Psychological Review
1967, Vol. 74, No. 6, 431–461

PERCEPTION OF THE SPEECH CODE[1]

A. M. LIBERMAN,[2] F. S. COOPER, D. P. SHANKWEILER, AND M. STUDDERT-KENNEDY[3]

Haskins Laboratories, New York, New York

Man could not perceive speech well if each phoneme were cued by a unit sound. In fact, many phonemes are encoded so that a single acoustic cue carries information in parallel about successive phonemic segments. This reduces the rate at which discrete sounds must be perceived, but at the price of a complex relation between cue and phoneme: cues vary greatly with context, and there are, in these cases, no commutable acoustic segments of phonemic size. Phoneme perception therefore requires a special decoder. A possible model supposes that the encoding occurs below the level of the (invariant) neuromotor commands to the articulatory muscles. The decoder may then identify phonemes by referring the incoming speech sounds to those commands.

Our aim is to identify some of the conditions that underlie the perception of speech. We will not consider the whole process, but only the part that lies between the acoustic stream and a level of perception corresponding roughly to the phoneme.[4] Even this, as we will try to show, presents an interesting challenge to the psychologist.

The point we want most to make is that the sounds of speech are a special and especially efficient code on the phonemic structure of language, not a cipher or alphabet. We use the term code,[5] in contrast to cipher or alpha-

[1] The research reported here was aided at its beginning, and for a considerable period afterward, by the Carnegie Corporation of New York. Funds have also come from the National Science Foundation and the Department of Defense. The work is currently supported by the National Institute of Child Health and Human Development, the National Institute of Dental Research, and the Office of Naval Research.

[2] Also at the University of Connecticut.

[3] Also at the University of Pennsylvania.

[4] For our purposes the phoneme is the shortest segment that makes a significant difference between utterances. It lies in the lowest layer of language, has no meaning in itself, and is, within limits, commutable. There are, for example, three such segments in the word "bad"—/b/, /æ/, and /d/—established by the contrasts with "dad," "bed," and "bat." As commonly defined, a phoneme is an abstract and general type of segment, represented in any specific utterance by concrete and specific tokens, called phones, that may vary noticeably as a function of context. The distinguishable variants so produced are referred to as allophones.

We do not mean to imply that every phoneme is necessarily perceived as one listens to speech. Linguistic constraints of various kinds make it possible to correct or insert segments that are heard wrongly or not at all. Phonemes can be perceived, however, and some number of them must be perceived if the listener is to discover which constraints apply.

[5] We borrow the terms cipher and code from cryptography. A cipher substitutes a symbol for each of the units (letters,

Psychological Review 74, 431–461 (1967);

bet, to indicate that speech sounds represent a very considerable restructuring of the phonemic "message." The acoustic cues for successive phonemes are intermixed in the sound stream to such an extent that definable segments of sound do not correspond to segments at the phoneme level. Moreover, the same phoneme is most commonly represented in different phonemic environments by sounds that are vastly different. There is, in short, a marked lack of correspondence between sound and perceived phoneme. This is a central fact of speech perception. It is, we think, the result of a complex encoding that makes the sounds of speech especially efficient as vehicles for the transmission of phonemic information. But it also poses an important question: by what mechanism does the listener decode the sounds and recover the phonemes?

In this paper we will (*a*) ask whether speech could be well perceived if it were an alphabet or acoustic cipher, (*b*) say how we know that speech is, in fact, a complex code, (*c*) describe some properties of the perceptual mode that results when speech is decoded, (*d*) consider how the encoding and decoding might occur, and (*e*) show that the speech code is so well matched to man as to provide, despite its complexity, a uniquely effective basis for communication.

usually) of the original message. In a code, on the other hand, the units of the original and encoded forms do not correspond in structure or number, the encoded message typically containing fewer units. Since these distinctions are relevant to our purpose here, we have adopted the terms cipher and code as a convenient way to refer to them. We should add, however, that the arbitrary relation between the original and encoded forms of a message, so usual in cryptography, is not a feature of the encoding of phonemes into syllables.

COULD SPEECH BE ALPHABETIC?

There are reasons for supposing that phonemes could not be efficiently communicated by a sound alphabet—that is, by sounds that stand in one-to-one correspondence with the phonemes. Such reasons provide only indirect support for the conclusion that speech is a code rather than an alphabet. They are important, however, because they indicate that the encoded nature of speech may be a condition of its effectiveness in communication. More specifically, they tell us which aspects of the code are likely to be relevant to that effectiveness.

Phoneme Communication and the Properties of the Ear

Of the difficulties we might expect to have with a sound alphabet, the most obvious concerns rate. Speech can be followed, though with difficulty, at rates as high as 400 words per minute (Orr, Friedman, & Williams, 1965). If we assume an average of four to five phonemes for each English word, this rate yields about 30 phonemes per second. But we know from auditory psychophysics (Miller & Taylor, 1948) that 30 sounds per second would overreach the temporal resolving power of the ear: discrete acoustic events at that rate would merge into an unanalyzable buzz; a listener might be able to tell from the pitch of the buzz how fast the speaker was talking, but he could hardly perceive what had been said. Even 15 phonemes per second, which is not unusual in conversation, would seem more than the ear could cope with if phonemes were a string of discrete acoustic events.

There is at least one other requirement of a sound alphabet that would be hard to satisfy: a sufficient number of identifiable sounds. The number of phonemes, and hence the number of

acoustic shapes required, is in the dozens. In English there are about 40. We should, of course, be able to find 40 identifiable sounds if we could pattern the stimuli in time, as in the case of melodies. But if we are to communicate as rapidly as we do, the phoneme segments could last no longer than about 50 milliseconds on the average. Though it is not clear from research on auditory perception how many stimuli of such brief duration can be accurately identified, the available data suggest that the number is considerably less than 40 (Miller, 1956a, 1956b; Nye, 1962; Pollack, 1952; Pollack & Ficks, 1954). We will be interested, therefore, to see whether any features of the encoding and decoding mechanisms are calculated to enhance the identifiability of the signals.

Results of Attempts to Communicate Phonemes by an Acoustic Alphabet

That these difficulties of rate and sound identification are real, we may see from the fact that it has not been possible to develop an efficient sound alphabet despite repeated and thoroughgoing attempts to do so. Thus, international Morse code (a cipher as we use the term here) works poorly in comparison with speech, even after years of practice. But Morse is surely not the best example; the signals are one-dimensional and therefore not ideal from a psychological standpoint. More interesting, if less well known, are the many sound alphabets that have been tested in the attempt to develop reading machines for the blind (Coffey, 1963; Cooper, 1950a; Freiberger & Murphy, 1961; Nye, 1964, 1965; Studdert-Kennedy & Cooper, 1966; Studdert-Kennedy & Liberman, 1963). These devices convert print into sound, the conversion being made typically by encipherment of the optical alphabet into one that is acoustic. The worth of these devices has been limited, not by the difficulty of converting print to sound, but by the perceptual limitations of their human users. In the 50-year history of this endeavor, a wide variety of sounds has been tried, including many that are multidimensional and otherwise appropriately designed, it would seem, to carry information efficiently. Subjects have practiced with these sounds for long periods of time, yet there has nowhere emerged any evidence that performance with these acoustic alphabets can be made to exceed performance with Morse, and that is little more than a tenth of what can be achieved with speech.

Reading and Listening

We hardly need say that language can be written and read by means of an alphabet, but we would emphasize how different are the problems of communicating phonemes by eye and by ear, and how different their relevance to a psychology of language. In contrast to the ear, the eye should have no great difficulty in rapidly perceiving ordered strings of signals. Given the eye's ability to perceive in space, we should suppose that alphabetic segments set side by side could be perceived in clusters. Nor is there reason to expect that it might be difficult to find identifiable optical signals in sufficient number. Many shapes are available, and a number of different alphabets are, indeed, in use. Thus, written language has no apparent need for the special code that will be seen to characterize language in its acoustic form. In writing and reading it is possible to communicate phonemes by means of a cipher or alphabet; indeed, there appears to be no better way.

Spoken and written language differ, then, in that the former must be a complex code while the latter can be a simple cipher. Yet perception of

speech is universal, though reading is not. In the history of the race, as in the development of the individual, speaking and listening come first; writing and reading come later, if at all. Moreover, the most efficient way of writing and reading—namely, by an alphabet—is nevertheless so unnatural that it has apparently been invented only once in all history (Gelb, 1963). Perceiving the complex speech code is thus basic to language, and to man, in a way that reading an alphabet is not. Being concerned about language, we are therefore the more interested in the speech code. Why are speech sounds, alone among acoustic signals and in spite of the limitations of the ear, perceived so well?

ACOUSTIC CUES: A RESTRUCTURING OF PHONEMES AT THE LEVEL OF SOUND

To know in what sense speech sounds are a code on the phonemes, we must first discover which aspects of the complex acoustic signal underlie the perception of particular phonemes. For the most part, the relevant data are at hand. We can now identify acoustic features that are sufficient and important cues for the perception of almost all the segmental phonemes.[6] Much remains to be learned, but we know enough to see that the phonemic message is restructured in the sound stream and, from that knowledge, to make certain inferences about perception.

[6] For the discussion that follows we shall rely most heavily on the results of experiments with synthetic speech carried out at Haskins Laboratories. The reader will understand that a review of the relevant literature would refer to many other studies, including, in particular, those that rest on analysis of real speech. For recent reviews and discussions, see Stevens and House, in press; Kozhevnikov and Chistovich, 1965, Chapter 6.

An Example of Restructuring

To illustrate the nature of the code, we will describe an important acoustic cue for the perception of the voiced stop /d/. This example is important in its own right and also broadly representative. The phoneme /d/, or something very much like it, occurs in all the languages of the world.[7] In English, and perhaps in other languages, too, it carries a heavy load of information, probably more than any other single phoneme (Denes, 1963); and it is among the first of the phonemelike segments to appear in the vocalizations of the child (Whetnall & Fry, 1964, p. 84). The acoustic cue we have chosen to examine—the second-formant transition[8]—is a major cue for all the consonants except, perhaps, the fricatives /s/ and /š/, and is probably the single most important carrier of linguistic information in the speech signal (Delattre, 1958, 1962; Delattre, Liberman, & Cooper, 1955, 1964; Harris, 1958; Liberman, 1957;

[7] Some form of a voiceless unaspirated stop having a place of production in the alveolar-dental region is universal (Hockett, 1955; Joseph Greenberg, personal communication, November, 1966). The particular example, /d/, used here is a member of that class in almost all important respects. Even in regard to voicing, it is a better fit than might at first appear, since it shares with the voiceless, unaspirated stop of, say, French or Hungarian, the same position on the dimension of voice-onset-time, which is a most important variable for phonemic differences in voicing. (Lisker & Abramson, 1964a, 1964b).

[8] A formant is a concentration of acoustic energy within a restricted frequency region. Three or four formants are usually seen in spectrograms of speech. In the synthetic, hand-painted spectrograms of Figures 1 and 2, only the lowest two are represented. Formants are referred to by number, the first being the lowest in frequency, the second, the next higher, and so on. A formant transition is a relatively rapid change in the position of the formant on the frequency scale.

Liberman, Delattre, Cooper, & Gerstman, 1954; Liberman, Ingemann, Lisker, Delattre, & Cooper, 1959; Lisker, 1957; O'Connor, Gerstman, Liberman, Delattre, & Cooper, 1957).

Context-conditioned variations in the acoustic cue. Figure 1 displays two highly simplified spectrographic patterns that will, when converted into sound, be heard as the syllables /di/ and /du/ (Liberman, Delattre, Cooper, & Gerstman, 1954). They exemplify the results of a search for the acoustic cues in which hand-drawn or "synthetic" spectrograms were used as a basis for experimenting with the complex acoustic signal (Cooper, 1950, 1953; Cooper, Delattre, Liberman, Borst, & Gerstman, 1952; Cooper, Liberman, & Borst, 1951). The steady-state formants, comprising approximately the right-hand two-thirds of each pattern, are sufficient to produce the vowels /i/ and /u/ (Delattre, Liberman, & Cooper, 1951; Delattre, Liberman, Cooper, & Gerstman, 1952). At the left of each pattern are the relatively rapid changes in frequency of the formants—the formant transitions —that are, as we indicated, important acoustic cues for the perception of the consonants. The transition of the first, or lower, formant, rising from a very low frequency to the level appropriate for the vowel, is a cue for the class of voiced stops /b,d,g/ (Delattre, Liberman, & Cooper, 1955; Liberman, Delattre, & Cooper, 1958). It would be exactly the same for /bi, bu/ and /gi, gu/ as for /di, du/. Most generally, this transition is a cue for the perception of manner and voicing. The acoustic feature in which we are here interested is the transition of the second formant, which is, in the patterns of Figure 1, a cue for distinguishing among the voiced stops, /b,d,g/; that is to say, the second-formant transition for /gi/ or /bi/, as well as /gu/ or /bu/, would be different from those for /di/ and /du/ (Liberman, Delattre, Cooper, & Gerstman, 1954). In general, transitions of the second formant carry important information about the place of production of most consonants (Delattre, 1958; Liberman, 1957; Liberman, Delattre, Cooper, & Gerstman, 1954).

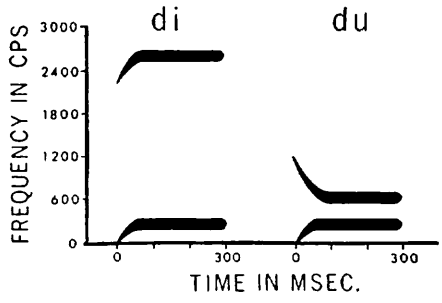

FIG. 1. Spectrographic patterns sufficient for the synthesis of /d/ before /i/ and /u/.

It is, then, the second-formant transitions that are, in the patterns of Figure 1, the acoustic cues for the perception of the /d/ segment of the syllables /di/ and /du/. We would first note that /d/ is the same perceptually in the two cases, and then see how different are the acoustic cues. In the case of /di/ the transition rises from approximately 2200 cps to 2600 cps; in /du/ it falls from about 1200 to 700 cps. In other words, what is perceived as the same phoneme is cued, in different contexts, by features that are vastly different in acoustic terms. How different these acoustic features are in nonspeech perception can be determined by removing them from the patterns of Figure 1 and sounding them in isolation. When we do that, the transition isolated from the /di/ pattern sounds like a rapidly rising whistle or glissando on high pitches, the one from /du/ like a rapidly falling whistle on low pitches.[9]

[9] This is true of the patterns shown in Figure 1 as converted to sound by the Pattern Playback. When the formants correspond

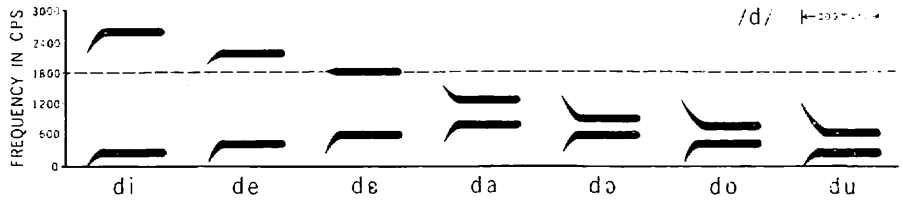

FIG. 2. Spectrographic patterns sufficient for the synthesis of /d/ before vowels. (Dashed line at 1800 cps shows the "locus" for /d/.)

These signals could hardly sound more different from each other. Furthermore, neither of them sounds like /d/ nor like speech of any sort.

The disappearance of phoneme boundaries: Parallel transmission. We turn now to another, related aspect of the code: the speech signal typically does not contain segments corresponding to the discrete and commutable phonemes. There is no way to cut the patterns of Figure 1 so as to recover /d/ segments that can be substituted one for the other. Nor can we make the commutation simply by introducing a physical continuity between the cut ends, as we might, for example, if we were segmenting and recombining the alphabetic elements of cursive writing.

Indeed, if we could somehow separate commutability from segmentability, we should have to say that there is no /d/ segment at all, whether commutable or not. We cannot cut either the /di/ or the /du/ pattern in such a way as to obtain some piece that will produce /d/ alone. If we cut progressively into the syllable from the right-hand end, we hear /d/ plus a vowel, or a nonspeech sound; at no point will we hear only /d/. This is so because the formant transition is, at every instant, providing information about two phonemes, the consonant and the vowel—that is, the phonemes are being transmitted in parallel.

more closely in their various constant features to those produced by the human vocal apparatus, the musical qualities described above may be harder to hear or may disappear altogether. So long as the second-formant transitions of /di/ and /du/ are not heard as speech, however, they do not sound alike.

The Locus: An Acoustic Invariant?

The patterns of Figure 2 produce /d/ in initial position with each of a variety of vowels, thus completing a series of which the patterns shown in Figure 1 are the extremes. If one extrapolates the various second-formant transitions backward in time, he sees that they seem to have diverged from a single frequency. To find that frequency more exactly, and to determine whether it might in some sense be said to characterize /d/, we can, as in Figure 3, pair each of a number of straight second formants with first formants that contain a rising transition sufficient to signal a voiced stop of some kind. On listening to such patterns, one hears an initial /d/ most strongly when the straight second formant is at 1800 cps. This has been called the /d/ "locus" (Delattre, Liberman, & Cooper, 1955). There are, correspondingly, second-formant loci for other consonants, the frequency position of the locus being correlated with the place of production. In general, the locus moves somewhat as a function of the associated vowel, but, except for a discontinuity in the case of /k, g, ŋ), the locus is more nearly invariant than the formant transition (Delattre, Liberman, & Cooper, 1955; Ohman, 1966; Stevens & House, 1956).

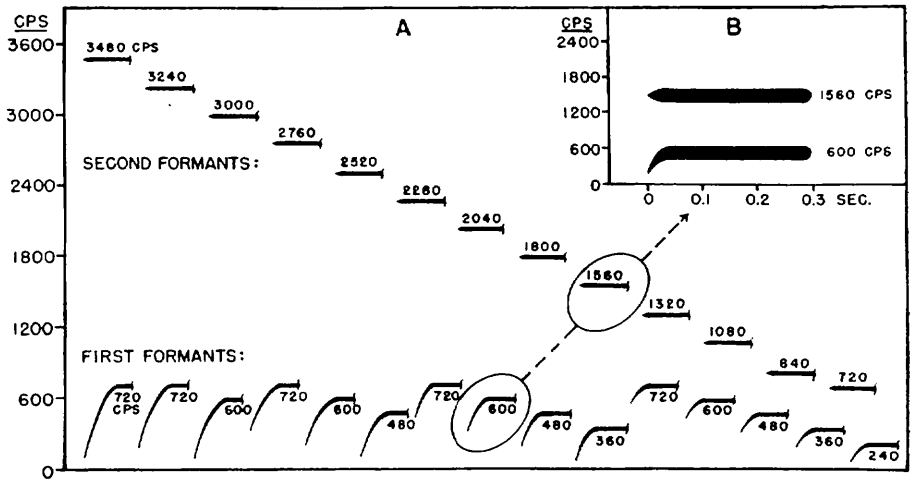

FIG. 3. Schematic display of the stimuli used in finding the second-formant loci of /b/, /d/, and /g/. (A—Frequency positions of the straight second formants and the various first formants with which each was paired. B—A typical test pattern, made up of the first and second formants circled in A. The best pattern for /d/ was the one with the straight second formant at 1800 cps. Figure taken from Delattre, Liberman, and Cooper, 1955.)

Is there, then, an invariant acoustic cue for /d/? Consider the various second formants that all begin at the 1800-cycle locus and proceed from there to a number of different vowel positions, such as are shown in Figure 4. We should note first that these transitions are not superimposable, so that if they are to be regarded as invariant acoustic cues, it could only be in the very special and limited sense that they start at the same point on the frequency

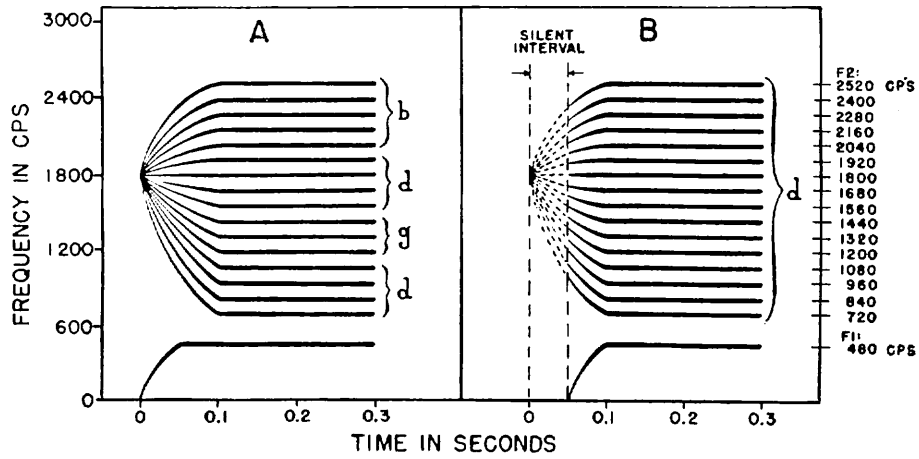

FIG. 4. A—Second-formant transitions that start at the /d/ locus and B—comparable transitions that merely "point" at it, as indicated by the dotted lines. (Those of A produce syllables beginning with /b/, /d/, or /g/, depending on the frequency-level of the formant; those of B produce only syllables beginning with /d/. Figure taken from Delattre, Liberman, and Cooper, 1955.)

scale. But even this very limited invariance is not to be had. If we convert the patterns of Figure 4 to sound, having paired each of the second formants in turn with the single first formant shown at the bottom of the figure, we do not hear /d/ in every case. Taking the second formants in order, from the top down, we hear first /b/, then /d/, then /g/, then once again /d/, as shown in the figure. In order to hear /d/ in every case, we must erase the first part of the transition, as shown in Figure 4B, so that it "points" at the locus but does not actually begin there. Thus, the 1800-cps locus for /d/ is not a part of the acoustic signal, nor can it be made part of that signal without grossly changing the perception (Delattre, Liberman, & Cooper, 1955).

Though the locus can be defined in acoustic terms—that is, as a particular frequency—the concept is more articulatory than acoustic, as can be seen in the rationalization of the locus by Stevens and House (1956). What is common to /d/ before all the vowels is that the articulatory tract is closed at very much the same point. According to the calculations of Stevens and House, the resonant frequency of the cavity at the instant of closure is approximately 1800 cps, but since no sound emerges until some degree of opening has occurred, the locus frequency is not radiated as part of the acoustic signal. At all events it seems clear that, though the locus is more nearly invariant with the phoneme than is the transition itself, the invariance is a derived one, related more to articulation than to sound. As we will see later, however, the locus is only a step toward the invariance with phoneme perception that we must seek; better approximations to that invariance are probably to be had by going farther back in the chain of articulatory events, beyond the cavity shapes that underlie the locus, to the commands that produce the shapes.

How General Is the Restructuring?

Having dealt with the restructuring of only one phoneme and one acoustic cue, we should say now that, with several interesting exceptions yet to be noted, it is generally true of the segmental phonemes that they are drastically restructured at the level of sound.[10] We will briefly summarize what is known in this regard about the various types of cues and linguistic classes or dimensions.

Transitions of the second formant. For the second-formant transition—the cue with which we have been concerned and the one that is, perhaps, the most important for the perception of consonants according to place of production—the kind of invariance lack we found with /d/ characterizes all the voiced stops, voiceless stops, and nasal consonants (Liberman, Delattre, Cooper, & Gerstman, 1954; Malecot, 1956). Indeed, the invariance problem is, if anything, further complicated in these other phonemes. In the case of /g, k, ŋ), for example, there is a sudden and considerable shift in the locus as between the unrounded and rounded vowels, creating a severe lack of correspondence between acoustic sig-

[10] Somewhat similar complications arise in the suprasegmental domain. For data concerning the relation between the acoustic signal and the perception of intonation see Hadding-Koch and Studdert-Kennedy, 1964a, 1964b.

Lieberman (1967) has measured some of the relevant physiological variables and, on the basis of his findings, has devised a hypothesis to account for the perception of intonation, in particular the observations of Hadding-Koch and Studdert-Kennedy. Lieberman's account of perception in the suprasegmental domain fits well with our own views about the decoding of segmental phonemes, discussed later in this paper.

nal and linguistic perception that we mentioned earlier in this paper and that we have dealt with in some detail elsewhere (Liberman, 1957).

With the liquids and semivowels /r, l, w, j/ the second-formant transition originates at the locus—as it cannot in the case of the stop and nasal consonants—so the lack of correspondence between acoustic signal and phoneme is less striking, but even with these phonemes the transition cues are not superimposable for occurrences of the same consonant in different contexts (Lisker, 1957; O'Connor, Gerstman, Liberman, Delattre, & Cooper, 1957).

Transitions of the third formant. What of third-formant transitions, which also contribute to consonant perception in terms of place of production? Though we know less about the third-formant transitions than about those of the second, such evidence as we have does not suggest that an invariant acoustic cue can be found here (Harris, Hoffman, Liberman, Delattre, & Cooper, 1958; Lisker, 1957b; O'Connor, Gerstman, Liberman, Delattre, & Cooper, 1957). In fact, the invariance problem is further complicated in that the third-formant transition seems to be more or less important (relative to the second-formant transition, for example) depending on the phonemic context. In the case of our /di, du/ example (Figure 1), we find that an appropriate transition of the third formant contributes considerably to the perception of /d/ in /di/ but not at all to /d/ in /du/. To produce equal—that is, equally convincing—/d/'s before both /i/ and /u/, we must use acoustic cues that are, if anything, even more different than the two second-formant transitions we described earlier in this section.

Constriction noises. The noises produced at the point of constriction—as in the fricatives and stops—are another set of cues for consonant perception according to place of production, the relevant physical variable being the frequency position of the band-limited noise. When these noises have considerable duration—as in the fricatives—the cue changes but little with context. Since the cue provided by these noises is of overriding importance in the perception of /s/ and /š/, we should say that these consonants show little or no restructuring in the sound stream (Harris, 1958; Hughes & Halle, 1956). (This may not be true when the speech is rapid.) They therefore constitute an exception to the usual strong dependence of the acoustic signal on its context—that is, to the effects of a syllabic coding operation.

On the other hand, the brief noise cues—the bursts, so called—of the stop consonants display as much restructuring as do the transitions of the second formant (Liberman, Delattre, & Cooper, 1952). Bursts of noise that produce the best /k/ or /g/ vary over a considerable frequency range depending on the following vowel. The range is so great that it extends over the domain of the /p, b/ burst, creating the curiosity of a single burst of noise at 1440 cps that is heard as /p/ before /i/ but as /k/ before /a/.[11] We should also note that the relative importance of the bursts, as of the transitions, varies greatly for the same stop in different contexts. For example, /g/ or /k/, which are powerfully cued by a second-formant transition before the vowel /æ/, will likely require a burst at an appropriate frequency if they are to be well perceived in front of /o/.[12]

[11] This result was obtained in the experiment (Liberman, Delattre, & Cooper, 1952) with synthetic speech referenced above and verified for real speech in a tape-cutting experiment by Carol Schatz (1954).

[12] This can be seen in the results of experiments on the bursts and transitions as

Thus, the same consonant is primarily cued in two different contexts by signals as physically different as a formant transition and a burst of noise.

Manner, voicing, and position. In describing the complexities of the relation between acoustic cue and perceived phoneme, we have so far dealt only with cues for place of production and only with consonants in initial position in the syllable. A comparable lack of regularity is also found in the distinctions of manner and voicing and in the cues for consonants in different positions (Abramson & Lisker, 1965; Delattre, 1958; Delattre, Liberman, & Cooper, 1955; Liberman, 1957; Liberman, Delattre, & Cooper, 1958; Liberman, Delattre, Cooper, & Gerstman, 1954; Liberman, Delattre, Gerstman, & Cooper, 1956; Lisker, 1957a, 1957b; Lisker & Abramson, 1964b; Ohman, 1966). We will not consider these cues in detail since the problems they present are similar to those already encountered. Indeed, the cues we have discussed as examples of encoding are merely a subset of those that show extensive restructuring as a function of context: it is the usual case that the acoustic cues for a consonant are different when the consonant is paired with different vowels, when it is in different positions (initial, medial, or final) with respect to the same vowels, and for all types of cues (manner or voicing, as well as place). Thus, for example, the cues for manner, place, and voicing of /b/ in /ba/ are acoustically different from those of /b/ in /ab/, the transitional cues for place being almost mirror images; further, the preceding set of cues differs from corresponding sets for /b/ with each of the other vowels.

The vowels. We should remark, finally, on the acoustic cues for the vowels. For the steady-state vowels of Figures 1 and 2, perception depends primarily on the frequency position of the formants (Delattre, Liberman, & Cooper, 1951; Delattre, Liberman, Cooper, & Gerstman, 1952). There is, for these vowels, no restructuring of the kind found to be so common among the consonant cues and, accordingly, no problem of invariance between acoustic signal and perception.[13]

However, vowels are rarely steady state in normal speech; most commonly these phonemes are articulated between consonants and at rather rapid rates. Under these conditions vowels also show substantial restructuring—that is, the acoustic signal at no point corresponds to the vowel alone, but rather shows, at any instant, the merged influences of the preceding or following consonant (Lindblom & Studdert-Kennedy, in press; Lisker, 1958; Shearme & Holmes, 1962; Stevens & House, 1963).

In slow articulation, then, the acoustic cues for the vowels—and, as we saw earlier, the noise cues for fricatives —tend to be invariant. In this respect they differ from the cues for the other phonemes, which vary as a function of context at all rates of speaking. However, articulation slow enough to permit the vowels and fricatives to avoid being encoded is probably artificial and rare.

cues for the stops (Liberman, Delattre, & Cooper, 1952; Liberman, Delattre, Cooper, & Gerstman, 1954). It is confirmed whenever one attempts, by using all possible cues, to synthesize the best stops.

[13] The absolute formant frequencies of the same vowel are different for men, women, and children, and for different individuals, partly as a consequence of differences in the sizes of vocal tracts. This creates an invariance problem very different from the kind we have been discussing and more similar, perhaps, to the problems encountered in the perception of nonspeech sounds, such as the constant ratios of musical intervals.

Phoneme segmentation. We return now to the related problem of segmentation, briefly discussed above for the /di, du/ example. We saw there that the acoustic signal is not segmented into phonemes. If one examines the acoustic cues more generally, he finds that successive phonemes are most commonly merged in the sound stream. This is, as we will see, a correlate of the parallel processing that characterizes the speech code and is an essential condition of its efficiency. One consequence is that the acoustic cues cannot be divided on the time axis into segments of phonemic size.[14]

The same general conclusion may be reached by more direct procedures. Working with recordings of real speech, Harris (1953) tried to arrive at "building blocks" by cutting tape recordings into segments of phoneme length and then recombining the segments to form new words. "Experiments indicated that speech based upon one building block for each vowel and consonant not only sounds unnatural but is mostly unintelligible . . . [p. 962]." In a somewhat similar attempt to produce intelligible speech by the recombination of parts taken from previously recorded utterances, Peterson, Wang, and Sivertsen (1958) concluded that the smallest segments one can use are of roughly half-syllable length. Thus, it has not been possible, in general, to synthesize speech from prerecorded segments of phonemic dimensions. Nor can we cut the sound stream along the time dimension so as to recover segments that will be perceived as separate phonemes. Of course, there are exceptions. As we might expect, these are the steady-state vowels and the long-duration noises of certain fricatives in which, as we have seen, the sounds show minimal restructuring. Apart from these exceptions, however, segments corresponding to the phonemes are not found at the acoustic level.

We shall see later that the articulatory gestures corresponding to successive phonemes—or, more precisely, their subphonemic features—are overlapped, or shingled, one onto another. This parallel delivery of information produces at the acoustic level the merging of influences we have already referred to and yields irreducible acoustic segments of approximately syllabic dimensions. Thus, segmentation also exhibits a complex relation between linguistic structure or perception, on the one hand, and the sound stream on the other.

PERCEPTION OF THE RESTRUCTURED PHONEMES: THE SPEECH MODE

If phonemes are encoded syllabically in the sound stream, they must be recovered in perception by an appropriate decoder. Perception of phonemes that have been so encoded might be expected to differ from the perception of those that have not and also, of course, from nonspeech. In this section we will suggest that such differences do, in fact, exist.

We have already seen one example of such a difference in the transition cues for /d/ in /di/ and /du/. Taken out of speech context, these transitions sound like whistles, the one rising through a range of high pitches and the other falling through low pitches: they do not sound like each other, nor

[14] This is not to say that the sound spectrogram fails to show discontinuities along the time axis. Fant (1962a, 1962b) discusses the interpretation to be given these abrupt changes and the temporal segments bounded by them. He warns: "Sound segment boundaries should not be confused with phoneme boundaries. Several adjacent sounds of connected speech may carry information on one and the same phoneme, and there is overlapping in so far as one and the same sound segment carries information on several adjacent phonemes [1962a, p. 9]."

even like speech. This example could be multiplied to include the transition cues for many other phonemes. With simplified speech of the kind already shown, the listener's perception is very different depending on whether he is, for whatever reason, in the speech mode or out of it (Brady, House, & Stevens, 1961).

Even on the basis of what can be heard in real speech, one might have suspected that the perception of encoded and unencoded phonemes [15] is somehow different. One has only to listen carefully to some of the latter in order to make reasonably accurate guesses about the auditory and acoustic dimensions that are relevant to their perception. The fricatives /s/ and /š/, for example, obviously differ in manner from other phonemes in that there is noise of fairly long duration; moreover, one can judge by listening that they differ from each other in the "pitch" of the noise—that is, in the way the energy is distributed along the frequency scale. Consider, on the other hand, the encoded phonemes /b,d,g/. No amount of listening, no matter how careful, is likely to reveal that an important manner cue is a rapidly rising frequency at the low end of the frequency scale (first formant), or that these stops are distinguished from each other primarily by the direction and extent of a rapid frequency glide in the upper frequency range (second and third formants).

[15] There is a need, in much that follows, for a convenient way to refer to classes of phonemes that show much—or little—restructuring of their acoustic cues as a function of context. The former are, indeed, encoded. We shall refer to the latter as "unencoded phonemes," implying only that they are found at the other end of a continuum on degree of restructuring; we do not wish to imply differences in the processes affecting these phonemes, whether or not such differences can be inferred from their perceptual characteristics.

These observations of perceptual differences between speech and nonspeech sounds, and even among classes of phonemes, do not stand alone. Controlled experiments can show more accurately, if sometimes less directly, the differences in perception. We will next consider some of these experiments.

Tendencies toward Categorical and Continuous Perception

Research with some of the encoded phonemes has shown that they are categorical, not only in the abstract linguistic sense, but as immediately given in perception. Consider, first, that in listening to continuous variations in acoustic signals, one ordinarily discriminates many more stimuli than he can absolutely identify. Thus, we discriminate about 1200 different pitches, for example, though we can absolutely identify only about seven. Perception of the restructured phonemes is different in that listeners discriminate very little better than they identify absolutely; that is to say, they hear the phonemes but not the intraphonemic variations.

The effect becomes clear impressionistically if one listens to simplified, synthetic speech signals in which the second-formant transition is varied in relatively small, acoustically equal steps through a range sufficient to produce the three stops, /b/, /d/, and /g/. One does not hear steplike changes corresponding to the changes in the acoustic signal, but essentially quantal jumps from one perceptual category to another.

To evaluate this effect more exactly, various investigators have made quantitative comparisons of the subjects' ability to identify the stimuli absolutely and to discriminate them on any basis whatsoever. For certain consonant distinctions it has been found that the

mode of perception is, in fact, nearly categorical: listeners can discriminate only slightly better than they can identify absolutely. In greater or lesser degree, this has been found for /b,d,g/ (Eimas, 1963; Griffith, 1958; Liberman, Harris, Hoffman, & Griffith, 1957; Studdert-Kennedy, Liberman, & Stevens, 1963, 1964); /d,t/ (Liberman, Harris, Kinney, & Lane, 1961)[16]; /b,p/ in intervocalic position (Liberman, Harris, Eimas, Lisker, & Bastian, 1961), and presence or absence of /p/ in *slit* vs. *split* (Bastian, Delattre, & Liberman, 1959; Bastian, Eimas, & Liberman, 1961; Harris, Bastian, & Liberman, 1961).

The perception of unencoded steady-state vowels is quite different from the perception of stops.[17] To appreciate this difference one need only listen to synthetic vowels that vary, as in the example of the stops, in relatively small and acoustically equal steps through a range sufficient to produce three adjacent phonemes—say /i/, /I/, and /ɛ/. As heard, these vowels change step-by-step, much as the physical stimulus changes: the vowel /i/ shades into /I/, and /I/ into /ɛ/. Immediate perception is more nearly continuous than categorical and the listener hears many intraphonemic variations. More precise measures of vowel perception indicate that, in contrast to the stops, listeners can discriminate many more stimuli than they can identify absolutely (Fry, Abramson, Eimas, & Liberman, 1962; Stevens, Ohman, & Liberman, 1963; Stevens, Ohman, Studdert-Kennedy, & Liberman, 1964). Similar studies of the perception of vowel duration (Bastian & Abramson, 1962) and tones in Thai (Abramson, 1961), both of which are phonemic in that language, have produced similar results. We should suppose that steady-state vowels, vowel duration, and the tones can be perceived in essentially the same manner as continuous variations in nonspeech signals. The results of a direct experimental comparison by Eimas (1963) suggest that this is so.

We emphasize that in speaking of vowels we have so far been concerned only with those that are isolated and steady state. These are, as we have said, unencoded and hence not necessarily perceived in the speech mode. But what of the more usual situation we described earlier, that of vowels between consonants and in rapid articulation? Stevens (1966) has supposed that the rapid changes in formant position characteristic of such vowels would tend to be referred in perception to the speech mode, and he has some evidence that this is so, having found that perception of certain vowels in proper dynamic context is more nearly categorical than that of steady-state vowels. Inasmuch as these rapidly articulated vowels are substantially restructured in the sound stream, Stevens' results may be assumed to reflect the operation of the speech decoder.

Lateral Differences in the Perception of Speech and Nonspeech

The conclusion that there is a speech mode, and that it is characterized by processes different from those underlying the perception of other sounds,

[16] Studies of this distinction (and of the corresponding ones for the other stops) in 11 diverse languages indicate that it tends to be categorical also in production. (Abramson & Lisker, 1965; Lisker & Abramson, 1964a, 1964b.)

[17] In experiments with mimicry, Ludmilla Chistovich and her colleagues have obtained differences between vowels and consonants that are consistent with the differing tendencies toward categorical and continuous perception described here. (Chistovich, 1960; Chistovich, Klaas, & Kuz'min, 1962; Galunov & Chistovich, 1965; Kozhevnikov & Chistovich, 1965.)

is strengthened by recent indications that speech and nonspeech sounds are processed primarily in different hemispheres of the brain. Using Broadbent's (1954) method of delivering competing stimuli simultaneously to the two ears, investigators have found that speech stimuli presented to the right ear (hence, mainly to the left cerebral hemisphere) are better identified than those presented to the left ear (hence, mainly to the right cerebral hemisphere), and that the reverse is true for melodies and sonar signals (Broadbent & Gregory, 1964; Bryden, 1963; Chaney & Webster, 1965; Kimura, 1961, 1964, 1967). In the terminology of this paper, the encoded speech signals are more readily decoded in the left hemisphere than in the right. This suggests the existence of a special left-hemisphere mechanism different from the right-hemisphere mechanism for the perception of sounds not similarly encoded. It is of interest, then, to ask whether the encoded stops and the unencoded steady-state vowels are, perhaps, processed unequally by the two hemispheres. An experiment was carried out (Shankweiler & Studdert-Kennedy, 1967b) designed to answer this question, using synthetic speech syllables that contrasted in just one phoneme. A significantly greater right-ear advantage was found for the encoded stops than for the unencoded steady-state vowels. The fact that the steady-state vowels are less strongly lateralized in the dominant (speech) hemisphere may be taken to mean that these sounds, being unencoded, can be, and presumably sometimes are, processed as if they were nonspeech. In another experiment (Shankweiler & Studdert-Kennedy, 1967a), the consonant and vowel comparisons were made with real speech. Different combinations of the same set of consonant-vowel-consonant syllables were used for both tests. As before, a decisive right-ear advantage was found for contrasting stop consonants, and again there was no difference for vowels, even though these were here articulated in dynamic context between consonants. We will be interested to determine what happens to lateral differences in vowel perception when the vowels are very rapidly articulated. Such vowels are, as has been said, necessarily restructured to some extent and may be correspondingly dependent on the speech decoder for their perception.

Perception of Speech by Machine and by Eye

If speech is, as we have suggested, a special code, its perception should be difficult in the absence of an appropriate decoder such as we presumably use in listening to speech sounds. It is relevant, therefore, to note how very difficult it is to read visual transforms of speech or to construct automatic speech recognizers.

Consider, first, a visual transform—for example, the spectrogram. As we have already seen, the spectrographic pattern for a particular phoneme typically looks very different in different contexts. Furthermore, visual inspection of a spectrogram does not reveal how a stretch of speech might be divided into segments corresponding to phonemes, or even how many phonemes it might contain: the eye sees the transformed acoustic signal in its undecoded form. We should not be surprised, then, to discover that spectrograms are, in fact, extremely difficult to read.

Some part of the difficulty may be attributed to inadequacies of the transform or to lack of training. But an improved transform, if one should be found, would not by itself suffice to make spectrograms readable, since it would not obviate the need to decode.

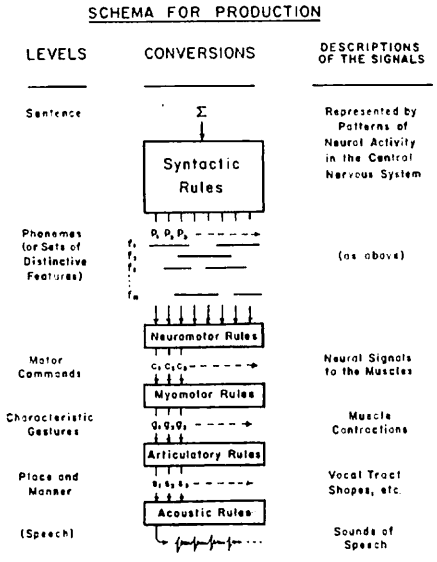

FIG. 5. Schematic representations of assumed stages in speech production.

Nor is training likely to overcome the difficulty. Many persons have had considerable experience with spectrograms, yet none has found it possible to read them well.[18] Ideally, training in "reading" spectrograms should cause the transitions for /d/ in /di/ and /du/ to look alike. The speech decoder, after all, makes them sound alike when speech is perceived by ear; moreover, it seems obvious that if the decoder did not do this, speech would be much more difficult to perceive. If, as we suspect, training alone cannot make the acoustic cues for the same phoneme *look* alike, then we should, perhaps, conclude that the speech decoder, which makes them *sound* alike, is biologically tied to an auditory input.

[18] "As a matter of fact I have not met one single speech researcher who has claimed he could read speech spectrograms fluently, and I am no exception myself [Fant, 1962a, p. 4]." Spectrograms are, even so, less difficult to read than oscillograms—whence their popular name, "visible speech," and much of the early enthusiasm for them (Potter, Kopp, & Green, 1947).

How, then, do machines fare in recognizing the encoded sounds of speech? If speech were a cipher, like print, it would be no more difficult to build a speech recognizer than a print reader. In fact, the speech recognizer has proved to be more difficult, and by a very wide margin, largely because it needs an appropriate decoder, as a print reader does not, and because the design of that decoder is not easily accomplished—a conclusion confirmed by two decades of intensive effort. We might repeat, in passing, that for human beings the difficulty is the other way around: perceiving speech is far easier than reading.

ENCODING AND DECODING: FROM PHONEME TO SOUND AND BACK

Conversions from Phoneme to Sound

Having considered the evidence that speech is a code on the phonemes, we must ask how the encoding might have come about. The schema in Figure 5 represents the various conversions that presumably occur as speech proceeds from sentence, through the empty segments at the phoneme level, to the final acoustic signal. The topmost box, labeled Syntactic Rules, would, if properly developed, be further broken down into phrase-structure rules, transformation rules, morphophonemic rules, and the like. (See, for example, Chomsky, 1964.) These processes would be of the greatest interest if we were dealing with speech perception in the broadest sense. Here, however, we may start with the message in the form of an ordered sequence of phonemes (or, as we will see, their constituent features) and follow it through the successive converters that yield, at the end, the acoustic waveform.

Subphonemic features: Their role in production. First, we must take account of the fact that the phonemes

are compounded of a smaller number of elements, and, indeed, shift the emphasis from the phoneme to these subphonemic features.[19] The total gesture in the articulation of /b/, for example, can be broken down into several distinctive elements: (*a*) closing and opening of the upper vocal tract in such a way as to produce the manner feature characteristic of the stop consonants; (*b*) closing and opening of the vocal tract specifically at the lips, thus producing the place feature termed bilabiality; (*c*) closing the velum to provide the feature of orality; and (*d*) starting vocal fold vibration simultaneously with the opening of the lips, appropriate to the feature of voicing. The phoneme /p/ presumably shares with /b/ features 1, 2, and 3, but differs as to feature 4, in that vocal fold vibration begins some 50 or 60 milliseconds after opening of the lips; /m/ has features 1, 2, and 4 in common with /b/, but differs in feature 3, since the velum hangs open to produce the feature of nasality; /d/ has features 1, 3, and 4 in common with /b/, but has a different place of articulation; and so on.

That subphonemic features are present both in production and perception has by now been quite clearly established.[20] Later in this paper we will discuss some relevant data on production. Here we merely note that we must deal with the phonemes in terms of their constituent features because the existence of such features is essential to the speech code and to the efficient production and perception of language. We have earlier remarked that high rates of speech would overtax the temporal resolving power of the ear if the acoustic signal were merely a cipher on the phonemic structure of the language. Now we should note that *speaking* at rates of 10 to 15 phonemes per second would as surely be impossible if each phoneme required a separate and distinct oral gesture. Such rates can be achieved only if separate parts of the articulatory machinery—muscles of the lips, tongue, velum, etc.—can be separately controlled, and if the linguistic code is such that a change of state for any one of these articulatory entities, taken together with the current state of the others, is a change to another element of the code—that is, to another phoneme. Thus, dividing the load among the articulators allows each to operate at a reasonable pace, and tightening the code keeps the information rate high. It is this kind of parallel processing that makes it possible to get high-speed performance

[19] The term "feature" has varied uses, even in this paper. Thus, in this paragraph, we describe a particular speech gesture in conventional phonetic terms, referring to how the articulators move and where the constrictions occur as the manner and place features that, taken together, characterize the gesture. Two paragraphs later, the description of a possible model for speech production identifies subphonemic features as implicit instructions to separate and independent parts of the motor machinery. Viewed in this way, the distinctive features of a phoneme are closely linked to specific muscles and the neural commands that actuate them.

Distinctive features of this kind are clearly different from those so well known from the work of Jakobson and his colleagues. See Jakobson, Fant, and Halle (1952) and the various revisions proposed by the several authors (Fant, 1967; Halle, 1964; Jakobson & Halle, 1956; Stevens & Halle, in press). We are, nevertheless, deeply indebted to them for essential points—in particular, that the phonemes can be characterized by sets of features which are few in number, limited to a very few states each, and maintained throughout several phonemes. Other important characteristics for an ideal system of distinctive features include mutual independence, clear correspondences with physiological—or acoustic—observables, and as much universality across languages as parsimony for a single language will allow.

[20] For discussions of the psychological reality of phoneme segments and subphonemic features, see Kozhevnikov and Chistovich, 1965; Stevens and House, in press; Wickelgren, 1966.

with low-speed machinery. As we will see, it also accounts for the overlapping and intermixing of the acoustic signals for the phonemes that is characteristic of the speech code.

A model for production. The simplest possible model of the production process would, then, have the phonemes of an utterance represented by sets of subphonemic features, and these in turn would be assumed to exist in the central nervous system as implicit instructions [21] to separate and independent parts of the motor machinery. The sequence of neural signals corresponding to this multidimensional string of control instructions may well require, for its actualization, some amplitude adjustments and temporal coordination, in the box labeled "Neuromotor Rules," in order to yield the neural impulses that go directly to the selected muscles of articulation and cause them to contract. If we are to continue to make the simplest possible assumption, however, we must suppose that reorganization of the instructions at this stage would be limited to such coordination and to providing supplementary neural signals to insure cooperative activity of the remainder of the articulatory apparatus; there would be no reorganization of the commands to the "primary" actuators for the selected features. In that case, the neural signals that emerge would bear still an essentially one-to-one correspondence with the several dimensions of the subphonemic structure. Indeed, this is a necessary condition for our model, since total reorganization at this stage would destroy the parallel processing, referred to earlier, on which high-speed reception depends, or else yield a syllabic language from which the phoneme strings could not be recovered.

The next conversion, from neural command (in the final common paths) to muscle contraction, takes place in the box labeled "Myomotor Rules." If muscles contract in accordance with the signals sent to them, then this conversion should be essentially trivial, and we should be able not only to observe the muscle contractions by looking at their EMG signals, but also to infer the neural signals at the preceding level.

It is at the next stage—the conversion from muscle contraction to vocal tract shapes by way of Articulatory Rules—that a very considerable amount of encoding or restructuring must surely occur. If we take into account the structure and function of the articulatory system, in particular the intricate linkages and the spatial overlap of the component parts, we must suppose that the relation between contraction and resulting shape is complex, though predictable. True encoding occurs as a consequence of two further aspects of this conversion; the fact that the subphonemic features can be, and are, put through in parallel means that each new set of contractions (*a*) starts from whatever configuration then exists (as the result of the preceding contractions) and (*b*) typically occurs before the last set has ended, with the result that the shape of the tract at any instant represents the merged effects of past and present instructions. Such merging is, in effect, an encoding operation according to our use of that term, since it involves an extensive restructuring of the output—in this case, the

[21] These instructions might be of two types, "on-off" or "go to," even in a maximally simple model. In the one case, the affected muscle would contract or not with little regard for its current state (or the position of the articulator it moves); in the other, the instruction would operate via the γ-efferent system to determine the degree of contraction (hence, the final position of the articulator, whatever its initial position). Both types of instruction—appropriate, perhaps, to fast and slow gestures, respectively—may reasonably be included in the model.

shape of the tract. The relation of message units to code becomes especially complex when temporal and spatial overlaps occur together. Thus, the conversion from muscle contraction to shape is, by itself, sufficient to produce the kinds of complex relation between phoneme and sound that we have found to exist in the overall process of speech production, that is, a loss of segmentability together with very large changes in the essential acoustic signals as a function of context. Given the structure of the articulatory apparatus, these complexities appear to be a necessary concomitant of the parallel processing that makes the speech code so efficient.

The final conversion, from shape to sound, proceeds by what we have in Figure 5 called Acoustic Rules. These rules, which are now well understood,[22] are complex from a computational standpoint, but they operate on an instant-by-instant basis and yield (for the most part) one-to-one relations of shape to sound.

Does the Encoding Truly Occur in the Conversion from Command to Shape?

We have supposed that, of the four conversions between phoneme and sound, one at least—the conversion from contractions to tract shape—is calculated to produce an encoding of the kind we found when earlier we examined the acoustic cues for phoneme perception. But does it, in fact, produce that encoding, and if so, does it account for all of it? Downstream from this level there is only the conversion from shape to sound, and that, though complex, does not appear to involve encoding. But what of the upstream conversions, particularly the one that lies between the neural representations of the phonemes and the commands to the articulatory muscles? We cannot at the present time observe those processes, nor can we directly measure their output—that is, the commands to the muscles. We can, however, observe some aspects of the contractions—for example, the electromyographic correlates—and if we assume, as seems reasonable, that the conversion from command to contractions is straightforward, then we can quite safely infer the structure of the commands.[23] By determining to what extent those inferred commands (if not the electromyographic signals themselves) are invariant with the phoneme we can, then, discover how much of the encoding occurs in the conversion from contraction to shape (Articulatory Rules) and how much at higher levels.

[22] The dependence of speech sound on vocal tract shape (and movement) has, of course, been studied by many workers and for many years. A landmark in this field, and justification for our statement, is the book by Fant (1960).

[23] For commands of the on-off type (see Footnote 21), we would expect muscle contractions—and EMG potentials—to be roughly proportional to the commands; hence, the commands will be mirrored directly by the EMG potentials when these can be measured unambiguously for the muscles of interest. All these conditions are realizable, at least approximately, for a number of phoneme combinations—for example, the bilabial stops and nasal consonants with unrounded vowels, as described in later paragraphs.

Commands of the "go to" type, which may well be operative in gestures that are relatively slow and precise, would presumably operate via the γ-efferent system to produce only so much contraction of the muscle as is needed to achieve a target length. The contraction—and the resulting EMG signal—would then be different for different starting positions, that is, for the same phoneme in different contexts. Even so, the significant aspects of the command can be inferred, since presence versus absence and sequential position (if not the relative timing) of the EMG signal persist despite even large changes in its magnitude.

For general discussions of the role of electromyography in research on speech, see the reviews by Cooper (1965) and Fromkin and Ladefoged (1966).

Motor commands as presumed invariants. Before discussing the electromyographic evidence, we should say what kind of invariance with the subphonemic features (hence, with the phonemes) we might, at best, expect to find. It should be clear from many of the data presented in this paper that language is no more a left-to-right process at the acoustic level than it is syntactically. As we have seen, the acoustic representations of the successive phonemes are interlaced. Some control center must therefore "know" what the syllable is to be in order that its component subphonemic features may be appropriately combined and overlapped. There are other grounds than the data we have cited for inferring such syllabic organization. Temporal relations between syllables may be adjusted for slow or rapid speech and for changes in syllable duration such as Lindblom has reported for a given syllable placed in different polysyllabic contexts. But these changes in syllable duration do not affect all portions of the syllable equally: timing relations within the syllable are also adjusted according to context (Lindblom, 1964), and this must entail variation in the relative timing of the component subphonemic features, as is suggested by Kozhevnikov and Chistovich's (1965) analysis of the articulatory structure of consonant clusters. Ohman (1964, 1966) has proposed a model for syllable articulation consistent with this analysis. Such contextual variations preclude the invariance of all the features that comprise a phoneme in a particular context. The most that we can expect is that some subset of these features, and so of the neural signals to the muscles (after operation of the Neuromotor Rules of Figure 5), will be invariant with the phoneme; there will then be for each subphonemic feature characteristic neuromotor "markers," implicating only one or a few component parts of the system, perhaps only the contraction of a single specific muscle. These characteristic components of the total neural activity we will refer to as "motor commands." [24] We should also emphasize that we refer here to the phoneme as perceived, not as an abstract linguistic entity serving primarily a classificatory function.

Indications from electromyography. What, then, do we find when we look at the electromyographic (EMG) correlates of articulatory activity? Recent research in our laboratory and several others has been directed specifically to the questions raised in the preceding paragraphs, but the data are, as yet, quite limited. As we see these data, they do, however, permit some tentative conclusions.

When two adjacent phonemes are produced by spatially separate groups of muscles, there are essentially invariant EMG tracings from the characteristic gestures for each phoneme, regardless of the identity of the other.[25]

[24] Thus, the motor commands are, in one sense, abstract "-eme" type entities, with invariance assumed and observation directed to their discovery and enumeration; in another sense, and to the degree that observation justifies the assumption about invariance, motor commands constitute the essential subset of real neural signals with which a general model of production and perception should be principally concerned.

[25] It is not, of course, necessary that the EMG signals be precisely the same since invariance is expected only of the motor commands inferred from them. In the minimally complicated cases cited in this paragraph, the commands do transform into EMG potentials that are essentially the same (in different contexts), aside from some differences in magnitude.

Magnitude differences do occur, however, variously for the individual and context, and may have linguistic significance attributed to them. Thus, Fromkin (1966) concludes, mainly from EMG data on the lip-closing gestures of her principal speaker:

This has been found for /f/, for example, followed by /s/, /t/, /ts/, /θ/, or /θs/ and preceded by /i/, /il/, or /im/ (MacNeilage, 1963). Similarly, invariance has been found for /b/, /p/, and /m/ regardless of the following vowel (Fromkin, 1966; Harris, Lysaught, & Schvey, 1965). Here it is easy to associate the place and manner features with the contractions of specific muscles, and to equate the EMG signals for a specific feature wherever it occurs, at least within this limited set of phonemes. We should emphasize that corresponding invariance is not to be found in the acoustic signal: for /f/ the duration of the noise varies over a range of two to one in the contexts listed above; the vast differences in acoustic cue for /b/ or /p/ before various vowels were described in an earlier section.

When the temporally overlapping gestures for successive phonemes involve more or less adjacent muscles that control the same structures, it is of course more difficult to discover whether there is invariance or not.[25] This is the situation for the /di, du/ examples we used earlier. In our own studies of such cases we find essentially identical EMG signals from tongue-tip electrodes for the initial consonant; however, the signal for a following phoneme that also involves the tongue tip may show substantial changes from its characteristic form (Harris, 1963; Harris, Huntington, & Sholes, 1966). Such changes presumably reflect the execution of position-type commands (see Footnotes 21 and 23) rather than reorganization at a higher level. True reorganization of the neural commands is not excluded by such data—indeed, we have some data that might be so interpreted (MacNeilage, DeClerk, & Silverman, 1966)—but the evidence so far is predominantly on the side of invariance.

If the commands—or, indeed, signals of any kind—are to be invariant with the phonemes, they must reflect the segmentation that is so conspicuously absent at the acoustic level. Do we, then, find such segmentation in the EMG records? Before answering that question, we should remind ourselves that the activity of a single muscle or muscle group does not, in any case, reflect a phoneme but only a subphonemic feature, and that a change from one phoneme to another may require a change in command for only one of several features. We should not expect, therefore, that all the subphonemic features will start and stop at each phoneme boundary, and, in fact, they do not. We do find, however, in the onsets and offsets of EMG activity in various muscles a segmentation like that of the several dimensions that constitute the phoneme. One can see, then, where the phoneme boundaries

The results of the present investigation do not support the hypothesis that a simple one-to-one correspondence exists between a phoneme and its motor commands. For the bilabial stops /b/ and /p/ different motor commands produce different muscular gestures for these consonants occurring in utterance initial and final positions [p. 195].

Harris, Lysaught, and Schvey (1965) report, on the contrary, that differences in EMG signal for the lip-closing gesture did not reach statistical significance for two groups of five subjects each in two experiments, one of which overlapped Fromkin's study. They noted, though, that each of the subjects showed an individual pattern of small but consistent variations of EMG signal with context. (For a discussion of these differences in the interpretation of basically similar data, see Cooper, 1966.)

[26] A practical difficulty exists in allocating observed EMG potentials to specific muscles when more than one muscle is close to a surface electrode. Needle electrodes will often resolve such ambiguities, but they pose other problems of a practical nature.

must be.[27] This is in striking contrast to what is found at the acoustic level. There, as we have earlier pointed out, almost any segment we can isolate contains information about—and is a cue for—more than one phoneme. Thus in the matter of segmentation, too, the EMG potentials—and even more the motor commands inferred from them—bear a simpler relation to the perceived phonemes than does the acoustic signal.

In summary we should say that we do not yet know precisely to what extent motor commands are invariant with the phonemes. It seems reasonably clear, however, that they are more nearly invariant than are the acoustic signals, and we may conclude, therefore, that a substantial part of the restructuring occurs below the level of the commands. Whether some significant restructuring occurs also at higher levels can only be determined by further investigation.

Decoding: From Sound to Phoneme

If speech were a simple cipher one might suppose that phoneme perception could be explained by the ordinary principles of psychophysics and discrimination learning. On that view, which is probably not uncommon among psychologists, speech consists of sounds, much like any others, that happen to signal the phoneme units of the language. It is required of the listener only that he learn to connect each sound with the name of the appropriate phoneme. To be sure, the sounds must be discriminably different, but this is a standard problem of auditory psychophysics and poses no special difficulty for the psychologist. It would follow that perception of speech is not different in any fundamental way from perception of other sounds (assuming sufficient practice), and that, as Lane (1965) and Cross, Lane, and Sheppard (1965) suppose, no special perceptual mechanism is required.

The point of this paper has been, to the contrary, that speech is, for the most part, a special code that must often make use of a special perceptual mechanism to serve as its decoder.[28] On the evidence presented here we should conclude, at the least, that most phonemes can not be perceived by a straightforward comparison of the incoming signal with a set of stored phonemic patterns or templates, even if one assumes mechanisms that can extract simple stimulus invariances such as, for example, constant ratios. To find acoustic segments that are in any reasonably simple sense invariant with linguistic (and perceptual) segments—that is, to perceive without decoding—one must go to the syllable level or higher. Now the number of syllables, reckoned simply as the number of permissible combinations and permutations of phonemes, is several thousand in English. But the number of acoustic patterns is far greater than that, since the acoustic signal varies considerably with the speaker and with the stress, intonation, and tempo of his speech

[27] In some cases the boundaries are sharply marked by a sudden onset of EMG signal for a particular muscle. Usually, though, the onsets are less abrupt and activity persists for a few hundred milliseconds. This might seem inadequate to provide segmentation markers for the rapid-fire phoneme sequences; indeed, precise time relationships may be blurred in normal articulation, but without obscuring the sequential order of events and the separateness of the channels carrying information about the several features.

[28] We referred earlier to perception in the speech mode and indicated that it was not operative—or, at least, not controlling—for some speech sounds as well as all nonspeech sounds. It need hardly be said, then, that the speech decoder provides only one pathway for perception; nonspeech sounds obviously require another pathway, which may serve adequately for the unencoded aspects of speech as well.

(Cooper, Liberman, Lisker, & Gaitenby, 1963; Liberman, Ingemann, Lisker, Delattre, & Cooper, 1959; Lindblom, 1963, Peterson & Sivertsen, 1960; Peterson, Wang, & Sivertsen, 1958; Sivertsen, 1961). To perceive speech by matching the acoustic signal to so many patterns would appear, at best, uneconomical and inelegant. More important, it would surely be inadequate, since it must fail to account for the fact that we do perceive phonemes. No theory can safely ignore the fact that phonemes are psychologically real.

How, then, is the phoneme to be recovered? We can imagine, at least in general outline, two very different possibilities: one, a mechanism that operates on a purely auditory basis, the other, by reference to the processes of speech production.

The case for an auditory decoder. The relevant point about an auditory decoder is that it would process the signal in *auditory* terms—that is, without reference to the way in which it was produced—and successfully extract the phoneme string. On the basis of what we know of speech, and of the attempts to build an automatic phoneme recognizer, we must suppose such a device would have to be rather complex. There is, to be sure, some evidence for the existence of processing mechanisms in the auditory system that illustrate at a very simple level one kind of thing that an auditory decoder might have to do. Thus, Whitfield and Evans (1965) have found single cells in the auditory cortex of the cat that respond to frequency changes but not to steady tones. Some of these cells responded more to tones that are frequency modulated upward and others to tones modulated downward. Such specificity of response also exists at lower levels in the auditory nervous system (Nelson, Erulkar, & Bryant, 1966). These findings are examples of a kind of auditory processing that might, perhaps, be part of a speech decoder, but one that would have to go beyond these processes—to more complex ones—if it were to deal successfully with the fact that very different transitions will, in different contexts, cue the same phoneme. But if this could be accomplished, and if it were done independently of motor parameters, we should have a purely auditory decoder.

The case for mediation by production. Though we cannot exclude the possibility that a purely auditory decoder exists, we find it more plausible to assume that speech is perceived by processes that are also involved in its production. The most general and obvious motivation for such a view is that the perceiver is also a speaker and must be supposed, therefore, to possess all the mechanisms for putting language through the successive coding operations that result eventually in the acoustic signal.[29] It seems unparsimonious to assume that the speaker-listener employs two entirely separate processes of equal status, one for encoding language and the other for decoding it. A simpler assumption is that there is only one process, with appropriate linkages between sensory and motor components.[30]

[29] We have noted that training in reading speech spectrograms has so far not succeeded in developing in the trainee a visual decoder for speech comparable to the one that works from an auditory input. This may well reflect the existence of special mechanisms for the speech decoder that are lacking for its visual counterpart. In general, a theory about the nature of the speech decoder—and, in particular, a motor theory—must be concerned with the nature of the mechanism, though not necessarily with the question of how it was acquired in the history of the individual or the race. This is an interesting, but separate, question and is not considered here.

[30] We should suppose that the links between perception and articulation exist at

Apart from parsimony, there are strong reasons for considering this latter view. Recall, for example, the case of /di, du/. There we saw that the acoustic patterns for /d/ were very different, though the perception of the consonantal segment was essentially the same. Since it appears that the /d/ gesture—or, at least, some important, "diagnostic" part of it—may also be essentially the same in the two cases, we are tempted to suppose that one hears the same /d/ because perception is mediated by the neuromotor correlates of gestures that are the same. This is basically the argument we used in one of our earliest papers (Liberman, Delattre, & Cooper, 1952) to account for the fact that bursts of sound at very different frequencies are required to produce the perception of /k/ before different vowels. Extending the argument, we tried in that same paper to account also for the fact that the same burst of sound is heard as /p/ before /i/ but as /k/ before /a/, the point being that, because of temporal overlap (and consequent acoustic encoding), very different gestures happen to be necessary in these different vowel environments in order to produce the same acoustic effect (the consonant burst). The argument was applied yet again to account for a finding about second-formant transitions as cues: that the acoustic cues for /g/ can be radically different even when the consonant is paired with closely related vowels (i.e., in the syllables /ga/ and /gɔ/), yet the perception and the articulation are essentially the same (Liberman, 1957). But these are merely striking examples of what must be seen now as a general rule: there is typically a lack of correspondence between acoustic cue and perceived phoneme, and in all these cases it appears that perception mirrors articulation more closely than sound.[31] If this were not so, then for a listener to hear the same phoneme in various environments, speakers would have to hold the acoustic signal constant, which would, in turn, require that they make drastic changes in articulation. Speakers need not do that—and in all probability they cannot—yet listeners nevertheless hear the same phoneme. This supports the assumption that the listener uses the inconstant sound as a basis for finding his way back to the articulatory gestures that produced it and thence, as it were, to the speaker's intent.

The categorical perception of stop consonants also supports this assumption.[32] As described earlier in this paper, perception of these sounds is categorical, or discontinuous, even when the acoustic signal is varied continuously. Quite obviously, the required articulations would also be discontinuous. With /b,d,g/, we can vary the acoustic cue along a continuum, which corresponds, in effect, to closing the vocal tract at various points along its length. But in actual speech the closing is accomplished by discontinuous or categorically different gestures: by the lips for /b/, the tip of the tongue for /d/, and the back of the tongue for /g/. Here, too, perception appears to be tied to articulation.

relatively high levels of the nervous system. For information about, or reference to, motor activity, the experienced organism need not rely—at least not very heavily and certainly not exclusively—on proprioceptive returns from the periphery, for example, from muscular contractions. (See von Holst & Mittelstadt, 1950.)

[31] For further discussion of this point, see Liberman, 1957; Cooper, Liberman, Harris, and Grubb, 1958; Lisker, Cooper, and Liberman, 1962; Liberman, Cooper, Harris, MacNeilage, and Studdert-Kennedy, in press.

[32] For further discussion of this point, see Liberman, Cooper, Harris, and MacNeilage, 1962.

The view that speech is perceived by reference to production is receiving increased attention. Researchers at the Pavlov Institute in Leningrad have adduced various kinds of evidence to support a similar hypothesis (Chistovich, 1960; Chistovich, Klaas, & Kuz'min, 1962). Ladefoged (1959) has presented a motor-theoretical interpretation of some aspects of speech perception. More recently, Ladefoged and McKinney (1963) have found in studies of stress that perception is related more closely to certain low-level aspects of stress production than it is to the acoustic signal. And in a very recent study Lieberman (1967) has found interesting evidence for a somewhat similar conclusion in regard to the perception of some aspects of intonation.

The role of productive processes in models for speech perception. In its most general form, the hypothesis being described here is not necessarily different in principle from a model for speech perception called "analysis-by-synthesis" that has been advanced by Stevens (1960) and by Halle and Stevens (1962; Stevens & Halle, in press) following a generalized model proposed by MacKay (1951). In contrast, perhaps, to the computer-based assumptions of analysis-by-synthesis, we would rather think in terms of overlapping activity of several neural networks—those that supply control signals to the articulators and those that process incoming neural patterns from the ear—and to suppose that information can be correlated by these networks and passed through them in either direction. Such a formulation, in rather general terms, has been presented elsewhere (Liberman, Cooper, Studdert-Kennedy, Harris, & Shankweiler, in press).

The most general form of the view that speech is perceived by reference to production does not specify the level at which the message units are recovered. The assumption is that at some level or levels of the production process there exist neural signals standing in one-to-one correspondence with the various segments of the language—phoneme, word, phrase, etc. Perception consists in somehow running the process backward, the neural signals corresponding to the various segments being found at their respective levels. In phoneme perception—our primary concern in this paper—the invariant is found far down in the neuromotor system, at the level of the commands to the muscles. Perception by morphophonemic, morphemic, and syntactic rules of the language would engage the encoding processes at higher levels. The level at which the encoding process is entered for the purposes of perceptual decoding may, furthermore, determine which shapes can and cannot be detected in raw perception. The invariant signal for the different acoustic shapes that are all heard as /d/, for example, may be found at the level of motor commands. In consequence, the listener is unaware, even in the most careful listening, that the acoustic signals are, in fact, quite radically different. On the other hand, a listener can readily hear the difference between tokens of the morphophoneme {S} when it is realized as /s/ in cats and as /z/ in dogs, though he also "knows" that these two acoustic and phonetic shapes are in some sense the same. If the perception of that commonality is also by reference to production, the invariant is surely at a level considerably higher than the motor commands.

THE EFFICIENCY OF THE SPEECH CODE

Speech can be produced rapidly because the phonemes are processed in parallel. They are taken apart into their constituent features, and the fea-

tures belonging to successive phonemes are overlapped in time. Thus the load is divided among the many largely independent components of the articulatory system. Within the phonological constraints on combinations and sequences that can occur, a speaker produces phonemes at rates much higher than those at which any single articulatory component must change its state.[33]

In the conversion of these multidimensional and overlapping articulatory events into sound, a complex encoding occurs. The number of dimensions is necessarily reduced, but the parallel transmission of phonemic information is retained in that the cues for successive phonemes are imprinted on a single aspect of the acoustic signal. Thus, the movement of articulators as independent as the lips and the tongue both affect the second formant and its transitions: given an initial labile consonant overlapped with a tongue shape appropriate for the following vowel, one finds, as we have already shown, that the second-formant transition simultaneously carries information about both phonemic segments.[34]

If the listener possesses some device for recovering the articulatory events from their encoded traces in the sound stream, then he should perceive the phonemes well and, indeed, evade several limitations that would otherwise apply in auditory perception. As we pointed out early in this paper, the temporal resolving power of the ear sets a relatively low limit on the rate at which discrete acoustic segments can be perceived. To the extent that the code provides for parallel processing of successive phonemes, the listener can perceive strings of phonemes more rapidly than he could if the acoustic signals for them were arranged serially, as in an alphabet. Thus, the parallel processing of the phonemes is as important for efficient perception as for production.

We also referred earlier to the difficulty of finding a reasonable number of acoustic signals of short duration that can be readily and rapidly identified. This limitation is avoided if the decoding of the acoustic signal enables the listener to recover the articulatory events that produced it, since perception then becomes linked to a system of physiological coordinates more richly multidimensional—hence more distinctive—than the acoustic (and auditory) signal.

Having said of the speech code that it seems particularly well designed to circumvent the shortcomings of the ear, we should consider whether its ac-

[33] The phonological constraints may, in fact, play an essential part in making the decoding operation fast and relatively free of error. Since the set of phonemes and phoneme sequences that is used in any particular language is far smaller than the possible set of feature combinations—smaller even than the combinations that are physiologically realizable—information about the allowable combinations of features could be used in the decoding process to reestablish temporal relationships that may have been blurred in articulation (see Footnote 27). Such a "recutting" operation would ease the requirements on precision of articulation and so allow faster communication; also, it would make unambiguous the segmentation into phonemes, thereby qualifying them as units of immediate perception. One may speculate, further, that the serial string of reconstituted phonemes is useful—perhaps essential—in the next operation of speech reception, namely, gaining immediate access to an inventory of thousands of syllable- or word-size units.

[34] Fant (1962a) makes essentially the same point:

> The rules relating speech waves to speech production are in general complex since one articulation parameter, e.g., tongue height, affects several of the parameters of the spectrogram. Conversely, each of the parameters of the spectrogram is generally influenced by several articulatory variables [p. 5].

complishments stop there. As we pointed out earlier, these shortcomings do not apply to the eye, for example, and we do indeed find that the best way to communicate language in the visual mode is by means of an alphabet or cipher, not a code. But we also noted that the acoustic code is more easily and naturally perceived than is the optical alphabet. Perhaps this is due primarily to the special speech decoder, whose existence we assumed for the conversion of sound to phoneme. We would suggest an additional possibility: the operations that occur in the speech decoder—including, in particular, the interdependence of perceptual and productive processes—may be in some sense similar to those that take place at other levels of grammar. If so, there would be a special compatibility between the perception of speech sounds and the comprehension of language at higher stages. This might help to explain why, so far from being merely one way of conveying language, the sounds of speech are, instead, its common and privileged carriers.

REFERENCES

ABRAMSON, A. S. Identification and discrimination of phonemic tones. *Journal of the Acoustical Society of America*, 1961, 33, 842. (Abstract)

ABRAMSON, A. S., & LISKER, L. Voice onset time in stop consonants: Acoustic analysis and synthesis. *Reports of the Fifth International Congress on Acoustics*, 1965, Ia, Paper A51.

BASTIAN, J., & ABRAMSON, A. S. Identification and discrimination of phonemic vowel duration. *Journal of the Acoustical Society of America*, 1962, 34, 743. (Abstract)

BASTIAN, J., DELATTRE, P. C., & LIBERMAN, A. M. Silent interval as a cue for the distinction between stops and semivowels in medial position. *Journal of the Acoustical Society of America*, 1959, 31, 1568. (Abstract)

BASTIAN, J., EIMAS, P. D., & LIBERMAN, A. M. Identification and discrimination of a phonemic contrast induced by silent interval. *Journal of the Acoustical Society of America*, 1961, 33, 842. (Abstract)

BRADY, P. T., HOUSE, A. S., & STEVENS, K. N. Perception of sounds characterized by a rapidly changing resonant frequency. *Journal of the Acoustical Society of America*, 1961, 33, 1337–1362.

BROADBENT, D. E. The role of auditory localization in attention and memory span. *Journal of Experimental Psychology*, 1954, 47, 191–196.

BROADBENT, D. E., & GREGORY, M. Accuracy of recognition for speech presented to the right and left ears. *Quarterly Journal of Experimental Psychology*, 1964, 16, 359–360.

BRYDEN, M. P. Ear preference in auditory perception. *Journal of Experimental Psychology*, 1963, 65, 103–105.

CHANEY, R. B., & WEBSTER, J. C. Information in certain multidimensional acoustic signals. Report No. 1339, 1965. United States Navy Electronics Laboratory Reports, San Diego, Calif.

CHISTOVICH, L. A. Classification of rapidly repeated speech sounds. *Akusticheskii Zhurnal*, 1960, 6, 392–398. (Trans. in *Soviet Physics-Acoustics*, New York, 1961, 6, 393–398).

CHISTOVICH, L. A., KLAAS, Y. A., & KUZ'MIN, Y. I. The process of speech sound discrimination. *Voprosy Psikhologii*, 1962, 8, 26–39. (Research Library, Air Force Cambridge Research Laboratories, Bedford, Mass. TT-64-13064 35-P)

CHOMSKY, N. Current issues in linguistic theory. In J. A. Fodor & J. J. Katz (Eds.), *The structure of language*. Englewood Cliffs, N. J.: Prentice Hall, 1964. Pp. 50–118.

COFFEY, J. L. The development and evaluation of the Battelle Aural Reading Device. *Proceedings of the International Congress on Technology and Blindness I*. New York: American Foundation for the Blind, 1963. Pp. 343–360.

COOPER, F. S. Research on reading machines for the blind. In P. A. Zahl (Ed.), *Blindness: Modern approaches to the unseen environment*. Princeton, N. J.: Princeton University Press, 1950. Pp. 512–543. (a)

COOPER, F. S. Spectrum analysis. *Journal of the Acoustical Society of America*, 1950, 22, 761–762. (b)

COOPER, F. S. Some instrumental aids to research on speech. *Report on the Fourth Annual Round Table Meeting on Lin-*

guistics and Language Teaching. Monograph Series No. 3. Washington: Georgetown University Press, 1953. Pp. 46–53.

COOPER, F. S. Research techniques and instrumentation: EMG. *American Speech and Hearing Association Reports,* 1965, 1, 153–158.

COOPER, F. S. Describing the speech process in motor command terms. *Journal of the Acoustical Society of America,* 1966, 39, 1221 (Abstract) (*Status Report of Speech Research, Haskins Laboratories, SR-5/6,* 1966, 2.1–2.27— text.)

COOPER, F. S., DELATTRE, P. C., LIBERMAN, A. M., BORST, J., & GERSTMAN, L. J. Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America,* 1952, 24, 597–606.

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. *Proceedings of the National Academy of Sciences,* 1951, 37, 318–325.

COOPER, F. S., LIBERMAN, A. M., HARRIS, K. S., & GRUBB, P. M. Some input-output relations observed in experiments on the perception of speech. *Proceedings of the Second International Congress of Cybernetics, 1958.* Namur, Belgium: Association Internationale de Cybernetique. Pp. 930–941.

COOPER, F. S., LIBERMAN, A. M., LISKER, L., & GAITENBY, J. Speech synthesis by rules. *Proceedings of the Speech Communication Seminar,* Stockholm, 1962. Stockholm: Royal Institute of Technology, 1963. F2.

CROSS, D. V., LANE, H. L., & SHEPPARD, W. C. Identification and discrimination functions for a visual continuum and their relation to the motor theory of speech perception. *Journal of Experimental Psychology,* 1965, 70, 63–74.

DELATTRE, P. C. Les indices acoustiques de la parole: Premier rapport. *Phonetica,* 1958, 2, 108–118, 226–251.

DELATTRE, P. C. Le jeu des transitions des formants et la perception des consonnes. *Proceedings of the Fourth International Congress of Phonetic Sciences, Helsinki, 1961.* s-Gravenhage: Mouton, 1962. Pp. 407–417.

DELATTRE, P. C., LIBERMAN, A. M., & COOPER, F. S. Voyelles synthétiques à deux formants et voyelles cardinales. *Le Maître Phonétique,* 1951, 96, 30–37.

DELATTRE, P. C., LIBERMAN, A. M., & COOPER, F. S. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America,* 1955, 27, 769–773.

DELATTRE, P. C., LIBERMAN, A. M., & COOPER, F. S. Formant transitions and loci as acoustic correlates of place of articulation in American fricatives. *Studia Linguistica,* 1964, 18, 104–121.

DELATTRE, P. C., LIBERMAN, A. M., COOPER, F. S., & GERSTMAN, L. J. An experimental study of the acoustic determinants of vowel color: Observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word,* 1952, 8, 195–210.

DENES, P. B. On the statistics of spoken English. *Journal of the Acoustical Society of America,* 1963, 35, 892–904.

EIMAS, P. D. The relation between identification and discrimination along speech and non-speech continua. *Language and Speech,* 1963, 6, 206–217.

FANT, C. G. M. *Acoustic theory of speech production.* 's-Gravenhage: Mouton, 1960.

FANT, C. G. M. Descriptive analysis of the acoustic aspects of speech. *Logos,* 1962, 5, 3–17. (a)

FANT, C. G. M. Sound spectrography. *Proceedings IV International Congress on Phonetic Sciences, Helsinki, 1961.* 's-Gravenhage: Mouton, 1962. Pp. 14–33. (b)

FANT, C. G. M. Theory of distinctive features. *Speech Transmission Laboratory Quarterly Progress and Status Report,* Royal Institute of Technology (KTH), Stockholm, January 15, 1967. Pp. 1–14.

FREIBERGER, J., & MURPHY, E. F. Reading machines for the blind. *IRE Professional Group on Human Factors in Electronics,* 1961, HFE-2, 8–19.

FROMKIN, V. A. Neuro-muscular specification of linguistic units. *Language and Speech,* 1966, 9, 170–199.

FROMKIN, V. A., & LADEFOGED, P. Electromyography in speech research. *Phonetica,* 1966, 15, 219–242.

FRY, D. B., ABRAMSON, A. S., EIMAS, P. D., & LIBERMAN, A. M. The identification and discrimination of synthetic vowels. *Language and Speech,* 1962, 5, 171–189.

GALUNOV, V. I., & CHISTOVICH, L. A. Relationship of motor theory to the general problem of speech recognition (review). *Akusticheskii Zhurnal,* 1965, 11, 417–426. (Trans. in *Soviet Physics-Acoustics,* New York, 1966, 11, 357–365.)

GELB, I. J. *A study of writing.* Chicago: University of Chicago Press, 1963.

GRIFFITH, B. C. A study of the relation between phoneme labeling and discrimina-

bility in the perception of synthetic stop consonants. Unpublished doctoral dissertation, University of Connecticut, 1958.

HADDING-KOCH, K., & STUDDERT-KENNEDY, M. An experimental study of some intonation contours. *Phonetica*, 1964, 11, 175–185. (a)

HADDING-KOCH, K., & STUDDERT-KENNEDY, M. Intonation contours evaluated by American and Swedish test subjects. *Proceedings of the Fifth International Congress of Phonetic Sciences*, Munster, August 1964. (b)

HALLE, M. On the bases of phonology. In J. A. Fodor & J. J. Katz (Eds.), *The structure of language*. Englewood Cliffs, N. J.: Prentice-Hall, 1964. Pp. 324–333.

HALLE, M., & STEVENS, K. N. Speech recognition: A model and a program for research. *IRE Transactions on Information Theory IT-8*, 1962, 2, 155–159.

HARRIS, C. M. A study of the building blocks in speech. *Journal of the Acoustical Society of America*, 1953, 25, 962–969.

HARRIS, K. S. Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech*, 1958, 1, 1–7.

HARRIS, K. S. Behavior of the tongue in the production of some alveolar consonants. *Journal of the Acoustical Society of America*, 1963, 35, 784. (Abstract)

HARRIS, K. S., BASTIAN, J., & LIBERMAN, A. M. Mimicry and the perception of a phonemic contrast induced by silent interval: Electromyographic and acoustic measures. *Journal of the Acoustical Society of America*, 1961, 33, 842. (Abstract)

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. *Journal of the Acoustical Society of America*, 1958, 30, 122–126.

HARRIS, K. S., HUNTINGTON, D. A., & SHOLES, G. N. Coarticulation of some disyllabic utterances measured by electromyographic techniques. *Journal of the Acoustical Society of America*, 1966, 39, 1219. (Abstract)

HARRIS, K. S., LYSAUGHT, G., & SCHVEY, M. M. Some aspects of the production of oral and nasal labial stops. *Language and Speech*, 1965, 8, 135–147.

HOCKETT, C. F. *A manual of phonology*. Baltimore: Waverly Press, 1955.

HUGHES, G. W., & HALLE, M. Spectral properties of fricative consonants. *Journal of the Acoustical Society of America*, 1956, 28, 303–310.

JAKOBSON, R., FANT, G., & HALLE, M. *Preliminaries to speech analysis. The distinctive features and their correlates*. Technical Report No. 13, 1952, Acoustics Laboratory, M.I.T. (Republished, Cambridge, Mass.: M.I.T. Press, 1963.)

JAKOBSON, R., & HALLE, M. *Fundamentals of language*. 's-Gravenhage: Mouton, 1956.

KIMURA, D. Cerebral dominance and perception of verbal stimuli. *Canadian Journal of Psychology*, 1961, 15, 166–171.

KIMURA, D. Left-right differences in the perception of melodies. *Quarterly Journal of Experimental Psychology*, 1964, 16, 355–358.

KIMURA, D. Functional asymmetry of the brain in dichotic listening. *Cortex*, 1967, 3, in press.

KOZHEVNIKOV, V. A., & CHISTOVICH, L. A. *Rech' Artikuliatsia i vospriiatie*. Moscow-Leningrad, 1965. (Trans. in *Speech: Articulation and perception*. Washington: Joint Publications Research Service, 1966, 30, 543.)

LADEFOGED, P. The perception of speech. In *mechanization of thought processes*, 1959. London: H. M. Stationery Office. Pp. 397–409.

LADEFOGED, P., & MCKINNEY, N. P. Loudness, sound pressure, and sub-glottal pressure in speech. *Journal of the Acoustical Society of America*, 1963, 35, 454–460.

LANE, H. Motor theory of speech perception: A critical review. *Psychological Review*, 1965, 72, 275–309.

LIBERMAN, A. M. Some results of research on speech perception. *Journal of the Acoustical Society of America*, 1957, 29, 117–123.

LIBERMAN, A. M., COOPER, F. S., HARRIS, K. S., & MACNEILAGE, P. F. A motor theory of speech perception. *Proceedings of the Speech Communication Seminar*, Stockholm, 1962. Stockholm: Royal Institute of Technology, 1963, D3.

LIBERMAN, A. M., COOPER, F. S., HARRIS, K. S., MACNEILAGE, P. F., & STUDDERT-KENNEDY, M. Some observations on a model for speech perception. *Proceedings of the AFCRL Symposium on Models for the Perception of Speech and Visual Form*, Boston, November 1964. Cambridge: Massachusetts Institute of Technology Press, in press.

LIBERMAN, A. M., COOPER, F. S., STUDDERT-KENNEDY, M., HARRIS, K. S., & SHANKWEILER, D. P. Some observations on the efficiency of speech sounds. Paper presented at the XVIII International Congress

of Psychology, Moscow, August 1966. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikations-forschung,* 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. *American Journal of Psychology,* 1952, 65, 497–516.

LIBERMAN, A. M., DELATTRE, P. C., & COOPER, F. S. Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech,* 1958, 1, 153–167.

LIBERMAN, A. M., DELATTRE, P. C., COOPER, F. S., & GERSTMAN, L. J. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs,* 1954, 68(8, Whole No. 379).

LIBERMAN, A. M., DELATTRE, P. C., GERSTMAN, L. J., & COOPER, F. S. Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology,* 1956, 52, 127–137.

LIBERMAN, A. M., HARRIS, K. S., EIMAS, P. D., LISKER, L., & BASTIAN, J. An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. *Language and Speech,* 1961, 4, 175–195.

LIBERMAN, A. M., HARRIS, K. S., HOFFMAN, H. S., & GRIFFITH, B. C. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology,* 1957, 54, 358–368.

LIBERMAN, A. M., HARRIS, K. S., KINNEY, J. A., & LANE, H. The discrimination of relative onset time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology,* 1961, 61, 379–388.

LIBERMAN, A. M., INGEMANN, F., LISKER, L., DELATTRE, P. C., & COOPER, F. S. Minimal rules for synthesizing speech. *Journal of the Acoustical Society of America,* 1959, 31, 1490–1499.

LIEBERMAN, P. *Intonation, perception and language.* Cambridge: Massachusetts Institute of Technology Press, 1967.

LINDBLOM, B. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America,* 1963, 35, 1773–1781.

LINDBLOM, B. Articulatory activity in vowels. *Journal of the Acoustical Society of America,* 1964, 36, 1038. (Abstract)

LINDBLOM, B., & STUDDERT-KENNEDY, M. On the role of formant transitions in vowel recognition. *Speech transmission laboratory quarterly progress and status report,* Royal Institute of Technology (KTH), Stockholm, in press. (Also Status report of speech research. Haskins Laboratories, in press.)

LISKER, L. Closure duration and the voiced-voiceless distinction in English. *Language,* 1957, 33, 42–49. (a)

LISKER, L. Minimal cues for separating /w,r,l,j/ in introvocalic production. *Word,* 1957, 13, 257–267. (b)

LISKER L,. Anatomy of unstressed syllables. *Journal of the Acoustical Society of America,* 1958, 30, 682. (Abstract)

LISKER, L., & ABRAMSON, A. S. A cross-language study of voicing in initial stops: Acoustical measurements. *Word,* 1964, 20, 384–422. (a)

LISKER, L., & ABRAMSON, A. S. Stop categories and voice onset time. *Proceedings of the Fifth International Congress of Phonetic Sciences,* Munster, August, 1964. (b)

LISKER, L., COOPER, F. S., & LIBERMAN, A. M. The uses of experiment in language description. *Word,* 1962, 18, 82–106.

MACKAY, D. M. Mindlike behavior in artefacts. *British Journal for the Philosophy of Science,* 1951, 2, 105–121.

MACNEILAGE, P. F. Electromyographic and acoustic study of the production of certain final clusters. *Journal of the Acoustical Society of America,* 1963, 35, 461–463.

MACNEILAGE, P. F., DECLERK, J. L., & SILVERMAN, S. I. Some relations between articulator movement and motor control in consonant-vowel-consonant monosyllables. *Journal of the Acoustical Society of America,* 1966, 40, 1272. (Abstract)

MALECOT, A. Acoustic cues for nasal consonants. *Language,* 1956, 32, 274–284.

MILLER, G. A. The magical number seven, plus or minus two, or, some limits on our capacity for processing information. *Psychological Review,* 1956, 63, 81–96. (a)

MILLER, G. A. The perception of speech. In M. Halle (Ed.), *For Roman Jakobson.* 's-Gravenhage: Mouton, 1956. Pp. 353–359. (b)

MILLER, G. A., & TAYLOR, W. G. The perception of repeated bursts of noise. *Journal of the Acoustical Society of America,* 1948, 20, 171–182.

NELSON, P. G., ERULKAR, S. D., & BRYAN, S. S. Responses of units of the inferior-colliculus to time-varying acoustic stimuli. *Journal of Neurophysiology,* 1966, 29, 834–860.

NYE, P. W. Aural recognition time for multidimensional signals. *Nature*, 1962, 196, 1282–1283.

NYE, P. W. Reading aids for blind people —a survey of progress with the technological and human problems. *Medical Electronics and Biological Engineering*, 1964, 2, 247–264.

NYE, P. W. An investigation of audio outputs for a reading machine. February, 1965. Autonomics Division, National Physical Laboratory, Teddington, England.

O'CONNOR, J. D., GERSTMAN, L. J., LIBERMAN, A. M., DELATTRE, P. C., & COOPER, F. S. Acoustic cues for the perception of initial /w,j,r,l/ in English. *Word*, 1957, 13, 25–43.

OHMAN, S. E. G. Numerical model for coarticulation, using a computer-simulated vocal tract. *Journal of the Acoustical Society of America*, 1964, 36, 1038. (Abstract)

OHMAN, S. E. G. Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, 1966, 39, 151–168.

ORR, D. B., FRIEDMAN, H. L., & WILLIAMS, J. C. C. Trainability of listening comprehension of speeded discourse. *Journal of Educational Psychology*, 1965, 56, 148–156.

PETERSON, G. E., & SIVERTSEN, E. Objectives and techniques of speech synthesis. *Language and Speech*, 1960, 3, 84–95.

PETERSON, G. E., WANG, W. S.-Y., & SIVERTSEN, E. Segmentation techniques in speech synthesis. *Journal of the Acoustical Society of America*, 1958, 30, 739–742.

POLLACK, I. The information of elementary auditory displays. *Journal of the Acoustical Society of America*, 1952, 24, 745–749.

POLLACK, I., & FICKS, L. Information of elementary multidimensional auditory displays. *Journal of the Acoustical Society of America*, 1954, 26, 155–158.

POTTER, R. K., KOPP, G. A., & GREEN, H. C. *Visible speech*. New York: Van Nostrand, 1947.

SCHATZ, C. The role of context in the perception of stops. *Language*, 1954, 30, 47–56.

SHANKWEILER, D., & STUDDERT-KENNEDY, M. An analysis of perceptual confusions in identification of dichotically presented CVC syllables. *Journal of the Acoustical Society of America*, 1967, in press. (Abstract) (a)

SHANKWEILER, D., & STUDDERT-KENNEDY, M. Identification of consonants and vowels presented to left and right ears. *Quarterly Journal of Experimental Psychology*, 1967, 19, 59–63. (b)

SHEARME, J. N., & HOLMES, J. N. An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1-formant 2 plane. In *Proceedings of the Fourth International Congress of Phonetic Sciences*, Helsinki, 1961. 's-Gravenhage: Mouton, 1962. Pp. 234–240.

SIVERTSEN, E. Segment inventories for speech synthesis. *Language and Speech*, 1961, 4, 27–61.

STEVENS, K. N. Toward a model for speech recognition. *Journal of the Acoustical Society of America*, 1960, 32, 47–55.

STEVENS, K. N. On the relations between speech movements and speech perception. Paper presented at the meeting of the XVIII International Congress of Psychology, Moscow, August, 1966. (*Zeitschrift fur Phonetik, Sprachwissenschaft und Kommunikations-forschung*, in press.)

STEVENS, K. N., & HALLE, M. Remarks on analysis by synthesis and distinctive features. *Proceedings of the AFCRL Symposium on Models for the Perception of Speech and Visual Form*, Boston, November 1964. Cambridge: M.I.T. Press, in press.

STEVENS, K. N., & HOUSE, A. S. Studies of formant transitions using a vocal tract analog. *Journal of the Acoustical Society of America*, 1956, 28, 578–585.

STEVENS, K. N., & HOUSE, A. S. Perturbation of vowel articulations by consonantal context: An acoustical study. *Journal of Speech and Hearing Research*, 1963, 6, 111–128.

STEVENS, K. N., & HOUSE, A. S. Speech perception. In J. Tobias & E. Schubert (Eds.), *Foundations of modern auditory theory*. New York: Academic Press, in press.

STEVENS, K. N., OHMAN, S. E. G., & LIBERMAN, A. M. Identification and discrimination of rounded and unrounded vowels. *Journal of the Acoustical Society of America*, 1963, 35, 1900. (Abstract)

STEVENS, K. N., OHMAN, S. E. G., STUDDERT-KENNEDY, M., & LIBERMAN, A. M. Cross-linguistic study of vowel discrimination. *Journal of the Acoustical Society of America*, 1964, 36, 1989. (Abstract)

STUDDERT-KENNEDY, M., & COOPER, F. S. High-performance reading machines for the blind; psychological problems, technological problems, and status. Paper pre-

sented at the meeting of St. Dunstan's International Conference on Sensory Devices for the Blind, London, June 1966.

STUDDERT-KENNEDY, M., & LIBERMAN, A. M. Psychological considerations in the design of auditory displays for reading machines. *Proceedings of the International Congress on Technology and Blindness,* 1963. Pp. 289–304.

STUDDERT-KENNEDY, M., LIBERMAN, A. M., & STEVENS, K. N. Reaction time to synthetic stop consonants and vowels at phoneme centers and at phoneme boundaries. *Journal of the Acoustical Society of America,* 1963, 35, 1900. (Abstract)

STUDDERT-KENNEDY, M., LIBERMAN, A. M., & STEVENS, K. N. Reaction time during the discrimination of synthetic stop consonants. *Journal of the Acoustical Society of America,* 1964, 36, 1989. (Abstract)

VON HOLST, E., & MITTELSTADT, H. Das reafferenzprinzip. *Naturwissenschaft,* 1950, 37, 464–476.

WHETNALL, E., & FRY, D. B. *The deaf child.* London: Heinemann, 1964.

WHITFIELD, I. C., & EVANS, E. F. Responses of auditory cortical neurons to stimuli of changing frequency. *Journal of Neurophysiology,* 1965, 28, 655–672.

WICKELGREN, W. A. Distinctive features and errors in short-term memory for English consonants. *Journal of the Acoustical Society of America,* 1966, 39, 388–398.

(Received June 19, 1967)