PERCEIVING PHONETIC EVENTS*

Michael Studdert-Kennedy+

In her report on the auditory processing of speech, prepared for the Ninth International Congress of Phonetic Sciences in Copenhagen, Chistovich wrote of herself and her colleagues at the Pavlov Institute in Leningrad: "We believe that the only way to describe human speech 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" (Chistovich, 1980, p. 71). Chistovich went on to doubt that psychologists would agree with her, but I suspect that many may find her view quite reasonable. However, they would probably not find the view reasonable if we were to replace the words "speech perception" and "artificial speech understanding system" with the words "speech production" and "speech synthesis system." Perhaps that is because even an articulatory synthesizer does not look like a vocal tract, while our image of what goes on in the head is so vague that we can seriously entertain the notion that a network of inorganic plastic and wire might be made to operate on the same general principles as an organic network of blood and nerves.

Of course, this is impossible, not only because the physics and chemistry of organic and inorganic substances are different, but also because machines and animals have different origins. A machine is an artifact. Its maker designs the parts for particular functions and assembles them according to a plan. The machine then operates on principles that its maker knows and has made explicit in the plan. The development of an animal is just the reverse. There is no plan. The animal exists before its parts and the parts emerge by differentiation. In the human fetus, a hand (say) buds from the emerging arm, swells and gradually, by cell-death and other processes, differentiates into digits. There is no reason to suppose that the principles of behavioral development are different from those of morphological development. On the contrary, structure and function are deeply intertwined in both evolution and

---

+
Also Queens College and Graduate Center, City University of New York
Acknowledgment. Preparation of this chapter was supported in part by NICHD Grant No. HD-01994 to Haskins Laboratories. I thank my teachers in these matters: Carol Fowler, Scott Kelso, Alvin Liberman, Björn Lindblom, Sverker Runeson, Quentin Summerfield, and Michael Turvey. My thanks also to Bruno Repp for much stimulating discussion, and to Nancy Frishberg for providing a simultaneous translation in American Sign Language of the first portion of the paper, as it was read at the conference.

[HASKINS LABORATORIES: Status Report on Speech Research SR-74/75 (1983)]
ontogeny. Behavior emerges by differentiation, according to principles implicit in the animal's form and substance.

In short, the appropriate constraints on a model of human speech perception are biological. The model must be compatible with what we know not only of speech perception and production, but also of speech acquisition. What the infant hears determines, in part, what the infant says; and if perception is to guide production, the two processes must be, in some sense, isomorphic.

An artificial speech understanding system is therefore of limited interest to the student of human speech perception. Such a device necessarily develops in the opposite direction to the human that it is intended to mimic. For while the human infant must discover the segments of its language—words, syllables, phonemes—from their specification in the signal, the machine is granted these segments a priori by its makers. As a model of speech perception, the machine is tautologous and empty of explanatory content, because it necessarily contains only what its makers put in. Unfortunately, all our models of speech perception are essentially machine models.

What theories of event perception have to offer to the study of language, in general, and of speech perception, in particular, is a framework for a biological alternative to such models. Three aspects of the approach seem promising. First is the commitment to discovering the physical invariances that support perception, with an emphasis on the time-varying properties of events. Second is the view of event perception as amodal, independent of the sensory system by which information is gathered. This is important for several reasons, not least for the light it may throw on the bases of imitation and on the underlying capacities common to the perception of signed and spoken language. The third aspect is the general commitment to deriving cognitive process from physical principles and thus, for language, to understanding how its structure emerges from and is constrained by its modes of production and perception.

None of these viewpoints is entirely new to the study of speech perception. What is new is their possible combination in a unified approach. I will briefly discuss each aspect, but before I do, I must lay out certain general properties of language and central problems of speech perception.

**LANGUAGE STRUCTURE**

As a system of animal communication, language has the distinctive property of being open, that is, fitted to carrying messages on an unlimited range of topics. Certainly, human cognitive capacity is greater than that of other animals, but this may be a consequence as much as a cause of linguistic range. Other primate communication systems have a limited referential scope—sources of food or danger, personal and group identity, sexual inclination, emotional state, and so on—and a limited set of no more than 10 to 40 signals (Wilson, 1975, p. 183). In fact, 10 to 40 holistically distinct signals may be close to the upper range of primate perceptual and motor capacity. The distinctive property of language is that it has finessed that upper limit, by developing a double structure, or dual pattern (Hockett, 1958).
The two levels of patterning are phonology and syntax. The first permits us to develop a large lexicon, the second permits us to deploy the lexicon in predicating relations among objects and events (Liberman & Studdert-Kennedy, 1978; Studdert-Kennedy, 1981). My present concern is entirely with the first level. A six-year-old middle-class American child already recognizes some 13,000 words (Templin, 1957), while an adult’s recognition vocabulary may be well over 100,000. Every language, however primitive the culture of its speakers by Western standards, deploys a large lexicon. This is possible because the phonology, or sound pattern, of a language draws on a small set (roughly between 20 and 100 elements) of meaningless units—consonants and vowels—to construct a very large set of meaningful units, words (or morphemes). These meaningless units may themselves be described in terms of a smaller set of recurrent, contrasting phonetic properties or distinctive features. Evidently, there emerged in our hominid ancestors a combinatorial principle (later, perhaps, extended into syntax) by which a finite set of articulatory gestures could be repeatedly permuted to produce a very large number of distinctively different patterns.

Let me note, in passing, that manual sign languages have an analogous dual structure. I do not have the space to discuss this matter in any detail. However, we have learned over the past 10 to 15 years that American Sign Language (ASL) (the first language of over 100,000 deaf persons, and the fourth most common language in the United States [Mayberry, 1978]) is a fully independent language with its own characteristic formational (“phonological”) structure and syntax (Klima & Bellugi, 1979). Whether signed language is merely an analog of spoken language (related as the bat’s wing to the bird’s) or a true homolog, drawing on the same underlying neural structures, we do not know. But there can be no doubt that as we come to understand the structure, function, acquisition, and neuropsychological underpinnings of sign language, what we learn will profoundly condition our view of the biological status of language, in general.

Here, returning to my theme, I note simply that each ASL sign is formed by combining four intrinsically meaningless components: a hand configuration, a palm orientation, a place in the body space where it is formed, and a movement. There are some fifty values, or “primes,” distributed across these four dimensions; their combination in a sign follows “phonological rules,” analogous to those that constrain the structure of a syllable in spoken languages. In short, both spoken and signed languages exploit combinatorial principles of lexical formation. Their sublexical structures seem to “...provide a kind of impedance match between an open-ended set of meaningful symbols and a decidedly limited set of signaling devices” (Studdert-Kennedy & Lane, 1980, p. 35).

THE ANISOMORPHISM PARADOX

If words are indeed formed from strings of consonants and vowels, and signs from simultaneous combinations of primes, we must suppose that the listener, or viewer, somehow finds these elements in the signal. Yet from the first spectrographic descriptions of speech (Joos, 1948), two puzzling facts have been known. First, the signal cannot be divided into a neat sequence of units corresponding to the consonants and vowels of the message; at every instant, the form of the signal is determined by gestures associated with
several neighboring elements. Second, as an automatic consequence of this, the acoustic patterns associated with a particular segment vary with their phonetic context. The apparent lack of invariant segments in the signal matching the invariant segments of perception constitutes the anisomorphism paradox.

The recalcitrance of the problem is reflected in the current states of the arts of speech synthesis and automatic speech recognition. Weaving a coherent, continuous pattern from a set of discrete instructions is evidently easier than recovering the discrete instructions from a continuous pattern. Speech synthesis has thus developed to a point where a variety of systems, taking a sequence of discrete phonetic symbols as input and offering a coherent, perceptually tolerable sequence of words as output, is already in use. By contrast, automatic speech recognition is still, after thirty years of research, at its beginning. Current devices recognize limited vocabularies of no more than about a thousand words. Moreover, the words must be spoken carefully, usually by a single speaker, in a small set of syntactic frames, and be confined to a limited topic of discourse. None of these devices approaches within orders of magnitude the performance of a normal human listener.

We may gain insight into why automatic speech recognition has so far failed from the corollary fact that no one has yet succeeded in devising an acceptable acoustic substitute for speech. In the burst of technological enthusiasm that followed World War II, a characteristic endeavor was to construct a sound alphabet that might substitute for spoken sounds in a reading machine for the blind. Of the dozens of codes tested, none was more successful than Morse Code, which a highly skilled operator can follow at a rate of about 35 words a minute, as against the 150-200 words a minute of normal speech. Yet with a visual alphabet, reading rates of 300-400 words a minute are commonplace. Why should this be?

Part of the answer perhaps lies in differences between seeing and hearing. Eyes comfortably scan a spatial array of static, discrete objects for information; ears are attuned to dynamic patterns of spectral change over time rather than to the abrupt "dots and dashes" of an arbitrary code. Speech has evidently evolved to distribute the acoustic information that specifies its discrete phonetic segments in patterns of change that match the ear's capacities. Yet, ironically, theories of speech perception, like the models implicit in automatic speech recognition devices, have all assumed that the signal is a collection of more or less discrete cues or properties. Not surprisingly, with this crypto-alphabetic assumption, these theories then have difficulty in recovering an integrated percept.

RESOLVING THE PARADOX

There are two possible lines of resolution of the paradox. We may reformulate our definition of the perceptual units or we may recast our description of the acoustic signal. In what follows, I will briefly sketch two current approaches that, extended and combined, may lead toward a resolution along both these lines.
Studdert-Kennedy, M.: Perceiving Phonetic Events

Note, first, that we cannot abandon the concept of the phoneme-sized phonetic segment, and the features that describe it, without abandoning the sound structure and dual pattern on which language is premised. Moreover, there is ample evidence from historical patterns of sound change (e.g., Lehmann, 1973), errors in production (Fromkin, 1980), errors in perception (Bond & Garnes, 1980), aphasic deficit (Blumstein, 1978) and, not least, the existence of the alphabet, that the phoneme is a functional element in both speaking and listening (for fuller discussion, see Liberman & Studdert-Kennedy, 1978). What we can abandon, however, is the notion of the phoneme-sized phonetic segment as a static, timeless unit. We can attempt to recast it as a synergistic pattern of articulatory gesture, specified in the acoustic signal by spectrally and temporally distributed patterns of change.

Here, it may be useful to distinguish between the information in a spoken utterance and in its written counterpart (a similar distinction is drawn in another context by Carello, Turvey, Kugler, & Shaw, in press). Both speech and writing may serve to control a speaker's output: We may ask a subject either to repeat the words he/she hears or to read aloud their alphabetic transcription, and the two spoken outcomes will be essentially identical. But the information that subjects use to control their output is quite different in the two cases.

The form of the spoken utterance is not arbitrary: Its acoustic structure is a necessary consequence of the articulatory gestures that shaped it. In other words, its acoustic structure specifies those gestures, and the human listener has no difficulty in reading out the specifications, and thus organizing his own articulations to accord with those of the utterance. By contrast, the form of the written transcription is an arbitrary convention that specifies nothing. Rather, it is a set of instructions that indicate to the reader what he is to do, but do not specify how he is to do it (Carello, et al., in press; Turvey, personal communication). A road sign indicates "Stop," a tennis coach instructs us, "Keep your eye on the ball," but neither tells us how to do it. Their instructions are chosen to symbolize actions presumed to be in the repertoires of motorists and tennis players. If these actions were not in their repertoires, the instructions would be useless. Similarly, the elements of a transcription—whether words, syllables, or phonemes—are chosen to symbolize actions presumed to be in the repertoires of speakers. If they were not in their repertoires, the instructions would be useless. Our task is therefore to describe those actions and to understand how they are specified in the flow of speech.

Thirty years of research with synthetic speech have demonstrated that the speech signal is replete with independently manipulable "cues," which, if varied appropriately, change the phonetic percept. Two puzzling facts emerge from this work. (See Repp, 1982, for an extensive review.) First, every phonetic distinction seems to be signaled by many different cues. Therefore, to demonstrate that a particular cue is effective, we must set other cues in the synthesis program at neutral (that is, ambiguous) values. We then discover the second puzzle, namely, that equivalent, indistinguishable percepts may arise from quite different combinations of contexts and cues. Thus, Bailey and Summerfield (1980) showed that perceived place of articulation of an English stop consonant /p, t, k/, induced by a brief silence between /s/ and a following vowel (as in /spu/ or /ski/), depends on the length of the
silence, on spectral properties at the offset of /s/, and on the relation between those properties and those of the following vowel. How are we to understand the perceptual equivalence of variations in the spectral structure of a vowel and in the duration of the silence that precedes it? More importantly, how are we to understand the integration of many spectrally and temporally scattered cues into a unitary percept?

The quandary was recognized and a rationale for its solution proposed some years ago by Lisker and Abramson (1964, 1971). They pointed out that the diverse array of cues that separate so-called voiced and voiceless initial stop consonants in many languages—plosive release energy, aspiration energy, first formant onset frequency—were all consequences of variations in timing of the onset of laryngeal vibration with respect to plosive release, that is, voice onset time (VOT).

"Laryngeal vibration provides the periodic or quasi-periodic carrier that we call voicing. Voicing yields harmonic excitation of a low frequency band during closure, and of full formant pattern after release of the stop. Should the onset of voicing be delayed until some time after the release, however, there will be an interval between release and voicing onset when the relatively unimpeded air rushing through the glottis will provide the turbulent excitation of a voiceless carrier commonly called aspiration. This aspiration is accompanied by considerable attenuation of the first formant, an effect presumably to be ascribed to the presence of the tracheal tube below the open glottis. Finally, the intensity of the burst, that is, the transient shock excitation of the oral cavity upon release of the stop, may vary depending on the pressures developed behind the stop closure. Thus it seems reasonable to suppose that all these acoustic features, despite their physical dissimilarities, can be ascribed ultimately to actions of the laryngeal mechanism." (Abramson & Lisker, 1965, p. 1).

If, now, we extend this principle of articulatory coherence to other collections of cues for other phonetic features—for which, to be sure, the details have not yet been worked out—we can, at least, see how the cues may originate, and may even cohere perceptually as recurrent acoustic patterns. Moreover, we have a view of the perceptual object—consistent with Gibson's (1966, 1979) principles—as an event that modulates acoustic energy. In other words, the perceptual object is a pattern of gesture perceived directly by means of its radiated sound, or, if we are watching the movements of a signing hand, by means of a pattern of reflected light. This view, developed at Haskins Laboratories over the past thirty years, takes a step toward resolving the anisomorphism paradox by treating the perceptual object as a dynamic event rather than a static unit, but does nothing to address the problems of invariance and segmentation in the acoustic signal. For this we must turn to the work of Stevens (1972, 1975) and of Stevens and Blumstein (1978; Blumstein & Stevens, 1979, 1980).

Stevens' (1972, 1975) approach is entirely consistent with Gibson's view that "Phonemes are in the air" (Gibson, 1966, p. 94), in other words, that the acoustic signal carries invariant segments isomorphic with our phonetic percepts. For Stevens, the perceptual elements are the features of distinc-
tive feature theory (Jakobson, Fant, & Halle, 1963). He has adopted an explicitly evolutionary approach to the link between production and perception by positing that features have come to occupy those acoustic spaces where, by calculations from a vocal tract model, relatively large articulatory variations have little acoustic effect, and to be bounded by regions where small articulatory changes have a large acoustic effect. (As a simple example, the reader might test the acoustic consequences of whispering the word east, moving slowly from the high front vowel [i] through the alveolar fricative [s] to the alveolar stop [t].)

Most of Stevens' work in recent years has been concerned with acoustic properties that specify place of articulation in stop consonants, for the good reason that the acoustic correlates of this feature have seemed particularly labile and subject to contextual variation (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). For example, in a well-known series of studies (Stevens & Blumstein, 1978; Blumstein & Stevens, 1979, 1980), Stevens and Blumstein derived by acoustic analysis a set of three "templates," characterizing the gross spectral structure at onset, integrated over the first 26 ms after stop release, for the three syllable-initial, English stop consonants, [b,d,g]. They described the templates in the terminology of distinctive feature theory as diffuse-falling for [b], diffuse-rising for [d], compact for [g]. They tested the perceptual effectiveness of these brief, static spectra by synthesis, before or as part of either steady or moving formant transitions in three vowel environments, [i,a,u]. The studies are too complex and subtly devised for summary here, but the general outcome was that most subjects were able to identify the stops with 80%-100% accuracy from the first 20-30 ms after consonant onset. Nonetheless, accuracy did vary with vowel environment and, in some syllables, subjects evidently made use of what Blumstein and Stevens term "secondary" properties, such as formant transitions, to identify the consonants.

Before we examine the implications of this last fact, we should note three important aspects of this approach to the invariance problem. First, in accord with distinctive feature theory and with the acoustic analyses of Fant (1960, 1973), Stevens and Blumstein assume that phonetic information is primarily given in the entire spectral array. "Cues" are not extracted; rather, the phonetic segment is directly specified by the signal. Second, the weight assigned to the spectrum at onset is justified by recent evidence from auditory physiology (cf. Chistovich et al., 1982; e.g., Delgutte, 1982; Kiang, 1980) that the (cat) ear is particularly sensitive to abrupt spectral discontinuities, and that the number of fibers responding to the input is increased immediately following such a discontinuity. Third, Stevens and Blumstein acknowledge the role of "secondary"--and potentially context-dependent--sources of information in patterns of spectral change (i.e., formant transitions), but attempt to exclude them by positing innate property detectors. These detectors filter out the secondary properties, it is said, and enable an infant to extract the "primary" invariances, leaving the secondary properties to be learned from their co-occurrence with the primary (Stevens & Blumstein, 1978, p. 1367).

Here, in this third aspect, we see that Stevens and Blumstein have not, in fact, completely freed their theory of perceptual atomism. By dividing the properties into "primary" and "secondary," they slip back into requiring some
process of perceptual integration, accomplished, they propose, by the tautological process of "co-occurrence" or association. Moreover, the detectors themselves are purely ad hoc, tautologous entities (or processes) for which there is no independent evidence: Their existence is inferred from the fact that infants and adults respond in a particular way to stimuli that may be described as having certain properties. If we have learned nothing else from behaviorist philosophy, we should at least have learned to eschew the "Conceptual Nervous System."

Yet the detectors are supererogatory to the enterprise that Stevens and Blumstein are launched upon. The importance of their work is that they have taken the first systematic, psycholinguistically motivated, steps toward describing the invariant acoustic properties of a notoriously context-dependent class of phonetic segments. What is missing from their approach is not an imaginary physiological device, but a recognition that the signal is no more a sequence of static spectral sections than it is a collection of isolated cues. Rather the signal reflects a dynamic articulatory event of which the invariances must lie in a pattern of change.

And, indeed, moves toward this recognition have already begun. Kewley-Port (1980, 1983) has shown that an invariant pattern may be found in running spectra at stop consonant onset, and that identification accuracy for synthetic stop syllables improves, if they are synthesized from running spectra, updated at 5 ms intervals, rather than from static spectra sustained over 26 ms (Kewley-Port, Pisoni, & Studdert-Kennedy, 1983). Blumstein, Isaacs, and Mertus (1982) have found that the perceptually effective invariant may lie, not in the gross spectral shape, as originally hypothesized, but in the pattern of formant frequencies at onset. This suggests that characteristic formant shifts of the kind described in the earliest synthetic speech studies (e.g., Liberman, Cooper, Delattre, & Gerstman, 1954) may yet prove to play a role: for example, an upward shift in the low frequencies for labials, a downward shift in the high frequencies for alveolars. In fact, Lahiri and Blumstein (1982) report a cross-language (English, French, Malayalam) acoustic analysis of labial, dental, and alveolar stops that seems to be consistent with this hypothesis. The distinctions were carried by maintenance or shift in the relative weights of high and low frequencies from consonant release over the first three glottal pulses at voicing onset. All these studies move toward a dynamic rather than a static description of speech invariants.

We may see then, in (distant) prospect, a fruitful merger, consistent with theories of event perception, by which invariances in the acoustic signal are discovered as coherent patterns of spectral change, specifying a synergism of underlying articulatory gestures. From such a resolution of the invariance paradox there would follow a resolution of the segmentation paradox. For implicit in a view of the perceptual object as a coherent event is a view of "cues," "features," and, indeed, "phonemes" as descriptors rather than substantive categories of speech. The utility of features and phonemes for describing the structure of spoken languages would remain, as would—in some not yet clearly formulated sense—the functional role of the phoneme-sized phonetic segment in the organization of an utterance. But phonemes and features in perception would be seen, in origin, not as substantive categories, formed by specialized categorical mechanisms, but as emergent properties of recurrent acoustic pattern. As we will see later, this view of perception
is coordinate with current research into the origins of phonological systems.

IMITATION AND THE AMODALITY OF SPEECH PERCEPTION

Let us turn now to another body of research that encourages a view of speech perception as a particular type of event perception: research on lip reading in adults and infants. The importance of this work is that it promises to throw light on imitation, a process fundamental to the acquisition of speech.

The story begins with the discovery by McGurk and MacDonald (1976; MacDonald & McGurk, 1978) that subjects' perceptions of a spoken syllable often change, if they simultaneously watch a video display of a speaker pronouncing a different syllable. For example, if subjects hear the syllable /ba/ repeated four times, while watching a synchronized video display of a speaker articulating /ba, va, da, da/, they will typically report the latter sequence. This is not simply a matter of visual dominance in a sensory hierarchy, familiar from many intermodal studies (Marks, 1978). Nor is it a matter of combining phonetic features independently extracted from acoustic and optic displays—for example, voicing from the acoustic, place of articulation from the optic. For, although voicing is indeed specified acoustically, place of articulation may be specified both optically and acoustically, as when subjects report a consonant cluster or some merged element. Thus, presented with acoustic /ba/ and optic /ga/, subjects often report /b'ga/, /g'ba/ or a merger, /da/. (See Summerfield, 1979, for fuller discussion).

The latter effect was used in an ingenious experiment by Roberts and Summerfield (1981) to demonstrate that speech adaptation is an auditory not a phonetic process, and, more importantly, for the present discussion, to show that auditory and phonetic processes in perception can be dissociated. The standard adaptation paradigm, devised by Eimas and Corbit (1973), asks listeners to classify syllables drawn from a synthetic acoustic continuum, stretching from, say [ba] to [da], or [ba] to [pa], both before and after repeated exposure to (that is, adaptation with) one or other of the endpoint syllables. The effect of adaptation, reported in several dozen studies (see Eimas & Miller, 1978, for review), is that listeners perceive significantly fewer tokens from the continuum as instances of the syllable with which they have been adapted.

Roberts and Summerfield (1981) followed this paradigm with a series of synthetic syllables ranging from [bs] to [ds]. Their novel twist was to include a condition in which subjects were adapted audiovisually by an acoustic [bs], synchronized with an optic [gs], intended to be perceived phonetically as [ds]. In the event, six of their twelve subjects reported the adapting syllable as either [ds] or [gs], four as [ks], one as [fs], one as [ma]. Not a single subject reported the phonetic event corresponding to the adapting acoustic syllable actually presented, [bs]. Yet, after adaptation, every subject displayed a drop in the number of tokens identified as [bs], roughly equal to the drop for the control condition in which acoustic [bs] was presented alone. Thus, while subjects' auditory systems were normally adapted by the acoustic input, their conscious phonetic percepts were specified intermodally by a blend of acoustic and optic information.
We might extend the demonstration that phonetic perception is intermodal (or, better, amodal) by citing the Tadoma method in which the deaf-blind learn to perceive speech by touch, with fingers on the lips and neck of the speaker. Tactile information may even help to guide a deaf-blind individual's own articulation (Norton, Schultz, Reed, Braida, Durlich, Rabinowitz, & Chomsky, 1977). But the lip-reading studies alone suffice to raise the question of the dimensions of the phonetic percept. The acoustic information is presumably carried by the familiar pattern of formants, friction noise, plosive release, harmonic variation and so on; the optic information is carried by varying configurations of the lips and, perhaps, of the tongue and teeth (Summerfield, 1979). But how are these qualitatively distinct patterns of light and sound combined to yield an integrated percept? What we need is some underlying metric common to both the light reflected and the sound radiated from mouth and lips (Summerfield, 1979). Such a notion will hardly surprise students of action and of event-perception (e.g., Fowler, Rubin, Remez, & Turvey, 1980; Runeson & Frykholm, 1981; Summerfield, 1980). But, as I have already suggested, it is worth pursuing a little further for the light that it may throw on the bases of imitation.

Consider, first, that infants are also sensitive to structural correspondences between the acoustic and optic specifications of an event. Spelke (1976) showed that 4-month-old infants preferred to watch the film (of a woman playing "peekaboo," or of a hand rhythmically striking a wooden block and a tambourine with a baton) that matched the sound track they were hearing. Dodd (1978) showed that 4-month-old infants watched the face of a woman reading nursery rhymes more attentively when her voice was synchronized with her facial movements than when it was delayed by 400 ms. If these preferences were merely for synchrony, we might expect infants to be satisfied with any acoustic-optic pattern in which moments of abrupt change are arbitrarily synchronized. Thus, in speech they might be no less attentive to an articulating face whose closed mouth was synchronized with syllable amplitude peaks and open mouth with amplitude troughs than to the (natural) reverse. However, Kuhl and Meltzoff (1982) showed that 4- to 5-month-old infants looked longer at the face of a woman articulating the vowel they were hearing (either [i] or [a]) than at the same face articulating the other vowel in synchrony. Moreover, the preference disappeared when the signals were pure tones, matched in amplitude and duration to the vowels, so that the infant preference was evidently for a match between a mouth shape and a particular spectral structure. Similarly, MacKain, Studdert-Kennedy, Spieker, and Stern (1983) showed that 5- to 6-month-old infants preferred to look at the face of a woman repeating the disyllable they were hearing (e.g., [zu-zul]) than at the synchronized face of the same woman repeating another disyllable (e.g., [vava]). In both these studies, the infants' preferences were for natural structural correspondences between acoustic and optic information.

Interestingly, in the study by MacKain et al. (1983), the infants' preferences were only statistically significant when the infants were looking to their right sides. Kinsbourne (1973) has proposed that attention to one side of the body activates the contralateral hemisphere and facilitates processes for which that hemisphere is specialized. Given the well-known specialization of the left hemisphere for motor control of speech, we might suspect that these infants were displaying a left-hemisphere sensitivity to intermodal correspondences that could play a role in learning to speak. This
hypothesis would gain support if we could establish that the underlying metric of auditory–visual correspondence was the same as that of the auditory–motor correspondence required for an individual to repeat or "imitate" the utterances of another.

To this end we may note, first, the visual–motor link evidenced in the capacity to imitate facial expression and, second, the association across many primate species between facial expression and pattern of vocalization (Hoooff, 1976; Marler, 1975; Ohala, in press). Recently, Field, Woodson, Greenberg, and Cohen (1982) reported that 36-hour-old infants could imitate the "happy, sad and surprised" expressions of a model. However, these are relatively stereotyped emotional responses that might be evoked without recourse to the visual–motor link required for imitation of novel movements. More striking is the work of Meltzoff and Moore (1977) who showed that 12- to 21-day-old infants could imitate both arbitrary mouth movements, such as tongue protrusion and mouth opening, and (of particular interest for the acquisition of ASL) arbitrary hand movements, such as opening and closing the hand by serially moving the fingers. Here mouth opening was elicited without vocalization; but had vocalization occurred, its structure would, of course, have reflected the shape of the mouth. Kuhl and Meltzoff (1982) do, in fact, report as an incidental finding of their study of intermodal preferences, that 10 of their 32 4- to 5-month-old infants "...produced sounds that resembled the adult female's vowels. They seemed to be imitating the female talker, 'taking turns' by alternating their vocalizations with hers" (p. 1140). If we accept the evidence that the infants of this study were recognizing acoustic–optic correspondences, and add to it the results of the adult lip-reading studies, calling for a metric in which acoustic and optic information are combined, then we may conclude that the perceptual structure controlling the infants' imitations was specified in this common metric.

Evidently, the desired metric must be "...closely related to that of articulatory dynamics" (Summerfield, 1979, p. 329). Following Runeson and Frykholm (1981) (see also Summerfield, 1980), we may suppose that in the visual perception of an event we perceive not simply the surface kinematics (displacement, velocity, acceleration), but also the underlying biophysical properties that define the structure being moved and the forces that move it (mass, force, momentum, elasticity, and so on). Similarly, in perceiving speech, we do not simply perceive its "kinematics," that is, the changes and rates of change in spectral structure, but the underlying dynamic forces that produce these changes. Some such formulation is demanded by the facts of imitation on which the learning of speech and language rests.

ORIGINS OF THE SOUND PATTERN OF LANGUAGE

We come finally to a third aspect of current phonetic study, compatible with theories of action and event perception. The goal of the work to be discussed may be simply stated: to derive language from non-language. The topic is broad and complex. My comments here are brief, no more than a sketch of the approach.

As we have seen, every language builds its words or signs from a small set of meaningless elements, its phonemes or primes. These elements are themselves constructed from a small set of contrasting properties or distinc-
tive features. For modern phonology, phonemes (or syllables) and their constitutive features are axiomatic primitives that require no explanation (Chomsky & Halle, 1968; Jakobson, Fant, & Halle, 1963). A central goal of linguistic study is to describe a small set of 15–20 "given" or "universal" features that will serve to describe the phonological systems of every known language. The goal has proved difficult to achieve, in large part because the various sets of features that have been proposed as potential systemic components have lacked external constraints—for example, physiological constraints on their combination (Ladefoged, 1971).

If there is indeed a universal set of linguistic features that owes nothing to the non-linguistic capacities of talkers and listeners, their biological origin must be due to some quantal evolutionary jump, a structure-producing mutation. While modern biologists may look more favorably on evolutionary discontinuities than did Darwin (e.g., Gould, 1982), we are not justified in accepting discontinuity until we have ruled continuity out. This has not been done. On the contrary, the primacy of linguistic form has been a cardinal, untested assumption of modern phonology—with the result that phonology is sustained in grand isolation from its surrounding disciplines (Lindblom, 1980).

An alternative approach is to suppose that features and phonemes reflect prior organismic constraints from articulation, perception, memory, and learning. Thus, F. S. Cooper proposed that features were shaped by the articulatory machinery. Typical speaking rates of 10 to 15 phonemes per second could "...be achieved only if separate parts of the articulatory machinery—muscles of the lips, tongue, velum, etc.—can be separately controlled, and if...a change of state for any one of these articulatory entities, taken together with the current state of others, is a change to...another phoneme...It is this kind of parallel processing that makes it possible to get high-speed performance with low-speed machinery" (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967, p. 446). A similar view was elaborated by Studdert-Kennedy and Lane (1980) for both signed and spoken language.

The most concerted attack along these lines has been developed over the past decade by Lindblom and his colleagues (e.g., Liljencrantz & Lindblom, 1972; Lindblom, 1972, 1980, in press). Their goal has been not simply to specify the articulatory and acoustic correlates of certain distinctive features (as in the work of Stevens and Blumstein, discussed above), but to show how a self-organizing system of features and phonemes may arise from perceptual and motoric constraints.

The early work (Lindblom, 1972) began by specifying a possible vowel as a point in acoustic space, defined by the set of formant frequencies associated with states of the lips, tongue, jaw, and larynx. A computer was programmed to search the space for k maximally distinct vowels according to a least squares criterion. The vowels found were then compared with those observed in languages having k vowels: Despite certain obvious deficiencies, the fit of the predicted to the observed data was remarkably good. Later studies (e.g., Lindblom, 1983) have improved the fit by incorporating the results of work in auditory psychophysics (cf. Bladon & Lindblom, 1981), together with certain articulatory constraints, and by relaxing the search criterion to one of "sufficient" rather than maximum distinctness. The last move permits more
than one solution for a k-vowel system, as indeed the observed language data require. For the present discussion, the most interesting outcome is that the derived sets of vowels form systems that invite description in terms of standard features, despite the fact that the notion "feature" was never at any point introduced into the derivation.

Recently, Lindblom has extended the procedure to derive the phoneme from sets of consonant-vowel trajectories through the acoustic space between consonant and vowel loci (Lindblom, MacNeilage, & Studdert-Kennedy, forthcoming). This work brings to bear both talker constraints (sensory discriminability, preference for less extreme articulation) and listener constraints (perceptual distance, perceptual salience) to select the syllable trajectories. Again, the interesting outcome is that when a set of trajectories is selected from a large number of possible trajectories, the syllables are not, as they might well have been, holistically distinct: Each chosen syllable does not differ from every other chosen syllable in both consonant and vowel. Rather, a few consonants and a slightly larger number of vowels occur repeatedly, while other consonants and vowel combinations do not occur at all. Thus, just as the feature emerges as a byproduct of phoneme selection, so the phoneme emerges as a byproduct of syllable selection.

This work rests on a number of assumptions that might be challenged (for example, the precise nature of talker- and listener-based constraints) and on a wealth of phonetic detail that might be questioned. Its importance does not rest on the correctness of its assumptions nor on the accuracy of its predictions—both may, and surely will, be improved in the future. Its importance lies in its style of approach: substance-based rather than formal. For if we are to do the biology of language at all, it will have to be done by tracing language to its roots in the anatomy, physiology, and social environment of its users. Only in this way can we hope to arrive at an account of language perception and production fitted to animals rather than machines.

REFERENCES


Studdert-Kennedy, M.: Perceiving Phonetic Events


