument springs from the suspicion that the study of speech perception has been tied to the isolated syllable and its components at the expense of attention to the overall acoustic pattern of running speech. This overall pattern, or prosody, certainly conveys important information. SVENSSON (1974), for example, has shown that the perceived form of hummed speech (that is, speech lacking all the acoustic cues for its phonetic segments) is often syntactically correct. MARTIN (1975) has argued that speech rhythm may enable listeners to predict upcoming stresses. And DARWIN (1975) has even induced listeners to reveal a preference in some circumstances for good prosody over good syntax and meaning. These and other studies (for example, COHEN and NOOTEBOOM, 1975, passim) do suggest that the role of prosody in speech perception may have been underestimated. In fact, if we combine these studies with recent work on possible invariant acoustic correlates of distinctive features in the speech stream (STEVENS, 1975), we may be tempted to propose once again the "novel theory of speech perception," first put forward by CHOMSKY and MILLER (1963, p. 311) and elaborated by CHOMSKY and HALLE (1968, p. 24), by which a few more-or-less invariant acoustic properties give the listener access to his lexicon and so precipitate a plausible syntactic and semantic analysis of an utterance.

In short, a fair body of evidence suggests that the acoustic structure of spoken utterances may be sufficient to gain access to the listener's lexicon, or at least his syllabary, without an intermediate stage of phonologic analysis. However, we do not believe that this view is correct and in the following sections we will try to explain why.

III. Evidence in Favor of Segments Smaller Than the Syllable

1. Experimental Evidence

There is a great weight of evidence for the psychological reality of every level of phonetic analysis, from feature to phone to syllable. We have reviewed much of this evidence elsewhere (STUDDERT-KENNEDY, 1976). Here we do no more than remark that studies of speaking errors (BOOMER and LAVER, 1968; FROMKIN, 1971), perceptual confusions (MILLER and NICELY, 1955; MITCHELL, 1973), synthetic speech continua (LIBERMAN et al., 1967), dichotic listening (SHANKWEILER and STUDDERT-KENNEDY, 1967; STUDDERT-KENNEDY and SHANKWEILER, 1970) and "verbal transformations" (GOLDSTEIN and LACKNER, 1973; WARREN, 1976a) leave little room for doubt that both phones and features have some form of psychological reality. To this experimental evidence we may add the testimony of linguistic analysis (for example, GLEASON, 1955), including studies of language change (for example, LEHMAN, 1975), not to mention the very existence of alphabetic writing.

2. Structure of the Syllable

As the etymology of its name implies, the syllable is a compound, the vehicle of a natural acoustic contrast between consonant constriction and vowel opening, a contrast frequently claimed as a phonologic universal (e.g., POSTAL, 1968). The contrast is clearly reflected in perception, as evidenced by a long series of studies over the past 15 years. These studies, employing a variety of experimental paradigms—identification and discrimination of synthetic speech sounds, short-term memory, reaction time, dichotic listening, backward masking and others—converge on the conclusion that consonants and vowels perform distinct perceptual functions. Once again, we have reviewed this matter elsewhere (STUDDERT-KENNEDY, 1975b, 1976) and will not do so here. We simply remark that none of the varied evidence for the perceptual contrast between consonants and vowels could exist if the syllable were not analyzed in perception.

We may note, in passing, one further point. In all languages, the syllable is the unit of poetic meter. Except where syllable and morpheme normally coincide (for example, Japanese haiku), metrical rules are specified in terms of syllables and their expected "length" or degree of stress. Of particular interest in the present context is the fact that length is frequently specified not by lexical form but by neighboring phonetic segments. In both Latin and ancient Greek verse, for example, the length assigned to a word-final CVC syllable varies as a function of the initial phone of the following word. Thus a famous ode of Horace (Book III, Ode XXVI) begins: "Vixi puellis nuper idoneus . . ." Here the third word scans as a trochee because the following word begins with a vowel, but would have scanned as a spondee had the following word begun with a consonant. This simple rule of ancient verse obviously required that the singer (and presumably the listener who could detect a singer's error) be aware of the phonologic structure of the syllable.

3. Perceptual Function of Phonologic Categories

A crucial process in the perception of fluent speech must be short-term storage of early portions of an utterance pending final interpretation. What is the form of this store? Clearly it cannot be simply auditory, since a precategorical auditory store (CROWDER and MORTON, 1969; CROWDER, 1972) is sensitive to overwriting from immediately following items (CROWDER, 1971). Nor, given our sensitivity to phonetic structure (most obviously in listening to poetry) can the store be purely functional or semantic, with all phonetic detail stripped away.

In fact, we wish to argue that, to fulfill this linguistic function, a general perceptual process is invoked, namely, division into "stages." Among the likely functions of "perceptual stages"—whether defined in time or in neural locus—is to isolate one process from another and to store energy or information for later use. We may see this most clearly at the periphery. Every sensory system integrates energy: If the system were infinitely damped, threshold for activation would never be reached. Accumulation of energy over some finite period permits the mechanical response of the ear, for example, to develop.

On the other hand, the period of integration must be finite to prevent physical destruction of the system: Mechanical energy becomes bioelectricity. Analogous cycles of integration and transformation presumably recur, as energy or information progresses through the system. Activity in afferent fibers gives rise to more central neural activity and, ultimately (jumping levels of discourse), to a preperceptual "image" (MASSARO, 1972). The image, in turn, must have some finite duration, long enough to institute further processing, short enough to prevent "babble."

Returning with this metaphor to language, we note that speech is arrayed in time, and that both syntax and meaning demand some minimum quantity of information before linguistic structure can emerge. The perceptual function of phonologic categories may then be, on the one hand, to forestall auditory babble, on the other, to store information derived from the signal until such time as it can be granted a linguistic interpretation. In other words, the perceptual function of phonologic categories is that of a buffer between acoustic signal and meaningful message.

4. Recovery of the Morpheme

We come, finally, to the phonologic function without which, we believe, linguistic communication would not be possible—namely, to provide a code for lexical storage.

Notice first that if lexical items are coded according to overall acoustic structure, the form must be sufficiently stylized, stripped of acoustic detail, for the word to be accessed despite a wide variety of surface forms. For example, the duration of a single monosyllabic word, spoken by a single speaker at a conversational rate in a random list or in a sentence, may vary by a factor of 2 to 1 (GAITENBY, 1965; KOZHEVNIKOV and CHISTOVICH, 1965; LACKNER and LEVINE, 1975), and yet be fully intelligible in both contexts. Furthermore, the durational variants are not related by a simple scale factor: Most of the variation occurs over the syllable nucleus rather than over its edges (GAITENBY, 1965; LEHISTE, 1970; HUGGINS, 1972), so that an algorithm for generalizing two extreme acoustic variants could hardly succeed without at least some analysis of the overall acoustic shape.

If we add to durational variations, other within-speaker variations in fundamental frequency (which, coupled with duration, is the primary acoustic correlate of variations in linguistic stress) and in formant structure (due to cross-morphemic effects of coarticulation in running speech), not to mention acoustically similar across-speaker variations due to age, sex, and dialect, we are confronted with a formidable array of acoustic forms each of which—if unanalyzed acoustic structure is to give access to the lexicon—will have to be reduced to canonical acoustic form.

Now, it is true that the invariance problem is scarcely less serious if the message units to be recovered from the signal are phonologic entities, such as features or phones, than if they are morphemes or words, and even a cursory survey of the literature of speech perception will show that, as in the earlier sections of our chapter, this is a recurrent preoccupation. However, we should

note that the "audile" listener, consigned to lexical search with nothing but overall acoustic shape (and a few syntactic-semantic hints derived from prosody and context) to guide him, is deprived of at least one valuable aid, namely the systematic phonologic and phonotactic constraints of his language. He will not be permitted to resolve uncertainty by drawing on his knowledge that a particular portion of the acoustic pattern, or a particular sequence of acoustic segments, cannot occur in his language. Rather, every morphemic sound pattern will be distinct, and access to its semantic and syntactic attributes will be direct. In other words, the vast and subtle array of systematic phonology that linguistic studies have brought into view over the past 150 years will be no more than epiphenomenal froth, communicatively vacuous, at least for the listener, if not for the speaker.

Nonetheless, let us set the problem of invariance aside. Let us assume, for the moment, that it has been solved and that we are able to specify for every word or morpheme a unique canonical acoustic form apt for every context and every speaker. We shall then be confronted with the deeper problem of how the listener segments an utterance into its constituent morphemes or words.

The heart of the problem is simply that speakers freely coarticulate across word and morpheme boundaries. A consequence is that dividing the speech stream by use of an acoustic (or auditory) criterion will yield segments that bear a random relation (in size) to the words or morphemes. In that circumstance, the audile listener would have to store, not merely the 20,000–30,000 canonical auditory patterns that would represent the words in his vocabulary, but rather a number unimaginably greater than that (LIBERMAN and PISONI, in press). Even if he had a reliable acoustic criterion for dividing an utterance into syllables (see MERMELSTEIN, 1975), he would not be able to assign the syllables to their appropriate morphemes without analyzing them into their phonetic segments. For example, syllabification of the simple phrase, "He's a repeated offender," will yield eight CV syllables, four of which cross morpheme boundaries and two of which cross word boundaries. In other words, syllable boundaries in fluent speech are frequently random with respect to words or morphemes.

The problem is exacerbated for inflectional languages where changes in a single phoneme (initial, medial, or final) often suffice to signal changes in tense, mood, person, number, or case. Simple suffix changes, such as English plurals, might pose no problem for the audile listener, despite the lawful [s], [z] and [ɨz] alterations, and the absence of an acoustically marked morpheme boundary, for we need only suppose that the perceptual "morpheme detector" is automatically sprung as soon as a recognizable acoustic unit enters the system. We might even suppose that tense contrasts signaled by a change in medial vowel (as in "win"-"won") are learned as special cases. But we will find it a good deal more difficult to explain, for example, the formation of the Greek perfect tense by duplication of the initial consonant of the present, a fact presumably not lost on the listener. Indeed, as we multiply examples (and ad hoc solutions for the imaginary audile listener), we cannot but wonder why the various forms of a lexical item bear any relation to one another at all. Are we to suppose that these variations are lawful for the speaker, but merely adventitious for the listener?

Surely not. For, quite apart from the general lack of parsimony in positing totally independent input and output lexicons, we would be reduced to the absurdity of supposing that a listener consults a lexicon of auditory segments which bear no more than a random relation to the articulatory segments he deploys as a speaker. We are forced to conclude that only by extracting the phonetic segments—or, more properly, their underlying phonologic forms—can the listener discover most of what is said to him.

Acknowledgments

The preparation of this chapter was supported by a grant from the National Institute of Child Health and Human Development. Having freely borrowed the ideas of our colleagues at Haskins Laboratories, we are unable to say exactly which are owed to whom. We do especially acknowledge the contributions of QUENTIN SUMMERFIELD and PETER BAILEY, who were guests at the Laboratories while this chapter was being prepared, and we are indebted to Mark Liberman for valuable suggestions.

References

Ades, A.E.: Source assignment and feature extraction in speech. J. Exp. Psychol. [Human Percept.] (1977) (in press).

Ainsworth, W.A.: Automatic speech recognition. In: Mechanisms of Speech Recognition. Oxford: Pergamon Pr. 1976, pp. 104–119.

Bailey, P.J.: Perceptual adaptation for acoustical features in speech. Speech Perception. (Prog. Rep., Univ. Belfast) 2.2, 29–34 (1973).

Bailey, P.J., Dorman, M.F., Summerfield, A.Q.: Identification of sine-wave analogues of CV syllables in speech and non-speech modes. J. Acoust. Soc. Am. 61 S(A) (1977).

Bever, T.G.: The influence of speech performance on linguistic structure. In: Advances in Psycholinguistics. Flores D-Arcais, G.B., Levelt, W.J.M., (eds.). Amsterdam: North Holland 1970.

Boomer, D.S., Laver, J.D.M.: Slips of the tongue. Br. J. Disord. Commun. 3, 1–12 (1968).

Chomsky, N., Halle, M.: The Sound Pattern of English. New York: Harper & Row 1968.

Chomsky, N., Miller, G.A.: Introduction to the formal analysis of natural languages. In: Handbook of Mathematical Psychology. Luce, R.D., Bush, R.R., Galauker, E.E., (eds.). New York: Wiley 1963, Vol. I, pp. 269–321.

Cohen, A., Nooteboom, S.G., (eds.): Structure and Process in Speech Perception. New York: Springer-Verlag 1975.

Cole, R.A., Scott, B.: Toward a theory of speech perception. Psychol. Rev. 81, 348–374 (1974).

Cooper, F.S.: Spectrum analysis. J. Acoust. Soc. Am. 22, 761–762 (1950).

Cooper, F.S.: Some instrumental aids to research on speech. In: Proceedings of the Fourth Annual Round Table Meeting on Linguistics and Language Teaching. Washington: Georgetown Univ. 1953, pp. 46–53.

Cooper, F.S.: How is language conveyed by speech? In: Language by Ear and by Eye. Kavanagh, J.F., Mattingly, I.G., (eds.). Cambridge, Mass.: MIT Pr. 1972, pp. 25–45.

Cooper, F.S., Liberman, A.M., Borst, J.M.: The interconversion of audible and visible patterns as a basis for research in the perception of speech. Proc. Nat. Acad. Sci. USA 37, 318–325 (1951).

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M., Gerstman, L.J.: Some experiments on the perception of synthetic speech sounds. J. Acoust Soc. Am. 24, 597–606 (1952).

Cooper, W.E.: Selective adaptation to speech. In: Cognitive Theory. Restle, F., Shiffrin, R.M., Castellan, J.N., Lindman, H., Pisoni, D.B., (eds.). Hillsdale, N.J.: Erlbaum 1975, Vol. I, pp. 23–54.

Crowder, R.G.: Waiting for the stimulus suffix: decay, delay, rhythm and readvent in immediate memory. Q. J. Exp. Psychol. 23, 324–340 (1971).
Crowder, R.G.: Visual and auditory memory. In: Language by Ear and by Eye. Kavanagh, J.F., Mattingly, I.G., (eds.). Cambridge: MIT Pr. 1972.
Crowder, R.G., Morton, J.: Precategorical acoustic storage (PAS). Percept. Psychophys. 5, 365–373 (1969).
Darwin, C.J.: Ear differences in the recall of fricatives and vowels. Q. J. Exp. Psychol. 23, 46–62 (1971).
Darwin, C.J.: On the dynamic use of prosody in speech perception. In: Structure and Process in Speech Perception. Cohen, A., Nooteboom, J.G., (eds.). New York: Springer-Verlag 1975, pp. 178–194.
Darwin, C.J.: The perception of speech. In: Handbook of Perception. Carterette, E., Friedman, M., (eds.). New York: Academic Pr. 1976, Vol. VII, pp. 175–226.
Day, R.S.: Temporal order perception of reversible phoneme cluster. J. Acoust. Soc. Am. 48, 95(A) (1970).
Delattre, P.C., Liberman, A.M., Cooper, F.S.: Acoustic *loci* and transitional cues for consonants. J. Acoust. Soc. Am. 27, 769–773 (1955).
Denes, P.B.: Effect of duration on the perception of voicing. J. Acoust. Soc. Am. 27, 761–764 (1955).
Dorman, M., Cutting, J.E., Raphael, L.J.: Perception of temporal order in vowel sequences with and without formant transitions. J. Exp. Psychol. [Human Percept.] 104, 121–129 (1975a).
Dorman, M.F., Raphael, L.J., Liberman, A.M., Repp, B.: Masking-like phenomena in speech perception. J. Acoust. Soc. Am. 57 (Suppl. 1) S48(A) (1975b). [Full text in Haskins Lab. Status Rep. Speech Res. 42/43, 265–276.]
Dorman, M.F., Raphael, L.J., Liberman, A.M.: Further observations on the role of silence in the perception of stop consonants. Haskins Lab. Status Rep. Speech Res. 48, 199–207 (1976).
Eimas, P.D., Corbit, J.D.: Selective adaptation of linguistic feature detectors. Cog. Psychol. 13, 247–252 (1973).
Erickson, D.M., Fitch, H.L., Halwes, T.G., Liberman, A.M.: Trading relation in perception between silence and spectrum. J. Acoust. Soc. Am. 61, S46(A) (1977).
Fant, C.G.M.: Descriptive analysis of the acoustic aspects of speech. Logos 5, 3–17 (1962).
Fant, C.G.M.: Analysis and synthesis of speech processes. In: Manual of Phonetics. Malmberg, B., (eds.). Amsterdam: North-Holland 1968, pp. 173–277.
Foss, D.J., Swinney, D.A.: On the psychological reality of the phoneme: perception, identification and consciousness. J. Verbal Learn. Verbal Behav. 12, 246–257 (1973).
Fromkin, V.A.: The non-anomalous nature of anomalous utterances. Language 47, 27–52 (1971).
Fujimura, O.: Syllable as a unit of speech recognition. IEEE Trans. Acoust. Sp. Sig. Proc. 23, 82–87 (1975).
Fujimura, O.: A look into the effects of context: some articulatory and perceptual findings. In: Proc. of VIIIth Int. Congr. of Phonetic Sciences, Leeds, 1975 (1977) (in press).
Gaitenby, J.H.: The elastic word. Haskins Lab. Status Rep. Speech Res. 2, 3.4 (1965).
Ganong, W.F.: An experiment on "phonetic adaptation." Prog. Rep. (Res. Lab. Electr. MIT) 116, 206–210 (1975).
Ganong, W.F.: Amplitude contingent selective adaptation to speech. J. Acoust. Soc. Am. 59 S(A) (1976).
Gleason, H.A., Jr.: Workbook in Descriptive Linguistics. New York: Holt, Rinehart & Winston 1955.
Goldstein, L.M., Lackner, J.R.: Alterations of the phonetic coding of speech sounds during repetition. Cognition 2, 279–297 (1973).
Haggard, M.P., Ambler, S., Callow, M.: Pitch as a voicing cue. J. Acoust. Soc. Am. 47, 613–617 (1970).
Halwes, T., Jenkins, J.J.: Problem of serial order in behavior is not resolved by context-sensitive associative memory models. Psychol. Rev. 78, 122–129 (1971).
Harris, C.M.: A study of the building blocks of speech. J. Acoust. Soc. Am. 25, 962–969 (1953).
Harris, K.S.: Cues for the discrimination of American English fricatives in spoken syllables. Lang. Speech 1, 1–7 (1958).

Harris, K.S., Hoffman, H.S., Liberman, A.M., Delattre, P.C., Cooper, F.S.: Effect of third-formant transitions on the perception of the voiced stop consonants. J. Acoust Soc. Am. 30, 122–126 (1958).

Hoffman, H.S.: Study of some cues in the perception of the voiced stop consonants. J. Acoust. Soc. Am. 30, 1035–1041 (1958).

Huggins, A.W.F.: On the perception of temporal phenomena in speech. J. Acoust. Soc. Am. 51, 1279–1290 (1972).

Hyde, S.R.: Automatic speech recognition: a critical survey and discussion of the literature. In: Human Communication: A Unified View. David, E.E., Jr., Denes, P.B., (eds.). N.Y.: McGraw-Hill 1972, pp. 399–438.

Ingemann, F.: Speech synthesis by rule. J. Acoust. Soc. Am. 29, 1255(A) (1957).

Kelly, J.L., Gerstman, L.J.: An artificial talker driven from a phonetic input. J. Acoust. Soc. Am. 33, 835(A) (1961).

Kelly, J.L., Lochbaum, C.: Speech synthesis. Proc. of the Speech Commun. Sem., paper F7. Stockholm: Speech Transmission Lab. R. Inst. of Technology 1962.

Klatt, D.H.: Linguistic uses of segmental duration in English: acoustic and perceptual evidence. J. Acoust. Soc. Am. 59, 1208–1221 (1976).

Klima, E.S.: Sound and its absence in the linguistic symbol. In: The Role of Speech in Language. Kavanagh, J.F., Cutting, J.E. (eds.). Cambridge: MIT Press 1975, pp. 247–270.

Kozhevnikov, V.A., Chistovich, L.A.: Rech' Artikuliatsia i Vospriatie. Moscow-Leningrad 1965 [Transl. as Speech: Articulation and Perception. Washington: Clearinghouse for Fed. Sci. and Techn. Inf. JPRS, 30]

Lackner, J.R., Levine, B.K.: Speech production: evidence for syntactically and phonologically determined units. Percept. Psychophys. 17, 107–113 (1975).

Ladefoged, P.: Preliminaries to Linguistic Phonetics. Chicago: U. Chicago Pr. 1971.

Lehiste, I.: Suprasegmentals. Cambridge: MIT Pr. 1970.

Lehman, W.: Historical Linguistics, 2nd ed. New York: Holt, Rinehart & Winston 1975.

Liberman, A.M.: The grammars of speech and language. Cog. Psychol. 1, 301–323 (1970).

Liberman, A.M.: The specialization of the language hemisphere. In: The Neurosciences: Third Study Program. Schmitt, F.O., Worden, F.G., (eds.). Cambridge, Mass.: MIT Pr. 1974, pp. 43–56.

Liberman, A.M.: How abstract must a motor theory of speech perception be? In: Proc. of the VIIIth Int. Congr. of Phonetic Sci., Leeds, England, 17–23 August, 1975. [Also in: Haskins Lab. Status Rep. on Speech Res. 44, 1–15 (1975).]

Liberman, A.M.: Discussion paper. In: Origins and Evolution of Language and Speech. Harnad, S.R., Steklis, H.D., Lancaster, J., (eds.). New York: N.Y. Academy of Sci. 1976, pp. 718–724. [Annals of the New York Academy of Sciences 280, 718–724 (1976).]

Liberman, A.M., Pisoni, D.B.: Evidence for a special speech-perceiving subsystem in the human. In: The Recognition of Complex Acoustic Signals. Bullock, T.H., (ed.). Berlin: Dahlem Konferenzen 1977 (in press).

Liberman, A.M., Delattre, P.C., Cooper, F.S.: The role of selected stimulus-variables in the perception of the unvoiced stop consonants. Am. J. Psychol. 55, 497–516 (1952).

Liberman, A.M., Ingemann, F., Lisker, L., Delattre, P., Cooper, F.S.: Minimal rules for synthesizing speech. J. Acoust. Soc. Am. 31, 1490–1499 (1959).

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M.: The perception of the speech code. Psychol. Rev. 74, 431–461 (1967).

Liberman, A.M., Mattingly, I.G., Turvey, M.T.: Language codes and memory codes. In: Coding Processes in Human Memory. Melton, A.W., Martin, E., (eds.). Washington, D.C.: V.H. Winston 1972, pp. 307–334.

Lisker, L.: Some cues to the voiced-voiceless distinction among the intervocalic stops in English. Haskins Lab. 24th Q. Prog. Rep., Appendix 5, 1957a.

Lisker, L.: Closure duration and the intervocalic voiced-voiceless distinction in English. Language 33, 42–49 (1957b).

MacNeilage, P.F.: Motor control of serial ordering of speech. Psych. Rev. 77, 182–196 (1970).

Martin, J.G.: Rhythmic expectancy in continuous speech preception. In: Stimulus and Process in Speech Perception. New York: Springer-Verlag 1975, pp. 161–177.

Massaro, D.W.: Preperceptual images, processing time and perceptual units in auditory perception. Psych. Rev. 79, 124–145 (1972).
Massaro, D.W., Cohen, M.M.: The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction. J. Acoust. Soc. Am. 60, 704–717 (1976).
Mattingly, I.G.: Speech cues and sign stimuli. Am. Sci. 60, 327–337 (1972).
Mattingly, I.G.: Speech synthesis for phonetic and phonological models. In: Current Trends in Linguistics, Vol. XII: Linguistics and Adjacent Arts and Sciences. Sebeok, T.A., (ed.). The Hague: Mouton 1974, pp. 2451–2488.
Mattingly, I.G.: Syllable synthesis. J. Acoust. Soc. Am. 60, S75(A) (1976). [Also in: Haskins Lab. Status Rep. on Speech Res. 49, (1977).]
Mattingly, I.G., Liberman, A.M.: The speech code and the physiology of language. In: Information Processing in the Nervous System. Leibovic, K.N., (ed.). New York: Springer-Verlag 1969, pp. 97–117.
Mattingly, I.G., Liberman, A.M., Syndal, G.K., Halwes, T.: Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131–157 (1971).
McNeil, D., Lindig, L.: The perceptual reality of phonemes, syllables, words and sentences. J. Verbal Learn. Verbal Behav. 12, 419–430 (1973).
Mermelstein, P.: Automatic segmentation of speech into syllabic units. J. Acoust. Soc. Am. 58, 880–883 (1975).
Miller, G.A., Nicely, P.: An analysis of some perceptual confusions among some English consonants. J. Acoust. Soc. Am. 27, 338–352 (1955).
Miller, J.D.: Perception of speech sounds by animals: evidence for speech processing by mammalian auditory mechanisms. In: Recognition of Complex Acoustic Signals. Bullock, T.H., (ed.). Berlin: Dahlem Konferenzen 1977 (in press).
Mitchell, P.D.: A test of differentiation of phonemic feature contrasts. Unpubl. Ph.D. thesis, City Univ. New York, 1973.
Ochiai, K., Fujimura, O.: Vowel identification and phonetic contexts. Rep. Univ. Electro-Commun. (Chofu, Tokyo) 22-2, 103–111 (1971).
Ogden, C.K.: Opposition. Bloomington: Indiana U. Pr. 1967.
Peterson, G.E., Barney, H.L.: Control methods used in a study of vowels. J. Acoust. Soc. Am. 24, 175–184 (1952).
Peterson, G.E., Wang, W.S.-Y., Sivertsen, E.: Segmentation techniques of speech synthesis. J. Acoust. Soc. Am. 30, 739–742 (1958).
Pickett, J.M., Decker, L.R.: Time factors in perception of a double consonant. Lang. Speech 3, 11–17 (1960).
Pisoni, D.B.: Speech perception. In: Handbook of Learning and Cognitive Processes. Estes, W., (ed.). Hillsdale, N.J.: Lawrence Erlbaum 1977, Vol. XVI (in press).
Port, R.F.: The influence of speaking tempo on the duration of stressed vowel and medial stop in English trochee words. Unpubl. Ph.D. thesis, Univ. of Connecticut, 1967.
Postal, M.: Aspects of Phonological Theory. New York: Harper & Row 1968.
Rand, T.C.: Vocal tract size normalization in the perception of stop consonants. Haskins Lab. Status Rep. on Speech Res. 25/26, 141–146 (1971).
Rand, T.C.: Dichotic release from masking for speech. J. Acoust. Soc. Am. 55, 678–680 (1974).
Raphael, L.J.: Preceding vowel duration as a cue to the perception of voicing characteristic of word-final consonants in American English. J. Acoust. Soc. Am. 51, 1296–1303 (1972).
Raphael, L.J., Dorman, M.F., Liberman, A.M.: Perception of vowel and syllable duration in VC and CVC syllables. J. Acoust. Soc. Am. 58, S57(A) (1975).
Raphael, L.J., Dorman, M.F., Liberman, A.M.: Some ecological constraints on the perceptions of stops and affricates. J. Acoust. Soc. Am. 59, Suppl. 1: S25(A) (1976).
Rubin, P., Turvey, M.T., Gelder, P. van: Initial phonemes are detected faster in spoken words than in spoken nonwords. Percept. Psychophys. 19, 394–398 (1976).
Savin, H.B., Bever, T.G.: The non-perceptual reality of the phoneme. J. Verbal Learn. Verbal Behav. 9, 295–302 (1970).
Schatz, C.D.: The role of context in the perception of stops. Language 30, 47–56 (1954).
Shankweiler, D.P., Studdert-Kennedy, M.: Identification of consonants and vowels presented to left and right ears. Q. J. Exp. Psychol. 19, 59–63 (1967).

Shattuck, S.R., Klatt, D.H.: The perceptual similarity of mirror-image acoustic patterns in speech. Percept. Psychophys. **20**, 470–474 (1976).

Stevens, K.N.: The potential role of property detectors in the perception of consonants. In: Auditory Analysis and Perception of Speech. Fant, C.G.M., Tatham, M.A.A., (eds.). New York: Academic Pr. 1975, pp. 303–330.

Stevens, K.N., House, A.S.: The perception of speech. In: Foundations of Modern Auditory Theory. Tobias, J., (ed.). New York: Academic Pr. 1972, pp. 3–62.

Strange, W., Verbrugge, R., Shankweiler, D.P., Edman, T.R.: Consonant environment specifies vowel identity. J. Acoust. Soc. Am. **60**, 198–212 (1976).

Studdert-Kennedy, M.: The perception of speech. In: Current Trends in Linguistics. Sebeok, T.A., (ed.). The Hague: Mouton 1974, Vol. XII (4), pp. 2349–2385.

Studdert-Kennedy, M.: From acoustic signal to phonetic message. J. Commun. Discord. **8**, 181–188 (1975a).

Studdert-Kennedy, M.: From continuous signal to discrete message: syllable to phoneme. In: The Role of Speech in Language. Kavanagh, J.F., Cutting, J.E., (eds.). Cambridge: MIT Pr. 1975b, pp. 113–125.

Studdert-Kennedy, M.: Speech perception. In: Contemporary Issues in Experimental Phonetics. Lass, N.J., (ed.). New York: Academic Pr. 1976, pp. 243–293.

Studdert-Kennedy, M.: Universals in phonetic structure and their role in linguistic communication. In: The Recognition of Complex Acoustic Signals. Bullock, T.H., (ed.). Berlin: Dahlem Konferenzen 1977 (in press).

Studdert-Kennedy, M., Shankweiler, D.P.: Hemispheric specialization for speech perception. J. Acoust. Soc. Am. **48**, 579–594 (1970).

Summerfield, A.Q., Bailey, P.J.: On the dissociation of spectral and temporal cues for stop consonant manner. J. Acoust. Soc. Am. **61**, S(A) (1977).

Summerfield, A.Q., Haggard, M.P.: On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. Haskins Lab. Status Rep. on Speech Res. **49** (1977) (in press).

Svennson, S.-G.: Prosody and grammar in speech perception. MILUS (Inst. of Linguistics, Univ. of Stockholm) **2** (1974).

Warren, R.M.: Auditory sequence and classification. In: Contemporary Issues in Experimental Phonetics. Lass, N.J., (ed.). New York: Academic Pr. 1976a, pp. 389–417.

Warren, R.M.: Auditory perception and speech evolution. In: Origins and Evolution of Language and Speech. Harnad, S.R., Steklis, H.D., Lancaster, J., (eds.). New York: Academy of Sci. 708–717 (1976b) [Ann. N.Y. Acad. Sci. **280**, 708–717].

Warren, R.M., Obusek, C.J., Farmer, R.M., Warren, R.P.: Auditory sequence: confusion of patterns other than speech and music. Science **164**, 568–587 (1969).

Whitfield, I.C.: 'Edges' in auditory information processing. In: Proc. of XXIIIrd Int. Congr. of Physiological Sci. Tokyo, September 1965, pp. 245–247.

Whitfield, I.C., Evans, E.F.: Responses of auditory cortical neurons to stimuli of changing frequency. J. Neurophysiol. **28**, 655–672 (1965).

Wickelgren, W.A.: Context-sensitive coding, associative memory and serial order in (speech) behavior. Psychol. Rev. **76**, 1–15 (1969).