LANGUAGE AND SPEECH, Vol. 23, Part 1, 1980

SPEECH PERCEPTION

MICHAEL STUDDERT-KENNEDY, REPORTER
Queens College, City University of New York
and
Haskins Laboratories

The paper reviews selected studies in speech perception, most of them published in the past five years. Topics include the contributions of prosody to segmental perception, the problems of segmentation and invariance, categorical perception of speech and non-speech, the role of feature detectors, the scaling of speech sounds to an auditory-articulatory space, acoustic phonetic dependencies within the syllable, the contributions of higher order (non-phonetic) factors to the comprehension of fluent speech, and cerebral specialization. The bias of the paper is toward viewing phonetic segments as abstract processes that link sound and articulation, and that become available to the listener through specialized sensorimotor mechanisms.

The past few years of research in speech perception have been very active. The old questions are still there – What are the units? How do we segment? Where are the invariants? – but some old answers have turned out to be wrong and some new ones are beginning to emerge. The intricate articulatory and acoustic structure of the syllable is still at the center of the maze, but other sources of information for the listener – prosody, syntax, semantics – have begun to receive experimental attention. Studies of fluent speech are taking their place beside the established methods of syllable analysis and synthesis. Theory has dropped into the background (or perhaps the back room) and no one seems very eager to argue the merits of analysis-by-synthesis or the "motor theory" any more. Certainly, theory continues to guide research, but a refreshing atheoretical breeze has been blowing in from artifical speech understanding research (Klatt, 1977, 1978) and from developmental psychology (Aslin and Pisoni, 1978). In the latter regard, I shall not have much to say directly about infant speech perception, but much of what I have to say will bear on it. The infant is a listener, a very attentive one, because by learning to listen it learns to speak. In my opinion, only by carefully tracking the infant through its first two years of life shall we come to understand adult speech perception and, in particular, how speaking and listening establish their links at the base of the language system. This said, let us begin, as infants do, with prosody.

PROSODY

Prosody means the melody, rhythm, rate, amplitude, quality and temporal organization of speech. There has been an upsurge of interest in these factors in recent years, partly because they seem to hold a key to improved speech synthesis, partly because

prosodic contributions to speech perception have been unjustly neglected (Cohen and Nooteboom, 1975; Nooteboom, Brokx and de Rooij, 1976). To say that prosody "contributes" to speech perception may seem to imply that speech perception is confined to segmental processes, of which prosody is a mere subsidiary, conveying no distinctive information of its own. This, of course, is false. Prosody carries much of that important indexical information (Abercrombie, 1967) without which, if it is dark, you don't know who is talking to you or whether he means what he says. However, it is with the adjutant function – contributions to segmental perception – that I am concerned here.

One prosodic function is to maintain a coherent auditory signal. Darwin (1975) asked listeners to shadow a sentence in one ear, while a competing sentence was led into the other. At some arbitrary point, prosodic contours were suddenly switched across ears, while syntactic and semantic sequences were maintained. Prosodic continuity then often overrode syntax, semantics and ear of entry, leading to the intrusion of words from the supposedly unattended ear. Evidently, listeners were tracking the prosodic contour, a process that Nooteboom *et al.* (1976) suggest may be necessary to maintain "perceptual integrity."

What physical dimensions of the signal sustain this integrity? Rate is probably not important, because quite sharp rate variations are regularly used to convey syntactic information (e.g., Klatt, 1976). Of course, rate *can* affect segmental classification (Ainsworth, 1972), but listeners adjust rapidly, within less than a second (Fujisaki, Nakamura and Imoto, 1975; Summerfield, 1975; Nooteboom *et al.*, 1976). Amplitude changes, within limits, are also probably of little importance (Darwin and Bethell-Fox, 1977). In fact, the principal determinants of prosodic continuity seem to be fundamental frequency (F_0) and spectrum: Nooteboom *et al.* (1976) showed that when pitches, alternating over a 2–6 Hz range, are imposed on a sequence of three vowels, repeated at intervals of less than 150 msec, the vowels split into two streams, as though from two speakers. The effect is reduced, if the vowels are granted a degree of spectral continuity by being placed into consonantal context. This work, taken with similar studies by Dorman, Cutting and Raphael (1975) and by Darwin and Bethell-Fox (1977), leads to the conclusion that continuity of both formant structure and F_0 underlies the perceptual integrity of running speech.

A second prosodic function is to facilitate phrasal grouping. Here the main variables seem to be F_0 and segment duration. Several studies have documented syntactic control of timing and segment duration in production (e.g., Cooper, 1976; Klatt, 1976). Klatt and Cooper (1975) show, further, that listeners expect segment duration to vary with the syntactic position of a word in a sentence. For example, they judge lengthened syllables to be more natural at the end of a clause than at the beginning or middle. Similarly, Nooteboom *et al.* (1976) report that listeners judge a vowel of a particular length to be shorter if it occurs at the end of a word than if it occurs at the beginning. Presumably, such observations reflect listeners' habitual use of final lengthening as an aid to parsing.

The role of F_0 has been more extensively studied. For example, Collier and 't Hart (1975) constructed synthetic utterances consisting of 13 or 15 200-msec steady-state, vowel-like "syllables," separated by 50-msec silent intervals. They imposed ten theoretically derived F_0 contours ('t Hart and Cohen, 1973) on these syllables, deploying charac-

teristic "continuation rises" and "non-final falls" to delimit the ends and the beginnings, respectively, of possible syntactic constituents. Finally, following Svensson (1974) and Kozhevnikov and Chistovich (1965), they asked listeners to write down syntactically acceptable sentences to match each contour in number of syllables, location of stresses and overall intonation. Of the resulting sentences, 72% matched the predicted syntactic structures. Since two hypotheses were under test here – both the correctness of the theoretically derived contours and the listeners' capacity to infer syntactic structure from intonation – this is a remarkably high score.

Finally, a third perceptual function of prosody has aroused a great deal of interest in recent years. This is the function – nobody knows what it is – supposedly fulfilled by rhythm. Martin (1972) wrote a persuasive paper in which he argued that speaking involves more than a simple concatenation of motor elements: like other motor behaviors speech is compelled, by natural constraints on the relative timing of components, to be rhythmic. Moreover, some components (syllables) are "accented," and these are predictable: accent level (or stress) covaries with timing and the main accents are equidistant (i.e., isochronous). Finally, since ". . . speaking and listening are dynamically coupled rhythmic activities . . ." (p. 489), listeners can predict the main stresses and can use that fact to "cycle" their attention, saving it, as it were, for the more important words.

There is, in fact, evidence from phoneme-monitoring experiments that reaction time (RT) is shorter to initial phonemes in stressed words than in unstressed words (Shields, McHugh and Martin, 1974). This is apparently not due to the greater energy of the stressed words, since, if the words are presented in isolation, no RT difference appears (Shields *et al.*, 1974). Moreover, Cutler (1976) has found that the RT difference holds, even if stress, or the lack of it, is merely "predicted" by prior prosodic contour and if the actual target is acoustically identical in both conditions. Cutler and Foss (1977) demonstrate, further, that RT advantage is not due to syntactic form class, since it is found for stressed function words as well as for stressed content words. They conclude that the reduced reaction time may reflect heightened attention to the semantic focus of a sentence, and they cite (p. 10) unpublished evidence from Allen and O'Shaugnessy that ". . . reliable correlates of semantic focus are to be found in the fundamental frequency contour."

By this last point Cutler and Foss seem to be cutting themselves free from Martin's (1972) claim for isochrony, whether wisely or not remains to be seen. Lehiste (1977) has recently reopened the isochrony issue in a paper summarizing much of her research on the topic. She concludes that although isochrony is "primarily a perceptual phenomenon" (p. 253), it does have some basis in production and is therefore available for communicative uses. Lehiste shows that English interstress intervals are often lengthened to signal a syntactic boundary.

Isochrony has also come under experimental scrutiny. Morton, Marcus and Frankish (1976), recording a list of spoken digits for experimental use, discovered that acoustically (onset to onset) isochronous sequences sounded anisochronous. Moreover, listeners, asked to adjust a sequence to perceptual isochrony, made it acoustically anisochronous. Morton *et al.* (1976) coined the term "perceptual centers" ("P-centers") to refer to those points in a sequence of words that are equidistant when the words sound isochronous.

But they were unable to locate the points or specify their acoustic correlates. Surprisingly, the P-center does not correspond to any obvious acoustic marker, such as sound onset, vowel onset or syllable peak. However, Fowler (1979, p. 375) has recently discovered that "... when asked to *produce* isochronous sequences, talkers generate precisely the acoustic anisochronies that listeners require in order to *hear* a sequence as isochronous." The acoustic anisochronies apparently arise because the articulatory onsets of words beginning with sounds from different manner classes have acoustic consequences at different relative points in time. From a review of her own and related studies (e.g., Allen, 1972; Lindblom and Rapp, 1973), Fowler (1979, p. 375) concludes that "... listeners judge isochrony based on acoustic information about *articulatory* timing rather than on some articulation-free acoustic basis." Finally, although this work seems to be a thread that might unravel isochrony, Fowler is cautious in her claims. Most of the relevant experimental studies have used monosyllables and artificially repetitive utterances. What inroads this approach can make into the apparent isochrony of phonetically heterogeneous running speech remains to be seen.

SEGMENTATION AND INVARIANCE

We turn now from the broad questions of prosody to the narrower puzzle of the syllable on which the prosody is carried. In what follows, I assume (together with most other investigators) that our task is to understand the process by which phonemes or features are extracted from the signal. Let us begin with a question raised by Myers, Zhukova, Chistovich and Mushnikov (1975): Is segmentation an auditory process, preceding phonetic classification, or an automatic consequence of classification itself? Several studies from the Pavlov Institute in Leningrad speak to the question. Chistovich, Fyodorova, Lissenko and Zhukova (1975) showed that a sudden amplitude drop, roughly in the middle of a 460-msec steady-state vowel, caused listeners to hear either two vowels or a VCV sequence, depending on the magnitude and rate of the amplitude decrease. Subsequently, Myers *et al.* (1975) used an ingenious dichotic technique to suggest that such amplitude decreases are registered by the peripheral auditory system; they inferred that, since classification is presumably central, segmentation must precede classification. Finally, Zhukov, Zhukova and Chistovich (1974) reported on the use of a similar technique to study the effects of spectral variation at segment boundaries. The investigators presented a time-varying value of the second formant (roughly 2200 to 800 Hz over 200 msec), to one ear, steady-state values of F1 and F3 to the other. The latter were interrupted by a 12–15 msec pause, of which the position could be set by the subject so as to vary the fused percept from hard to soft [r], that is, from [iru] to [iɻu]. Subjects reliably set the pause so that its endpoint coincided with an F2 value of roughly 1600 Hz. Since this value is close to that of the hard-soft boundary previously determined for the steady-state isolated consonants [s] and [ʂ], the authors infer that listeners were also judging the soft consonant [ɻ] by its F2 value at onset. They conclude (p. 237) that "the auditory system interprets the acoustic flow as a sequence of time segments between instants of variation," and that it derives consonantal information by sampling formant

frequencies at these instants.

However, this conclusion does not seem to be forced by the data. On the one hand, the presumed peripheral segment boundary, determined by a sharp amplitude drop, seems to have something in common with the boundary proposed by certain automatic recognition procedures for isolating syllables rather than phonemes (e.g., Mermelstein, 1975). On the other hand, an invariant formant onset is not incompatible with the use of formant movement into the following vowel as a consonantal cue (see Dorman, Studdert-Kennedy and Raphael, 1977). My inclination, therefore, is to suppose that the preliminary auditory segmentation (if any) is syllabic rather than phonemic, and that within-syllable segmentation may often be synonymous with classification. I will return to this point below.

The view of the perceptual process, proposed by the Russian group, as a succession of brief time slices (rather than as the active continuous tracking suggested by studies of prosody), is close to that currently being explored by K.N. Stevens. In a succession of publications over recent years, Stevens (e.g., 1975) has elaborated on the "quantal nature of speech." He points out that, although the vocal apparatus is capable of producing a wide variety of sounds, relatively few are actually used in the languages of the world. He attributes this restriction to a nonlinear relation between articulatory and acoustic parameters: some articulatory configurations are acoustically stable, in the sense that small changes in articulation have little acoustic effect; others are unstable in the sense that equally small changes have a substantial effect. The universal set of phonetic features is drawn from those articulatory configurations that generate acoustically stable, invariant "properties." The properties, it should be stressed, are higher order spectral configurations, rather than isolated cues such as F2 onset frequency. To define these configurations, Stevens has largely relied on computations from a vocal tract model. Finally, to assure quantal (or categorical) perception of the invariant properties and to afford the human infant a mechanism for netting them in the speech stream, Stevens postulates a matching set of innate "property detectors."

Empirical tests of the quantal theory have been few. But a recent study of English stops (Blumstein and Stevens, 1979; see also Stevens and Blumstein, 1978) is a good illustration of the approach, since it deals with a notoriously context-dependent set of sounds. The goal was to demonstrate the presence of invariant properties in the acoustic signal, sufficient for recognition by fixed templates. The first step was to record two male speakers reading random lists of the voiced stops [b d g], followed by each of five vowels [i e a o u]. Short-time spectra were then determined, integrated over a 26-msec window at onset. The spectra were used to construct, by trial and error, a template fitted to each place of articulation, such that it either correctly accepted or correctly rejected the majority of utterances. Descriptions of the templates ("diffuse-rising" for alveolar, "diffuse-falling" for labial and "compact" for velar) recall the terminology of distinctive feature theory.

In the second part of the study, a corpus of utterances was collected for classification by the templates. Six subjects (4 males, 2 females) recorded five repetitions each of the voiced and voiceless stop consonants [b d g p t k], followed by each of the vowels [æ e i o u], or preceded by each of the vowels [i ɛ æ ʌ u]. The resulting 1800 utterances

were then analyzed spectrally in the same way as the original utterances, and compared with the templates. The results were at least 80% (and often higher) correct rejection and correct acceptance for initial stops and a slightly lower performance for released final stops, although for some unreleased final stops scores dropped as low as 40%. Analysis of variance revealed significant differences in template-matching performance as a function of vowel context, but performance was significantly above chance in every case. Quite similar results have been reported by Searle, Jacobson and Rayment (1979) using a very much longer time slice (100-200 msec) and deriving their invariant patterns from a running sequence of spectra.

Where then does this leave us? A score of 80% or better is good, although, as A.M. Liberman has suggested to me, we might do almost as well with the binary recipe proposed by Cooper, Delattre, Liberman, Borst and Gerstman (1952); high burst, falling F2 transition for alveolar; low burst, falling F2 transition for velar; low burst, rising F2 transition for labial.

The question, of course, is: Is this really the way that humans do it? Dorman *et al.* (1977), modeling their study on the work of Fischer-Jørgensen (1972), edited release bursts and/or formant transitions out of English voiced stop consonants ([b d g]), spoken before nine different vowels. Acoustic analysis of the bursts for a given place of articulation showed them to be largely invariant (cf. Zue, 1976). However, the bursts were not invariant in their effect: for the most part, listeners only perceived the bursts correctly if their main spectral weight lay close to the main formant of the following vowel, as Stevens himself has remarked (1975, pp. 312-313). Kuhn (1975) has shown that the main vowel formant varies with the length of the cavity in front of the point of maximum tongue constriction. Since front cavity length is a function of place of articulation, an estimate of front cavity resonance is tantamount to an estimate of place of articulation. Thus, proximity on the frequency scale may facilitate perceptual integration of the burst with the vowel, enabling the listener to track the changing cavity shape characteristic of a particular place of articulation followed by a particular vowel.

Stevens (see especially, 1975) does not deny that contextually variable cues – such as formant transitions, voice onset time, vowel formant structure – can be used by the human listener. However, he regards them as "secondary," learned cues, acquired by repeated association with the "primary" invariant properties, and used only as safety devices when invariant cues fail. Given the many knotty questions concerning the possible mechanisms for extracting and interpreting these "secondary" context-dependent cues, one may wonder how an organism whose primary endowment is a set of passive templates learns to use them at all.

The question becomes even more pressing when one considers that there is no independent evidence for the existence of the hypothesized templates or property detectors. To understand this we must briefly review recent findings in the study of categorical perception.

CATEGORICAL PERCEPTION

As is well known, early work with speech synthesizers showed that a useful procedure

for defining the acoustic properties of a phoneme was to construct tokens of opponent categories, distinguished on a single phonological feature, by varying a single acoustic parameter along a continuum (e.g., [ba] to [da], [da] to [ta], etc.). If listeners were asked to identify these tokens, they tended to identify any particular stimulus in the same way every time they heard it; there were few ambiguous tokens. Moreover, if they were asked to discriminate between neighboring tokens, they tended to do very badly if they assigned the two tokens to the same class, very well if they assigned them to different classes – even though the acoustic distance between tokens was identical in the two cases. This phenomenon was dubbed "categorical perception" (Liberman, Harris, Hoffman and Griffith, 1957). Although there were usually no grounds for supposing that the acoustic variations along synthetic continua mimicked the intrinsic allophonic variations of natural speech, categorical perception in the laboratory was taken to reflect a necessary aspect of normal speech perception, namely, the rapid transfer of speech sounds into a phonetic or phonological code. The phenomenon was also believed by some people, including myself, to be peculiar to speech (Studdert-Kennedy, Liberman, Cooper and Harris, 1970).

However, we now know that categorical perception, as observed in the laboratory, is neither peculiar nor necessary to speech. Demonstrations that it is not peculiar we owe to Cutting and Rosner (1974) (rise-time at the onset of sawtooth waves, analogous to a fricative-affricate series); to Miller, Wier, Pastore, Kelly and Dooling (1976) (noise-buzz sequences analogous to the aspiration-voice sequence of a voice onset time (VOT) series); to Pisoni (1977) (relative onset time of two tones); and to Pastore, Ahroon, Baffuto, Friedman, Puleo and Fink (1977). These last investigators extended their work into vision, demonstrating categorical perception of flicker, with a sharp boundary at the critical flicker-fusion threshold. They also induced clearly categorical perception of a sine-wave intensity series by providing listeners with a constant-reference tone, or "pedestal," at the center of the series. Pastore *et al.* (1977) conclude that a continuum may be categorically divided either by a sensory threshold (as in flicker-fusion) or by a stimulus-internal reference (as in the intensity series). Presumably, the portion of the signal with the earlier onset serves as a reference in a VOT series, while in a place-of-articulation series, cued by direction and extent of formant transitions, a reference is provided by the fixed vowel. If this last point is correct, we perceive a place series categorically precisely because the consonants are judged relationally rather than absolutely – an interpretation not compatible with the notion of invariant property detectors.

Just how a stimulus-internal reference suppresses discrimination within categories is not clear, but the results of Carney, Widin and Viemeister (1977) suggest that it may simply serve to divert the listener's attention from detailed variations within the series. To Carney *et al.* (1977) (see also Pisoni and Lazarus, 1974; Samuel, 1977), we owe the demonstration that a VOT continuum *need* not be perceived categorically. Each of their subjects displayed good within-category discrimination after moderate training on a bilabial VOT continuum. Indeed discrimination was so good that subjects were able to shift category boundaries on request and assign consistent labels to arbitrary subsets of the stimuli. The outcome suggests that ". . . utilization of acoustic differences between speech stimuli may be determined primarily by attentional factors, . . . distinct from the

perceptual *capacities* of the organism" (Carney *et al.*, p. 969).

This is precisely what is suggested by the numerous instances in which speakers of different languages perceive an acoustic continuum in different ways. (For a review, see Strange and Jenkins, 1978). For example, while American English speakers perceive an [r] to [l] continuum categorically, Japanese speakers do not (Miyawaki, Strange, Verbrugge, Liberman, Jenkins and Fujimura, 1975). For another example, not only do Spanish and American English speakers place their category boundaries at different points along the VOT continuum (Abramson and Lisker, 1973; Williams, 1977), but also Spanish-English bilinguals can be induced to shift their boundaries by a shift in language set within a single test (Elman, Diehl and Buchwald, 1977). Not unrelated, perhaps, is the recent demonstration by Ganong (1978) that listeners have a bias for words over nonwords: offered a continuum of which one end is a word (e.g., *gift*) and the other not (e.g., *kift*), they shift their normal boundary away from the word, thus increasing the number of words they hear.

Presumably there are limits to this sort of thing. With adequate synthesis, the range of uncertainty must be limited and we may still use synthetic continua to assess "the auditory tolerance of phonological categories" (Brady and Darwin, 1978, p. 1556) – precisely the use for which they were first designed over 25 years ago.

Feature or Property Detectors

The demonstration that listeners can be trained to hear a supposedly categorical continuum noncategorically undercuts the original evidence for acoustic feature, or property, detectors in speech perception, namely, categorical perception itself. Moreover, it throws into doubt the interpretation of a substantial body of work on selective adaptation of speech sounds that has appeared in the past five years.

The series began with a paper by Eimas and Corbit (1973). They asked listeners to categorize members of a synthetic voice onset time (VOT) continuum (Lisker and Abramson, 1964) and demonstrated that the perceptual boundary between voiced and voiceless categories along that continuum was shifted by repeated exposure to (that is, adaptation with) either of the endpoint stimuli: there was a decrease in the frequency with which stimuli close to the original boundary were assigned to the adapted category and a consequent shift of the boundary toward the adapting stimulus. Since the effect could be obtained on a labial VOT continuum after adaptation with a syllable drawn from an alveolar VOT continuum, and vice versa, adaptation was clearly neither of the syllable as a whole nor of the unanalyzed phoneme, but of a feature within the syllable. Eimas and Corbit therefore termed the adaptation "selective" and attributed their result to the fatigue of specialized detectors and to the relative "sensitization" of opponent detectors. Subsequent studies replicated the results for VOT and extended them to other feature oppositions, such as place and manner of articulation. These studies have been reviewed by Cooper (1975), Ades (1976), and Eimas and Miller (1978).

Unfortunately, there are many grounds for doubting the opponent detector model. First, as already remarked, is the demonstration that listeners can be trained to dis-

criminate at least some speech continua within categories. Second, the model lacks behavioral or neurological motivation. For, while the facts of additive color mixture make an opponent detector account of after-effects entirely plasuible, the facts of laryngeal timing or spectral scatter at stop consonant onset certainly do not. Third, the hypothesis is rendered implausible by dozens of reports of contextual effects: adaptation of consonantal features is apparently specific to following vowel, to syllable position, to syllable structure (Hall and Blumstein, 1978) and even to fundamental frequency (Ades, 1977). As Simon and Studdert-Kennedy (1978) remark, ". . . the theoretical utility of selectively tuned feature detectors goes down as the number of contexts to which they must be tuned goes up." Moreover, the degree of adaptation varies quite generally with the acoustic distance between adaptor and test syllables, an effect typical of psychophysical contrast studies. In fact, Simon and Studdert-Kennedy (1978), drawing on their own work and that of Sawusch (1977), marshall evidence to argue that selective adaptation along speech continua reflects a combination of peripheral auditory fatigue and central auditory contrast. They do not deny that selective adaptation has possible fruitful use in isolating functional channels of analysis. But if their argument is correct, we now have no evidence at all for specialized speech detector mechanisms tuned to the acoustic correlates of abstract linguistic features.

Scaling Studies and Feature Interactions

This conclusion sits nicely with the results of many studies in which phoneme confusions or similarity judgments have been used to characterize the psychological representation of speech sounds. Although results vary widely with experimental method (van den Broecke, 1976), these studies typically find that vowels (e.g., Terbeek, 1977) and consonants (e.g., Singh, Woods and Becker, 1972) fall readily into low-confusion/high-similarity groups isomorphic with some standard phonological feature set. However, as Goldstein (1977) has pointed out, relations within these feature groups are usually not random. Rather, the psychological space is structured in such a way as to suggest a continuous auditory representation within feature groups. Presumably, since the continuous auditory representation derives from an acoustic structure shaped by articulation, we could describe an analogous articulatory space by scaling articulatory errors. It was Goldstein's (1977) insight to hypothesize that the variance common to the auditory and articulatory spaces would then prove to be categorical. His study – too complicated for summary here – largely supported that hypothesis. We may fairly conclude that our models of perception should allow for continuous auditory and articulatory representations from which categories can only be derived by some abstract metric common to both.

The idea that speech sounds (perhaps unsegmented syllables) may be internally represented in a continuous auditory space (at some point before classification) is compatible with the repeated finding of interaction between features during perceptual processing (e.g., Sawusch and Pisoni, 1974; Miller, 1977). There is, in fact, no good reason to refer to these auditory processes as "featural" at all (Parker, 1977). Repp

(1977) and Oden and Massaro (1978), for example, have already proposed specific models of integration based on a continuous spatial representation.

STEPS TOWARD AN AUDITORY-ARTICULATORY SPACE

The view of speech perception that seems to be emerging from the studies we have reviewed is of an active, continuous process. We turn now to several studies of perceptual integration across the syllable which seem to call for just such an interpretation.

Perhaps the most familiar example is provided by voicing cues for stops in initial position. The concept of voice onset time (VOT) originally offered an *articulatory* account of how a range of disparate and incommensurable acoustic cues (including, as it happens, the interval between release burst and the onset of voicing) comes to signal the voiced-voiceless distinction. In fact, as Abramson (1977) has recently reminded us, VOT is itself simply a special case of the laryngeal timing mechanisms by which voicing distinctions are, in general, implemented.

To illustrate the underlying articulatory rationale, consider the suggestion by Stevens and Klatt (1974) that the duration of the first formant voiced transition might be a more potent cue than VOT itself. The motivation for the proposal seems to have been to coordinate the voicing cue with Stevens' hypothesized cues to place of articulation (rapid spectral scatter), and perhaps to avoid saddling the infant with a delicate timing mechanism. As it happens, Simon and Fourcin (1978) have shown that English speaking children do not learn to use the F1 cue until they are five years old, while French-speaking children never use it at all. In any event, careful analysis by Lisker (1975) and by Summerfield and Haggard (1977) has shown that the principal first formant cue is not transition duration, but frequency at onset: the higher the frequency, the less likely is a sound to be judged voiced. Listeners apparently take a high first formant onset as a cue that the mouth was relatively wide open (and release therefore well past) when voicing began.

A less familiar set of cues to another distinction has recently been studied by Repp, Liberman, Eccardt and Pesetsky (1978). They recorded the utterance: "Did anybody see the *gray ship?*" Then, by varying the durations of fricative noise at the onset of *ship* and of the silent interval between *gray* and *ship*, they explored the conditions under which the utterance was heard as ending with "gray chip," "great ship" or "great chip." Among their results was the finding that whether or not a syllable final stop was heard (*gray* v. *great*) depended not only on the duration of the silence, but also on the duration of the noise following the silence. Just such an equivalence between a spectral property and silence emerges from an analysis of the trading relation between silence and formant transition in the cues for the medial [p] of [splɪt] (Liberman and Pisoni, 1977). How are we to rationalize such an equivalence? Repp *et al.* (1978) point out that neither a single feature detector nor a set of feature detectors, integrated by some higher level decision mechanism (as proposed by Massaro and Cohen, 1976), nor, it would seem, any purely auditory principle can explain why such phenomenologically diverse cues can be traded off and integrated into a unitary percept.

As a final example, consider a positively daedalian series of experiments by Bailey and Summerfield (1978). They explored the conditions under which a particular voiceless stop ([p], [t] or [k]) is perceived if a silence is introduced between [s] and a following vowel. Whether a stop is heard at all depends, of course, on the duration of the silence, but the effect of that duration itself depends on the onset frequency of F1, while the perceived place of articulation depends on the duration of the closure, on spectral properties at the offset of [s] and on the relation between those properties and the following vowel (cf. Dorman *et al.*, 1977). Bailey and Summerfield (1978, p. 55) suggest that, "... given sufficiently precise stimulus control, perceptual sensitivity could be demonstrated to every difference between two articulations" (cf. Haggard, 1977). Again, the problem is to understand the principles by which such heterogeneous collections of spectral and temporal cues are combined into a percept. What rationalizes their integration?

The answer, explicitly proposed by the authors of these several studies, is that the cues are held together by their origin in the integral, articulatory gesture. We should be absolutely clear that this is *not* a form of motor theory. Rather, it is a description of what the perceptual system appears to do. The system follows the moment-to-moment acoustic flow, apprehending an auditory "motion picture," as it were, of the articulation, in a manner totally analogous to that by which the visual system might follow the optic flow to apprehend the articulation by reflected light rather than by radiated sound (cf. Fowler, 1979; Studdert-Kennedy, 1977).

Reading Lips and Reading Spectrograms

The argument is clarified, and developed, in a recent study of lip reading by Summerfield (1979). Subjects were asked to write down a series of sentences spoken over an audio system, but simultaneously masked by the talker's own voice reading another text. There were three conditions of interest to the present discussion: (1) audio alone; (2) audio with full video of the speaker's face; (3) audio with a display of the speaker's lips. Without any training, naïve subjects scored 23%, 65% and 54% correct, respectively. In a second experiment, Summerfield analyzed errors made against deliberately conflicting video. He found, as did McGurk and McDonald (1976), that subjects frequently made judgments reflecting a compound between the auditory and visual information. Summerfield (as also Haggard, 1977) points out that such instantaneous interplay between modalities seems to require a common metric by which the two streams of information can be combined (cf. Campbell and Dodd, in press). (The problem, incidentally, is quite general and may apply to any sound-producing visual event.)

It is instructive to compare the ease with which naïve subjects used the visual display of face or lips with the obvious difficulty experienced by even the most skilled spectrogram reader. Cole, Rudnicky, Reddy and Zue (1978) report a systematic study of subject VZ who has been studying acoustic phonetics for more than seven years and has logged some 2000–2500 hours reading spectrograms – perhaps as many hours as a child of two years has spent listening to speech. Despite the fact that VZ is free to use the ample

context of vision (rather than the narrow window of audition) and that he reports conscious, acoustic-phonetic interpretation of visual context at least 18% of the time; despite the fact that he came to the spectrograms knowing that their visual segments were not isomorphic with phonetic segments (a crucial piece of knowledge that cannot be derived from the spectrograms themselves); despite the fact that, in the hours devoted to spectrograms, he could probably have learned to read several foreign languages with fair proficiency, VZ now transcribes spectrograms at a rate some 20 to 40 times real time (Cole, personal communication).

One is not surprised. There are, after all, biological constraints on learning (see Hinde and Stevenson-Hinde, 1973): pigeons learn more readily to peck plastic keys for grain and to jump to avoid shock than vice versa. The visual display of talking lips and face is natural and its code is known to every sighted speaker of a natural language, as the code of a spectrographic display is not. Watching its mother's face and listening to her speak, the infant learns to perceive articulation directly, whether by light or by sound.

Extracting Information from the Syllable

The primary unit of perception is evidently the unsegmented syllable (the rhythmic unit of nursery rhymes), and there is ample evidence for perceptual interaction among its components (see Studdert-Kennedy, 1976, for a review). For a recent example, Hasegawa and Daniloff (1976) synthesized two fricative continua, /s/ – /ʃ/, after two different vowels, /i/ and /u/, and found a significant shift in the phoneme boundary as a function of preceding vowel. Kunisaki and Fujisaki (1977) developed the finding by showing that contextual dependency in perception corrects for a mirror-image contextual dependency in production: just as the frequencies of fricative poles and zeros are lower before /u/ than before /ɑ/, so, in perception, the frequencies of the poles and zeros at the synthetic boundary between /s/ and /ʃ/ are higher before /ɑ/ than before /u/. These results mesh neatly with our earlier conclusion that consonantal onset is judged as part of a dynamic, temporal pattern.

Just such a process has recently been shown to play an important role also in vowel perception. Strange, Jenkins and Edman (1978) recorded tokens of /b/–vowel–/b/ syllables with ten different medial vowels, spoken by several speakers. They edited out the steady-state syllable nuclei (50% to 65% of the entire syllable, depending on the vowel) and presented various fragments of the syllables for identification. The results varied with both speaker and vowel, but overall, for three speakers of the same dialect as the listeners, error rates on the original syllables, on the syllables without their centers ("silent centers") and on the isolated centers were 4%, 10% and 18% respectively. The error rates for either the initial or the final transitions alone were approximately 60%. Evidently, the dynamic sweep of the spectral information and its temporal distribution across the syllable was the principal source of listener information in identifying these vowels, even when that portion usually said to characterize a vowel (namely, its steady state) was completely missing.

Results such as these return us to the segmentation issue. Clearly, there was little

basis for peripheral segmentation in these syllables. In fact, one is tempted to suppose that listeners recognized syllables (Massaro, 1975) or perhaps phone-pairs (Klatt, 1978) rather than phonemes. Mermelstein (1978) reports a subtle experiment that speaks to this issue. He varied the duration and first formant frequency of the steady-state nucleus of synthetic syllables to yield /bɛd/, /bæd/, /bɛt/, /bæt/. Notice that exactly the same acoustic information (namely, duration of the steady-state nucleus) controls both vowel and final consonant decision. Accordingly, if subjects are asked to determine duration boundaries for both consonant voicing and vowel quality as a function of F1 frequency, and if the boundaries prove to be correlated, then we can conclude that listeners made a single – presumably syllabic – decision. However, if the boundary values prove independent, we can conclude that listeners recognized phonemes rather than syllables, and that they made two phonetic decisions on the basis of a single piece of acoustic information. This was, in fact, the outcome. If this is the normal mode of speech perception, it would seem that, even if syllabic segmentation is peripheral (cf. Myers *et al.*, 1975), phonemic segmentation may be a central process consequent upon classification. Usually, this process is facilitated by auditory contrast within the syllable (cf. Bondarko, 1969).

CONTINUOUS SPEECH

We now come full circle to continuous speech with its prosody, syntax and "real world" constraints. Here, the main question is whether the perceptual processes we have been discussing up to this point have any bearing at all. Is it possible, for example, that, given the contextual aids of prosody, syntax, semantics, the listener needs no more than the "auditory contour" of a word (Nooteboom *et al.*, 1976; cf. Morton and Long, 1976) or perhaps a few "invariant features" (Cole and Jakimik, 1978) to gain access to his lexicon?

I have no space for a full discussion of this issue (see Liberman and Studdert-Kennedy, 1978). But a good place to start is with a paper by Shockey and Reddy (1975) who studied speech recognition in the absence of phonological and all other higher order constraints. They recorded some 50 short utterances, spoken by native speakers of 11 different languages, and presented them to four phoneticians for transcription. The transcriptions were then compared with a "target" description, determined from native speakers and spectral analysis. The average "correct" score for the four transcribers was 56% and their average agreement, 50%. Comparable scores for transcription of a familiar language, without contextual or syntactic constraints, would be roughly 90% – the level reached by the three transcribers of Cole *et al.* (1978), in their spectrogram-reading study, cited above, and, moreover, a level close to that of VZ himself when reading spectrograms. The difference of roughly 40% is evidently due to the transcribers' knowledge of the phonology of the language being transcribed.

The point of this example is that the main difference between listening to continuous speech in a familiar language and to isolated words in a foreign one may not be in the syntax, semantics or real world constraints so much as in the phonology. This is a simpli-

fication, since phonology and syntax are not independent. But it serves to emphasize that phonology makes linguistic communication possible by setting limits on how a speaker is permitted to articulate and what a listener can expect to hear (Liberman and Studdert-Kennedy, 1978). The problem of how the listener extracts and combines information from the signal to arrive at a unitary percept is, of course, exactly the same for continuous speech as for isolated words.

The function of the other higher order constraints – syntax, context, semantics – is facilitative. They serve to delimit the sampling space from which the listener's percepts may be drawn. This is well illustrated by several experiments of Cole and Jakimik (1978), using the ingenious "listening for mispronunciations" technique, devised by Cole (1973). Subjects are asked to listen to a recorded story into which mispronunciations have been systematically introduced. Their accuracy and speed of detection is then measured as a function of different variables. Mispronunciations prove to be more rapidly reported for high than for low transitional probability words (cf. Morton and Long, 1976), for words appropriate to a theme than for words inappropriate, for words implied by previous statements than for words not implied, and so on. Presumably the more rapid reports reflect the varied ways in which thresholds for words are lowered by contextual factors. Of course, the fact that listeners recover the words at all means that they can do so without a full phonetic analysis. But this does not mean that they regularly do so without any phonetic analysis at all.

By far the fullest and most careful account of the interactive processes of word recognition in continuous speech is offered by Marslen-Wilson (1975; Marslen-Wilson and Welsh, 1978). His experimental procedure also involves mispronunciations, but the subjects' task is to shadow the text as rapidly as possible. Marslen-Wilson examines the effects of context on the frequency of fluent restorations. These restorations are often so fast that the shadower begins to say the correct word (e.g., "company") before the second syllable of the mispronounced word (e.g., "compsiny") has begun (cf. Kozhevnikov and Chistovich, 1965). Since such restorations only occur when the disrupted word is syntactically and semantically apt, it is evident that these higher order factors have facilitated recovery of the correct word. However, they cannot do so for an entire utterance in the absence of all phonetic information. It is reassuring to read as the conclusion of a lengthy and subtle discussion of these matters: ". . . word-recognition in continuous speech is fundamentally data-driven, in the specific sense that the original selection of word-candidates is based on the acoustic-phonetic properties of the initial segment of the incoming word" (Marslen-Wilson and Welsh, 1978, p. 60; cf. Nakatani and Dukes, 1977). Perhaps all these years of studying CV syllables have not been wasted after all.

CEREBRAL SPECIALIZATION

Nonetheless, opposition between the two modes of lexical access – holistic, from "auditory contour," analytic, from phonetic segments – should not be too sharply drawn. The work of Zaidel (1978a,b) with "split-brain" patients has demonstrated that holistic access is certainly possible. The cerebral hemispheres of such patients have

been surgically separated by section of the connecting pathways (corpus callosum) for relief of epileptic seizure. The separation permits an investigator to assess the linguistic capacities of each hemisphere independently. Zaidel (1978a,b) has shown that the isolated right hemisphere of such a patient, though totally mute, can recognize a sizeable auditory lexicon and has a rudimentary syntax sufficient for understanding phrases of up to three or four words in length. However, it is incapable of identifying nonsense syllables or of performing tasks that call for phonetic analysis, such as recognizing rhyme (cf. Levy, 1974). This phonetic deficit evidently precludes short-term verbal store, thus limiting the right hemisphere's capacity for syntactic analysis of lengthy utterances, and forces organization of language around meaning. Whether we assume a similar, subsidiary organization in the left hemisphere or some process of interhemispheric collaboration, it is clear that normal language comprehension could, at least in principle, draw on both holistic and analytic mechanisms.

At the same time, Zaidel's work provides striking support for the hypothesis, originally derived from dichotic studies, that the distinctive linguistic capacity of the left hemisphere is for phonological analysis of auditory pattern (Studdert-Kennedy and Shankweiler, 1970). Further support has come from electroencephalography (Wood, 1975) and, quite recently, from studies of the effects of electrical stimulation during craniotomy (Ojemann and Mateer, 1979). The latter work isolated, in four patients, left frontal, temporal and parietal sites, surrounding the final cortical motor pathway for speech, in which stimulation blocked both sequencing of oro-facial movements and phoneme identification.

This fascinating discovery meshes neatly with a growing body of data and theory that has sought, in recent years, to explain the well-known link between lateralizations for hand control and speech. Semmes (1968) offered a first account of the association by arguing, from a lengthy series of gunshot lesions, that the left hemisphere is focally organized for fine motor control, the right hemisphere diffusely organized for broader control. Subsequently, Kimura and her associates reported that skilled manual movements (Kimura and Archibald, 1974) and non-verbal oral movements (Mateer and Kimura, 1977) tend to be impaired in cases of non-fluent aphasia. These impairments are specifically for the sequencing of fine motor movements and are consistent with other behavioral evidence that motor control of the hands and of the speech apparatus is vested in related neural centers (Kinsbourne and Hicks, 1979). In fact, Kimura (1976, p. 154) has proposed that ". . the left hemisphere is particularly well adapted, not for symbolic function *per se*, but for the execution of some categories of motor activity which happened to lend themselves readily to communication." Among these categories we must, incidentally, include those that support the complex "phonological" and morphological processes of manual sign languages, now being discovered by the research of Klima, Bellugi and their colleagues (Klima and Bellugi, 1979).

The drift of all this work is toward a view of the left cerebral hemisphere as the locus of interrelated sensorimotor centers, essential to the development of language, whether spoken or signed. To understanding of the speech sensorimotor system, perceptual studies of dichotic listening will doubtless contribute. Indeed, important dichotic studies have recently found evidence for the double dissociation of left and right hemisphere, speech

and music, in infants as young as two or three months (Entus, 1977; Glanville, Best and Levenson, 1977). However, dichotic work has not fulfilled its early promise, largely because it has proved extraordinarily difficult to partial out the complex of factors, behavioral and neurological, that determine the degree of observed ear advantage (cf. Shankweiler and Studdert-Kennedy, 1975). For the future, we may increasingly rely on instrumental techniques for monitoring brain activity, such as the blood-flow studies of Lassen and his colleagues (Lassen, Ingvar and Skinhøj, 1978), induced reversible lesions by focal cooling (Zaidel, 1978b), improved methods of electroencephalographic analysis, auditory evoked potentials (Molfese, Freeman and Palermo, 1975) and, perhaps infrequently, direct brain stimulation.

References

ABERCROMBIE, D. (1967). *Elements of General Phonetics* (Chicago).

ABRAMSON, A.S. (1977). Laryngeal timing in consonant distinctions. *Phonetica*, **34**, 295-303.

ABRAMSON, A.S. and LISKER, L. (1973). Voice timing perception in Spanish word-initial stops. *Journal of Phonetics*, **1**, 1-8.

ADES, A.E. (1976). Adapting the property detectors for speech perception. In R.J. Wales and E. Walker (eds.), *New Approaches to Language Mechanisms* (Amsterdam).

ADES, A.E. (1977). Source assignment and feature extraction in speech. *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 673-685.

AINSWORTH, W.A. (1972). Duration as a cue in the recognition of synthetic vowels. *Journal of the Acoustical Society of America*, **15**, 72-100.

ALLEN, G. (1972). The location of rhythmic stress beats in English: An experimental study. *Language and Speech*, **15**, 72-100.

ASLIN, R.N. and PISONI, D.B. (1978). Some developmental processes in speech perception. Paper presented at NICHD Conference on Child Phonology: Perception, Production and Deviation, Bethesda, Maryland, May 28-31, 1978.

BAILEY, P.J. and SUMMERFIELD, Q. (1978). Some observations on the perception of [s] + stop clusters. *Haskins Laboratories Status Report on Speech Research*, SR-53, 25-60.

BLUMSTEIN, S.E. and STEVENS, K.N. (1979). Acoustic invariance in speech production. *Journal of the Acoustical Society of America*, **66**, 1001-1017.

BONDARKO, L.V. (1969). The syllable structure of speech and distinctive features of phonemes. *Phonetica*, **20**, 1-40.

BRADY, S.A. and DARWIN, C.J. (1978). Range effect in the perception of voicing. *Journal of the Acoustical Society of America*, **63**, 1556-1558.

CAMPBELL, R. and DODD, B. (in press). Hearing by eye. *Quarterly Journal of Experimental Psychology*.

CARNEY, A.E., WIDIN, G.P. and VIEMEISTER, N.F. (1977). Noncategorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America*, **62**, 961-970.

CHISTOVICH, L.A., FYODOROVA, N.A., LISSENKO, D.M. and ZHUKOVA, M.G. (1975). Auditory segmentation of acoustic flow and its possible role in speech processing. In G. Fant and M.A.A. Tathem (eds.), *Auditory Analysis and the Perception of Speech* (New York), pp. 221-232.

COHEN, A. and NOOTEBOOM, S.G. (eds.) (1975). *Structure and Process in Speech Perception* (New York).

COLE, R.A. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception and Psychophysics*, **1**, 153-156.

COLE, R.A. and JAKIMIK, J. (1978). Understanding speech: How words are heard. In G. Underwood (ed.), *Strategies of Information Processing* (New York).

COLE, R.A. and SCOTT, B. (1973). Perception of temporal order in speech: The role of vowel transitions. *Canadian Journal of Psychology*, **27**, 441-449.

COLE, R.A., RUDNICKY, A., REDDY, R. and ZUE, V.W. (1978). Speech as patterns on paper. In R.A. Cole (ed.), *Perception and Production of Fluent Speech* (Hillsdale, New Jersey).

COLLIER, R. and 't HART, J. (1975). The role of intonation in speech perception. In A. Cohen and S.G. Nooteboom (eds.), *Structure and Process in Speech Perception* (New York), pp. 107-123.

COOPER, W.E. (1975). Selective adaptation of speech. In F. Restle, R.M. Shiffrin, N.J. Castellan, H. Lindman and D.B. Pisoni (eds.), *Cognitive Theory* (Hillsdale, New Jersey).

COOPER, W.E. (1976). Syntactic control of timing in speech production: A study of complement clauses. *Journal of Phonetics*, **4**, 151-171.

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

CUTLER, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics*, **20**, 55-60.

CUTLER, A. and FOSS, D.J. (1977). On the role of sentence stress in sentence processing. *Language and Speech*, **20**, 1-10.

CUTTING, J.E. and ROSNER, B.S. (1974). Categories and boundaries in speech and music. *Perception and Psychophysics*, **16**, 564-570.

DARWIN, C.J. (1975). On the dynamic use of prosody in speech perception. In A. Cohen and S.G. Nooteboom (eds.), *Structure and Process in Speech Perception* (New York).

DARWIN, C.J. and BETHELL-FOX, C.E. (1977). Pitch continuity and speech source attribution. *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 665-672.

DORMAN, M.F., CUTTING, J.E. and RAPHAEL, L. (1975). Perception of temporal order in vowel sequences with and without formant transitions. *Journal of Experimental Psychology: Human Perception and Performance*, **1**, 121-129.

DORMAN, M.F., STUDDERT-KENNEDY, M. and RAPHAEL, L.J. (1977). Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception and Psychophysics*, **22**, 109-122.

EIMAS, P.D. and CORBIT, J.D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, **4**, 99-109.

EIMAS, P.D. and MILLER, J.L. (1978). Effects of selective adaptation on the perception of speech and visual patterns: Evidence for feature detectors. In R.D. Walk and H.L. Pick, Jr. (eds.), *Perception and Experience* (New York).

ELMAN, J.L., DIEHL, R.L. and BUCHWALD, S.E. (1977). Perceptual switching in bilinguals. *Journal of the Acoustical Society of America*, **62**, 971-974.

ENTUS, A.K. (1977). Hemispheric asymmetry in processing dichotically presented speech and nonspeech stimuli by infants. In S.J. Segalowitz and F.A. Gruber (eds.), *Language Development and Neurological Theory* (New York), pp. 64-73.

FISCHER-JØRGENSEN, E. (1972). Tape cutting experiments with Danish stop consonants in initial position. *Annual Report of the Institute of Phonetics, University of Copenhagen*, 7.

FOWLER, C.A. (1979). Perceptual centers in speech production and perception. *Perception and Psychophysics*, **25**, 375-388.

FUJISAKI, H., NAKAMURA, K. and IMOTO, T. (1975). Auditory perception of duration of speech and non-speech stimuli. In G. Fant and M.A.A. Tatham (eds.), *Auditory Analysis and Perception of Speech* (New York), pp. 197-220.

GANONG, F. (1978). A word advantage in phoneme boundary experiments. *Journal of the Acoustical Society of America*, **63**, 520(A).

GLANVILLE, B.B., BEST, C.T. and LEVENSON, R. (1977). A cardiac measure of asymmetries in

infant auditory perception. *Developmental Psychology*, **13**, 54-59.

GOLDSTEIN, L. (1977). Categorical features in speech perception and production. To appear in V. Fromkin (ed.), *Proceedings of the Workshop on Slips of the Tongue and Ear* (Vienna). (Also *UCLA Working Papers in Phonetics*, 39).

HAGGARD, M.P. (1977). Do we want a theory of speech perception? Paper presented to the Research Conference on Speech-Processing Aids for the Deaf, Gallaudet College, Washington, D.C., May 23-26.

HALL, L.L. and BLUMSTEIN, S.E. (1978). The effect of syllabic stress and syllabic organization on the identification of speech sounds. *Perception and Psychophysics*, **24**, 137-144.

't HART, J. and COHEN, A. (1973). Intonation by rule: A perceptual quest. *Journal of Phonetics*, **1**, 309-327.

HASEGAWA, A. and DANILOFF, R.G. (1976). Effects of vowel context upon labeling the /s/ – /ʃ/ continuum. *Journal of the Acoustical Society of America*, **59**, 525(A).

HINDE, R.A. and STEVENSON-HINDE, J. (1973). *Constraints on Learning* (New York).

KIMURA, D. (1976). The neural basis of language *qua* gesture. In H. Whitaker and H.A. Whitaker (eds.), *Studies in Neurolinguistics*, vol. III (New York).

KIMURA, D. and ARCHIBALD, Y. (1974). Motor functions of the left hemisphere. *Brain*, **97**, 337-350.

KINSBOURNE, M. and HICKS, R.E. (1979). Mapping cerebral functional space: Competition and collaboration in human performance. In M. Kinsbourne (ed.), *Asymmetrical Function of the Brain* (New York), pp. 267-273.

KLATT, D.H. (1978). Speech perception: A model of acoustic phonetic analysis and lexical access. In R.A. Cole (ed.), *Perception and Production of Fluent Speech* (Hillsdale, New Jersey).

KLATT, D.H. (1977). Review of the ARPA Speech Understanding Project. *Journal of the Acoustical Society of America*, **62**, 1345-1366.

KLATT, D.H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, **59**, 1208-1221.

KLATT, D.H. and COOPER, W.E. (1975). Perception of segment duration in sentence context. In A. Cohen and S.G. Nooteboom (eds.), *Structure and Process in Speech Perception* (New York), pp. 69-89.

KLIMA, E.S. and BELLUGI, U. (1979). *The Signs of Language* (Cambridge, Mass.).

KOZHEVNIKOV, V.A. and CHISTOVICH, L.A. (1965). Rech 'artikuliatsiia i vospriiatiie. Transl. as *Speech Articulation and Perception*. Clearinghouse for Federal Scientific and Technical Information, JPRS 30-543 (Washington, D.C.).

KUHN, G.M. (1975). On the front cavity resonance and its possible role in speech perception. *Journal of the Acoustical Society of America*, **58**, 428-433.

KUNISAKI, O. and FUJISAKI, H. (1977). On the influence of context upon perception of voiceless fricative consonants. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics, University of Tokyo*, **11**, 85-91.

LASSEN, N.A., INGVAR, D.H. and SKINHØJ, E. (1978). Brain function and blood flow. *Scientific American*, **239**, 62-71.

LEHISTE, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, **5**, 253-264.

LEVY, J. (1974). Psychobiological implications of bilateral asymmetry. In S.J. Dimond and J.G. Beaumont (eds.), *Hemisphere Function in the Human Brain* (London).

LIBERMAN, A.M. and PISONI, D.B. (1977). Evidence for a special speech-perceiving subsystem in the human. In T.H. Bullock (ed.), *Recognition of Complex Acoustic Signals* (Berlin), pp. 59-76.

LIBERMAN, A.M. and STUDDERT-KENNEDY, M. (1978). Phonetic perception. In R. Held, H. Leibowitz and H.L. Teuber, *Handbook of Sensory Physiology*, Vol. VIII (Heidelberg).

LIBERMAN, A.M., HARRIS, K.S., HOFFMAN, H.S. and GRIFFITH, B.C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, **53**, 358-368.

LINDBLOM, B. and RAPP, K. (1973). Some temporal regularities of spoken Swedish. *Papers from the Institute of Linguistics, University of Stockholm*, **21**, 1-59.

LISKER, L. (1975). Is it VOT or a first-formant transition detector? *Journal of the Acoustical Society of America*, **57**, 1547-1551.

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

MARSLEN-WILSON, W.D. (1975). Sentence perception as an interactive parallel process. *Science*, **189**, 226-228.

MARSLEN-WILSON, W.D. and WELSH, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.

MARTIN, J.G. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological Review*, **79**, 487-509.

MASSARO, D.W. (1975). Preperceptual images, processing time, and perceptual units in speech perception. In D.W. Massaro (ed.), *Understanding Language* (New York).

MASSARO, D.W. and COHEN, N.M. (1976). The contribution of fundamental frequency and voice onset time to the /zi/–/si/ distinction. *Journal of the Acoustical Society of America*, **60**, 704-717.

MATEER, C. and KIMURA, D. (1977). Impairment of non-verbal oral movements in aphasia. *Brain and Language*, **4**, 262-276.

MCGURK, H. and MCDONALD, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.

MERMELSTEIN, P. (1975). A phonetic context-controlled strategy for segmentation and phonetic labeling of speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-23, 79-82.

MERMELSTEIN, P. (1978). On the relationship between vowel and consonant identification when cued by the same acoustic information. *Perception and Psychophysics*, **23**, 331-336.

MILLER, J.L. (1977). Nonindependence of feature processing in initial consonants. *Journal of Speech and Hearing Research*, **20**, 519-528.

MILLER, J.D., WIER, C.C., PASTORE, R., KELLY, W.J. and DOOLING, D.J. (1976). Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America*, **60**, 410-417.

MIYAWAKI, K., STRANGE, W., VERBRUGGE, R., LIBERMAN, A.M., JENKINS, J.J. and FUJIMURA, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception and Psychophysics*, **18**, 331-340.

MOLFESE, D.L., FREEMAN, R.B. and PALERMO, D.S. (1975). The ontogeny of brain lateralization for speech and nonspeech stimuli. *Brain and Language*, **2**, 356-368.

MORTON, J. and LONG, J. (1976). Effect of word transition probability on phoneme identification. *Journal of Verbal Learning and Verbal Behavior*, **15**, 43-51.

MORTON, J., MARCUS, S. and FRANKISH, C. (1976). Perceptual centers (P-centers). *Psychological Review*, **83**, 405-408.

MYERS, T.F., ZHUKOVA, M.G., CHISTOVICH, L.A. and MUSHNIKOV, V.N. (1975). Auditory segmentation and the method of dichotic stimulation. In G. Fant and M.A.A. Tatham (eds.), *Auditory Analysis and the Perception of Speech* (New York), pp. 243-274.

NAKATANI, L.H. and DUKES, K.D. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America*, **62**, 714-719.

NOOTEBOOM, S.G., BROKX, J.P.L. and DE ROOIJ, J.J. (1976). Contributions of prosody to speech perception. Institute for Perception Research Preprint, Eindhoven, Netherlands. (To be published in W.J.M. Levelt and G.B. Flores d'Arcais, *Studies in Language Perception* (New York).

ODEN, G.C. and MASSARO, D.W. (1978). Integration of featural information in speech perception. *Psychological Review*, **85**, 172-191.

ÖHMAN, S.E.G. (1975). What is it that we perceive when we perceive speech? In A. Cohen and S.G. Nooteboom (eds.), *Structure and Process in Speech Perception* (New York), pp. 36-48.

OJEMANN, G. and MATEER, C. (1979). Human language cortex: Localization of memory, syntax and sequential motor-phoneme identification systems. *Science*, **205**, 1401-1403.

PARKER, F. (1977). Distinctive features and acoustic cues. *Journal of the Acoustical Society of America*, **62**, 1051-1054.

PASTORE, R.E., AHROON, W.A., BAFFUTO, K.J., FRIEDMAN, C., PULEO, J.S. and FINK, E.A. (1977). Common factor model of categorical perception. *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 686-696.

PISONI, D.B. (1977). Identification and discrimination of the relative onset of two component tones: Implications for the perception of voicing in stops. *Journal of the Acoustical Society of America*, **61**, 1352-1361.

PISONI, D.B. and LAZARUS, J.H. (1974). Categorical and non-categorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, **55**, 328-333.

REPP, B.H. (1977). Dichotic competition of speech sounds: The role of acoustic stimulus structure. *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 37-50.

REPP, B.H., LIBERMAN, A.M., ECCARDT, T. and PESETSKY, D. (1978). Perceptual integration of cues for stop, fricative and affricate manner. *Haskins Laboratories Status Report on Speech Research*, SR-53(2), 61-83.

SAMUEL, A.G. (1977). The effect of discrimination training on speech perception: Noncategorical perception. *Perception and Psychophysics*, **22**, 321-330.

SAWUSCH, J.R. (1977). Peripheral and central processing in speech perception. *Journal of the Acoustical Society of America*, **62**, 738-750.

SAWUSCH, J.R. and PISONI, D.B. (1974). On the identification of place and voicing features in synthetic stop consonants. *Journal of Phonetics*, **2**, 181-194.

SEARLE, C.L., JACOBSON, J.Z. and RAYMENT, S.G. (1979). Stop consonant discrimination based on human audition. *Journal of the Acoustical Society of America*, **65**, 799-809.

SEMMES, J. (1968). Hemispheric specialization: A possible clue to mechanism. *Neuropsychologia*, **6**, 11-26.

SHANKWEILER, D.P. and STUDDERT-KENNEDY, M. (1975). A continuum of lateralization for speech perception? *Brain and Language*, **2**, 212-225.

SHIELDS, J.L., MCHUGH, A. and MARTIN, J.G. (1974). Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. *Journal of Experimental Psychology*, **102**, 250-255.

SHOCKEY, L. and REDDY, R. (1975). Quantitative analysis of speech perception. In G. Fant (ed.), *Proceedings of the Speech Communication Seminar (Stockholm, Sweden)* (New York).

SIMON, C. and FOURCIN, A.J. (1978). Cross-language study of speech-pattern learning. *Journal of the Acoustical Society of America*, **63**, 925-935.

SIMON, H.J. and STUDDERT-KENNEDY, M. (1978). Selective anchoring and adaptation of phonetic and nonphonetic continua. *Journal of the Acoustical Society of America*, **64**, 1338-1368.

SINGH, S. (1978). Distinctive features: A measurement of consonant perception. In S. Singh (ed.), *Measurement Procedures in Speech, Hearing and Language* (Baltimore), pp. 93-155.

SINGH, S., WOODS, D.R. and BECKER, G.M. (1972). Perceptual structure of 22 prevocalic English consonants. *Journal of the Acoustical Society of America*, **52**, 1698-1713.

STEVENS, K.N. (1975). The potential role of property detectors in the perception of consonants. In G. Fant and M.A.A. Tatham (eds.), *Auditory Analysis and Perception of Speech* (New York), pp. 303-330.

STEVENS, K.N. and BLUMSTEIN, S. (1978). Invariant cues for place of articulation. *Journal of the Acoustical Society of America*, **64**, 1358-1368.

STEVENS, K.N. and KLATT, D.H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *Journal of the Acoustical Society of America*, **55**, 653-659.

STRANGE, W. and JENKINS, J.J. (1978). The role of linguistic experience in the perception of speech. In H.L. Pick, Jr. and R.D. Walk (eds.), *Perception and Experience* (New York).

STRANGE, W., JENKINS, J.J. and EDMAN, T.R. (1978). Dynamic information specifies vowel identity. *Journal of the Acoustical Society of America*, **63**, S5 (A).

STRANGE, W., VERBRUGGE, R., SHANKWEILER, D.P. and EDMAN, T.R. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, **60**, 213-224.

STUDDERT-KENNEDY, M. (1976). Speech perception. In N.J. Lass (ed.), *Contemporary Issues in Experimental Phonetics* (New York).

STUDDERT-KENNEDY, M. (1977). Universals in phonetic structure and their role in linguistic communication. In T.H. Bullock (ed.), *Recognition of Complex Acoustic Signals* (Berlin), pp. 37-48.

STUDDERT-KENNEDY, M. and SHANKWEILER, D.P. (1970). Hemispheric specialization for speech perception. *Journal of the Acoustical Society of America*, **48**, 579-594.

STUDDERT-KENNEDY, M., LIBERMAN, A.M., HARRIS, K.S. and COOPER, F.S. (1970). Motor theory of speech perception: A reply to Lane's critical review. *Psychological Review*, **77**, 234-249.

SUMMERFIELD, Q. (1975). How a full account of segmental perception depends on prosody and vice versa. In A. Cohen and S.G. Nooteboom (eds.), *Structure and Process in Speech Perception* (New York), pp. 51-68.

SUMMERFIELD, Q. (1979). Use of visual information for phonetic perception. *Phonetica*, **36**, 314-331.

SUMMERFIELD, Q. and HAGGARD, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America*, **62**, 436-448.

SVENSSON, S.G. (1974). *Prosody and Grammar in Speech Perception.* University of Stockholm, MILUS 2.

TERBEEK, D. (1977). Across-language multi-dimensional scaling of study of vowel perception. *UCLA Working Papers in Phonetics*, 37.

VAN DEN BROECKE, M.P.R. (1976). *Hierarchies and Rank Orders in Distinctive Features* (Utrecht).

WILLIAMS, L. (1977). The perception of stop consonant voicing by Spanish-English bilinguals. *Perception and Psychophysics*, **21**, 289-297.

WOOD, C.C. (1975). Auditory and phonetic levels of processing in speech perception: Neurophysiological and information-processing analysis. *Journal of Experimental Psychology: Human Perception and Performance*, **104**, 3-20.

ZAIDEL, E. (1978a). Lexical organization in the right hemisphere. In P.A. Buser and A. Rougeul-Buser (eds.), *Cerebral Correlates of Conscious Experience*, (Amsterdam), pp. 177-197.

ZAIDEL, E. (1978b). Concepts of cerebral dominance in the split-brain. In P.A. Buser and A. Rougeul-Buser (eds.), *Cerebral Correlates of Conscious Experience* (Amsterdam), pp. 263-285.

ZHUKOV, S.Ya., ZHUKOVA, M.G. and CHISTOVICH, L.A. (1974). Some new concepts in the auditory analysis of acoustic flow. *Soviet Physics and Acoustics*, **20**, 237-240 [*Akust. Ah.*, **20**, 386-392].

ZUE, V.W. (1976). Acoustic characteristics of stop consonants: A controlled study. *Lincoln Laboratory Technical Report, 523.*

SUMMARY OF THE REPORTER'S ADDITIONAL REMARKS ON SPEECH PERCEPTION

Michael Studdert-Kennedy summarized his report. He said that he might have misunderstood the aims of the Leningrad group to some extent. He had thought that they were looking for phonetic segments in the acoustic signal, i.e. for acoustic segments that would be isomorphic with phonetic segments, but it appears from Ludmilla Chistovich's report that they are, in fact, looking primarily for acoustic segmentation, which will, e.g., be essential for the estimation of durational events.

Discussing the problem of feature detectors, he mentioned that animals that have feature detectors and templates (e.g., the bullfrog and birds), have them because they need them, having to get along very soon after birth without parental help, but that this is not the case with the human infant, who has a long period of parental care.

Concerning the problem of the perception of sounds by means of an integration of a variety of cues, he emphasized that the idea that these cues may be held together by an underlying gesture should not be understood as a claim for a motor theory of perception, which implies that perception requires reference to the production system. The idea is rather that you perceive the production gesture directly, as you perceive the movement of a hand by the light reflected from it. If the hand were moved inside a resonating chamber that had a source exciting it, you might hear the gesture instead of seeing it.

Studdert-Kennedy added material on cerebral specialization not found in the original report. A written form of it has been added to his report in the present edition.

Studdert-Kennedy concluded by quoting L. Chistovich, who concludes her report, "We (our group) believe that the only way to describe human perception is to describe not the perception itself but the artificial speech understanding system which is most compatible with the experimental data obtained in speech perception research." He found that this was a very good statement of a heuristic programme, but emphasized that what is required is a constant interplay between the psycho-biological facts of human behaviour and whatever robotic facsimile the engineers have managed to construct.