Speech Perception*

Michael Studdert-Kennedy
Queens College and the Graduate Center of the City University of New York, and Haskins Laboratories, New Haven, Conn.

> "The understanding of speech involves essentially the same problems as the production of speech.... The processes...have too much in common to depend on wholly different mechanisms" (Lashley, 1951:120).

INTRODUCTION

We can listen to speech at many levels. We can listen selectively for meaning, sentence structure, words, phones, intonation, chatter, or even, at a distance, Auden's "high, thin, rare, continuous hum of the self-absorbed." This paper is concerned solely with phonetic perception, the transformation of a more-or-less continuous acoustic signal into what may be transcribed as a sequence of discrete, phonetic symbols. The study of speech perception, in this sense, has in recent years begun to adopt the aims, and often the methods, of the information-processing models of cognitive psychology which have proved fruitful in the study of vision (Neisser, 1967; Haber, 1969; Reed, 1973). The underlying assumption is that perception has a time-course, during which information in the sensory array is "transformed, reduced, elaborated" (Neisser, 1967:4) and brought into contact with long-term memory (recognized). The experimental aim is to intervene in this process (either directly or by inference) at various points between sensory input and final percept, in order to discover what transformations the original information has undergone. The ultimate objective is to describe the process in terms specific enough for neurophysiologists to search for neural correlates.

Let us begin by considering how speech perception differs from general auditory perception. It does so in both stimulus and percept. First, the sounds of speech constitute a distinctive class, drawn from the set of sounds that can be produced by the human vocal mechanism. They can be described, to an approximation, as the output of a filter excited by an independent source. The source is the flow of air from the lungs, modulated at the glottis to produce a quasi-periodic sound, or above the glottis to produce a noisy turbulence. The filter

---

*Chapter prepared for Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass (in press).

Acknowledgment: I thank Alvin Liberman, Ignatius Mattingly, and Donald Shankweiler for their valuable comments and criticism, David Pisoni for much fruitful conversation and for drawing my attention to the work of Eleanor Rosch.

[HASKINS LABORATORIES: Status Report on Speech Research SR-39/40 (1974)]

is the supralaryngeal vocal tract, whose varying configurations give rise to varying resonances (formants). The resulting sound wave may be displayed as an oscillogram or, after spectral analysis, as a spectrogram. It is important to bear in mind that the spectrogram does not display the sensory input, but a transformation of that input, often presumed to represent the output at an early stage of auditory analysis. [For accounts of the speech signal and its mechanisms of production, see Fant, 1960; Stevens and House, 1972; Kent (in Lass, in press); Babcock (in Lass, in press).]

Here our main concern is to stress functional differences between speech and nonspeech acoustic structure in perception. Speech does not lie at one end of an auditory (psychological) continuum which we can approach by closer and closer acoustic (physical) approximation. The sounds of speech are distinctive. They form a set of "natural categories" similar to those described by Rosch (1973). She studied form and color perception among the Dani, a Stone-Age people of New Guinea, whose language contains "only two color terms which divide the color space on the basis of brightness rather than hue" (p. 331), and no words for the Gestalt "good forms" of square, circle, and equilateral triangle. She found that her subjects were significantly faster in learning arbitrary names for the four primary hue points than for other hues, and for the three "good forms" of Gestalt psychology than for others. She points to the possible physiological underpinnings of these "natural prototypes." Her work is reminiscent of a study by House, Stevens, Sandel, and Arnold (1962). They constructed several ensembles of sounds along an acoustic continuum from clearly nonspeech to speech. The time taken by subjects to learn associations between sounds and buttons on a box was least for the speech ensemble, and did not decrease with the acoustic approximation of the ensembles to speech. In short, a signal is heard as either speech or nonspeech, and once heard as speech, elicits characteristic perceptual functions that we shall discuss below.

The second peculiarity of speech perception, as we are viewing it, is in perceptual response. The final percept is a phonetic name, and the name (unlike those for "natural categories" of form and color) bears a necessary, rather than an arbitrary, relation to the signal. In other words, speech sounds "name themselves." Notice that this is not true of the visual counterparts of phonetic entities: the forms of the alphabet are arbitrary, and we are not concerned that, for example, the same visual symbol, P, stands for /p/ in the Roman alphabet, for /r/ in the Cyrillic. Nothing comparable occurs in the speech system: the acoustic correlates of [p] or [r] can be perceived as nothing other than [p] or [r]. A central problem for the student of speech perception is to define the nature of this inevitable percept.

LEVELS OF PROCESSING

Implicit in the foregoing is a distinction between auditory and phonetic perception. As a basis for future discussion, we will lay out a rough conceptual model of the perceptual process (cf. Studdert-Kennedy, 1974; also Day, 1968, 1970). We can conceive the signals of running speech as climbing a hierarchy through at least these successive transformations: (1) auditory, (2) phonetic, (3) phonological, (4) lexical, syntactic, and semantic. The levels must be at least partially successive, to preserve aspects of temporal order in the signal. They must also be at least partially parallel, to permit higher decisions to guide and correct lower decisions [cf. Turvey's (1973) discussion of peripheral and central processes in vision].

The auditory level is itself a series of processes (Fourcin, 1972). Early work (Licklider and Miller, 1951) showed that the speech waveform could be vastly distorted without serious loss of intelligibility. Spectrographic analysis (Potter, Kopp, and Green, 1947; Joos, 1948) and speech synthesis (Liberman, 1957) showed that patterns of speech important to its perception lay not in its wave-form, but in its time-varying spectrum as revealed by the spectrogram. We may imagine, therefore, an early stage of the auditory display, soon after cochlear analysis, as the neural correlate of a spectrogram. Notice in Figure 1: regions of high energy concentration (formants, usually labeled from the bottom up as F1, F2, F3); different formant patterns associated with the vowels of read and book, for example; intervals of silence during stop consonant closure; a sharp scatter of energy (noise burst) upon release of the voiceless stop in to, and fainter bursts following release of the voiced stops in began; rapid formant movements (transitions) as articulators move into and out of vowels; a nasal formant (between F1 and F2) at the end of began; a broad band of noise associated with the fricative of she; and finally, regular vertical striations, reflecting a series of glottal pulses, from which fundamental frequency can be derived. A later, perhaps cortical, stage of auditory analysis may entail detection of just such features in the spectrographic display. Whether there are acoustic feature analyzers specially tuned to speech is an open question that we consider below. In any event, the signal has not yet been transformed into the message, and may indeed have passed through the same processes as any other auditory input.

The phonetic level is abstract in the sense that its output is a set of properties not inherent in the signal. They derive from the auditory display by processes that must be peculiar to humans, since they can only be defined by reference to the human vocal mechanism. These properties correspond to the linguistic entities of distinctive feature (Jakobson, Fant, and Halle, 1963) and phoneme. For the psychological reality of these units, there is ample evidence, discussed below. There is also evidence that extraction of these units from the auditory display calls upon specialized decoding mechanisms (Studdert-Kennedy and Shankweiler, 1970). In any event, the output from this level is now speech, although much variability remains to be resolved.

Resolution is accomplished at the phonological level, where processes peculiar to the listener's language are engaged. Here, the listener merges phonetic variations that have no function in his language, treating, for example, both the initial segment of [p^hIt] and the second segment of [spIt] as instances of /p/. Here, too, the listener may shift distinctions across segments, interpreting English vowel length before a final stop, for example, as a phonetic cue to the voicing value of the stop. In short, this is the level at which phonetic variability is transformed into phonological system. Of course, for untrained listeners all of the time, and for phoneticians most of the time, the distinction between phonetic and phonological levels has little import. Listeners usually hear speech in terms of the categories of their native language (e.g., Lotz, Abramson, Gerstman, Ingemann, and Nemser, 1960; Scholes, 1968; Day, 1968, 1969, 1970a, 1970b). However, since they may learn (at some pain) to make phonetic distinctions, we must assume that phonetic information is available in the system, though unattended in normal listening. Most of the research to be discussed has concerned itself with a single language and has not distinguished between phonetic and phonological levels. (For extended discussion of experimental paradigms that serve to reflect several levels of processing from auditory to phonological, see Cutting, 1973, in press-a.)

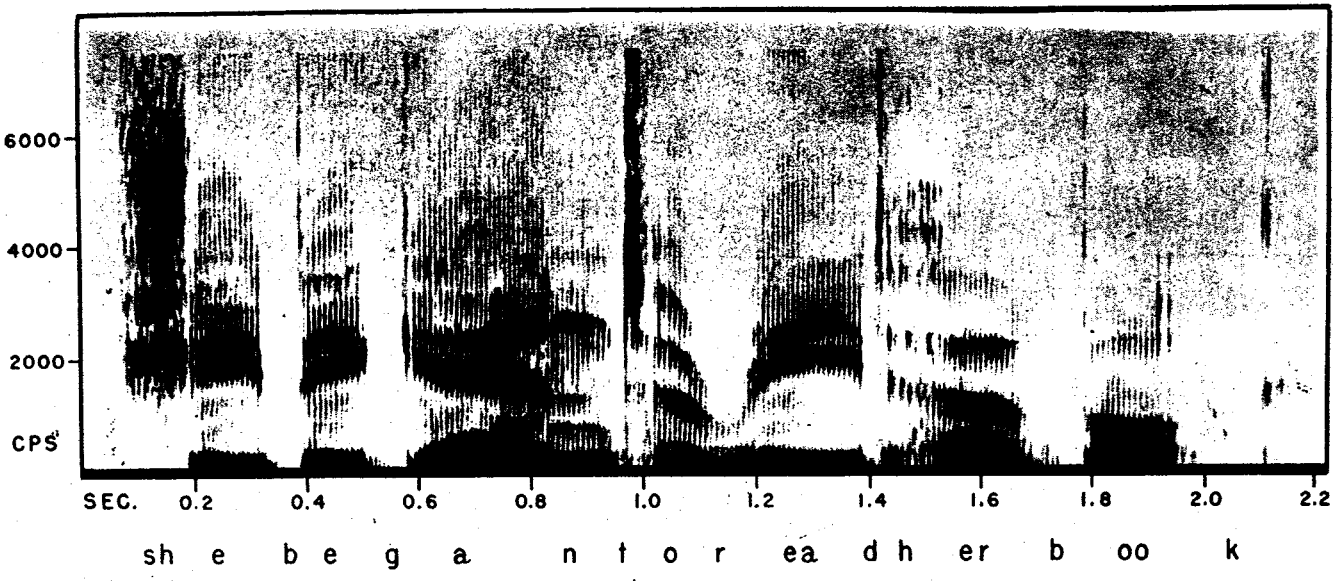

Figure 1: Spectrogram of a natural utterance: She began to read her book. Frequency is plotted against time, with relative intensity represented by degree of blackness. For fuller description, see text.

The upper levels of lexical, syntactic, and semantic processing complete the normal process of speech perception. There is good evidence that outputs from these levels can affect phonological and phonetic perception. Miller, Heise, and Lichten (1951), for example, showed that words were more intelligible in a sentence than in a list. Pollack and Pickett (1963) and Lieberman (1963) found that words excised from sentences and presented to listeners without syntactic and semantic context were often not recognized. Several writers (e.g., Jones, 1948; Chomsky and Miller, 1963; Chomsky and Halle, 1968) have placed a heavy load on the syntactic structure and semantic content of an utterance in their accounts of speech perception. However, while these higher levels may serve to "clean" the message when phonetic lapse is slight (cf. Warren, 1970; Warren and Obusek, 1971; Cole, 1973a), and may even be deliberately brought to bear while conversing with a foreigner in a railway tunnel, their control is not sufficient to disguise all slips of the tongue (cf. Fromkin, 1971). Unambiguous perception is possible in spite of context, and, as will be seen, presents sufficient theoretical problems. Bearing in mind our primary distinction between auditory and phonetic levels, we turn now to a brief review of acoustic cues and of the problems that emerge for perceptual theory.

THE ACOUSTIC CUES

Many of the acoustic cues to the phonetic message have been uncovered over the past twenty years by the complementary processes of analysis and synthesis. Spectrographic analysis of natural speech suggests likely candidates, such as formant frequency, formant movement, silent interval, or burst of noise. Synthesis then permits these "minimal cues" (Liberman, 1957) to be checked for perceptual validity. Results of this work are described elsewhere (Liberman, 1957; Fant, 1960, 1968; Mattingly, 1968, 1974; Flanagan, 1972; Stevens and House, 1972). Here, we do no more than summarize its outcome and frame the problems it raises for speech perception.

The problems are those of invariance and segmentation. The speech signal carries neither invariant acoustic cues nor isolable segments that reliably correspond to the invariant segments of linguistic analysis and perception. The speech signal can certainly be segmented. Fant (1968) and his colleagues have outlined a procedure for dividing the signal in both frequency and time, and have developed a terminology to describe its segments. But these do not correspond to the phonetic segments of distinctive feature or phoneme. There are exceptions: fricatives and stressed vowels, for example, may present stable and more-or-less isolable patterns. But, in general, as Fant (1962) has remarked, a single segment of sound contains information concerning several neighboring segments of the message, and a single segment of the message may draw upon several neighboring segments of sound. In short, the sounds of speech are not physically discrete, like letters of the alphabet, but rather are shingled into an intricate, continuously changing pattern (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967).

Whether the source of this shingled pattern is to be found in mechanical constraints, neuromuscular inertia, and temporal overlap of successive commands to the articulators (Öhman, 1967), or in elegantly controlled, yet variable responses to fixed articulatory instructions (MacNeilage, 1970), the result is not only a loss of segmentation, but also a loss of acoustic invariance. The cues to a given phonetic segment display enormous variability as a function of phonetic context, stress, and speaking rate (e.g., Kozhevnikov and Chistovich, 1965; Stevens, House, and Paul, 1966).

As a simple instance, consider the acoustic structure of a mirror image consonant-vowel-consonant (CVC) syllable such as [bæ b]. Experiments with synthetic speech have demonstrated the importance of second and third formant transitions as cues for distinguishing among labial, alveolar, and velar stops (Liberman, Delattre, Cooper, and Gerstman, 1954). Here, the two formants rise rapidly over the first 40 msec or so into the vowel, and then, after a relatively sustained formant pattern for, say, 200 msec, drop rapidly back to their starting points. The acoustic cues to initial and final allophones of [b] are mirror images, and, separated from the syllable, are heard as distinct nonspeech sounds. Experiments with tone glissandos matching such patterns in duration and frequency range reveal no psychoacoustic basis for the perceived phonetic identity (Klatt and Shattuck, 1973).

Similar discrepancies occur as a function of vowel context. Initial formant transitions in a CV syllable reflect the changing resonances of the vocal tract as the articulators move from consonant closure, or constriction, into a more open position for the following vowel. Since vowels are distinguished by the positions of their first two or three formant centers on the frequency scale (Delattre, Liberman, Cooper, and Gerstman, 1952; Peterson and Barney, 1952), consonantal approach varies with vowel: for example, both second and third formants fall in the syllable [dæ]; the second rises and the third falls in the syllable [de]. Yet listeners fail to detect these acoustic differences, and phonetic identity of the initial segments is preserved.

As a final example, consider vowels. Each stressed vowel, spoken in isolation, has its characteristic set of formant frequencies. However, in running speech, these values are seldom reached, particularly if speech is rapid and vowels unstressed (Lindblom, 1963). If vowel portions are excised from running speech and presented without their surrounding formant transitions, identifications shift (Fujimura and Ochiai, 1963). This suggests (as do the consonantal examples given above) that listeners track formants over at least a syllable in order to make their phonetic decisions. (For other examples of phonetic identity in face of acoustic variance, see Shearme and Holmes (1962), Lindblom (1963), Öhman (1966), Liberman et al. (1967), and Stevens and House (1972).)

A different class of acoustic variability is instanced by interspeaker variations. Here differences in acoustic quality can be clearly heard, but are disregarded in phonetic perception. Center frequencies of vowel formants vary widely among men, women, and children (Peterson and Barney, 1952), with the result that acoustically identical patterns may be judged phonetically distinct, while acoustically distinct patterns may be judged phonetically identical. "Normalization" probably cannot be accomplished by application of a simple scale factor (Peterson, 1961) because male-female formant ratios are not constant across the vowel quadrilateral (Fant, 1966).

A favored belief is that listeners judge vowels by reference to other vowels uttered by the same speaker. This notion originated with Joos (1948) and was tested by Ladefoged and Broadbent (1957). They demonstrated that the same synthetic vowel pattern could be judged differently, depending on the formant pattern of a precursor phrase. Gerstman (1968) developed an algorithm, derived from the formant frequencies of [i,a,u] for each speaker, that correctly identifies 97.5 percent of the Peterson and Barney (1952) vowels. And Lieberman (1973) claims that unless a listener has heard "calibrating signals," such as the vowels [i,a,u] or the glides [y] and [w], from which to assess the size of a

particular speaker's vocal tract, "it is impossible to assign a particular acoustic signal into the correct class" (p. 91).

However, an algorithm is not a perceptual model, and remarkably little is actually known in this area: there is a dearth of data on how listeners judge the varied vowel patterns of different speakers. Furthermore, the phenomenon of normalization is not confined to vowels. Fourcin (1968) demonstrated that a synthetic "whispered" syllable with a constant formant pattern could be heard as a token of [d] if preceded by a man's hallo, of [b] if preceded by a child's. Rand (1971) showed a similar systematic shift, without benefit of precursor, when formant frequencies of synthetic CV syllables were increased by 20 percent above the "male" base. Evidently, normalization can be accomplished within a syllable, presumably from information provided by formant structure and fundamental frequency (cf. Fujisaki and Nakamura, 1969). This is precisely what is suggested by recent work of Strange, Verbrugge, and Shankweiler (1974) and Verbrugge, Strange, and Shankweiler (1974). They find that a speaker's precursor vowels, whether [i,a,u] or [I,æ,ʌ], do little to reduce listener error in judging following vowels spoken by a panel of men, women, and children. Far more effective in reducing error is presentation of the vowel within a consonantal frame. Of course, formant reference is clearly involved in studies where consonantal context is held constant (Summerfield and Haggard, 1973). However, the results again suggest perceptual tracking of an entire syllable, and emphasize that invariant acoustic segments matching the invariants of perception are not readily found. [For a recent review of the normalization problem, see Shankweiler, Strange, and Verbrugge (in press).]

Nonetheless, the search for acoustic invariance has not been abandoned. A main reason for this is the obvious worth of some form of feature theory in linguistic description and, incidentally, in the description of listener behavior (see next section). Distinctive-feature theorists have always maintained that correlates of the features are to be found at every level of the speech process--articulatory, acoustic, auditory--(Jakobson and Halle, 1956; Jakobson, Fant, and Halle, 1963; Chomsky and Halle, 1968), and a good deal of current research is directed toward grounding features in acoustics and physiology (cf. Ladefoged, 1971a, 1971b; Lindblom, 1972).

Before giving examples, we should emphasize the redundancy of the speech signal. A given feature may be signaled by several different cues. Studies of synthetic speech have tended to emphasize "sufficient" cues and to disregard their interaction. Harris (1958) provides an exception, in her study of noise bands and formant transitions as cues to English fricatives. So, too, do Harris, Hoffman, Liberman, Delattre, and Cooper (1958) and Hoffman (1958), who examined the relative weights of second and third formant transitions in the perception of English voiced stops.

Finally, exceptions are also provided by Lisker and Abramson (1964, 1967, 1970, 1971) and by Abramson and Lisker (1965, 1970; see also Zlatin, 1974) in an extensive series of studies of voicing in many languages. Noting that voicing in initial stops may be cued by explosion energy, degree of aspiration, and first formant intensity, they sought a cover variable that would encompass all these cues. They found it in voice onset time (VOT), the interval between release of stop closure and the onset of laryngeal vibration. Figure 2 displays spectrograms of synthetic stops in which VOT is a sufficient cue for the distinction between [ba] and [pa]. Notice that VOT is not a simple variable, either articulatorily or acoustically: it refers to a temporal relation between

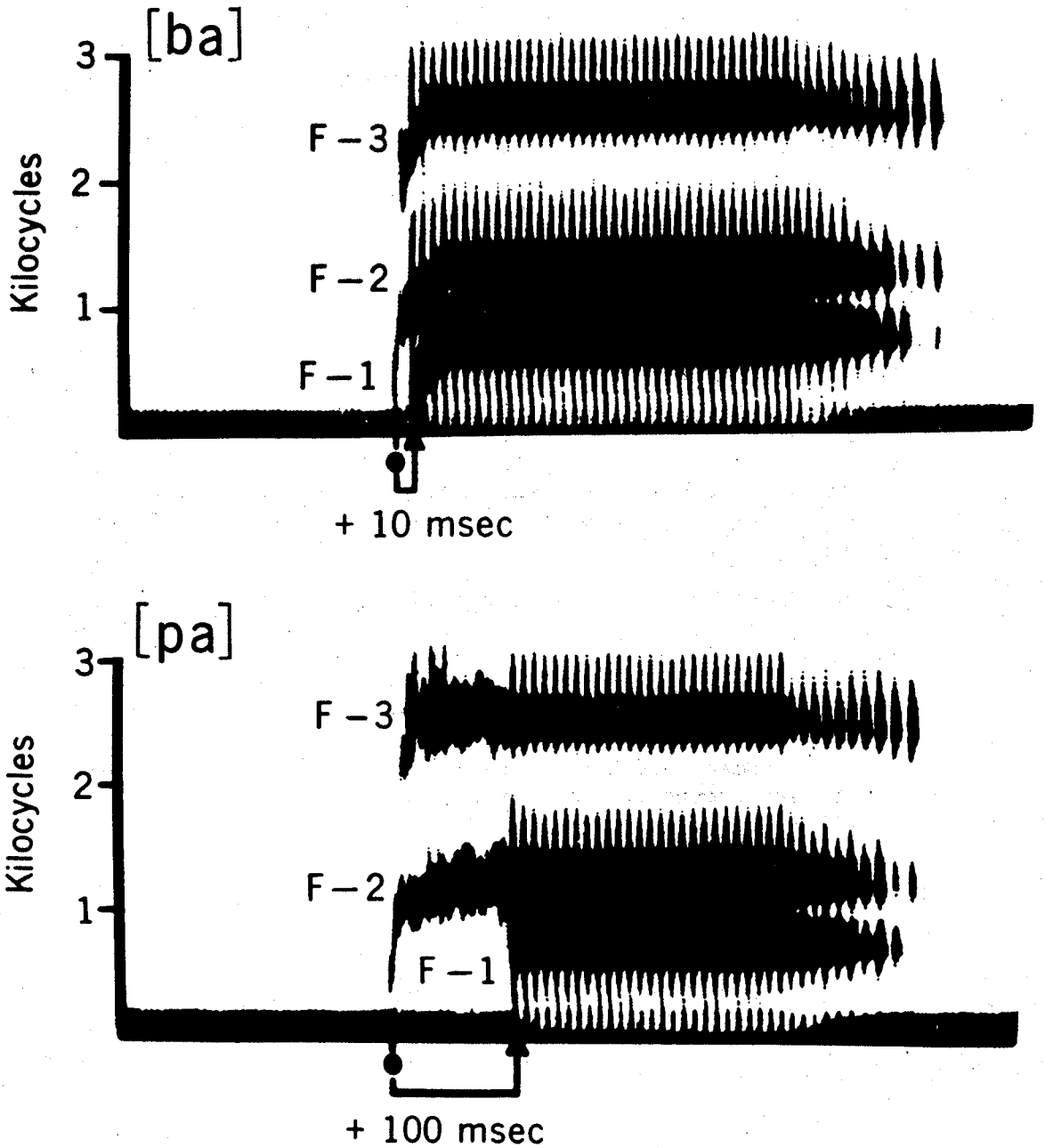

Figure 2: Spectrograms of synthetic syllables, [ba] and [pa]. The interval between release and voicing (vertical striations) (VOT) is 10 msec for [ba], 100 msec for [pa]. During this interval, F1 is absent and the regions of F2 and F3 are occupied by "aspirated" noise. [After Lisker and Abramson, with permission.]

two distinct events. In production, it calls for precise timing of a laryngeal gesture (approximation of the vocal cords) in relation to supralaryngeal release; in perception, it calls for judgment of a complex acoustic pattern arrayed in time. Nonetheless, within these limits, VOT offers a relatively invariant physical display and a relatively invariant sequence of coordinated articulatory gestures that might serve to define a feature, albeit not a feature within the generally accepted system (Chomsky and Halle, 1968). [For a full account of the underlying rationale, the reader is referred to the publications cited above and, for discussions of the approach, to Stevens and Klatt (1974) and Summerfield and Haggard (1972); see also Haggard, Ambler, and Callow (1970).]

A second example of the search for feature invariants is provided by the work of Stevens (1967, 1968a, 1968b, 1972a, 1972b, 1973). In a recent paper, for example (Stevens, 1973), he approaches an acoustic definition of [+ Consonantal], describing consonants as displaying "a rapid change in the acoustic spectrum" (p. 157) in the region of F2, following release (cf. Fant, 1962). He develops this description, emphasizing the entire spectrum rather than individual formants, into an acoustic account of place features [+ Coronal], [+ Labial], and [+ Velar], for which he posits "property detectors." The acoustic description is based on spectrographic analysis and computations from an idealized vocal tract model. The model reveals certain "quantal places of articulation which are optimal from the point of view of sound generation" (Stevens, 1968a: 200) since they permit relatively imprecise articulation without serious perturbation of the signal. Obviously, these tract shapes can be correlated with articulatory gestures to provide the needed feature correlates.

Finally, less ambitious attempts to discover feature invariants are instanced by tape-cutting experiments with natural speech, in which consonantal portions of a syllable are removed and presented for identification alone or with vowels other than the original (Fischer-Jørgensen, 1972; Cole and Scott, 1974). If this approach leads to precise definition of acoustic invariants, it will have proved valuable. However, if experiments merely demonstrate that transposing initial portions of two CV syllables, for example, yields no change in perception of initial consonant, we have not advanced. The transposed patterns remain different both acoustically and, if removed from the speech stream, psychoacoustically, and the demonstrated source of invariance is still the listener. The ultimate test of all these attempts will be in control of a speech synthesizer from a set of invariant articulatory or acoustic feature specifications (Mattingly, 1971).

THE PHONETIC PERCEPT

Up to this point we have simply assumed the units of speech perception. However, research has sporadically puzzled over their definition for the past 25 years. The puzzle arises, as we have seen, from the mismatch between the acoustic signal and the abstract entities of linguistic analysis, distinctive features, and phonemes. Nonetheless, each of these units has been shown to have psychological reality. Perhaps the most direct evidence comes from studies of speaking errors. Fromkin (1971) has analyzed many utterances for errors of metathesis (spoonerism). She finds that speakers may metathesize not only words and phrases, but syllables (clarinet and viola → clarinola), phonemes (far more → mar fore), and features (clear blue → glear plue) (cf. Boomer and Laver, 1968; MacKay, 1970; Cairns, Cairns, and Williams, 1974). Of particular interest is her observation that speakers may exchange consonant for consonant and vowel for

vowel, but never consonant for vowel. This reflects a distinction in production between phonetic elements of the syllable that are, as we shall see, repeatedly distinguished in perception. In any event, errors of metathesis logically require that the speaker have independent control over the unit of error. And if these units are independently produced, it is reasonable to believe that they are independently perceived.

Evidence from perceptual studies is not lacking. Errors of subjects listening to speech through noise (Miller and Nicely, 1955; Mitchell, 1973) or dichotically (Studdert-Kennedy and Shankweiler, 1970; Studdert-Kennedy, Shankweiler, and Pisoni, 1972; Blumstein, 1974), are patterned according to some form of feature system. Scaling studies, in which the experimenter attempts to determine the psychological space occupied by a set of consonants or vowels, repeatedly reveal a structure parsimoniously described by feature theory (Greenberg and Jenkins, 1964; Singh, 1966; Hanson, 1967; Singh and Woods, 1970; Shepard, 1972). A new paradigm has recently provided further evidence. Goldstein and Lackner (in press), adapting a technique devised by Warren and Gregory (1958; also Warren, 1968, and in Lass, in press), played a 200 msec nonsense syllable over and over (200 times per minute), asking listeners to report what they heard. After a few repetitions, listeners began to hear different words (verbal transformation). The new words were systematically related to the originals: they entailed changes in value of only one or two distinctive features, and reflected phonological constraints of English as described by distinctive feature theory. Finally, errors in short-term memory studies also follow a feature pattern (Sales, Cole, and Haber, 1969; Wickelgren, 1965, 1966). Several of these studies have used their perceptual data to compare the predictive power (and so the validity) of different feature systems. Such work is particularly important if linguistics is to be regarded as a branch of human psychology (Chomsky, 1972), and if the abstract units of phonology are to be grounded in human articulatory and perceptual capacities (Ladefoged, 1971a, 1971b; Liljencrants and Lindblom, 1972; Lindblom, 1972).

The perceptual status of the columns in a feature matrix has proved more controversial. Functionally, the column (phone) represents the grouping of distinctive features within a syllable, specifying the domain within which a particular feature is to apply. We recognize this perceptually in alliteration (big boy) and in rhyme (bee and see), where two syllables are perceived as identical at their beginning, but not at their end, or vice versa. Listeners reveal this function when asked to judge similarities among words. Vitz and Winkler (1973) found, in fact, that the number of phones shared by a pair of words was a more satisfactory predictor of their judged similarity than the number of shared features. In the verbal transformation study described above (Goldstein and Lackner, in press), transformations were best described in terms of phones and features rather than syllables and features: consonant transforms and vowel transforms, for example, were independent, reflecting feature shifts within, but not across, phones. Finally, several studies (Kozhevnikov and Chistovich, 1965; Savin and Bever, 1970; Day and Wood, 1972) have shown reaction time differences in identification of consonants and vowels within the same syllables. These differences would not occur if the syllable were an unanalyzed perceptual entity.

Despite such evidence and despite the clear role of phoneme-size phonetic segments in speaking and in writing systems, students have been tempted to regard these segments as "nonperceptual" (Savin and Bever, 1970) or as "fictitious units" based on the historical accident of alphabet invention [Warren (in Lass,

in press)]. Among the arguments for this conclusion seem to be three solid facts, two (or more) pieces of ambiguous evidence, and one false belief. The facts are: first, that no phoneme-size segment can be isolated in the acoustic signal; second, that some phonemes (stop consonants) cannot be spoken in isolation; third, that we do speak in syllables and that syllables are the carriers of stress and speech rhythm. The ambiguous evidence comes from reaction time studies suggesting that syllables, and even higher order units, may be identified before the elements of which they are composed. Savin and Bever (1970) and Warren (1971) showed that the reaction time of listeners monitoring a monosyllabic list for syllables is faster than their reaction time when monitoring the same list for the initial phoneme of the syllable. Subsequently, Foss and Swinney (1973) showed that, under similar conditions, listeners responded more rapidly to words than to their component syllables, while Bever (1970) revealed that listeners responded more rapidly to three-word sentences than to their component words. It was left to McNeill and Lindig (1973) to release us from this "Looking Glass" world, in which the trial precedes the crime, by demonstrating that reaction time was always fastest to the largest elements of which a list was composed. In other words, listeners' response is most rapid at the level of linguistic analysis to which context has directed their attention.

Finally, the false belief is that invariance and segmentation problems would disappear if the syllable were an unanalyzed unit of perception. This belief is no better founded than Wickelgren's (1969) attempt to solve the invariance problem by positing context-sensitive allophones, and is open to many of the same objections. These objections have been well summarized by Halwes and Jenkins (1971), and we will not review them here. However, it is worth adding that the syllable has resisted acoustic definition only somewhat less than the phoneme-size phonetic segment. Its nucleus may be detected by amplitude and fundamental frequency peak picking (Lea, 1974), and Malmberg (1955) drew attention to the possible role of formant transitions in defining syllable boundaries, but no fully satisfactory definition has yet emerged. Furthermore, coarticulation and perceptual context effects across syllables, though less marked than across phones, still occur. Öhman (1966), for example, found drastic variations in vowel formant transitions on either side of stop closure, as a function of the vowel on the opposite side of the closure. And Treon (1970) has demonstrated contextual effects in perception extending across two to three syllables. In fact, as Fodor, Bever, and Garrett (1974) hint, an account of syllable perception may well require the same theoretical apparatus as an account of phone perception.

Much of the confusion over units of speech perception might be resolved if the distinctions between signal and message, and among acoustic, phonetic, and higher levels were strictly maintained. There is wide agreement among writers, whose views may otherwise diverge, that the basic <u>acoustic</u> unit of speech perception (and production) is of roughly syllabic length [e.g., Liberman, Delattre, and Cooper, 1952; Liberman, 1957; Kozhevnikov and Chistovich, 1965; Öhman, 1966; Ladefoged, 1967; Liberman et al., 1967; Savin and Bever, 1970; Massaro, 1972; Stevens and House, 1972; Cole and Scott, 1973; Kirman, 1973; McNeill and Repp, 1973; Warren (in Lass, in press); Studdert-Kennedy, in press]. This is not to deny that there are longer stretches of the signal over which the perceptual apparatus must compute relations, but simply to say that the smallest stretch of signal on which it goes to work is produced by the articulatory syllabic gesture (Stetson, 1952). This does not mean [as Massaro (1972), for example, seems to suppose] that the syllable is the basic linguistic and perceptual unit.

We may clarify by conceptualizing the process of constructing an utterance from a lexicon of morphemes. The abstract entity of the morpheme is the fundamental unit in which semantics, syntax, and phonology converge. Each morpheme is constructed from phonemes and distinctive features. At this level, the syllable does not exist. But morphemic structure is matched to (and must ultimately derive from) the articulatory capacities of the speaker. Both universal and language-specific phonotactic constraints ensure that a morpheme will eventuate in pronounceable sequences of consonants and vowels. Under the control of a syntactic system governing their order and prosody, the morphemes pass through the phonetic transform into a sequence of coarticulated gestures. These gestures give rise to a sequence of acoustic syllables, into which the acoustic correlates of phoneme and distinctive feature are woven. The listener's task is to recover the features and their phonemic alignment, and so the morpheme and meaning. In short, perception entails the analysis of the acoustic syllable, by means of its acoustic features, into the abstract perceptual structure of features and phonemes that characterize the morpheme. We now turn to some theoretical accounts of how this might proceed.

MODELS OF PHONETIC PERCEPTION

We have no models specified in enough detail for serious test. But a brief account of two approaches that have influenced recent research may serve to summarize the discussion up to this point. The two approaches are those of the Haskins Laboratories investigators and of Stevens and his colleagues at the Massachusetts Institute of Technology. Both groups are impressed, in varying degrees, by the invariance and segmentation problem. Both have therefore rejected a passive template- or pattern-matching model in favor of an active or generative model. (For a review, see Cooper, 1972.)

Liberman et al. (1967), reformulating a theme that had appeared in many earlier papers from the Haskins group, proposed a "motor theory of speech perception." The crux of their argument was that an articulatory description of speech is not merely simpler, but is the only description that can rationalize the temporally scattered and contextually variable patterns of speech. They argue that phonetic segments undergo, in their passage through the articulatory system, a process of "encoding." They are restructured acoustically in the syllabic merger, so that cues to phonetic identity lose their alignment and are distributed over the entire syllable (Liberman, 1970). Not all phonetic segments undergo the same degree of restructuring: there is a hierarchy of encodedness, from the highly encoded stop consonants, through nasals, fricatives, glides, and semivowels, to the relatively unencoded vowels. Nonetheless, recovery of phonetic segments from the syllable calls for parallel processing of both consonant and vowel; neither can be decoded without the other. And this demands a specialized decoding mechanism, in which reference is somehow made to the articulatory gestures that gave rise to the encoded syllables.

Liberman et al. (1967) assume, reasonably enough, that "at some level...of the production system there exist neural signals standing in one-to-one correspondence with the various segments of the language," and that for the phoneme "the invariant is found far down in the neuromotor system, at the level of the commands to the muscles" (p. 454). It is important to note that actual motor engagement is not envisaged. Liberman (1957) has written: "We must assume that the process is somehow short-circuited--that is, that the reference to articulatory movements and their sensory consequences must somehow occur in the brain without getting out into the periphery" (p. 122).

A virtue of the model is that it accounts for a fair amount of data and has generated a steady stream of research. Also, the concept of encoding, though descriptive rather than explanatory, draws attention to a process at the base of language analogous to syntactic processes suggested by generative grammar, and hints at formal similarities in the physiological processes underlying phonetic and syntactic performance (Mattingly and Liberman, 1969; Liberman, 1970; Mattingly, 1973, 1974). Conspicuously absent is any account of first-language acquisition. The child may be presumed to be born with some "knowledge" of vocal tract physiology and an incipient capacity to interpret the output of an adult tract in relation to that of its own (Mattingly, 1973), but a detailed account of the process is lacking.

Stevens (1973) has concerned himself with this problem, and addresses it in the most recent version of his analysis-by-synthesis model (Stevens, 1972a; cf. Stevens, 1960; Stevens and Halle, 1967). The model is far more explicit than that of the Haskins group. The perceptual process is conceived as beginning with some form of peripheral spectral analysis, acoustic feature and pitch extraction. Pitch and spectral information, over a stretch of several syllables, is placed in auditory store. Acoustic feature information undergoes preliminary analysis by which a rough matrix of phonetic segments and features is extracted and passed to a control system. On occasion, this matrix may provide sufficient information for the control (which knows the possible sequences of phonetic segments and has access to the phonetic structure of earlier sections of the utterance) simply to pass the description on to higher levels. If this is not possible, the control guesses at a phonetic description on the basis of its inadequate information and sends the description to a generative rule system, the same that in speaking directs the articulatory mechanism. The rule system generates a version of the utterance and passes it to a comparator for comparison with the spectral description in temporary auditory store. The comparator computes a difference measure and feeds it back to the control. If the "error" is small enough, the control system accepts its original phonetic description as correct. If not, it makes a second guess and the cycle repeats until an adequate match is reached.

This rough account does no justice to the model's elegance and subtlety, but it may serve to focus attention on several points. First, the solution to the invariance problem is a more abstract and more carefully specified version of a motor theory. Second, the model emphasizes the necessity of at least a preliminary feature analysis, to ensure that the system is not doomed to an infinity of bad guesses, and that the child, given a set of innate "property detectors," can latch onto the utterance. At the same time, no account is offered of how the invariant acoustic properties are transformed into phonetic segments and features (the process is simply consigned to "a preliminary analysis"), nor of the precise form that the phonetic description takes. Finally, the model emphasizes the need for a short-term auditory store. As we shall see, the form and duration of such a store is currently the focus of a great deal of research.

THE PROCESSING OF CONSONANTS AND VOWELS

Preliminary

To brace ourselves for a fairly prolonged discussion of consonants and vowels, let us consider why they are interesting. For theory, the answer is that they lie at the base of all phonological systems. All languages are

syllabic, and all languages constrain syllabic structure in terms of consonants and vowels. If we are to ground phonological theory in human physiology, we must understand why this path was taken. Lieberman (1970) has argued that phonological features may have been selected through a combination of articulatory constraints and "best matches" to perceptual capacity. One purpose of current research is to understand the nature and basis of the best match between syllables, constructed from consonants and vowels, and perceptual capacity.

For experiment, the interest of consonants and vowels is that they are different. If all speech sounds were perceived in the same way, we would have no means of studying their underlying relations. Just as the biologist could not study the genetics of eye-color in Drosophila melanogaster until he had found two flies with different eyes, so the student of speech had no means of analyzing syllable perception until he had found portions of the syllable that reflected different perceptual processes (cf. Stetson, 1952). Fortunately, the interests of theory and research converge.

Categorical Perception

Study of sound spectrograms reveals that portions of the acoustic patterns for related phonetic segments (segments distinguished from one another by a single feature) often lie along an apparent acoustic continuum. For example, center frequencies of the first two or three formants of the front vowels /i,I, ɛ,æ / form a monotonic series; syllable-initial voice-voiceless pairs /b,p/, /d,t/, /k,g/ differ systematically in voice onset time; voiced stops /b,d,g/ before a particular vowel, differ primarily in the extent and direction of their formant transitions.

To establish the perceptual function of such variations speech synthesis is used. Figure 3 sketches a schematic spectrogram of a synthetic series in which changes of slope in F2 transition effect perceptual changes from /b/ through /d/ to /g/. Asked to identify the dozen or so sounds along such a continuum, listeners divide it into distinct categories. For example, a listener might consistently identify stimuli -6 through -3 of Figure 3 as /b/, stimuli -1 through +3 as /d/, and stimuli +5 through +9 as /g/. In other words, he does not, as might be expected on psychophysical grounds, hear a series of stimuli gradually changing from one phonetic class to another, but rather a series of stimuli, each of which (with the exception of one or two boundary stimuli) belongs unambiguously in a single class. The important point to note is that, although steps along the continuum are well above nonspeech auditory discrimination threshold, listeners disregard acoustic differences within a phonetic category, but clearly hear equal acoustic differences between categories.

To determine whether listeners can, in fact, hear the acoustic differences belied by their identifications, discrimination tests are carried out, usually in ABX format. Here, on a given trial, the listener hears three stimuli, separated by a second or so of silence: the first (A) is drawn from a point on the continuum two or three steps removed from the second (B), and the third (X) is a repetition of either A or B. The listener's task is to say whether the third stimulus is the same as the first or the second. The typical outcome for a stop consonant continuum, is that listeners hear few more auditory differences than phonetic categories: they discriminate very well between stimuli drawn from different phonetic categories, and very poorly (somewhat better than chance) between stimuli drawn from the same category. The resulting function displays peaks at

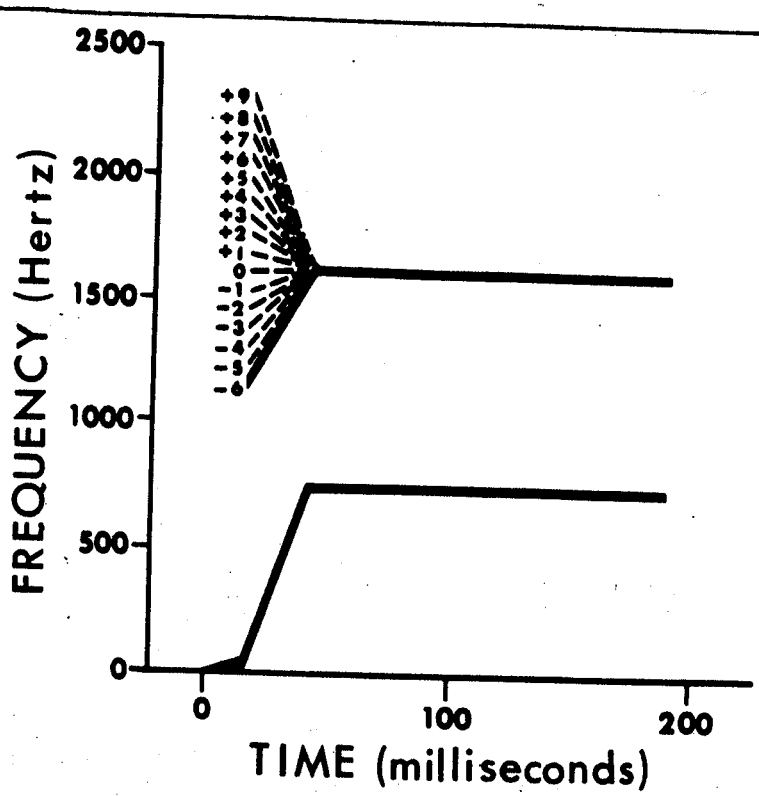

Figure 3: Schematic spectrogram for a series of synthetic stop-vowel syllables varying only in F2 transition. F2 steady-state, F1 transition, and steady-state remain constant. As F2 transition changes from -6 to +9, perception of initial consonant shifts from [b] through [d] to [g].

phonetic boundaries, troughs within phonetic categories. In fact, discriminative performance can be predicted with fair accuracy from identifications: the probability that acoustically different syllables are correctly discriminated is a positive function of the probability that they are differently identified (Liberman, Harris, Kinney, and Lane, 1961). This close relation between identification and discrimination has been termed "categorical perception": that is to say, perception by assignment to category. Figure 4 (left side) illustrates the phenomenon. Note that, although prediction from identification to discrimination is good, it is not perfect: listeners can sometimes discriminate between different acoustic tokens of the same phonetic type. Note, further, that neither identification nor discrimination functions display quantal leaps across category boundaries. This is not a result of data averaging, since the effect is given by individual subjects. Evidently auditory information about consonants is slight, but not entirely lacking.

We may now contrast categorical perception of stop consonants with "continuous perception" of vowels. Figure 4 (right side) illustrates the effect. There are two points to note. First, the vowel identification function is not as clear-cut as the consonant. Vowels, particularly those close to a phonetic boundary, are subject to context effects: for example, a token close to the /i-I/ boundary will tend to be heard, by contrast, as /i/, if preceded by a clear /I/, as /I/, if preceded by a clear /i/. The second point to note is that vowel discrimination is high across the entire continuum. Phonetic class is not

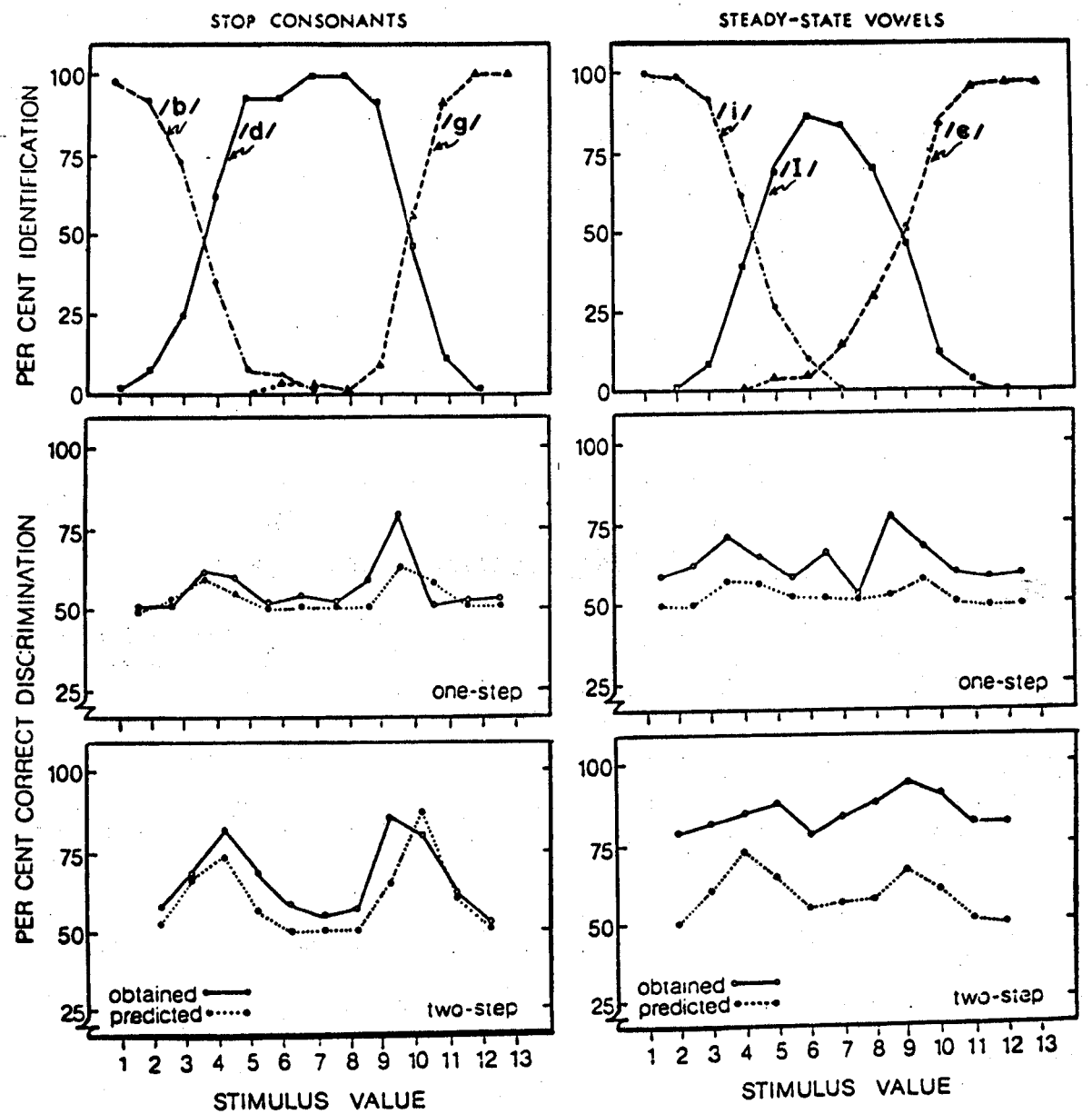

Figure 4: Average identification functions for synthetic series of stop consonants and vowels (top). Average one-step (middle) and two-step (bottom) predicted and obtained ABX discrimination functions for the same series. [After Pisoni (1971), with permission of the author.]

totally irrelevant (there is a peak in the discrimination function at the category boundary), but both within and between categories listeners discriminate many more differences than they identify. Their perception is said to be "continuous." [For fuller discussion, see Studdert-Kennedy, Liberman, Harris, and Cooper (1970a) and Pisoni (1971).]

Continuous perception is typical not only of vowels, but also of many nonspeech psychophysical continua along which we can discriminate more steps than we can identify (Miller, 1956). This fact has been taken as evidence both that categorical perception is peculiar to speech, and that stop consonants and vowels engage fundamentally different perceptual processes (Liberman et al., 1967; Studdert-Kennedy et al., 1970a). In fact, an early account of the phenomenon invoked a motor theory of speech perception (Liberman et al., 1967). As we have seen, there are independent grounds for hypothesizing that speech is perceived by reference to its articulatory origin. Here seemed to be additional evidence: the discrete articulatory gestures of stop consonants yielded discrete perceptual categories; the more variable gestures of vowels, more variable categories. But this account has several weaknesses, and recent work has largely eroded it. For one thing, we now know that categorical perception is not confined to speech (Locke and Kellar, 1973; Miller, Pastore, Wier, Kelly, and Dooling, 1974; Cutting and Rosner, in press).

However, this discovery in no way diminishes the importance of the phenomenon, as will become clear in the following sections. Here, we merely note two facts. First, the acoustic patterns distributed along a speech continuum are not arbitrary. They have been selected from the range of patterns that the articulatory apparatus can produce and that the auditory system can analyze. The categories are therefore natural, in the sense that they reflect physiological constraints on both production and perception. As Stevens (1972b) has pointed out, our task is to define the joint auditory and articulatory origin of phonetic categories.

Second, categorical perception reflects a functionally important property of certain speech sounds. The initial sound of /da/, for example, is difficult, if not impossible, to hear: the sound escapes us and we perceive the event, almost instantly, as phonetic. Rapid sensory decay and transfer into a nonsensory code is probably crucial to an efficient linguistic signaling system. Study of categorical perception has, in fact, revealed functional differences between stop consonants and vowels that are central to the syllabic structure of speech. At the same time it has provided basic evidence for the distinction between auditory and phonetic levels of processing.

In the following sections we consider two main aspects of categorical perception: first, the division of a physical continuum into sharply defined categories, and the assignment of names to the categories; second, listeners' apparent inability to discriminate among members of a category.

The Bases of Phonetic Categories

Phonetic categories do not arise from simple discriminative training, as proposed by Lane (1965). Subjects may certainly learn to divide a sensory continuum into clear-cut categories, with a resultant small peak in the discrimination function at the category boundary. But discrimination within categories remains high (Parks, Wall, and Bastian, 1969; Studdert-Kennedy et al., 1970a;

Pisoni, 1971): training may increase, but not obliterate discriminative capacity. Furthermore, the learned boundary is likely to be unstable. The process is familiar to the psychophysicist. For example, if we present a subject with a series of weights and ask him to judge each weight as either heavy or light, he will, with a minimum of practice, divide the range cleanly around its balance point (see Woodworth and Schlosberg, 1954:Ch. 8). However, the boundary between heavy and light can be readily shifted by a change in experimental procedure. If an extreme token is presented for judgment with a probability several times that of other stimuli along the continuum, it comes to serve as an anchor with which other stimuli contrast: the result is a shift in category boundary toward the anchoring stimulus. Pisoni and Sawusch (in press) have shown that such shifts occur for a series of tones, differing in intensity, and for vowels, but not for stop consonants distributed along a voice-onset time continuum. They suggest that response criteria for voicing categories are mediated by internal rather than external references. By thus reframing the observation that stop consonant categories are not subject to context effects, they invite us to consider the nature of the internal reference.

Such a reference must be some distinctive perceptual quality shared by all members, and by no nonmembers, of a category. There is, of course, no reason to suppose that distinctive perceptual qualities are confined to speech continua. They will emerge from any physical continuum for which sensitivity is low within restricted regions and, by corollary, high between these regions. However, while the distinctive perceptual quality of a nonspeech event (such as a click, a musical note, or a flash of light) has the character of its sensory mode, the distinctive perceptual quality of a speech sound is phonetic. It is into a phonetic code that speech sounds are rapidly and automatically transferred for storage and recall.

With this in mind, we turn to several studies of nonspeech continua. We begin with Cutting and Rosner (in press), who determined an auditory boundary between rapid and slow stimulus onsets. Variations in stimulus onset, or rise time, are known to contribute to the affricate/fricative distinction, /tʃa/ versus /ʃa/ (Gerstman, 1957). The authors varied rise time from 0 to 80 msec for sawtooth wave trains, generated by a Moog synthesizer, and for synthetic affricate/fricatives. The rapid-onset sawtooth waves sounded like a plucked guitar string, the slow-onset waves like a bowed string. Cutting and Rosner presented their two classes of stimuli for identification (pluck - bow, /tʃa/ - /ʃa/) and for ABX discrimination. Both speech and nonspeech yielded category boundaries at a 40-50 msec rise time, with appropriate peaks and troughs in the discrimination functions.

A second instance of nonspeech categorical perception is provided by Miller et al. (1974). These investigators constructed a rough nonspeech analog of the voice-onset time continuum. They varied the relative onset times of bursts of noise and periodic buzz, over a range of noise-leads from -10 to +80 msec, and presented them to subjects for labeling and discrimination. Listeners divided the continuum around an average noise-lead of approximately 16 msec, displaying clear discrimination troughs within no noise-lead and noise-lead categories, and a discrimination peak at the category boundary. The boundary value agrees remarkably well with that reported by Abramson and Lisker (1970) for the English labial VOT continuum, though not with the systematically longer perceptual boundaries associated with English apical and velar VOT continua (Lisker and Abramson, 1970). The authors conclude that the categories of their experiment (and,

presumably, of at least the English labial VOT continuum) lie on either side of a "difference limen" for duration of the leading noise. While possibly correct, their conclusion places a misleading emphasis on the boundary between categories rather than on the categories themselves.

The emphasis is reversed in a recent study of Stevens and Klatt (1974). Following Liberman, Delattre, and Cooper (1958), they examined auditory discrimination of two acoustic variables along the stop consonant voice-voiceless continuum: delay in formant onset and presence/absence of F1 transition. For their first experiment they constructed a nonspeech analog of plosive release and following vowel: a 5 msec burst of noise separated from a vowel-like buzz by between 0 and 40 msec of silence. Listeners' "threshold" for detection of silence between noise and buzz was approximately 20 msec, a close match with the value for detection of noise lead found by Miller et al. (1974). Stevens and Klatt (1974) imply that the unaspirated/aspirated stop consonant perceptual boundary in the 20-40 msec VOT range may represent "a characteristic of the auditory processing of acoustic stimuli independent of whether the stimuli are speech or nonspeech" (p. 654).

We will not pursue the details of their second experiment. However, they were able to confirm the contribution of a detectable F1 transition to the voice-voiceless distinction. Furthermore, by hewing to the articulated speech signal and by focusing on acoustic properties within categories rather than on acoustic differences between them, Stevens and Klatt were able to offer a fully plausible account of systematic increases in the voice-voiceless perceptual boundary that are associated with shifts from labial to apical to velar stop consonants (Lisker and Abramson, 1970; Abramson and Lisker, 1973).

If the argument of the last few pages has given the impression that auditory boundaries between phonetic categories are readily determined, the impression must be dispelled. The criterion for such boundaries is that they be demonstrated in a nonspeech analog, a feat that has proved peculiarly difficult for the voiced stop consonants. The typical outcome of studies in which formant patterns controlling consonant assignments are removed from context and presented for discrimination is that they are perceived continuously (e.g., Mattingly, Liberman, Syrdal, and Halwes, 1971). A striking instance is provided by the work of Popper (1972). He manipulated F2 transitions within a three-formant pattern (cf. Figure 3) to yield a synthetic series from /ab/ to /ad/. He then measured energy passed by a 300 Hz band-width filter, centered around the F2 steady-state frequency, and noted a sharp drop at the /b-d/ boundary both for isolated F2 and for the full formant pattern. However, subjects evinced the expected discrimination peak only for the full pattern: the isolated F2, despite its acoustic discontinuity, was continuously perceived.

In short, no simple notion of fixed regions of auditory sensitivity serves to account for categorical division even of the /ba,da,ga/ continuum, let alone for perceptual invariance across phonetic contexts, for the normalizing shifts in category boundary associated with speaker variation (cf. Fourcin, 1968; Rand, 1971), or for cross-language differences in boundary placement. The problem is not confined to articulatory place distinctions. Consider, for example, the fact that Spanish speakers typically yield a somewhat shorter labial VOT boundary than do English (Lisker and Abramson, 1964) and that their perceptual boundary shows a corresponding reduction (Lisker and Abramson, 1970). We can hardly account for the perceptual shift by appeal to an inherently sharp threshold. Precise

category position along a continuum is clearly a function of linguistic experience (see also Stevens, Liberman, Studdert-Kennedy, and Öhman, 1969). Popper (1972) proposes, in fact, that "people who speak different languages may tune their auditory systems differently" (p. 218). Differential "tuning" could result from cross-language differences in selective attention to aspects of the signal, and in criterion levels for particular phonetic decisions. Given the close match between perception and production (Stevens et al., 1969; Abramson and Lisker, 1970; Lisker and Abramson, 1970), it seems plausible that such differences should arise from complex interplay between speaking and listening during language acquisition (see below, From Acoustic Feature to Phonetic Percept).

The notion of "tuning" presupposes the existence of acoustic properties to which the auditory system may be attuned. The first steps toward definition of these properties have been taken by Stevens (see especially 1972b, 1973). As earlier remarked, Stevens has used spectrographic analysis and computations from an idealized vocal tract model to describe possible acoustic correlates of certain phonetic features. He finds, for example, that the spectral patterns associated with continuous changes in place of articulatory constriction along the vocal tract do not themselves change continuously. Rather, there are broad plateaux, within which changes in point of constriction have little acoustic effect, bounded by abrupt acoustic discontinuities. These acoustic plateaux tend to correlate with places of articulation in many languages. In short, Stevens is developing the preliminaries to a systematic acoustic account of phonetic categories and their boundaries. His work is important for its emphasis on the origin of phonetic categories in the peculiar properties of the human vocal tract. Furthermore, as will be seen below, his approach meshes neatly with recent work on auditory feature analyzing systems as the bases of phonetic categories.

Auditory and Phonetic Processes in Categorical Perception

We turn now to the second main aspect of categorical perception--listeners' failure to discriminate among members of a category--and to the contrast between continuously perceived vowels and categorically perceived stop consonants. A long series of experiments over the past few years has shown that listeners' difficulty in discriminating among members of a category is largely due to the low energy transience of the acoustic signal on the basis of which phonetic categories are assigned. Lane (1965) pointed to the greater duration and intensity of the vowels and showed that they were more categorically perceived if they were degraded by being presented in noise. Stevens (1968b) remarked the brief, transient nature of stop consonant acoustic cues, and showed, as did Sachs (1969) (1969), that vowels were more categorically perceived if their duration and acoustic stability were reduced by placing them in CVC syllables.

The role of auditory memory, implicit in the work just cited, was made explicit by Fujisaki and Kawashima (1969, 1970) in a model of the decision process during the ABX trial. If a listener assigns A and B to different phonetic categories (i.e., if A and B lie on opposite sides of a phonetic boundary), his only task is to determine whether X belongs to the same category as A or as B: his performance is then good and a discrimination peak appears in the function for both consonants and vowels. However, if a listener assigns A or B to the same phonetic category, he is forced to compare X with his auditory memory of A and B: his performance is then slightly reduced for vowels, for which auditory memory is presumed to be relatively strong, but sharply reduced for consonants, for

which auditory memory is presumed to be weak. Evidence for the operation of such a two-step process within phonetic categories in man, but not in monkey, has recently been reported by Sinnott (1974).

Before we proceed, let us spell out some distinctions between auditory and phonetic memory stores. The auditory store, or trace, is usually assumed to be rather like an echo: a faint simulacrum, if not of the waveform, at least of its neural correlates at an early stage of processing. Like an echo, the trace is an analog of its original, decays rapidly, and may be displaced if another sound arrives to interfere before decay is complete. The phonetic store, on the other hand, is a set of discrete features, its decay is a good deal slower, and interference can only be accomplished by another phonetic entity with similar phonetic features.

With this in mind, we turn to several experiments by Pisoni (1971, 1973a, 1973b) in which he tested and supported Fujisaki and Kawashima's hypothesis concerning auditory memory for consonants and vowels. In the first (Pisoni, 1973a) he varied the A-to-X delay interval from zero to two seconds in an AX same - different task for vowel and stop consonant continua. Between-category performance (presumably based on phonetic store) was high and independent of delay interval for both consonants and vowels; within-category performance (presumably based on auditory store) was low and independent of delay interval for consonants, but for vowels was high and declined systematically as delay interval increased. In subsequent experiments, Pisoni (1973b) demonstrated that the degree of categorical (or continuous) perception of vowels can be manipulated by the memory demands of the discrimination paradigm and by the amount of interference from neighboring stimuli (Glanzman and Pisoni, 1973).

Changing tack, Pisoni and Lazarus (1974) sought methods of increasing apparent auditory memory for stop consonants. This is more difficult, but by a particular combination and sequence of experimental conditions, they were able to demonstrate improved within-category discrimination on a voice-voiceless continuum. The same continuum (/ba-pa/) also elicited reaction time differences in a pair-matching task (Pisoni and Tash, 1974; cf. Posner, Boies, Eichelman, and Taylor, 1969). Here, listeners were asked to respond same or different to pairs of stimuli drawn from the continuum. Same reaction times were faster for identical pairs than for acoustically distinct pairs, drawn from the same phonetic category; different reaction times decreased as acoustic differences between pairs from different categories increased. This last result recalls Barclay's (1972) finding that listeners can correctly and reliably judge acoustic variants of /d/, drawn from a synthetic continuum, as more similar to /b/ or /g/. If we add these studies to our earlier observation that listeners always display a margin of within-category discrimination for consonants, and that discrimination functions do not display a quantal leap between categories, we must conclude that the auditory system does retain at least some trace of consonantal passage. At the same time, there is little question that this trace is fainter than that for vowels.

The conclusion of all these studies is pointed up by the work of Raphael (1972). He studied voice-voiceless VC continua, manipulating initial vowel duration as the acoustic cue to voicing of the final stop. Here, where the perceptual object was consonantal, but the acoustic cue vocalic, perception was continuous. In short, consonants and vowels are distinguished in the experiments we have been considering, not by their phonetic class or the processes of

assignment to that class, but by their acoustic characteristics and by the duration of their auditory stores. If the longer store of the vowels is experimentally reduced, their membership in the natural class of segmental phonetic entities is revealed by their categorical perception.

Stages of Auditory Memory

Several independent lines of research, drawing on different experimental paradigms, have recently begun to converge on perceptual and memorial processes below the level of phonetic classification. Experimenters often share neither terminology nor theoretical framework, but we can discern two, not entirely overlapping, lines of division in the perceptual process. The first divides short-term memory into a brief store lasting some hundreds of milliseconds, and a longer store lasting several seconds. The second divides peripheral from central processes; this is important, but we will not consider it in detail here, since the cut cannot be as surely made in audition as in vision (due to incomplete decussation of auditory pathways), and most of the processes to be discussed are certainly central.

Short-Term Auditory Stores

Store I. As a step toward further analysis of auditory memory for speech, consider the concept of parallel processing. Liberman et al. (1967) used this term to describe the decoding of a CV syllable, in which acoustic correlates of consonant and vowel are distributed over an entire syllable (Liberman, 1970). Obviously, the process requires a store at least as long as the syllable to register auditory information, and presumably somewhat longer to permit transfer into phonetic code.

Direct evidence of this type of parallel processing comes from several sources. Liberman et al. (1952) showed that the phonetic interpretation of a stop release burst varied with its following vowel, and concluded that we perceive speech over stretches of roughly syllabic length (cf. Schatz, 1954). Lindblom and Studdert-Kennedy (1967) demonstrated that the phonetic boundary for a series of synthetic vowels shifted as a function of the slope and direction of initial and final formant transitions: listeners judged vowels in relation to their surrounding consonantal frames (cf. Fujimura and Ochiai, 1963; Strange et al., 1974). More recently, Pisoni and Tash (1974) have studied reaction time to CV syllables: they called for same - different judgments on vowels or consonants of syllable pairs in which nontarget portions of the syllables were also either the same or different. Whether comparing consonants or vowels, listeners were consistently faster when target and nontarget portions of the syllable were redundant (i.e., both same, or both different). In other words, information from an entire syllable contributed to listeners' decisions concerning "segments" of the syllable. In a related study by Wood and Day (in press), listeners identified either the vowel or the consonant of synthetic CV syllables, /ba,da,bæ,dæ/. If all test items were identical on the nontarget dimension (i.e., if all had the same vowel on a consonant test, or all the same consonant on a vowel test), subjects' reaction times were significantly faster than if both target and nontarget dimensions varied. In the latter case, the unattended vowel (consonant) retarded listeners' decisions on the attended consonant (vowel). In short, we have a variety of evidence that, for at least some syllables, consonant and vowel recognition are interdependent, parallel processes, requiring a short-term auditory store of at least syllabic duration.

Massaro (1972) made the functional distinction between such a "perceptual auditory image" and a longer "synthesized" auditory store; he initiated attempts to estimate duration of the "image" by backward masking studies. First discovered in visual experiments (Werner, 1935), the paradigm takes advantage of the fact that perception of a stimulus may be blocked if a second stimulus is presented some hundreds of milliseconds later; it has been used to good effect in vision to separate and describe peripheral and central processes (Turvey, 1973). However, the belief that the critical interstimulus interval (ISI), at which the first stimulus is freed from interference by the second, may be taken as an estimate of the duration of primary auditory display (Massaro, 1972) is difficult to sustain, and application of the technique to the study of speech perception has proved problematic for several reasons.

To begin with, auditory information is displayed over time, so that perception of a target CV syllable of natural duration (say, 200-300 msec) can be interrupted only by a masking syllable that begins before the first syllable is complete. Temporal relations between syllables must then be expressed in terms of stimulus onset asynchrony (SOA) rather than in terms of ISI, and the effectiveness of the mask is reduced because it is itself masked by the first syllable (forward masking). For example, Studdert-Kennedy, Shankweiler, and Schulman (1970b) found that the first syllable was completely freed from masking by the second at a SOA of 50 msec, certainly an underestimate of display time, since it is no more than the duration of the critical consonant information in the formant transitions of the target CV syllable.

There are two solutions to this impasse: make the syllables unnaturally short, or present target and mask to opposite ears (dichotically), thus evading peripheral masking of the second syllable. Several investigators (Massaro, 1972; Pisoni, 1972; Dorman, Kewley-Port, Brady-Wood, and Turvey, 1973) have attempted the first solution. Results are difficult to interpret because both the degree of masking and the critical ISI for release from masking vary with target (consonant or vowel), size and range (acoustic or phonetic) of target set, target and mask energy, relations between target and mask structure (acoustic or phonetic), and individual listeners, many of whom show no masking whatever even for brief (15.5 msec) vowels (Dorman et al., 1973). Where masking could be obtained, the shortest critical ISI observed in these studies (80 msec) was for 40 msec steady-state vowels, and the longest (250 msec) for 40 msec CV syllables (Pisoni, 1972). Note, incidentally, that complete absence of masking has been observed only with vowels, and just as categorical perception of vowels can be induced by degrading them with noise, so too can their masking (Dorman, Kewley-Port, Brady, and Turvey, 1974).

In any event, these variable results do not encourage one to believe that critical ISI is measuring the fixed duration of auditory display. And the case is no better when we turn to dichotic masking paradigms. Pisoni and McNabb (1974), for example, observed a critical SOA for release from dichotic backward masking of between 20 and 150 msec, depending upon target and mask vowel relations. A somewhat longer estimate of 200-250 msec can be extrapolated from the data of Studdert-Kennedy et al. (1970b). A narrower estimate comes from McNeill and Repp (1973b). They studied forward masking of dichotically presented CV syllables, determining the SOA necessary for features of the leading syllable to have no further effect on errors in the lagging syllable, and so presumably to have passed out of the phonetic processor. Their estimate of 80-120 msec may be more realistic for running speech than others, since their procedure eliminated

a component present in all previous studies, namely, time taken to prepare a response, a period during which effective interruption may still occur (Repp, 1973).

However, it is more likely that the entire endeavor is misguided. It seems intuitively plausible that syllable processing time is not constant, but varies, under automatic attentional control, with speaking rate and other factors. The studies reviewed are simply measuring time required for release from masking under a variety of more or less adverse conditions. This is certainly not without interest, particularly if we can show it to be a function of well-specified target-mask relations. But we shall then be turning attention away from the notion of a primary auditory store, and toward the more important question of what acoustic dimensions are extracted in the very earliest stage of processing, and how they interact to determine the phonetic percept.

Store II. Nonetheless, some form of auditory store is clearly necessary. We would otherwise be unable to interpret the prosody of running speech, and there is ample experimental evidence of cross-syllabic auditory interaction (Hadding-Koch and Studdert-Kennedy, 1964; Studdert-Kennedy and Hadding, 1973; Atkinson, 1973). Detailed analysis of this longer store, perhaps lasting several seconds, was made possible by the work of Crowder and Morton (1969; see also Crowder, 1971a, 1971b, 1972, 1973). They were the first experimenters to undertake a systematic account of what they termed "precategorical acoustic storage" (PAS).

Evidence for the store comes from studies of immediate, ordered recall of span-length digit lists. Typically, error probability increases from beginning to end of list, with some slight drop on terminal items (recency effect). The terminal drop is significantly increased, if the list is presented by ear rather than by eye (modality effect). Crowder and Morton (1969) argue that these two effects reflect the operation of distinct visual and auditory stores for precategorical (prelinguistic) information, and of an auditory store that persists longer than the visual. Support comes from demonstrations that the recency effect is significantly reduced, or abolished, if subjects are required to recall the list by speaking rather than by writing (Crowder, 1971a), or if an auditory list is followed by a redundant, spoken suffix (such as the word zero), as a signal for the subject to begin recall (suffix effect). That the suffix interferes with auditory, rather than linguistic, store is argued by the facts that the effect (1) does not occur if the suffix is a tone or burst of noise; (2) is unaffected if the spoken suffix is played backward; (3) is unaffected by degree of semantic similarity between suffix and list; (4) is reduced if suffix and list are spoken in different voices; and (5) is reduced if suffix and list are presented to opposite ears.

Of particular interest in the present context is that all three effects (modality, recency, suffix) are observed for CV lists, of which members differ in vowel alone, or in both vowel and consonant (spoken letter names), but not for voiced stop consonant CV or VC lists, of which members differ only in the consonant (cf. Cole, 1973b). Crowder (1971a:595) concludes that "vowels receive some form of representation in PAS while voiced stop consonants receive none." Liberman, Mattingly, and Turvey (1972:329) argue further that phonetic classification "strips away all auditory information" from stop consonants.

However, this last claim is unlikely to be true. First, there is no good reason why the process of categorization should affect vowels and consonants differently. Second, we have a variety of evidence that listeners retain at least some auditory trace of stop consonants (see previous section). Third, consonant and vowel differences in PAS can be reduced by appropriate manipulation of the signal array (Darwin and Baddeley, 1974). These investigators demonstrated a recency effect for tokens of a stop CV, /ga/, and two highly discriminable CV syllables in which the consonantal portion is of longer duration, /ʃa/, /ma/. They also demonstrated that the recency effect for vowels can be eliminated if the vowels are both very short (30 msec of a 60 msec CV syllable) and close neighbors on an F1-F2 plot. They conclude that "the consonant-vowel distinction is largely irrelevant" (p. 48) and that items in PAS cannot be reliably accessed if, like /ba,da,ga/ or /I,ɛ,æ/, they are acoustically similar. The effect of acoustic similarity is, of course, to confound auditory memory. As we shall see shortly (The Acoustic Syllable, below) and, as Darwin and Baddeley (1974) themselves argue, it is to the more general concept of auditory memory that we must have recourse, if we are to understand the full range of experiments in which consonant-vowel differences have been demonstrated.

We turn now to the duration of PAS and the mechanisms underlying its reflection in behavior. Notice, first, that if an eight-item list is presented at a rate of two per sec and is recalled at roughly the same rate, time between presentation and recall will be roughly equal for all items. Therefore, the recency effect cannot be attributed to differential decay across the list, but is due rather to the absence of "overwriting" or interference from succeeding items. Second, since the degree of interference (i.e., probability of recall error) decreases as the time between items increases, and since the suffix effect virtually disappears if the interval between the last item and suffix is increased to 2 sec, we are faced with the paradox that performance improves as time allowed for PAS decay increases. Crowder's (1971b) solution is to posit an active "read-out" or rehearsal process at the articulatory level. Time for a covert run through the list is "...a second or two" (p. 339). If a suffix occurs during this period, PAS for the last couple of items is spoiled before they are reached; if no suffix occurs, the subject has time to check his rehearsal of later items against his auditory store, and so to confirm or correct his preliminary decision. Crowder (1971b) goes on to show that there is, in fact, no evidence for any decay in PAS: in the absence of further input, PAS has an infinite duration. This is intuitively implausible, but we will not pursue the matter here.

Notice, however, that the term precategorical refers to the nature of the information stored, not to the period of time during which it is stored. A preliminary (or even final) articulatory, if not phonetic, decision must have been made before PAS is lost, if rehearsal is to permit cross-check with the store. We are thus reminded of the temporary auditory store hypothesized in the analysis-by-synthesis model of Stevens (1960, 1972a). Crowder's account, with its preliminary analysis and generative rehearsal loop, is so similar to Stevens' model that we may be tempted to identify the two, and to see evidence for PAS function as support for Stevens' hypothesis.

We may remark, however, one important difference. Stevens introduced a synthesis loop to handle the invariance problem, a problem at its most acute for stop consonants. But these are precisely the items excluded from PAS, and all our evidence for consonantal auditory memory suggests a store considerably less

than infinite, probably less than a second. We may, of course, assume that a synthesis loop goes into operation very early in the process, while consonant auditory information is still available, and that the PAS rehearsal loop is simply a sustention beyond the point at which stop consonantal auditory information can be accessed. We would then be forced to posit the decay of consonantal information from auditory store. Continuation of the loop might be automatic during running speech, enabling prosodic pattern to emerge, but under attentional control for special purposes, such as listening to poetry and remembering telephone numbers. But we have, at present, no direct evidence for the earlier stage of the loop.

Stages of Processing

Nor, as we have seen, do we have direct evidence for the primary auditory store inferred from parallel processing. We may, in fact, do well to dismiss division of the process into hypothetical stores, and concentrate attention on the types of information extracted during early processing, and their interactions. Several experimental paradigms have already been applied.

Day and Wood (1972) and Wood (1974) have reported evidence for parallel extraction of pitch (fundamental frequency) and spectral information bearing on segmental classification. For the first experiment they synthesized two CV syllables, /ba,da/, each at two pitches, and prepared two types of random test order. In one, they varied a single dimension, either fundamental frequency or phonetic class; in the other, they varied both dimensions independently. They then called on subjects to identify, with a reaction-time button, either pitch or phonetic class, each in its appropriate one-dimensional test and also in the two-dimensional test. Reaction times were longer for both tasks on the two-dimensional test than on the one-dimensional test, but the increase was significantly greater on the phonetic test than on the pitch task: unpredictable pitch differences interfered with phonetic decision more than the reverse. The authors took this finding as evidence for separate nonlinguistic and linguistic processes, the first mandatory, the second optional. In a follow-up experiment, Wood (1974) substituted a two-dimensional test in which fundamental frequency and phonetic class variations were correlated rather than independent. Reaction times were now significantly shorter for both tasks on the two-dimensional test: subjects drew on both pitch and phonetic information for either pitch or phonetic classification. Wood (1974) concludes that the two types of information are separately and simultaneously extracted [as required, incidentally, by Stevens' (1960) model].

There is more to these experiments. The phonetic task called for a decision on the consonant (/ba/ vs /da/), but pitch information was primarily carried by the vowel. In fact, had fundamental frequency differences been carried solely by initial formant transitions, it is doubtful whether they would have interacted with phonetic decision. Dorman (1974) has shown that listeners are unable to discriminate intensity differences carried by the 50 msec initial transitions of a voiced stop CV syllable, but are well able to discriminate identical differences carried by isolated transitions, or by the first 50 msec of a steady-state vowel. While the experiment has not been done, it seems likely that Dorman's results would have held had he used fundamental frequency instead of intensity. We would then be forced to conclude that, in Wood's (1973b) experiment, subjects were using adventitious pitch information carried by the vowel to facilitate judgment of the consonant, and vice versa. The experiments thus reflect parallel

processing, both of linguistic and nonlinguistic information and of consonant and vowel.

Experimental separation of auditory and phonetic processes has also been attempted in dichotic studies. Consider, for example, the following series. Shankweiler and Studdert-Kennedy (1967; also Studdert-Kennedy and Shankweiler, 1970) found that listeners were significantly better at identifying the consonants of dichotically competing CV or CVC syllables if the consonants shared a phonetic feature than if they did not. Since the effect was present both for pairs sharing vowel (e.g., /bi,di/, /du,tu/, etc.) and for pairs not sharing vowel (e.g., /bi,du/, /di,tu/, etc.), and since the latter pairs differ markedly in the auditory patterns by which the shared features are conveyed, Studdert-Kennedy, Shankweiler, and Pisoni (1972) concluded that the effect had a phonetic rather than an auditory basis. In another experimental paradigm, Studdert-Kennedy et al. (1970b) presented CV syllables at various values of SOA and demonstrated dichotic backward masking. They attributed the masking to interruption of central processes of speech perception, but left the level at which the interruption occurred uncertain (cf. Kirstein, 1971, 1973; Porter, 1971; Berlin, Lowe-Bell, Cullen, Thompson, and Loovis, 1973; Darwin, 1971a).

Recently, Pisoni and McNabb (1974) have combined and elaborated the two paradigms in a dichotic feature-sharing study, varying both masks and SOA. Their targets were /ba,pa,da,ta/; their masks were /ga,ka,gæ,kæ,gɛ,kɛ/. If target and mask consonants shared voicing, little or no masking was observed. If they did not share voicing, masking of the target consonant increased both as the masking-syllable vowel approached target-syllable vowel from /ɛ/ through /æ/ to /ɑ/, and as the mask intensity increased. In other words, identification of the target consonant was facilitated by similarity of the masking consonant, but, in the absence of facilitation, was impeded by similarity of the masking vowel, particularly if the vowel was of relatively high intensity. In a theoretical discussion of these results, Pisoni (in press) concludes that masking and facilitation occur at different stages of the perceptual process: masking reflects integration (rather than interruption) at the auditory level, while facilitation reflects integration at the phonetic level.

However, these results are also open to a purely auditory interpretation. They seem, in fact, to be consistent with a system that extracts the acoustic correlates of voice onset time separately for each vowel context (cf. Cooper, 1974b). We are thus led to consider the possible role of discrete acoustic feature analyzing systems, tuned to speech. This has proved among the most fruitful approaches to analysis of early processing, but we defer discussion to a later section (see below, Feature-Analyzing Systems).

The Acoustic Syllable

We have now touched on some half dozen paradigms--categorical perception, backward masking, short-term memory, reaction time studies, and others--in which consonant and vowel perception differ. As a final example, we may mention dichotic experiments [Berlin (in Lass, in press)]. Shankweiler and Studdert-Kennedy (1967; also Studdert-Kennedy and Shankweiler, 1970) showed a significant right-ear advantage for dichotically presented CV or CVC syllables differing in their initial or final consonants, but little for steady-state vowels or CVC syllables differing in their vowels. Day and Vigorito (1973) and Cutting (in press-b) reported a hierarchy of ear advantages in dichotic listening from a

right-ear advantage for stop consonants through liquids to a null or small left-ear advantage for vowels. Recently, Weiss and House (1973) have demonstrated that a right-ear advantage emerges for vowels, if they are presented at suitably unfavorable signal-to-noise ratios, while Godfrey (1974) has shown that the right-ear advantage for vowels may be increased by adding noise, reducing duration, or using a more confusable set of vowels (cf. Darwin and Baddeley, 1974).

The pattern is familiar. In virtually every instance, a consonant-vowel difference can be reduced or eliminated by taxing the listener's auditory access to the vowel, or by sensitizing his auditory access to the consonant. These qualifications only serve to emphasize the contrast between them, and to pinpoint its source in their acoustic structure. The consonant is transient, low in energy, and spectrally diffuse; the vowel is relatively stable, high in energy, and spectrally compact.

Together they form the syllable, each fulfilling within it some necessary function. Consider, first, vowel duration. Long duration is not necessary for recognition. We can identify a vowel quite accurately and very rapidly from little more than one or two glottal pulses, lasting 10 to 20 msec. Yet in running speech, vowels last ten to twenty times as long. The increased length may be segmentally redundant, but it permits the speaker to display other useful information: variations in fundamental frequency, duration, and intensity within and across vowels offer possible contrasts in stress and intonation, and increase the potential phonetic range (as in tone languages). Of course, these gains also reduce the rate at which segmental information can be transferred, increase the duration of auditory store, and open the vowel to contextual effects--the more so, the larger the phonetic repertoire. A language built on vowels, like a language of cries, would be limited and cumbersome.

Adding consonantal "attack" to the vowel inserts a segment of acoustic contrast between the vowels, reduces vowel context effects, and increases phonetic range. The attack, itself part of the vowel [the two produced by "...a single ballistic movement" (Stetson, 1952:4)], is brief, and so increases the rate of information transfer. Despite its brevity, the attack has a pattern arrayed in time, and the full duration of its trajectory into the vowel is required to display the pattern. To compute its phonetic identity, time is needed, and this is provided by the segmentally redundant vowel. Vowels are the rests between consonants.

Finally, rapid consonantal gestures cannot carry the melody and dynamics of the voice. The segmental and suprasegmental loads are therefore divided over consonant and vowel: the first, with its poor auditory store, taking the bulk of the segmental load; the second taking the suprasegmental load. There emerges the syllable, a symbiosis of consonant and vowel, a structure shaped by the articulatory and auditory capacities of its user, fitted to, defining, and making possible linguistic and paralinguistic communication.

SPECIALIZED NEURAL PROCESSES

Cerebral Lateralization

That the left cerebral hemisphere is, in most persons, specialized for language functions is among the most firmly established findings of modern neurology. That one of those functions may be to decode the peculiar acoustic

structure of the syllable into its phonetic components was first suggested by the results of dichotic studies. Kimura (1961a, 1961b, 1967) discovered that if different digit triads were presented simultaneously to opposite ears, those presented to the right ear were more accurately recalled than those presented to the left. She attributed the effect to functional prepotency of contralateral pathways under dichotic competition, and to left-hemisphere specialization for language functions. Later experiments have amply supported her interpretation.

Shankweiler and Studdert-Kennedy (1967) applied the technique to analysis of speech perception. They demonstrated a significant right-ear advantage for single pairs of nonsense syllables differing only in initial or final stop consonant, and separable advantages for place of articulation and voicing (Studdert-Kennedy and Shankweiler, 1970; cf. Halwes, 1969; Darwin, 1969; Haggard, 1971). Among the questions raised by these studies was whether the left hemisphere was specialized only for phonetic analysis, or also for extraction of speech-related acoustic properties, such as voice onset, formant structure, temporal relations among portions of the signal, and so on. We will not rehearse the argument here, but simply state the conclusion that "while the auditory system common to both hemispheres is equipped to extract the auditory parameters of a speech signal, the dominant hemisphere may be specialized for the extraction of linguistic features from those parameters" (Studdert-Kennedy and Shankweiler, 1970:594).

Striking evidence in support of this conclusion has recently been gathered by Wood (1975) and Wood, Goff, and Day (1971). This work deserves careful study, as an exemplary instance of the use of electroencephalography (EEG) in the study of language-related neurophysiological processes. Wood synthesized two CV syllables, /ba/ and /ga/, each at two fundamental frequencies, 104 Hz (low) and 140 Hz (high). From these syllables he constructed two types of random test order: in one, items differed only in pitch [e.g., /ba/ (low) vs /ba/ (high)]; in the other, they differed only in phonetic class [e.g., /ba/ (low) vs /ga/ (low)]. Subjects were asked to identify either the pitch or the phonetic class of the test items with reaction-time buttons. While they did so, evoked potentials were recorded from a temporal and a central location over each hemisphere. Records from each location were averaged and compared for the two types of test. Notice that both tests contained an identical item [e.g., /ba/ (low)], identified on the same button by the same finger. Since cross-test comparisons were made only between EEG records for identical items, the only possible source of differences in the records was in the task being performed, auditory (pitch) or phonetic. Results showed highly significant differences between records for the two tasks at both left-hemisphere locations, but at neither of the right-hemisphere locations. A control experiment, in which the "phonetic" task was to identify isolated initial formant transitions (50 msec), revealed no significant differences at either location over either hemisphere. Since these transitions carry all acoustic information by which the full syllables are phonetically distinguished, and yet are not recognizable as speech, we may conclude that the original left-hemisphere differences arose during phonetic, rather than auditory, analysis. We will discuss the adequacy of isolated formant transitions as control patterns in the next section. However, the entire set of experiments strongly suggests that different neural processes go on during phonetic, as opposed to auditory, perception in the left hemisphere, but not in the right hemisphere (cf. Molfese, 1972).

The distinctive processes of speech perception would seem then to lie in linguistic rather than acoustic analysis. Two other types of evidence suggest

the same conclusion. First, visual studies have repeatedly shown a right-field (left-hemisphere) advantage for tachistoscopically presented letters and, by contrast, a left-field (right-hemisphere) advantage for nonlinguistic geometric forms (for a review, see Kimura and Durnford, 1974). Second, Papçun, Krashen, Terbeek, Remington, and Harshman (1974) and Krashen (1972) have shown a right-ear advantage in experienced Morse code operators for dichotically presented Morse code words and letters. If the arbitrary patterns of both a visual and an auditory alphabet can engage left-hemisphere mechanisms, there might seem to be little ground for claiming special status for the speech signal.

However, alphabets are secondary, and while their interpretation may well engage specialized linguistic mechanisms, analysis of their arbitrary signal patterns clearly should not. The speech signal, on the other hand, is primary, its acoustic pattern at once the natural realization of phonological system and the necessary source of phonetic percept. Given its special status and peculiar structure, we should perhaps be surprised less if there were, than if there were not, specialized mechanisms adapted to its auditory analysis.

Hints of such processes have begun to appear. Halperin, Nachshon, and Carmon (1973), for example, showed a shift from left-ear advantage to right-ear advantage for dichotically presented tone sequences as a function of the number of alternations in the sequence. Their stimuli were patterned permutations of brief (200 msec) tone bursts, presumably not unlike those of Papçun et al. (1974), who showed a right-ear advantage in naive subjects for Morse code patterns up to seven units in length. Both studies suggest left-hemisphere specialization for assessing the sort of temporal relations important in speech. Both studies suffer from having called upon subjects to label the patterns, a process that might well invoke left-hemisphere mechanisms.

This weakness is avoided in recent work by Cutting (in press-b). He synthesized two normal CV syllables, /ba/ and /da/, and two phonetically impossible "syllables" identical with the former except that their first formant transitions fell rather than rose along the frequency scale, so that they were not recognized as speech. In a nonlabeling dichotic task, subjects gave equal right-ear advantages for both types of stimulus. The outcome suggests a left-hemisphere mechanism for extraction of formant transitions and is reminiscent of a study by Darwin (1971b), who found a right-ear advantage for synthetic fricatives when formant transitions from fricative noise into vowel were included, but no ear advantage when transitions were excluded.

There are, then, grounds for believing that the left hemisphere is specialized not only for phonetic interpretation of an auditory input, but also for extraction of auditory information from the acoustic signal. The evidence is tenuous, but systematic study of feature-analyzing systems--whether lateralized or not remains to be seen (cf. Ades, 1974a)--has opened up a new range of possibilities.

Feature-Analyzing Systems

Neurophysiological systems of feature detectors, selectively responsive to light patterns, were first reported by Lettvin, Maturana, McCulloch, and Pitts (1959). They found receptive fields in the visual ganglion cells of frog that responded, under specific conditions, to movement. The biological utility of the system to an animal that preys on flies is obvious. Moving up the nervous

system, and the evolutionary scale, Hubel and Wiesel (1962) reported yet more complex detectors: single cells in the visual cortex of cat that responded selectively to the orientation of lines, to edges, and to movement in a certain direction. Since then, work on visual feature-detecting systems has proliferated (see Julesz, 1971:58-68, for a review).

Complex auditory feature detectors in the cortex of cat were reported by Evans and Whitfield (1964): single cells responsive to specific gradients of intensity change, and others ("miaow" cells) to the rate and direction of frequency change (Whitfield and Evans, 1965). Similar cells were reported by Nelson, Erulkar, and Bryan (1966) in the inferior colliculus of cat. Other research has borne directly on acoustic signaling systems. Frishkopf and Goldstein (1963) and Capranica (1965) reported single units in the auditory nerve of bullfrog responsive only to the male bullfrog's mating call. Recently, Wollberg and Newman (1972) have described single cells in the auditory cortex of squirrel monkey which answer to that species' "isolation peep." Stimulus and response were isomorphic: presentation of the "peep" with portions gated out yielded a response in which corresponding portions were absent. Furthermore, the remaining portions were no longer normal: if a central portion of the signal was missing, the response pattern to the final portion changed. Interaction of this kind is particularly interesting in light of the contextually variant cues of speech, for which interpretation may demand details of a complete pattern, such as the syllable.

The relevance of all this to speech has not gone unnoticed. The possible role of feature-detecting systems in speech perception was scouted briefly by Liberman et al. (1967), by Studdert-Kennedy (1974), and, at considerable length, by Abbs and Sussman (1971). However, advance awaited a telling experimental procedure. This was found in "adaptation" studies, a method with a long history in visual research (Woodworth and Schlosberg, 1954). The paradigm is simple enough. For example, after prolonged fixation of a line curved from the median plane, a vertical line, presented as a test stimulus, appears curved in the opposite direction: there is a "figural after-effect" in which portions of the image are displaced (Köhler and Wallach, 1944). Related effects in color and tilt also occur. While none of these effects is understood in any detail, they are frequently interpreted in terms of specific receptors or of feature-analyzing systems. Prolonged stimulation "fatigues" or "adapts" one system, and relatively "sensitizes" a physically adjacent or related (perhaps opponent) system. On this interpretation, to demonstrate perceptual shifts upon prolonged exposure to a particular physical (or psychological) "feature" is to demonstrate the presence of analyzing systems for that feature, and its relative.

The method was first used by Warren and Gregory [1958; see also Warren, 1968 (in Lass, in press); Perl, 1970; Clegg, 1971; Lass and Golden, 1971; Lass and Gasperini, 1973; Lass, West, and Taft, 1973; Obusek and Warren 1973], yielding an effect that they termed "verbal transformation." Subjects listen to a meaningful word played repeatedly once or twice per second for several minutes, and are asked to report any changes in the word that they hear. They report a large number of transformations, usually meaningful words and not always closely related phonetically to the original. However, Goldstein and Lackner (in press) refined the method by using nonsense syllables to reduce semantic influence (CV, V, VC) and by presenting them monaurally. They analyzed transforms phonetically, and showed that each was confined to a single phone, usually on one or two distinctive features (as defined by Chomsky and Halle, 1968), and were independent

of their syllabic context. Furthermore, the right ear gave significantly more transforms than the left ear on consonants, but not on vowels, and the transforms followed the phonological constraints of English. These last two points are among the arguments that the authors present for suspecting that the effects result from adaptation of phonetic, rather than auditory, analyzing systems.

In a further refinement, Lackner and Goldstein (in press) used a natural CV syllable, repeated monaurally 36 times in 30 sec, and a final test item presented to either the same or the opposite ear. Both adapting and test items were drawn from the set of six English stop consonants, followed by the same vowel (either /i/ or /e/). Subjects reported the last adapting item and the test item. Transforms in the test item occurred on both cross-ear (30 percent) and same-ear (40 percent) trials. They were significantly more likely to occur if the final adapting item was also transformed, and to be on the same feature(s) (place and/or voice) as the adapting item transform, a result that again hints at phonetic feature-detecting systems. The authors conclude from the cross-ear trials that adaptation is central, rather than peripheral, but, unable in this study to distinguish phonetic effects from the acoustic effects that underlie them, they withhold judgment on whether the transforms are auditory or phonetic.

This last is, of course, the crucial question. It can be approached only by use of synthetic speech in which acoustic features can be specified precisely and, within limits, manipulated independently of phonetic category. Eimas, working independently of the previous authors, took this step in a series of experiments growing out of his work on infants (discussed below), and has concluded that the effect is phonetic. We will consider his work in some detail because it introduced a fruitful paradigm that has already been put to good use by others.

In the first experiment (Eimas and Corbit, 1973), the authors used two voice-voiceless series synthesized along the VOT continuum, one from /ba/ to /pa/, the other from /da/ to /ta/ (Lisker and Abramson, 1964). On the assumption of two voicing detectors, each differentially sensitive to VOT values that lie clearly within its phonetic category, and both equally sensitive to a VOT value at the phonetic boundary, the authors reasoned that adaptation with an acoustically extreme token of one phonetic type should desensitize its detector, and relatively sensitize (a metaphor, not an hypothesis) its opponent detector, to boundary values of VOT, with a resulting displacement of the identification function toward the adapting stimulus. They, therefore, collected unadapted and adapted functions for both labial and alveolar series. The adapting stimuli were drawn from the extremes of both series, and their effects were tested within and across series. Figure 5 shows the results for one of their three subjects (the experiment is taxing and prohibits large samples, but the other two subjects gave similar functions). The predicted results obtain. Furthermore, the effect is only slightly weaker across series than within. (This result was replicated in an experiment, briefly reported in their next paper, for which they used eight subjects to demonstrate boundary shifts on alveolar and velar stop consonant VOT continua after adaptation with labial stops.) In a supporting experiment the authors showed that, following adaptation, the peak in an ABX discrimination function is neatly shifted to coincide with the adapted phonetic boundary (cf. Cooper, 1974a).

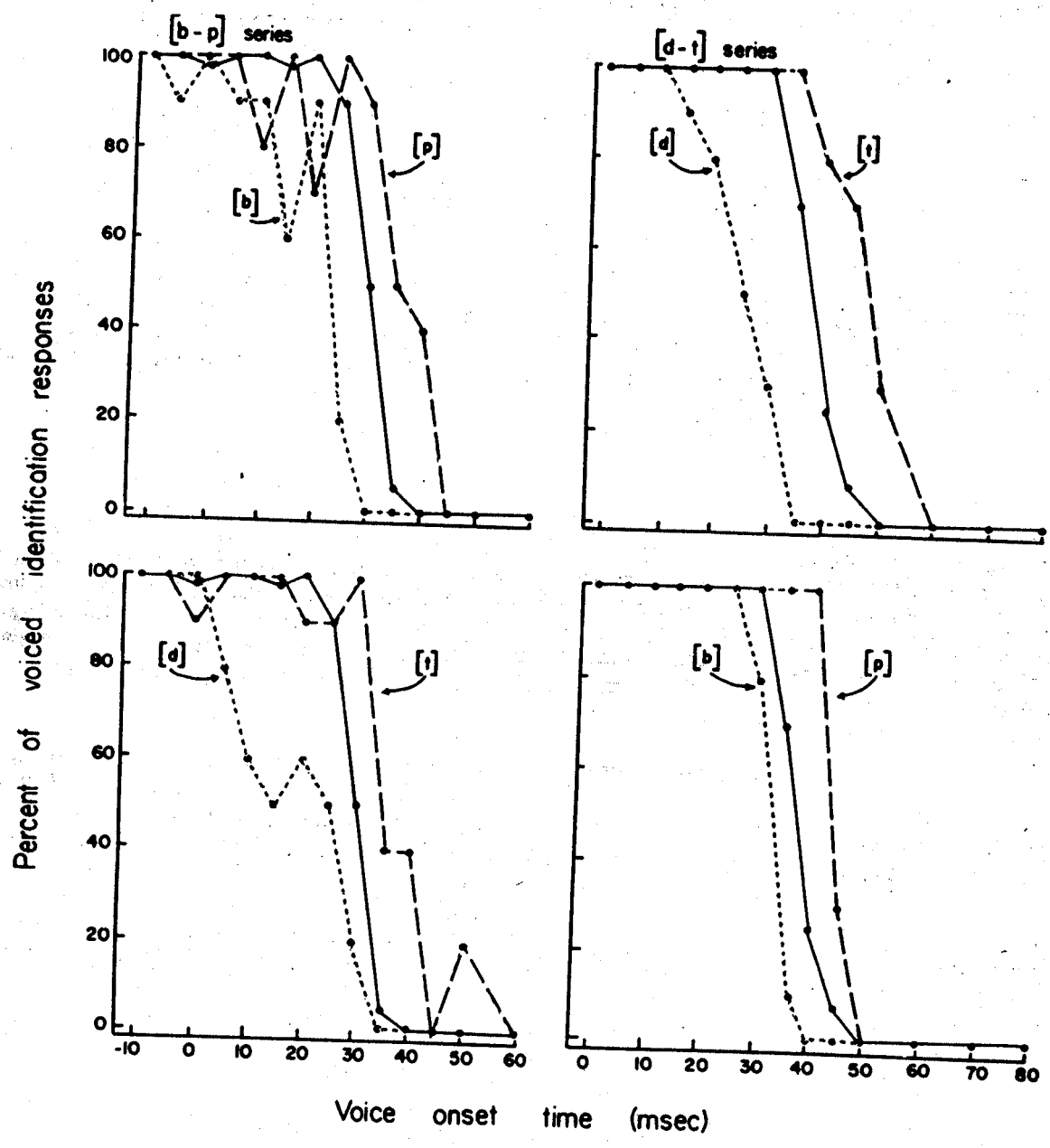

Figure 5: Percentages of voiced identification responses ([b or d]) obtained with and without adaptation, for a single subject. The functions for the [b,p] series are on the left and those for the [d,t] series are on the right. The solid lines indicate the unadapted identification functions; the dotted and dashed lines indicate the identification functions after adaptation. The phonetic symbols indicate the adapting stimulus. [From Eimas and Corbit (1973) with permission of authors and publishers.]

In a second study (Eimas, Cooper, and Corbit, 1973), the authors report three experiments. The first demonstrates that the site of the adaptation effect is probably central rather than peripheral: it obtains as strongly when the adapting stimulus is presented to one ear and the test stimulus to the other, as when both are presented binaurally (cf. Ades, 1974a). The second demonstrates that the effect is not obtained if the adapting stimulus is simply the first 50 msec of the syllable /da/, an acoustic pattern that contains all the voicing information, but is not heard as speech. The third experiment assesses the relative strengths of the two hypothesized detectors, finding that, as in the first study (see Figure 5), voiced stops tend to be more resistant to adaptation (yield smaller boundary shifts) than voiceless. The result encourages the hypothesis of separate detectors for each phonetic value along an acoustic continuum, a notion with obvious relevance to categorical perception. Additional support comes from the work of Cooper (1974a), who found evidence of three distinct detectors along a /b-d-g/ continuum: adaptation with /b/ shifted only the /b-d/ boundary, adaptation with /g/ shifted only the /d-g/ boundary, adaptation with /d/ shifted both neighboring boundaries.

Let us remark first the striking achievement of these studies. Whatever the underlying mechanism, Eimas and his colleagues have demonstrated in a novel, direct, and peculiarly convincing manner the operation of some form of feature-analyzing system in speech perception. The outcome was not foregone. There might, after all, have been no adaptation effect at all. Alternatively, the effect might have been on the whole syllable or on the unanalyzed phonemic segment. But these possibilities were ruled out by the cross-series results. The effect proved to be on a feature within the phonemic segment, and so has provided the strongest evidence to date of a physiologically grounded feature system (cf. Cooper and Blumstein, 1974).

What now is the evidence for phonetic rather than auditory adaptation? First, the cross-series effect: phonetic tokens drawn from labial, alveolar, or velar VOT continua differ acoustically in the extent and direction of their second and third formant transitions, yet they are mutually effective adaptors. If the effect were acoustic, the argument runs, the acoustic differences should eliminate the effect. Note, however, that the differences were in acoustic cues to place of articulation, while the feature being tested was voice onset time. The cues to this feature are complex and, as we have seen, relational. Furthermore, Cooper (1974b) has recently shown that VOT adaptation may be vowel-specific: simultaneous adaptation with [da] and [t^hi] produced opposite shifts on [ba-p^ha] and [bi-p^hi] series. Nonetheless, if outputs from such detectors funneled into acoustic analyzers, tuned to presence or absence of energy in the region of the first formant at syllable onset, we would expect precisely the results that were obtained (cf. Stevens and Klatt, 1974).

The second piece of evidence is the failure of the truncated /da/, not heard as speech, to "sensitize" the supposed /ta/ detector. Here the main problem is the status of the truncated /da/ as a control (cf. Wood, 1975). There are two possible types of design that may throw light on the auditory-phonetic issue. In one, control and test items are acoustically identical (on dimensions relevant to the phonetic dimension under test), but phonetically distinct; in the other, they are acoustically distinct, but phonetically identical. The first design, chosen by Eimas and his colleagues, may yield ambiguous results. If adaptation with the control item shifts the phonetic boundary, we have evidence for the existence of auditory detectors tuned to acoustic features of speech.

Precisely this outcome has, in fact, been reported by Ades (1973), using the first 38 msec of the extreme test stimuli to shift the /bæ/-/dæ/ boundary. If, on the other hand, the control item does not shift the boundary, the outcome is ambiguous. It may mean, as Eimas and his colleagues concluded, that the hypothetical detector is phonetic. But it may also mean that an acoustic detector tuned to features of speech is only adapted if stimulated by a complete (i.e., phonetically identifiable) signal (cf. Wollberg and Newman, 1972). It is not, after all, implausible to suppose that the human cortex contains sets of acoustic detectors tuned to speech and capable of mutual inhibition. Each detector may respond to a particular acoustic property, but may be inhibited from output to the phonetic system in the absence of a collateral response in other detectors. The auditory system would then be immune to adaptation by an incomplete signal.

The second type of design calls for control and test items that are acoustically distinct (on dimensions relevant to the phonetic dimension under study), but phonetically identical. This design rests, of course, on the fact that the speech signal may carry several acoustic cues, each a more or less effective determinant of a particular phonetic percept. The procedure is then to synthesize two acoustic continua, manipulating in each a different acoustic cue to the same phonetic distinction. If now the two series are mutually effective in shifting the phonetic boundaries of the other, we have some preliminary support for the hypothetical phonetic detector. This was the outcome of studies by Ades (1974b), Bailey (1973), and Cooper (1974a), all of whom demonstrated cross-series adaptations for /b-d/ continua with different vowels. The use of different vowels meant that formant transitions cueing a given phonetic type could be falling in one token (e.g., /dæ/), rising in another (e.g., /de/). Thus, adaptation of simple acoustic detectors responsive only to rising or only to falling formants (cf. Whitfield and Evans, 1965) was ruled out. Of course, a more complex "acoustic invariance," derived from some weighted ratio of F2 and F3 transitions, might be posited (Cooper, 1974a). But the conclusion that the detectors are phonetic was tempting enough for both Ades and Cooper to draw. Ades qualified his conclusion because, in a previous experiment (Ades, 1974b), he had found no cross-series adaptation of CV and VC continua (/bæ-dæ/, /æb-æd/): the phonetic detector, unlike phonetic listeners and phonological theory, evidently distinguishes between initial and final allophones. A funnel into a second level of phonetic analysis, possibly the point of contact with an abstract generative system, would be needed to account for the listener's inability to make this distinction.

For Bailey (1973), the phonetic conclusion was less compelling. He pointed to spectral overlap in the transitions of his two series, and suggested an acoustic system involving "...some generalizing balanced detectors of positive and negative transitions" (p. 31) (cf. Cooper, 1974a). To test for the effect of spectral overlap, he constructed two /ba-da/ series, one with a fixed F2 and all place cues in F3, the other with no F3 and all place cues in F2. This, by far the most stringent version of the phonetically identical-acoustically distinct design, yielded cross-adaptation from the F2 cues series to the fixed F2, but none from the F3 cues series to no F3. This argues strongly for auditory adaptation, and Bailey concluded that the system contains "...central feature extractors which process the phonetically relevant descriptors of spectral patterns" (p. 34).

Clearly, the issue of auditory versus phonetic detectors is not resolved. But let us consider implications of each possible resolution for speech perception theory and research. First, if discrete auditory detectors are being isolated by the adaptation technique, we may be in a position to begin more precise definition of the acoustic correlates of distinctive feature systems, ultimately essential if phonological theory is to be given a physical and physiological base. To the extent that this proved possible, we could be isolating invariants in the speech signal, thus aligning speech perception with that of other "natural categories," such as those of color and form (Rosch, 1973). But it is not inevitable that acoustic features be invariant correlates of phonetic features: both the work of Ades (1974b) on initial and final stop consonants and the work of Cooper (1974b) on vowel-specific VOT analyzers suggest that invariance may lie at some remove from the signal. And, in either event, to isolate acoustic features is not to define them phonetically, nor to explain how they are gathered from syllables of the signal into phonemes, each with its peculiar, nonarbitrary name: the auditory to phonetic transformation would remain obscure.

If, on the other hand, the adaptation technique isolates discrete phonetic detectors, its unequivocal achievement will have been to undergird the psychological and physiological reality of features in speech perception. Salutary though this may be for those of little faith, the outcome would be disappointing for research. For again, the process by which these features are drawn from the acoustic display and granted phonetic dimension will be hidden. To analysis of the analyzer a new technique must then be brought.

Finally, we should not discount the possibility that the auditory-phonetic distinction is misleading in this context, and that the adapted systems are both auditory and phonetic. If, for example, the output from auditory analyzers tuned to speech funneled directly into phonetic processors so that adaptation of one set entailed adaptation of the other, a convincing separation of the two would be difficult to demonstrate. Precisely, this is suggested by recent evidence (Cooper, in press-a, in press-b) that each system can be adapted selectively, yet is intimately related to the other. The closeness of the relation is revealed by Cooper's (1974c) extension of the adaptation technique to the study of relations between perceptual and motor aspects of speech. He has shown that adaptation on a [bi-pi] continuum yields not only shifts in the perceptual boundary, but correlated shifts in subjects' characteristic VOT values in production. If his findings are replicable, we have here clear evidence for the frequently hypothesized link between perception and production, and one that may supersede the auditory-phonetic distinctions we have been attempting to establish for these adaptation studies. To the origin of this link in the processes of language acquisition we turn in the final section.

From Acoustic Feature to Phonetic Percept

As we have seen, template-matching models of speech perception are not in good standing. Faced with gross acoustic variations as a function of phonetic context, rate, stress, and individual speaker, theorists have had recourse to motor, or analysis-by-synthesis, accounts of speech perception: they have sought invariance in the articulatory control system. Nonetheless, there are grounds for believing that some form of template-matching may operate in both speaking and listening, and there are more fundamental grounds than lack of acoustic invariance for positing a link between production and perception.

Consider the infant learning to speak. Several writers (e.g., Stevens, 1973; Mattingly, 1973) have pointed out that the infant must be equipped with some mechanism by which it plucks from the stream of speech just those acoustic

cues that convey the phonetic distinctions it will eventually learn to perceive and articulate. This fact motivates, in part, Stevens' (1973) pursuit of acoustic invariants and his hypothesized property detectors. Evidence for the existence of such detectors comes from the work of Eimas and his colleagues (Eimas, Siqueland, Jusczyk, and Vigorito, 1971; Eimas, 1974; for a review, see Cutting and Eimas, in press). They have investigated the capacity of infants as young as one month to discriminate synthetic speech sounds. We will not describe their method in detail, but broadly, it employs operant conditioning, a synthetic speech continuum, an adapting stimulus, and a test item. The results are reliable and striking: infants discriminate between pairs of stimuli drawn from different adult phonetic categories, but not between pairs drawn from the same phonetic category. The effect has been repeatedly demonstrated on both voicing and place of articulation continua (cf. Moffitt, 1971; Morse, 1972). Furthermore, the effect is absent for truncated control syllables, not heard by adults as speech, exactly as in the adult adaptation studies. Eimas and his colleagues interpret the effect as evidence for the operation of phonetic feature detectors, presumably innate. Unfortunately, the outcome is ambiguous for the same reasons as is the adult outcome: there is no way of assuring that the adapted detectors are phonetic rather than auditory (see Cutting and Eimas, 1974, for further discussion of this point). The more cautious, and perhaps more plausible, view is that they are auditory (cf. Stevens and Klatt, 1974:657-658).

We are then faced with two questions. First, do the acoustic features extracted by such detector systems bear an invariant relation to phonetic features? This is an empirical question and we will say no more here than that given the inconstancy of the speech signal, it is unlikely that they do. Second, and more importantly, how does the infant "know" that the extracted properties are speech? This, of course, is simply another version of the question: how are we to define the phonetic percept? But, asked in this form, an answer immediately suggests itself: the infant learns that sounds are speech by discovering that it can make them with its own vocal apparatus.

Before elaborating this point, let us consider the work of Marler (1970, in press). He has proposed a general model of the evolution of vocal learning, based on studies of the ontogenesis of male "song" in certain sparrows (see also Marler and Mundinger, 1971). Briefly, the hypothesis is that development of motor song-pattern is guided by sensory feedback matched to modifiable, innate auditory templates (cf. Mattingly, 1972). Marler describes three classes of birds. The first (for example, the dove or the chicken) needs to hear neither an external model nor its own voice for song to emerge: crowing and cooing develop normally, if the birds are reared in isolation and even if they are deafened shortly after birth. The second (for example, the song sparrow) needs no external model, but does need to hear its own voice: if reared in isolation, song develops normally, unless the bird is deafened in early life, in which case song is highly abnormal and insect-like.

An example of the third class of bird is the white-crowned sparrow, which needs both an external model and the sound of its own voice. Reared in isolation, the white-crown develops an abnormal song with "...certain natural characteristics, particularly the sustained pure tones which are one basic element in the natural song" (Marler and Mundinger, 1971:429). If the bird is deafened in early life, even this rudimentary song does not develop. There emerges instead a highly abnormal song "...rather like that of a deafened song sparrow...perhaps the basic output of the syringeal apparatus with a passive flow of air through

it" (Marler, in press). However, reared in isolation, but exposed to recordings of normal male song during a critical period (10-50 days after birth), the male (and the female, if injected with male hormone) develops normal song some 50 or more days after exposure. Exposure to the songs of other species will not serve, and deafening either before or after exposure to conspecific song prevents normal development [Konishi (1965), cited by Marler, in press].

Marler (in press) proposes that the rudimentary song of the undeafened, isolated white-crown reflects the existence of an auditory template, "...lying in the auditory pathway, embodying information about the structure of vocal sounds." The template matches certain features of normal song, and serves to guide development of the rudimentary song, as well as to "...focus...attention on an appropriate class of external models." Exposure to these models modifies and enriches the template, which then serves to guide normal development, through subsong and plastic song, as the bird gradually discovers the motor controls needed to match its output with the modified template. [Several studies have reported evidence for the "tuning" by experience of visual detecting systems in cat (Hirsch and Spinelli, 1970; Blakemore and Cooper, 1970; Pettigrew and Freeman, 1973) and man (Annis and Frost, in press), and of auditory detecting systems in rhesus monkey (Miller, Sutton, Pfingst, Ryan, and Beaton, 1972).]

Marler (in press) draws the analogy with language learning. He suggests that sensory control of ontogenetic motor development may have been the evolutionary change that made possible an elaborate communicative system as pivot of avian and human social organization. He argues that "new sensory mechanisms for processing speech sounds, applied first, in infancy, to analyzing sounds of others, and somewhat later in life to analysis of the child's own sounds, was a significant step toward achieving the strategy of speech development of Homo sapiens." On the motor side, he points out, vocal development must have become dependent on auditory feedback, and there must have developed "neural circuitry necessary to modify patterns of motor outflow so that sounds generated can be matched to preestablished auditory templates."

Certainly, human and avian parallels are striking. Deafened at birth, the human infant does not learn to speak: babbling begins normally, but dies away around the sixth month (Marvilya, 1972). Whether this is because the infant has been deprived of the sound of its own voice, of an external model, or of both, we do not know. But there does seem to be an (ill-defined) critical period during which exposure to speech is a necessary condition of normal development (Lenneberg, 1967; but see Fromkin, Krashen, Curtiss, Rigler, and Rigler, 1974). And the work of Eimas and his colleague has demonstrated the sensitivity of the infant to functionally important acoustic features of the speech signal. At least one of these features, the short VOT lag associated with stops in many languages (Lisker and Abramson, 1964), is known to be among the first to appear in infant babble (Kewley-Port and Preston, 1974). Finally, Sussman (1971) and his colleagues (Sussman, MacNeilage, and Lumbley, 1974; Sussman and MacNeilage, in press) have reported evidence for a speech-related auditory sensorimotor mechanism that may serve to modify patterns of motor outflow, so as to match sounds generated by the vocal mechanism against some standard. In short, Marler's account is consistent with a good deal of our limited knowledge of speech development. Its virtue is to emphasize sensorimotor interaction and to accord the infant a mechanism for discovering auditory-articulatory correspondences.

Paradoxically, if we are to draw on this account of motor development for insight into perceptual development, we must place more emphasis on the relatively rich articulatory patterns revealed in early infant babble. The infant is not born without articulatory potential. In fact, the work of Lieberman and his colleagues would suggest quite specific capacities (Lieberman, 1968, 1972, 1973; Lieberman and Crelin, 1971; Lieberman, Harris, Wolff, and Russell, 1971; Lieberman, Crelin, and Klatt, 1972). They have developed systematic evidence for evolution of the human vocal tract from a form with a relatively high larynx, opening almost directly into the oral cavity, capable of producing a limited set of schwa-like vowel sounds, to a form with a lowered larynx, a large pharyngeal cavity, and a right-angle bend in the supralaryngeal vocal tract, capable of producing the full array of human vowels. Lieberman (1973) argues that this development, taken with many other factors, including the capacity to encode and decode syllables, paved the way for development of language. Associated with changes in morphology must have come neurological changes to permit increasingly fine motor control of breathing and articulation, including in all likelihood, cerebral lateralization (cf. Lenneberg, 1967; Geschwind and Levitsky, 1968; Nottebohm, 1971, 1972). The outcome of these developments would have been a range of articulatory possibilities as determinate in their form as the patterns of manual praxis that gave rise to toolmaking. The inchoate forms of these patterns might then emerge in infant babble under the control of rudimentary articulatory templates.

In short, we hypothesize that the infant is born with both auditory and articulatory templates. Each embodies capacities that may be modified by, and deployed in, the particular language to which the infant is exposed. Presumably, these templates evolved more or less pari passu and are matched, in some sense, as key to lock. But they differ in their degree of specificity. For effective function in language acquisition the auditory template must be tuned to specific acoustic properties of speech. The articulatory template, on the other hand, is more abstract, a range of gestural control, potentially isomorphic with the segmented feature matrix of the language by which it is modified (cf. Chomsky and Halle, 1968:294).

Among the grounds for this statement are the results of several studies of adult speech production. Lindblom and Sundberg (1971), for example, found that, if subjects were thwarted in their habitual articulatory gestures by the presence of a bite block between their front teeth, they were nonetheless able to approximate normal vowel quality, even within the first pitch period of the utterance. Bell-Berti (1975) has shown that the pattern of electromyographic potentials associated with pharyngeal enlargement during medial voiced stop consonant closure varies from individual to individual and from time to time within an individual. Finally, Ladefoged, DeClerk, Lindau, and Papçun (1972) have demonstrated that different speakers of the same dialect may use different patterns of tongue height and tongue root advancement to achieve phonetically identical vowels. They do not report formant frequencies for their six speakers, so that the degree of acoustic variability associated with the varied vocal-tract shapes is not known. But since individuals obviously differ in the precise dimensions of their vocal tracts, it would be surprising if they accomplished a particular gesture and a particular acoustic pattern by precisely the same pattern of muscular action. In short, it seems likely that both infant and adult articulatory templates are control systems for a range of functionally equivalent vocal tract shapes rather than for specific patterns of muscular action. In fact, it is

precisely to exploration of its own vocal tract and to discovery of its own patterns of muscular action that the infant's motor learning must be directed.

We should emphasize that neither template can fulfull its communicative function in the absence of the other. Modified and enriched by experience, the auditory template may provide a "description" of the acoustic properties of the signal, but the description can be no different in principle than that provided by any other form of spectral analysis: alone, the output of auditory analysis is void. Similarly, babble without auditory feedback has no meaning. The infant discovers phonetic "meaning" (and linguistic function) by discovering auditory-articulatory correspondences, that is, by discovering the commands required by its own vocal tract to match the output of its auditory template. Since the articulatory template is relatively abstract, the infant will begin to discover these correspondences before it has acquired the detailed motor skills of articulation: perceptual skill will precede motor skill. In rare instances of peripheral articulatory pathology the infant (like the female white-crowned sparrow who learns the song without singing) may even discover language without speaking (cf. MacNeilage, Rootes, and Chase, 1967).

We hypothesize then that the infant is born with two distinct capacities, and that its task is to establish their links. Auditory feedback from its own vocalizations serves to modify the articulatory template, to guide motor development, and to establish the links. The process endows the communicatively empty outputs of auditory analysis and articulatory gesture with communicative significance. In due course the system serves to segment the acoustic signal and perhaps, as analysis-by-synthesis models propose, to resolve acoustic variability. But its prior and more fundamental function is to establish the "natural categories" of speech. To perceive these categories is to trace the sound patterns of speech to their articulatory source and recover the commands from which they arose. The phonetic percept is then the correlate of these commands.

REFERENCES

Abbs, J. H. and H. M. Sussman. (1971) Neurophysiological feature detectors and speech perception: A discussion of theoretical implications. J. Speech Hearing Res. 14, 23-36.

Abramson, A. S. and L. Lisker. (1965) Voice onset time in stop consonants: Acoustic analysis and synthesis. In Proceedings of the 5th International Congress of Acoustics, ed. by D. E. Commins. (Liege: Imp. G. Thone) A-51.

Abramson, A. S. and L. Lisker. (1970) Discriminability along the voicing continuum: Cross-language tests. In Proceedings of the 6th International Congress of Phonetic Science, Prague, 1967. (Prague: Academia) 569-573.

Abramson, A. S. and L. Lisker. (1973) Voice-timing perception in Spanish word-initial stops. J. Phonetics 1, 1-8.

Ades, A. E. (1973) Some effects of adaptation on speech perception. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 111, 121-129.

Ades, A. E. (1974a) Bilateral component in speech perception? J. Acoust. Soc. Amer. 56, 610-616.

Ades, A. E. (1974b) How phonetic is selective adaptation? Experiments on syllable position and vowel environment. Percept. Psychophys. 16, 61-66.

Annis, R. C. and B. Frost. (in press) Human visual ecology and orientation anisotropies in acuity. Science.

Atkinson, J. E. (1973) Aspects of intonation in speech: Implications from an experimental study of fundamental frequency. Unpublished Doctoral dissertation, University of Connecticut.

Bailey, P. (1973) Perceptual adaptation for acoustical features in speech. Speech Perception (Department of Psychology, The Queen's University of Belfast) Series 2, 2, 29-34.

Barclay, R. (1972) Noncategorical perception of a voiced stop. Percept. Psychophys. 11, 269-274.

Bell-Berti, F. (1975) Control of pharyngeal cavity size for English voiced and voiceless stops. J. Acoust. Soc. Amer. 57.

Berlin, C. I., S. S. Lowe-Bell, J. K. Cullen, C. L. Thompson, and C. F. Loovis. (1973) Dichotic speech perception: An interpretation of right-ear advantage and temporal offset effects. J. Acoust. Soc. Amer. 53, 699-709.

Bever, T. G. (1970) The influence of speech performance on linguistic structure. In Advances in Psycholinguistics, ed. by G. B. Flores D'Arcais and W. J. M. Levelt. (Amsterdam: North-Holland) 4-30.

Blakemore, C. and G. F. Cooper. (1970) Development of the brain depends on visual environment. Science 168, 477-478.

Blumstein, S. (1974) The use and theoretical implications of the dichotic technique for investigating distinctive features. Brain Lang. 4, 337-350.

Boomer, D. S. and J. D. M. Laver. (1968) Slips of the tongue. Brit. J. Dis. Communic. 3, 1-12.

Cairns, H. S., C. E. Cairns, and F. Williams. (1974) Some theoretical considerations of articulation substitution phenomena. Lang. Speech 17, 160-173.

Capranica, R. R. (1965) The Evoked Vocal Response of the Bullfrog. (Cambridge, Mass.: MIT Press).

Chomsky, N. (1972) Language and Mind, enlarged ed. (New York: Harcourt Brace Jovanovich).

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

Chomsky, N. and G. A. Miller. (1963) Introduction to the formal analysis of natural languages. In Handbook of Mathematical Psychology, ed. by R. D. Luce, R. R. Bush, and E. Galanter. (New York: Wiley) 269-321.

Clegg, J. M. (1971) Verbal transformations on repeated listening to some English consonants. Brit. J. Psychol. 62, 303-309.

Cole, R. A. (1973a) Listening for mispronunciations: A measure of what we hear during speech. Percept. Psychophys. 13, 153-156.

Cole, R. A. (1973b) Different memory functions for consonants and vowels. Cog. Psychol. 4, 39-54.

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

Cooper, F. S. (1972) How is language conveyed by speech? In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Cooper, W. E. (1974a) Adaptation of phonetic feature analyzers for place of articulation. J. Acoust. Soc. Amer. 56, 617-627.

Cooper, W. E. (1974b) Contingent feature analysis in speech perception. Percept. Psychophys. 16, 201-204.

Cooper, W. E. (1974c) Perceptuo-motor adaptation to a speech feature. Percept. Psychophys. 16, 229-234.

Cooper, W. E. (in press-a) Selective adaptation for acoustic cues of voicing in initial stops. J. Phonetics.

Cooper, W. E. (in press-b) Selective adaptation to speech. In Cognitive Theory, ed. by F. Restle, R. M. Shiffrin, J. N. Castellan, H. Lindman, and D. B. Pisoni. (Potomac, Md.: Erlbaum).

Cooper, W. E. and S. E. Blumstein. (1974) A "labial" feature analyzer in speech perception. Percept. Psychophys. 15, 591-600.

Crowder, R. G. (1971a) The sound of vowels and consonants in immediate memory. J. Verbal Learn. Verbal Behav. 10, 587-659.

Crowder, R. G. (1971b) Waiting for the stimulus suffix: Decay, delay, rhythm, and readout in immediate memory. Quart. J. Exp. Psychol. 23, 324-340.

Crowder, R. G. (1972) Visual and auditory memory. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Crowder, R. G. (1973) Precategorical acoustic storage for vowels of short and long duration. Percept. Psychophys. 13, 502-506.

Crowder, R. G. and J. Morton. (1969) Precategorical acoustic storage (PAS). Percept. Psychophys. 5, 365-373.

Cutting, J. E. (1973) Levels of processing in phonological fusion. Unpublished Doctoral dissertation, Yale University.

Cutting, J. E. (in press-a) Aspects of phonological fusion. Human Perception and Performance.

Cutting, J. E. (in press-b) Two left-hemisphere mechanisms in speech perception. Percept. Psychophys.

Cutting, J. E. and P. D. Eimas. (in press) Phonetic feature analyzers in the processing of speech by infants. In The Role of Speech in Language, ed. by J. F. Kavanagh and J. E. Cutting. (Cambridge, Mass.: MIT Press).

Cutting, J. E. and B. S. Rosner. (in press) Categories and boundaries in speech and music. Percept. Psychophys.

Darwin, C. J. (1969) Auditory perception and cerebral dominance. Unpublished Doctoral dissertation, University of Cambridge.

Darwin, C. J. (1971a) Dichotic backward masking of complex sounds. Quart. J. Exp. Psychol. 23, 386-392.

Darwin, C. J. (1971b) Ear differences in the recall of fricatives and vowels. Quart. J. Exp. Psychol. 23, 46-62.

Darwin, C. J. and A. D. Baddeley. (1974) Acoustic memory and the perception of speech. Cog. Psychol. 6, 41-60.

Day, R. S. (1968) Fusion in dichotic listening. Unpublished Doctoral dissertation, Stanford University.

Day, R. S. (1970a) Temporal order judgments in speech: Are individuals language-bound or stimulus-bound? Haskins Laboratories Status Report on Speech Research SR-21/22, 71-75.

Day, R. S. (1970b) Temporal order perception of reversible phoneme cluster. J. Acoust. Soc. Amer. 48, 95(A).

Day, R. S. and J. M. Vigorito. (1973) A parallel between encodedness and the ear advantage: Evidence from a temporal-order judgment task. J. Acoust. Soc. Amer. 53, 358(A).

Day, R. S. and C. C. Wood. (1972) Mutual interference between two linguistic dimensions of the same stimuli. Paper presented at the 83rd meeting of the Acoustical Society of America, April 18-21, Buffalo, N. Y.

Delattre, P. C., A. M. Liberman, F. S. Cooper, and L. J. Gerstman. (1952) An experimental study of the acoustic determinants of vowel color: Observations on one- and two-formant vowels synthesized from spectrographic patterns. Word 8, 195-210.

Dorman, M. (1974) Discrimination of intensity differences on formant transitions in and out of syllable context. Percept. Psychophys. 16, 84-86.

Dorman, M., D. Kewley-Port, S. Brady-Wood, and M. T. Turvey. (1973) Forward and backward masking of brief vowels. Haskins Laboratories Status Report on Speech Research SR-33, 93-100.

Dorman, M., D. Kewley-Port, S. Brady, and M. T. Turvey. (1974) Two processes in vowel perception: Inferences from studies of backward masking. Haskins Laboratories Status Report on Speech Research SR-37/38, 233-253.

Eimas, P. D. (1974) Speech perception in early infancy. In Infant Perception, ed. by L. B. Cohen and P. Salapatek. (New York: Academic Press).

Eimas, P. D., W. E. Cooper, and J. D. Corbit. (1973) Some properties of linguistic feature detectors. Percept. Psychophys. 13, 247-252.

Eimas, P. D. and J. D. Corbit. (1973) Selective adaptation of linguistic feature detectors. Cog. Psychol. 4, 99-109.

Eimas, P. D., E. R. Siqueland, P. Jusczyk, and J. M. Vigorito. (1971) Speech perception in infants. Science 171, 303-306.

Evans, E. F. and I. C. Whitfield. (1964) Classification of unit responses in the auditory cortex of the unanaesthetized and unrestrained cat. J. Physiol. 17, 476-493.

Fant, C. G. M. (1960) Acoustic Theory of Speech Production. (The Hague: Mouton).

Fant, C. G. M. (1962) Descriptive analysis of the acoustic aspects of speech. Logos 5, 3-17.

Fant, C. G. M. (1966) A note on vocal tract size factors and nonuniform F-pattern scalings. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) QPSR-4.

Fant, C. G. M. (1968) Analysis and systhesis of speech processes. In Manual of Phonetics, ed. by B. Malmberg. (Amsterdam: North-Holland).

Fischer-Jørgensen, E. (1972) Perceptual studies of Danish stop consonants. Annual Report (Institute of Phonetics, University of Copenhagen) 6, 75-176.

Flanagan, J. L. (1972) Speech Analysis, Synthesis, and Perception, 2nd ed. (New York: Academic Press).

Fodor, J. A., T. G. Bever, and M. F. Garrett. (1974) The Psychology of Language. (New York: McGraw Hill).

Foss, D. J. and D. A. Swinney. (1973) On the psychological reality of the phoneme: Perception, identification, and consciousness. J. Verbal Learn. Verbal Behav. 12, 246-257.

Fourcin, A. J. (1968) Speech source inference. IEEE Trans. Audio Electroacoust. AU-16, 65-67.

Fourcin, A. J. (1972) Perceptual mechanisms at the first level of speech processing. In Proceedings of the 7th International Congress of Phonetic Sciences. (The Hague: Mouton).

Frishkopf, L. and M. Goldstein. (1963) Responses to acoustic stimuli in the eighth nerve of the bullfrog. J. Acoust. Soc. Amer. 35, 1219-1228.

Fromkin, V. A. (1971) The nonanomalous nature of anomalous utterances. Language 47, 27-52.

Fromkin, V. A., S. Krashen, S. Curtiss, D. Rigler, and M. Rigler. (1974) The development of language in Genie: A case of language acquisition beyond the "Critical Period." Brain Lang. 1, 81-107.

Fujimura, O. and K. Ochiai. (1963) Vowel identification and phonetic contexts. J. Acoust. Soc. Amer. 35, 1889(A).

Fujisaki, H. and T. Kawashima. (1969) On the modes and mechanisms of speech perception. Annual Report of the Engineering Research Institute (University of Tokyo) 28, 67-73.

Fujisaki, H. and T. Kawashima. (1970) Some experiments on speech perception and a model for the perceptual mechanism. Annual Report of the Engineering Research Institute (University of Tokyo) 29, 207-214.

Fujisaki, H. and N. Nakamura. (1969) Normalization and recognition of vowels. Annual Report (Division of Electrical Engineering, Engineering Research Institute, University of Tokyo) 1.

Gerstman, L. J. (1957) Perceptual dimensions for the friction portion of certain speech sounds. Unpublished Doctoral dissertation, New York University.

Gerstman, L. J. (1968) Classification of self-normalized vowels. IEEE Trans. Audio Electroacoust. AU-16, 78-80.

Geschwind, N. and W. Levitsky. (1968) Human brain: Left-right asymmetries in temporal speech region. Science 161, 186-187.

Glanzman, D. L. and D. B. Pisoni. (1973) Decision processes in speech discrimination as revealed by confidence ratings. Paper presented at the 85th meeting of the Acoustical Society of America, April 10-13, Boston, Mass.

Godfrey, J. J. (1974) Perceptual difficulty and the right-ear advantage for vowels. Brain Lang. 4, 323-336.

Goldstein, L. M. and J. R. Lackner. (in press) Alterations of the phonetic coding of speech sounds during repetition. Cognition.

Greenberg, J. J. and J. J. Jenkins. (1964) Studies in the psychological correlates of the sound system of American English. Word 20, 157-177.

Haber, R. N. (1969) Information-Processing Approaches to Visual Perception. (New York: Holt, Rinehart, and Winston).

Hadding-Koch, K. and M. Studdert-Kennedy. (1964) An experimental study of some intonation contours. Phonetica 11, 175-185.

Haggard, M. (1971) Encoding and the REA for speech signals. Quart. J. Exp. Psychol. 23, 34-45.

Haggard, M. P., S. Ambler, and M. Callow. (1970) Pitch as a voicing cue. J. Acoust. Soc. Amer. 47, 613-617.

Halperin, Y., I. Nachshon, and A. Carmon. (1973) Shift of ear superiority in dichotic listening to temporally patterned verbal stimuli. J. Acoust Soc. Amer. 53, 46-50.

Halwes, T. (1969) Effects of dichotic fusion on the perception of speech. Unpublished Doctoral dissertation, University of Minnesota.

Halwes, T. and J. J. Jenkins. (1971) Problem of serial order in behavior is not resolved by context-sensitive associative memory models. Psychol. Rev. 78, 122-129.

Hanson, G. (1967) Dimensions in speech sound perception: An experimental study of vowel perception. Ericsson Tech. 23, 3-175.

Harris, K. S. (1958) Cues for the discrimination of American English fricatives in spoken syllables. Lang. Speech 1, 1-17.

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

Hirsch, H. V. B. and D. N. Spinelli. (1970) Visual experience modifies distribution of horizontally and vertically oriented receptive fields in cat. Science 168, 869-871.

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

House, A. S., K. N. Stevens, T. T. Sandel, and J. B. Arnold. (1962) On the learning of speech-like vocabularies. J. Verbal Learn. Verbal Behav. 1, 133-143.

Hubel, D. H. and T. N. Wiesel. (1962) Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. J. Physiol. 60, 106-154.

Jakobson, R., C. G. M. Fant, and M. Halle. (1963) Preliminaries to Speech Analysis. (Cambridge, Mass.: MIT Press).
Jakobson, R. and M. Halle. (1956) Fundamentals of Language. (The Hague: Mouton).
Jones, D. (1948) Differences between Spoken and Written Language. (London: Assn. Phonétique Internationale).
Joos, M. A. (1948) Acoustic phonetics. Language, Suppl. 24, 1-136.
Julesz, B. (1971) Foundations of Cyclopean Perception. (Chicago: University of Chicago Press).
Kewley-Port, D. and M. S. Preston. (1974) Early apical stop production: A voice onset time analysis. J. Phonetics 3, 195-210.
Kimura, D. (1961a) Some effects of temporal lobe damage on auditory perception. Canad. J. Psychol. 15, 156-165.
Kimura, D. (1961b) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.
Kimura, D. (1967) Functional asymmetry of the brain in dichotic listening. Cortex 3, 163-178.
Kimura, D. and M. Durnford. (1974) Normal studies on the function of the right hemisphere in vision. In Hemisphere Function in the Human Brain, ed. by S. J. Dimond and J. G. Beamont. [London: Paul Elek (Scientific Books)].
Kirman, J. H. (1973) Tactile communication of speech: A review and an analysis. Psychol. Bull. 80, 54-74.
Kirstein, E. (1971) Temporal factors in perception of dichotically presented stop consonants and vowels. Unpublished Doctoral dissertation, University of Connecticut.
Kirstein, E. (1973) The lag effect in dichotic speech perception. Haskins Laboratories Status Report on Speech Research SR-35/36, 81-106.
Klatt, D. H. and S. R. Shattuck. (1973) Perception of brief stimuli that resemble formant transitions. Paper presented at the 86th meeting of the Acoustical Society of America, October 30 - November 2, Los Angeles, Calif.
Köhler, W. and H. Wallach. (1944) Figural after-effects: An investigation of visual processes. Proc. Amer. Phil. Soc. 88, 269-357.
Konishi, M. (1965) The role of auditory feedback in the control of vocalization in the white-crowned sparrow. Z. f. Tierpsychol. 22, 770-783.
Kozhevnikov, V. A. and L. A. Chistovich. (1965) Rech' Artikuliatsia i vospriiatie. (Moscow-Leningrad). Transl. as Speech: Articulation and Perception. (Washington, D.C.: Clearinghouse for Federal Scientific and Technical Information) JPRS 30, 543.
Krashen, S. (1972) Language and the left hemisphere. Working Papers in Phonetics (University of California at Los Angeles, Phonetics Laboratory) 24.
Lackner, J. R. and L. M. Goldstein. (in press) The psychological representation of speech sounds. Cognition.
Ladefoged, P. (1967) Three Areas of Experimental Phonetics. (New York: Oxford University Press).
Ladefoged, P. (1971a) Preliminaries to Linguistic Phonetics. (Chicago: University of Chicago Press).
Ladefoged, P. (1971b) Phonological features and their phonetic correlates. Working Papers in Phonetics (University of California at Los Angeles) 21, 3-12.
Ladefoged, P. and D. E. Broadbent. (1957) Information conveyed by vowels. J. Acoust. Soc. Amer. 29, 98-104.
Ladefoged, P., J. DeClerk, M. Lindau, and G. Papçun. (1972) An auditory-motor theory of speech production. Working Papers in Phonetics (University of California at Los Angeles) 22, 48-75.

Lane, H. L. (1965) The motor theory of speech perception: A critical review. Psychol. Rev. 72, 275-309.
Lashley, K. S. (1951) The problem of serial order in behavior. In Cerebral Mechanisms in Behavior, ed. by L. A. Jeffress. (New York: Wiley) 112-136.
Lass, N. J., ed. (in press) Contemporary Issues in Experimental Phonetics. (Springfield, Ill.: C. C Thomas).
Lass, N. J. and R. M. Gasperini. (1973) The verbal transformation effect: A comparative study of the verbal transformations of phonetically trained and nonphonetically trained subjects. Brit. J. Psychol. 64, 183-192.
Lass, N. J. and S. S. Golden. (1971) The use of isolated vowels as auditory stimuli in eliciting the verbal transformation effect. Canad. J. Psychol. 25, 349-359.
Lass, N. J., L. K. West, and D. D. Taft. (1973) A non-verbal analogue to the verbal transformation effect. Canad. J. Psychol. 27, 272-279.
Lea, W. A. (1974) An algorithm for locating stressed syllables in continuous speech. J. Acoust. Soc. Amer. 55, 411(A).
Lenneberg, E. H. (1967) The Biological Foundations of Language. (New York: Wiley).
Lettvin, J. Y., H. R. Maturana, W. S. McCulloch, and W. H. Pitts. (1959) What the frog's eye tells the frog's brain. Proc. Inst. Rad. Engr. 47, 1940-1951.
Liberman, A. M. (1957) Some results of research on speech perception. J. Acoust. Soc. Amer. 29, 117-123.
Liberman, A. M. (1970) The grammars of speech and language. Cog. Psychol. 1, 301-323.
Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.
Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1952) The role of selected stimulus variables in the perception of the unvoiced stop consonants. Amer. J. Psychol. 65, 497-516.
Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1958) Some cues for the distinction between voiced and voiceless stops. Lang. Speech 1, 153-167.
Liberman, A. M., P. C. Delattre, F. S. Cooper, and L. H. Gerstman. (1954) The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monogr. 68, 1-13.
Liberman, A. M., K. S. Harris, J. Kinney, and H. Lane. (1961) The discrimination of relative onset time of the components of certain speech and nonspeech patterns. J. Exp. Psychol. 61, 379-388.
Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (New York: Wiley) 307-334.
Licklider, J. C. R. and G. A. Miller. (1951) The perception of speech. In Handbook of Experimental Psychology, ed. by S. S. Stevens. (New York: Wiley) 1040-1074.
Lieberman, P. (1963) Some effects of semantic and grammatical context on the production and perception of speech. Lang. Speech 6, 172-179.
Lieberman, P. (1968) Primate vocalizations and human linguistic ability. J. Acoust. Soc. Amer. 44, 1574-1584.
Lieberman, P. (1970) Toward a unified phonetic theory. Ling. Inq. 1, 307-322.
Lieberman, P. (1972) The Speech of Primates. (The Hague: Mouton).
Lieberman, P. (1973) On the evolution of language: A unified view. Cognition 2, 59-94.
Lieberman, P. and S. Crelin. (1971) On the speech of Neanderthal man. Ling. Inq. 2, 203-222.

Lieberman, P., E. S. Crelin, and D. H. Klatt. (1972) Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. Amer. Anthropol. 74, 287-307.
Lieberman, P., K. S. Harris, P. Wolff, and L. H. Russell. (1971) Newborn infant cry and nonhuman primate vocalizations. J. Speech Hearing Res. 14, 718-727.
Liljencrants, J. and B. Lindblom. (1972) Numerical simulation of vowel quality systems: The role of perceptual contrast. Language 48, 839-862.
Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Amer. 35, 1773-1781.
Lindblom, B. E. F. (1972) Phonetics and the description of language. In Proceedings of the 7th International Congress of Phonetic Sciences. (The Hague: Mouton) 63-97.
Lindblom, B. E. F. and M. Studdert-Kennedy. (1967) On the role of formant transitions in vowel recognition. J. Acoust. Soc. Amer. 42, 830-843.
Lindblom, B. E. F. and J. Sundberg. (1971) Neurophysiological representation of speech sounds. Paper presented at the 15th World Congress of Logopedics and Phoniatrics, August 14-19, Buenos Aires, Argentina.
Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word 20, 384-422.
Lisker, L. and A. S. Abramson. (1967) Some effects of context on voice onset time in English stops. Lang. Speech 10, 1-28.
Lisker, L. and A. S. Abramson. (1970) The voicing dimension: Some experiments in comparative phonetics. In Proceedings of the 6th International Congress of Phonetic Sciences. (Prague: Academia) 563-567.
Lisker, L. and A. S. Abramson. (1971) Distinctive features and laryngeal control. Language 47, 767-785.
Locke, S. and L. Kellar. (1973) Categorical perception in a nonlinguistic mode. Cortex 9, 355-369.
Lotz, J., A. S. Abramson, L. H. Gerstman, F. Ingemann, and W. J. Nemser. (1960) The perception of English stops by speakers of English, Spanish, Hungarian, and Thai: A tape-cutting experiment. Lang. Speech. 3, 71-77.
MacKay, D. G. (1970) Spoonerisms: The anatomy of errors in the serial order of speech. Neuropsychologia 8, 323-350.
MacNeilage, P. F. (1970) Motor control of serial ordering of speech. Psychol. Rev. 77, 182-196.
MacNeilage, P. F., T. P. Rootes, and R. A. Chase. (1967) Speech production and perception in a patient with severe impairment of somesthetic perception and motor control. J. Speech Hearing Res. 10, 449-467.
Malmberg, B. (1955) The phonetic basis for syllable division. Studia Linguistica 9, 80-87.
Marler, P. (1970) Bird song and speech development: Could there be parallels? Amer. Scient. 58, 669-673.
Marler, P. (in press) On the origin of speech from animal sounds. In The Role of Speech in Language, ed. by J. F. Kavanagh and J. E. Cutting. (Cambridge, Mass.: MIT Press).
Marler, P. and P. Mundinger. (1971) Vocal learning in birds. In Ontogeny of Vertebrate Behavior, ed. by. H. Moltz. (New York: Academic Press) 380-450.
Marvilya, M. P. (1972) Spontaneous vocalization and babbling in hearing-impaired infants. In Speech Communication Ability and Profound Deafness, ed. by C. G. M. Fant. (Washington, D.C.: A. G. Bell Association for the Deaf).
Massaro, D. W. (1972) Preperceptual images, processing time, and perceptual units in auditory perception. Psychol. Rev. 79, 124-145.
Mattingly, I. G. (1968) Synthesis by rule of General American English. Supplement to Haskins Laboratories Status Report on Speech Research, April.
Mattingly, I. G. (1971) Synthesis by rule as a tool for phonological research. Lang. Speech 14, 47-56.

Mattingly, I. G. (1972) Speech cues and sign stimuli. Amer. Scient. 60, 327-337.
Mattingly, I. G. (1973) Phonetic prerequisites for first-language acquisition. Haskins Laboratories Status Report on Speech Research SR-34, 65-69.
Mattingly, I. G. (1974) Speech synthesis for phonetic and phonological models. In Current Trends in Linguistics, Vol. 12, ed. by T. A. Sebeok. (The Hague: Mouton).
Mattingly, I. G. (in press) The human aspects of speech. In The Role of Speech in Language, ed. by J. F. Kavanagh and J. E. Cutting. (Cambridge, Mass.: MIT Press).
Mattingly, I. G. and A. M. Liberman. (1969) The speech code and the physiology of language. In Information Processing in the Nervous System, ed. by K. N. Leibovic. (New York: Springer Verlag) 97-114.
Mattingly, I. G., A. M. Liberman, A. K. Syrdal, and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.
McNeill, D. and L. Lindig. (1973) The perceptual reality of phonemes, syllables, words, and sentences. J. Verbal Learn. Verbal Behav. 12, 419-430.
McNeill, D. and B. Repp. (1973) Internal processes in speech perception. J. Acoust. Soc. Amer. 53, 1320-1326.
Miller, G. A. (1956) The magical number seven plus or minus two, or, some limits on our capacity for processing information. Psychol. Rev. 63, 81-96.
Miller, G. A., G. A. Heise, and W. Lichten. (1951) The intelligibility of speech as a function of the context of the test materials. J. Acoust. Soc. Amer. 41, 329-335.
Miller, G. A. and P. Nicely. (1955) An analysis of some perceptual confusions among some English consonants. J. Acoust. Soc. Amer. 27, 338-352.
Miller, J. D., R. E. Pastore, C. C. Wier, W. J. Kelly, and R. J. Dooling. (1974) Discrimination and labeling of noise-buzz sequences with varying noise-lead times. J. Acoust. Soc. Amer. 55, 390(A).
Miller, J. N., D. Sutton, B. Pfingst, A. Ryan, and R. Beaton. (1972) Single cell activity in the auditory cortex of rhesus monkeys: Behavioral dependency. Science 177, 449-451.
Mitchell, P. D. (1973) A test of differentiation of phonemic feature contrasts. Unpublished Doctoral dissertation, City University of New York.
Moffitt, A. R. (1971) Consonant cue perception by twenty- to twenty-four-week-old infants. Child Develop. 42, 717-731.
Molfese, D. L. (1972) Cerebral asymmetry in infants, children, and adults: Auditory evoked responses to speech and noise stimuli. Unpublished Doctoral dissertation, Pennsylvania State University.
Morse, P. A. (1972) The discrimination of speech and nonspeech stimuli in early infancy. J. Exp. Child Psychol. 14, 477-492.
Neisser, U. (1967) Cognitive Psychology. (New York: Appleton-Century-Crofts).
Nelson, P. G., S. D. Erulkar, and S. S. Bryan. (1966) Response units of the inferior colliculus to time-varying acoustic stimuli. J. Neurophysiol. 29, 834-860.
Nottebohm, F. (1971) Neural lateralization of vocal control in a passerine bird. I. Song. J. Exp. Zool. 177, 229-262.
Nottebohm, F. (1972) Neural lateralization of vocal control in a passerine bird. II. Subsong, calls, and theory of vocal learning. J. Exp. Zool. 179, 35-50.
Obusek, C. J. and R. M. Warren. (1973) A comparison of speech perception in senile and well-preserved aged by means of the verbal transformation effect. J. Gerontol. 28, 184-188.
Öhman, S. E. G. (1966) Coarticulation in VCV utterances: Spectrographic measurements. J. Acoust. Soc. Amer. 39, 151-168.

Öhman, S. E. G. (1967) Numerical model of coarticulation. J. Acoust. Soc. Amer. 41, 310-320.
Papçun, G., S. Krashen, D. Terbeek, R. Remington, and R. Harshman. (1974) Is the left hemisphere specialized for speech, language, and/or something else? J. Acoust. Soc. Amer. 55, 319-327.
Parks, T., C. Wall, and J. Bastian. (1969) Intercategory and intracategory discrimination for one visual continuum. J. Exp. Psychol. 81, 241-245.
Perl, N. T. (1970) The application of the verbal transformation effect to the study of cerebral dominance. Neuropsychologia 8, 259-261.
Peterson, G. E. (1961) Parameters of vowel quality. J. Speech Hearing Res. 4, 10-29.
Peterson, G. E. and H. L. Barney. (1952) Control methods used in a study of vowels. J. Acoust. Soc. Amer. 25, 175-184.
Pettigrew, J. D. and R. D. Freeman. (1973) Visual experience without lines: Effects on development of cortical neurons. Science 182, 599-600.
Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Unpublished Doctoral dissertation, University of Michigan.
Pisoni, D. B. (1972) Perceptual processing time for consonants and vowels. Haskins Laboratories Status Report on Speech Research SR-31/32, 83-92.
Pisoni, D. B. (1973a) Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, 253-260.
Pisoni, D. B. (1973b) The role of auditory short-term memory in vowel perception. Haskins Laboratories Status Report on Speech Research SR-34, 89-118.
Pisoni, D. B. (in press) Dichotic listening and the processing of phonetic features. In Cognitive Theory, Vol. 1, ed. by F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindam, and D. B. Pisoni. (Potomac, Md.: Erlbaum Associates).
Pisoni, D. B. and J. H. Lazarus. (1974) Categorical and noncategorical modes of speech perception along the voicing continuum. J. Acoust. Soc. Amer. 55, 328-333.
Pisoni, D. B. and S. D. McNabb. (1974) Dichotic interactions of speech sounds and phonetic feature processing. Brain Lang. 4, 351-362.
Pisoni, D. B. and J. R. Sawusch. (in press) Category boundaries for speech and nonspeech sounds. Percept. Psychophys.
Pisoni, D. B. and J. Tash. (1974) Reaction times to comparisons within and across phonetic categories. Percept. Psychophys. 15, 285-290.
Pollack, I. and J. M. Pickett. (1963) The intelligibility of excerpts from conversation. Lang. Speech 6, 165-172.
Popper, R. D. (1972) Pair discrimination for a continuum of synthetic voiced stops with and without first and third formants. J. Psycholing. Res. 1, 205-219.
Porter, R. J. (1971) The effect of delayed channel on the perception of dichotically presented speech and nonspeech sounds. Unpublished Doctoral dissertation, University of Connecticut.
Posner, M. I., S. J. Boies, W. H. Eichelman, and R. L. Taylor. (1969) Retention of visual and name codes of single letters. J. Exp. Psychol. Monogr. 79, 1-16.
Potter, R. K., G. A. Kopp, and H. C. Green. (1947) Visible Speech. (New York: van Nostrand).
Rand, T. C. (1971) Vocal tract size normalization in the perception of stop consonants. Haskins Laboratories Status Report on Speech Research SR-25/26, 141-146.

Raphael, L. J. (1972) Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. J. Acoust. Soc. Amer. 51, 1296-1303.

Reed, S. K. (1973) Psychological Processes in Pattern Recognition. (New York: Academic Press).

Repp, B. H. (1973) Dichotic forward and backward masking of CV syllables. Unpublished Doctoral dissertation, University of Chicago.

Rosch, E. H. (1973) Natural categories. Cog. Psychol. 4, 328-350.

Sachs, R. M. (1969) Vowel identification and discrimination in isolation vs word context. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 93, 220-229.

Sales, B. D., R. A. Cole, and R. N. Haber. (1969) Mechanisms of aural encoding: V. Environmental effects of consonants on vowel encoding. Percept. Psychophys. 6, 361-365.

Savin, H. B. and T. B. Bever. (1970) The nonperceptual reality of the phoneme. J. Verbal Learn. Verbal Behav. 9, 295-302.

Schatz, C. (1954) The role of context in the perception of stops. Language 30, 47-56.

Scholes, R. J. (1968) Phonemic interference as a perceptual phenomenon. Lang. Speech 11, 86-103.

Shankweiler, D. P., W. Strange, and R. Verbrugge. (in press) Speech and the problem of perceptual constancy. In Perceiving, Acting, and Comprehending: Toward an Ecological Psychology, ed. by. R. Shaw and J. Bransford. (Potomac, Md.: Erlbaum Associates).

Shankweiler, D. P. and M. Studdert-Kennedy. (1967) Identification of consonants and vowels presented to left and right ears. Quart. J. Exp. Psychol. 19, 59-63.

Shearme, J. N. and J. N. Holmes. (1962) 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 4th International Congress of Phonetic Sciences. (The Hague: Mouton) 234-240.

Shepard, R. N. (1972) Psychological representation of speech sounds. In Human Communication: A Unified View, ed. by E. E. David and P. B. Denes. (New York: McGraw Hill) 67-113.

Singh, S. (1966) Cross-language study of perceptual confusions of plosive phonemes in two conditions of distortion. J. Acoust. Soc. Amer. 40, 635-656.

Singh, S. and D. Woods. (1970) Multidimensional scaling of 12 American English vowels. J. Acoust. Soc. Amer. 48, 104(A).

Sinnott, J. M. (1974) A comparison of speech sound discrimination in humans and monkeys. Unpublished Doctoral dissertation, University of Michigan.

Stetson, R. H. (1952) Motor Phonetics. (Amsterdam: North-Holland).

Stevens, K. N. (1960) Toward a model for speech recognition. J. Acoust. Soc. Amer. 32, 47-55.

Stevens, K. N. (1967) Acoustic correlates of certain consonantal features. Paper presented at Conference on Speech Communication and Processing, MIT, November 6-8, Cambridge, Mass.

Stevens, K. N. (1968a) Acoustic correlates of place of articulation for stop and fricative consonants. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 89, 199-205.

Stevens, K. N. (1968b) On the relations between speech movements and speech perception. Z. Phon., Sprachwiss. u. Komm. Fschg. 213, 102-106.

Stevens, K. N. (1972a) Segments, features, and analysis by synthesis. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press) 47-52.

Stevens, K. N. (1972b) The quantal nature of speech: Evidence from articulatory-acoustic data. In Human Communication: A Unified View, ed. by E. E. David and P. B. Denes. (New York: McGraw Hill) 51-66.
Stevens, K. N. (1973) Potential role of property detectors in the perception of consonants. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 110, 155-168.
Stevens, K. N. and M. Halle. (1967) Remarks on analysis by synthesis and distinctive features. In Models for the Perception of Speech and Visual Form, ed. by W. Wathen-Dunn. (Cambridge, Mass.: MIT Press) 88-102.
Stevens, K. N. and A. S. House. (1972) The perception of speech. In Foundations of Modern Auditory Theory, Vol. 2, ed. by J. Tobias. (New York: Academic Press) 3-62.
Stevens, K. N., A. S. House, and A. P. Paul. (1966) Acoustical description of syllabic nuclei. J. Acoust. Soc. Amer. 40, 123-132.
Stevens, K. N. and D. H. Klatt. (1974) Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Amer. 55, 653-659.
Stevens, K. N., A. M. Liberman, M. Studdert-Kennedy, and S. E. G. Öhman. (1969) Cross-language study of vowel perception. Lang. Speech 12, 1-23.
Strange, W., R. Verbrugge, and D. P. Shankweiler. (1974) Consonantal environment specifies vowel identity. Paper presented at the 87th meeting of the Acoustical Society of America, April 23-26, New York City.
Studdert-Kennedy, M. (1974) The perception of speech. In Current Trends in Linguistics, ed. by T. A. Sebeok. (The Hague: Mouton). [Also in Haskins Laboratories Status Report on Speech Research SR-23 (1970) 15-48.]
Studdert-Kennedy, M. (in press) From continuous signal to discrete message: Syllable to phoneme. In The Role of Speech in Language, ed. by J. F. Kavanagh and J. E. Cutting. (Cambridge, Mass.: MIT Press).
Studdert-Kennedy, M. and K. Hadding. (1973) Auditory and linguistic processes in the perception of intonation contours. Lang. Speech 16, 293-313.
Studdert-Kennedy, M., A. M. Liberman, K. S. Harris, and F. S. Cooper. (1970a) Motor theory of speech perception: A reply to Lane's critical review. Psychol. Rev. 77, 234-249.
Studdert-Kennedy, M. and D. P. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.
Studdert-Kennedy, M., D. P. Shankweiler, and D. B. Pisoni. (1972) Auditory and phonetic processes in speech perception: Evidence from a dichotic study. Cog. Psychol. 2, 455-466.
Studdert-Kennedy, M., D. P. Shankweiler, and S. Schulman. (1970b) Opposed effects of a delayed channel on perception of dichotically and monotically presented CV syllables. J. Acoust. Soc. Amer. 48, 599-602.
Summerfield, A. and M. Haggard. (1972) Perception of stop voicing. Speech Perception (Department of Psychology, The Queen's University of Belfast) Series 2, 1, 1-14.
Summerfield, A. and M. Haggard. (1973) Vocal tract normalization as demonstrated by reaction times. Speech Perception (Department of Psychology, The Queen's University of Belfast) 2.
Sussman, H. (1971) The laterality effect in lingual-auditory tracking. J. Acoust. Soc. Amer. 49, 1874-1880.
Sussman, H. M. and P. F. MacNeilage. (in press) Studies of hemispheric specialization for speech production. Brain Lang.
Sussman, H. M., P. F. MacNeilage, and J. Lumbley. (1974) Sensorimotor dominance and the right-ear advantage in mandibular-auditory tracking. J. Acoust. Soc. Amer. 56, 214-216.

Treon, M. A. (1970) Fricative and plosive perception-identification as a function of phonetic context in CVCVC utterances. Lang. Speech 13, 54-64.
Turvey, M. (1973) On peripheral and central processes in vision. Psychol. Rev. 80, 1-52.
Verbrugge, R., W. Strange, and D. P. Shankweiler. (1974) What information enables a listener to map a talker's vowel space? Paper presented at the 87th meeting of the Acoustical Society of America, April 23-26, New York City.
Vitz, P. C. and B. S. Winkler. (1973) Predicting the judged similarity of sound of English words. J. Verbal Learn. Verbal Behav. 12, 373-388.
Warren, R. M. (1968) Verbal transformation effect and auditory perceptual mechanisms. Psychol. Bull. 70, 261-270.
Warren, R. M. (1970) Perceptual restoration of missing speech sounds. Science 167, 392-393.
Warren, R. M. (1971) Identification times for phonemic components of graded complexity and for spelling of speech. Percept. Psychophys. 9, 345-349.
Warren, R. M. and R. L. Gregory. (1958) An auditory analogue of the visual reversible figure. Amer. J. Psychol. 71, 612-613.
Warren, R. M. and C. J. Obusek. (1971) Speech perception and phonemic restorations. Percept. Psychophys. 9 (3B), 358-362.
Weiss, M. and A. S. House. (1973) Perception of dichotically presented vowels. J. Acoust. Soc. Amer. 53, 51-58.
Werner, H. (1935) Studies on contour: I. Qualitative analyses. Amer. J. Psychol. 47, 40-64.
Whitfield, I. C. and E. F. Evans. (1965) Responses of auditory cortical neurons to stimuli of changing frequency. J. Neurophysiol. 28, 655-672.
Wickelgren, W. A. (1965) Distinctive features and errors in short-term memory for English vowels. J. Acoust. Soc. Amer. 38, 583-588.
Wickelgren, W. A. (1966) Distinctive features and errors in short-term memory for English consonants. J. Acoust. Soc. Amer. 39, 388-398.
Wickelgren, W. A. (1969) Context-sensitive coding, associative memory, and serial order in (speech) behavior. Psychol. Rev. 76, 1-15
Wollberg, Z. and J. D. Newman. (1972) Auditory cortex of squirrel monkey: Response patterns of single cells to species-specific vocalizations. Science 175, 212-213.
Wood, C. C. (1974) Parallel processing of auditory and phonetic information in speech perception. Percept. Psychophys. 15, 501-508.
Wood, C. C. (1975) Auditory and phonetic levels of processing in speech perception: Neurophysiological and information-processing analyses. J. Exp. Psychol.: Human Percept. Perform. 104, 1-33.
Wood, C. C. and Day, R. S. (in press) Failure of selective attention to phonetic segments in consonant-vowel syllables. Percept. Psychophys.
Wood, C. C., W. R. Goff, and R. S. Day. (1971) Auditory evoked potentials during speech perception. Science 173, 1248-1251.
Woodworth, R. S. and H. Schlosberg. (1954) Experimental Psychology. (New York: Holt, Rinehart, and Winston).
Zlatin, M. A. (1974) Development of the voicing contrast: A psychoacoustic study of voice onset time. J. Acoust. Soc. Amer. 56, 981-994.