Journal of Phonetics (1986) 14, 3–28

An event approach to the study of speech perception from a direct–realist perspective

Carol A. Fowler

Dartmouth College, Hanover, New Hampshire, 03755, U.S.A.
and Haskins Laboratories, New Haven, Connecticut, 06510, U.S.A.

1. Introduction

There is, as yet, no developed event approach to a theory of speech perception and, accordingly, no body of research designed from that theoretical perspective. I will offer my view as to the form that the theory will take, citing relevant research findings where they are available. The theory places constraints on a theory of speech production too. Therefore, I will also have something to say about how talkers must talk for an event approach to be tenable. I will begin by defining the domain of the theory as I will consider it here.

An ecological event is an occurrence in the environment defined with respect to potential participants in it. Like most ecological events (henceforth, events), one in which linguistic communication takes place is highly structured and complex. Accordingly, it can be decomposed for study in many different ways. One way in which it is almost invariably decomposed by psycholinguists and linguists is into the linguistic utterance itself on the one hand, and everything else on the other. In ordinary settings in which communication takes place, this is almost certainly not a natural partitioning because it leaves out several aspects of the setting that contribute interactively with the linguistic utterance itself to the communication. These include the talker's gestures (McNeill, 1985), aspects of the environment that allow the talker to point rather than to refer verbally, and the audience, whose shared experiences with the talker affect his or her speaking style. The consequences of making this cut have not been worked out but, at least for purposes of studying language as communication, they may be substantial (cf. Beattie, 1983). For the present, however, I will preserve the partitioning and one within that as well.

The linguist, Hockett (1960), points out that languages have "duality of patterning": that is, they have words organized grammatically into sentences, and phonetic segments organized phonotactically into words. Both levels are essential to the communicative power of language.

Grammatical organization of words into sentences gives linguistic utterances two kinds of power. First the communicative content of an utterance is superadditive with respect to the contents of the words composing the sentences taken as individuals. Secondly, talkers can produce novel utterances that the audience has not heard before; and yet the utterance can convey the talkers' message to the audience. I will refer to a linguistic utterance at this level of description as a "linguistic event" and, having defined it, I will have little else to say about it until the final section of the paper.

0095-4470/86/010003 + 26 $03.00/0

The second structural tier, in which phonetic segments constitute words, support an indefinitely large lexicon. Were each word to consist of an holistic articulatory gesture rather than a phonotactically organized sequence of phonetic segments, our lexicons would be severely limited in size. Indeed, recent simulations by Lindblom (Lindblom, MacNeilage & Studdert-Kennedy, 1983) show that, as the size of the lexicon is increased (under certain constraints on how new word labels are selected), phonetic structure emerges almost inevitably from a lexicon consisting initially of holistic closing and opening gestures of the vocal tract. These simulations may show how and why phonetic structure emerged in the evolution of spoken language, and how and why it emerges in ontogeny.

I will refer to a talker's phonetically structured articulations as "speech events". It is the perception of these events that constitute the major topic of the paper. A speech event may also be defined as a linguistic utterance having phonetic structure as perceived by a listener. In defining speech event interchangeably from the perspectives of talkers and listeners, I am making the claim, following others (e.g. Shaw, Turvey & Mace, 1982) that a theory of event perception will adopt a "direct realist" stance. According to Shaw *et al.* (p. 159):

> [S]ome form of realism must be captured in any theory that claims to be a theory of perception. To do otherwise would render impossible an explanation of the practical success of perceptually guided activity.

That is, to explain the success of perceptually guided activity, perception is assumed to recover events in the real world. For this to be possible consistently (see Shaw & Bransford, 1977), perception must be direct and, in particular, unmediated by cognitive processes of inference or hypothesis testing, which introduce the possibility of error.[1]

By focusing largely on speech events, I will be discussing speech at a level at which it consists of phonetically structured syllables but not, necessarily, of grammatical, meaningful utterances. It is ironic, perhaps, that a presentation at a conference on event perception should focus on a linguistic level that does not have transparent ecological significance. However, speech events can be defended as natural partitionings of linguistic events—that is, they can be defended as ecological events—and there is important work to be done by event theorists even here.

The defense is that talkers produce phonetically structured speech, listeners perceive it as such and they use the phonetic structure they perceive to guide their subsequent behavior. Talkers reveal that they produce phonetically structured words when they make speech errors. Most submorphemic errors are misorderings or substitutions of single phonetic segments (e.g. Shattuck-Hufnagel, 1983). For their part, listeners can be shown to extract phonetic structure from a speech communication, at least in certain experimental settings. That they extract it generally, however, is suggested by the observation that they use phonetic variation to mark their identification with a social group, or to adjust their speaking style to the conversational setting. Of course, infant

[1]It may be useful to be explicit about the relationships among some concepts I will be referring to. Events are the primitive components of an "ecological" science; that is, of a study of actor/perceivers in contexts that preserve essential properties of their econiches. In the view of many theorists who engage in such studies (see, for example, the quotation from Shaw *et al.* above), the only viable version of a perceptual theory that can be developed within this domain is one that adopts a direct-realist perspective. I will take this as essential to the event (or ecological) approach, although, imaginably, a theory of the perception of natural events might be proposed from a different point of view.

perceivers must recover phonetic structure if they are to become talkers who make segmental speech errors.

This defense is not intended to suggest that perception of speech events is primary or privileged in any sense. It is only to defend it as one of the partitionings of an event involving linguistic communication that is perceived and used by listeners; therefore it is an event in its own right and requires explanation by a theory of perception.

I will discuss an event approach to phonetic perception in the next three major sections of the paper. In Sections 2 and 3 direct perception, first of local, short-term events, and then of longer ones is considered. In Section 4 some affordances of phonetically structured speech are considered.

Although there is a lot of work to be done at this more fine-grained of the dual levels of structure in language, there are also great challenges to an event theory offered by language considered as syntactically structured words that convey a message to a listener. I will discuss just two of these challenges briefly at the end of the paper, and I will suggest a perspective on linguistic events that an event theory might take.

2. Perception of speech events: a local perspective

There is a general paradigm that all instances of perception appear to fit. Perception requires events in the environment ("distal events"), and one or more "informational media"—that is, sources of information about distal events in energy media that can stimulate the sense organs—and a perceiver. As already noted, objects and occurrences in the environment are generally capable of multiple descriptions. Those that are relevant to a perceiver refer to "distal events". They have "affordances"; that is, sets of possibilities for interaction with them by the perceiver. (Affordances are "what [things] furnish, for good or ill" (Gibson, 1971/1982; see also, Gibson, 1979).) An informational medium, including reflected light, acoustic signals and the perceiver's own skin, acquires structure from an environmental event specific to certain properties of the event; because it acquires structure in this way, the medium can provide information about the event properties to a sensitive perceiver. A second crucial characteristic of an informational medium is that it can convey its information to perceivers by stimulating their sense organs and imparting some of its structure to them. By virtue of these two characteristics, informational media enable direct perception of environmental events. The final ingredient in the paradigm is a perceiver who actively seeks out information relevant to his or her current needs or concerns. Perceivers are active in two senses. They move around in the environment to intercept relevant sources of information. In addition, in ways not yet well understood, they "attune" their perceptual systems (e.g. Gibson, 1966/1982) to attend selectively to different aspects of available environmental structure.

In speech perception, the distal event considered locally is the articulating vocal tract. How it is best described to reflect its psychologically significant properties is a problem for investigators of speech perception as well as of speech production. However, I will only characterize articulation in general terms here, leaving its more precise description to Kelso, Saltzman & Tuller in their presentation. One thing we do know is that phonetic segments are realized as coordinated gestures of vocal-tract structures; that is, as coupled relationships among structures that jointly realize the segments (e.g. Kelso, Tuller, Vatikiotis-Bateson & Fowler, 1984). Therefore, studies of the activities of individual muscles or even individual articulators will not in themselves reveal the systems that constitute articulated phonetic segments.

The acoustic speech signal has the characteristics of an informational medium. It acquires structure from the activities of the vocal tract, and it can impart its structure to an auditory perceptual system thereby conveying its informational to a sensitive perceiver. In this way, it enables direct perception of the environmental source of its structure, the activities of the vocal tract. Having perceived an utterance, a listener has perceived the various "affordances" of the conversational event and can guide his or her subsequent activites accordingly.

This, in outline form, is a theory of the direct perception of speech events. The theory promotes a research program having four parts, three relating to the conditions supporting direct perception of speech events and the last relating to the work that speech events do in the environment. To assess the claim that speech events are directly perceived, the articulatory realizations of phonetic segments must be uncovered and their acoustic consequences identified. Next, the listener's sensitivity to, and use of, the acoustic information must be pinned down. Finally, the listener's use of the structure in guiding his or her activities must be studied. Although, of course, a great deal of research has been done on articulation and perception of speech, very little has been conducted from the theoretical perspective of an event theory, and very little falls within the research program just outlined.

Indeed my impression, based on publishing investigations of speech conducted from this perspective and on presentations of the theoretical perspective to other speech researchers, is that it has substantial face invalidity. There are several things seemingly true of speech production and perception that, in the view of many speech researchers, preclude development of a theory of direct perception of speech events. I will consider four barriers to the theory along with some suggestions concerning ways to surmount or circumvent them.

2.1. *The first barrier: if listeners recover articulation why do they not know it?*

A claim that perceivers see environmental events rather than the optic array that stimulates their visual systems seems far less radical than a claim that they hear phonetically structured articulatory gestures rather than the acoustic speech signal. Indeed, when Repp (1981) makes the argument that phonetic segments are "abstractions" and products of cognitive processes applied to stimulation, he says of them that "*they have no physical properties*—such as duration, spectrum and amplitude—and, therefore, *cannot be measured*" [p. 1463, italics in the original]. That is, he assumes that if phonetic segments were to have physical properties, the properties would be acoustic. Yet no-one thinks that, if the objects of visual perception—that is, trees, tables, people, etc.—do have physical properties, their properties are those of reflected light.

Somewhat compatibly, our phenomenal experience when we hear speech certainly is not of lips closing, jaws raising, velums lowering, and so on, although our visual experience is of the objects and events in the world. Of course, we do not experience surface features of the acoustic signal either; that is, silent gaps followed by stop-bursts, or formant patterns or nasal resonances.

I cannot explain the failure of our intuitions in speech to recognize that perceived phonetic events are articulatory, as compared to our intuitions about vision, which we do recognize that perceived events are environmental, but I can think of a circumstance that exacerbates the failure among researchers. If, in an experimental study, listeners do indeed recover articulatory events in perception, there is likely to be a large mismatch

between the level of description of an articulatory event that they recover and a researcher's description of the activities of the individual articulators. That is, speech researchers do not yet know what articulatory events consist of. If a perceiver does not experience "lips closing", for example, that is as it should be, because lip closure *per se* is not an articulatory speech event. Rather (see the contribution by KST), an articulatory event that is a phonetic event, for example, is a coordinated set of movements by vocal-tract structures.

By hypothesis, the percept [b] corresponds to extraction from the acoustic speech signal of information that the appropriate coordinated gestures occurred in the talker's vocal tract, just as the perceptual experience of a zooming baseball corresponds to extraction of information from the optic array that the event of zooming occurred in the environment.

The literature offers evidence from a wide variety of sources that listeners do extract information about articulation from the acoustic speech signal. Much of this evidence has recently been reviewed by Liberman & Mattingly (1985) in support of a motor theory.[2] I will select just a few examples.

2.1.1. *Perceptual equivalence of distinct acoustic "cues" specifying the same articulatory event*

In non-phonetic contexts, silence produces a very different perceptual experience from a set of formant transitions. However, interposed between frication for an [s] and a syllable sound like [It] in isolation, they may not (Fitch, Halwes, Erickson & Liberman, 1980). An appropriate interval of silence may foster perception of [p]; so may a lesser amount of silence, insufficient to cue a [p] percept in itself, followed by transitions characteristic of [p] release. Strikingly, a pair of syllables differing both in the duration of silence after the [s] frication, and in presence or absence of [p] transitions following the silence, are either highly discriminable (and more discriminable than a pair of syllables differing along just one of these dimensions) or nearly indiscriminable (and *less* discriminable than a pair differing in just one dimension) depending on whether the silence and transitions "co-operate" or "conflict". They co-operate if, within one syllable, both acoustic segments provide evidence for stop production and, within the other, they do not. They conflict if the syllable having a relatively long interval of silence appropriate to stop closure lacks the formant transitions characteristic of stop release, while the syllable with a short interval of silence has transitions. Depending on the durations of silence, these latter syllables may both sound like "split" or both like "slit".

[2]There are fundamental similarities between the view of speech perception from a direct—realist perspective and from the perspective of the motor theory. An important one is that both theories hold that the listener's percept corresponds to the talker's phonetic message, and that the message is best characterized in articulatory terms. There are differences as well. As Liberman & Mattingly (1985) note, one salient difference is that the direct—realist theory holds that the acoustic signal is, in a sense, transparent to the perceived components of speech, while the motor theory does not. According to the motor theory, achievement of a phonetic percept requires special computations on the signal that take into account both the physiological—anatomical and the phonetic constraints on the activities of the articulators. A second difference is more subtle and perhaps will disappear as the theories evolve. Liberman & Mattingly propose that the objects of speech perception (at the level of description under consideration) are the "control structures" for observed articulatory gestures. Due to coarticulatory smearing, these control structures are not entirely redundant with the collection of gestures as they occur. My own view is that the smearing is only apparent and, hence, the control structures are wholly redundant with the collections of articulatory gestures (properly described) constituting speech.

The important point is that very different acoustic properties sound similar or the same just when the information they convey about articulation is similar or the same. It should follow, and does, that when an articulation causes a variety of acoustic effects (for example, Lisker, 1978, has identified more than a dozen distinctions between voiced and voiceless stops intervocalically), the acoustic consequences individually tend to be sufficient to give rise to the appropriate perception, but none is necessary (see Liberman & Mattingly, 1985, for a review of those findings).

2.1.2. *Different perceptual experiences of the same acoustic segment just when it specifies different distal sources*

By the same token, the same acoustic segment in different contexts, where it specifies different articulations or none at all, sounds quite different to perceivers. In the experiment by Fitch *et al.* just described, a set of transitions characteristic of release of a bilabial stop will only give rise to a stop percept in that context if preceded by sufficient silence. This cannot be because, in the absence of silence, the [s] frication masks the transitions; other research demonstrates that transitions at fricative release themselves do contribute to fricative place perception (e.g. Harris, 1958; Whalen, 1981). Rather, it seems, release can only be perceived in this context given sufficient evidence for prior stop closure.

Similarly, if transitions are presented in isolation where, of course, they do not signal stop release, or even production by a vocal tract at all, they sound more or less the way that they look on a visual display; that is, like frequency rises and falls (e.g. Mattingly, Liberman, Syrdal & Halwes, 1971).

2.1.3. *"P centers"*

Spoken digits (Morton, Marcus & Frankish, 1976) or nonsense monosyllables (Fowler, 1979), aligned so that their onsets of acoustic energy are isochronous, do not sound isochronous to listeners. Asked to adjust the timing of pairs of digits (Marcus, 1981) or monosyllables (Cooper, Whalen & Fowler, 1984) produced repeatedly in alternation so that they sound isochronous, listeners introduce systematic departures from measured isochrony—just those that talkers introduce if they produce the same utterances to a real (Fowler & Tassinary, 1981; Rapp, 1971) or imaginary (Fowler, 1979; Tuller & Fowler, 1980) metronome. Measures of muscular activity supporting the talkers' articulations are isochronous in rhyming monosyllables produced to an imaginary metronome (Tuller & Fowler, 1980). Thus, talkers follow instructions to produce isochronous sequences, but due (in large part) to the different times after articulatory onset that different phonetic segments have their onsets of acoustic energy, acoustic measurements of their productions suggest a failure of isochrony. For their part, listeners appear to hear through the speech signal to the timing of the articulations.

2.1.4. *Lip reading*

Liberman & Mattingly (1985) describe a study in which an acoustic signal for a production of [ba] synchronized to a face mouthing [bɛ], [vɛ] and [ðɛ] may be heard as [ba], [va] and [ða], respectively (cf. McGurk & MacDonald, 1976). Listeners experience hearing syllables with properties that are composites of what is seen and heard, and they have no sense that place information is acquired largely visually, and vowel information auditorily. (This is reminiscent of the quotation from Hornbostel (1927), reprinted in Gibson (1966): "it matters little through which sense I realize that in the dark I have

blundered into a pigsty". Likewise, it seems, it matters little through what sense we realize what speech event has occurred.) Within limits anyway, information about articulation gives rise to an experience of hearing speech, whether the information is in the optic array or in the acoustic signal.

2.2. *The second barrier: linguistic units are not literally articulated*

A theory of perception of speech events is disconfirmed if the linguistic constituents of communications between talkers and listeners do not make public appearances. There are two kinds of reason for doubting that they do, both relating to an incommensurability that many theorists and researchers have identified between knowing and doing, between competence and performance, or even between the mental and the physical realizations of language.

One kind of incommensurability is graphically illustrated by Hockett's Easter egg analogy (Hockett, 1955). According to the analogy, articulation, and in particular, the coarticulation that inertial and other physical properties of the vocal tract requires, obliterates the discrete, context-free phonetic segments of the talker's planned linguistic message. Hockett suggests that the articulation of planned phonetic segments is analogous to the effects that a wringer would have on an array of (raw) Easter eggs. If the analogy is apt, and listeners nonetheless can recover the phonetic segments of the talker's plan, then direct detection of articulatory gestures in perception cannot fully explain perception, because the gestures themselves provide a distorted representation of the segments. To explain recovery of phonetic segments from the necessarily impoverished information in the acoustic signal, reconstructive processes or other processes involving cognitive mediation (Hammarberg, 1976, 1982; Hockett, 1955; Neisser, 1967; Repp, (1981) or non-cognitive mediation (Liberman & Mattingly, 1985) must be invoked.

Hockett is not the only theorist to propose that ideal phonetic segments are distorted by the vocal tract. For example, MacNeilage & Ladefoged (1976) describe planned segments as discrete, static, and context-free, whereas uttered segments are overlapped, dynamic, and context-sensitive.

A related view expressed by several researchers is that linguistic units are mental things that, thereby, cannot be identified with any set of articulatory or acoustic characteristics. For example:

> [Phonetic segments] are *abstractions*. They are the end result of complex perceptual and cognitive processes in the listener's brain. (Repp, 1981, p. 1462)
>
> They [phonetic categories] have no physical properties. (Repp, 1981, p. 1463)
>
> Segments cannot be objectively observed to exist in the speech signal nor in the flow of articulatory movements . . . [T]he concept of segment is brought to bear *a priori* on the study of physical–physiological aspects of language. (Hammarberg, 1976, p. 355)
>
> [T]he segment is internally generated, the creature of some kind of perceptual—cognitive process. (Hammarberg, 1976, p. 355)

This point of view, of course, requires a mentalist theory of perception.

For a realist event theory to be possible, what modifications to these views are required? The essential modification is to our conceptualization of the relation between

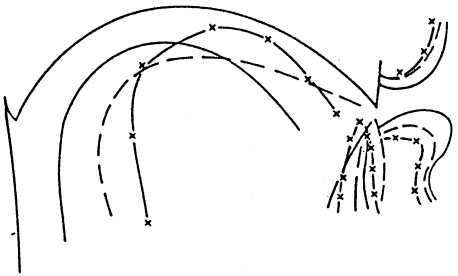

Figure 1. Cinefluorographic tracing of the vocal tract during three phases in production of /husi/ (redrawn from Carney & Moll, 1971). ——, /u/ in /husi/; – – – –, /s/ in /husi/; × – × – ×, /i/ in /husi/.

knowing and doing. First, phonetic segments as we know them can only have properties that can be realized in articulation. Indeed, from an event perspective, the primary reality of the phonetic segment is its public realization as vocal-tract activity. What we know of the segments, we know from hearing them produced by other talkers or by producing them ourselves. Secondly, the idea that speech production involves a translation from a mental domain into a physical, non-mental domain such as the vocal tract must be discarded.

With respect to the first point, we can avoid the metaphor of Hockett's wringer if we can avoid somehow ascribing properties to phonetic segments that vocal tracts cannot realize. In view of the fact that phonetic segments evolved to be spoken, and indeed, that we have evolved to speak them (Lieberman, 1984), this does not seem to be a radical endeavor.

Vocal tracts cannot produce a string of static shapes, so for an event theory to be possible, phonetic segments cannot be inherently static. Likewise, vocal tracts cannot produce the segments discretely, if discrete means "non-overlapping". However, neither of these properties is crucial to the work that phonetic segments do in a linguistic communication, and therefore can be abandoned without loss.

Phonetic segments do need to be separate from one another and serially ordered, however, and Hockett's Easter egg analogy suggests that they are not. But my own reading of the literature on coarticulation is that the Easter egg analogy is misleading and wrong. Figure 1 is a redrawing of a figure from Carney & Moll (1971): it is an outline drawing of the vocal tract with three tongue shapes superimposed. The shapes were obtained by cinefluorography at three points in time during the production of the disyllable [husi]. The solid line reflects the tongue shape during a central portion of the vowel [u]; the dashed line is the tongue shape during closure for [s]; the line of crosses is the tongue shape during a central portion of [i]. Thus, the figure shows a smooth vowel-to-vowel gesture of the tongue body taking place during closure of [s] (cf. Öhman, 1966). The picture these data reveal is much cleaner than the Easter egg metaphor would suggest. The sets of gestures for different segments overlap, but the separation and ordering of the segments is preserved.[3]

[3]This characterization may appear patently incorrect in cases where the same articulator is involved simultaneously in the production of more than one phonetic segment (for example, the tongue body during closure for [kh] in "key" and "coo" and the jaw during closure for [b] in "bee" and "boo"). However, Saltzman and Kelso (Saltzman, in press; Saltzman & Kelso, 1983) have begun to model this as overlapping, but separate demands of different control structures on the same articulator and my own findings on perceived segmentation of speech (Fowler, 1984; see also Fowler & Smith, 1986) suggest that perceivers extract exactly that kind of parsing of the speech signal.

With respect to the second point, Ryle (1949) offers a way of conceptualizing the relation between the mental and the physical that avoids the problems consequent upon identifying the mental with covert processes taking place inside the head:

> When we describe people as exercising qualities of mind, we are not referring to occult episodes of which their overt acts and utterances are effects, we are referring to those overt acts and utterances themselves. (p. 25)
>
> When a person talks sense aloud, ties knots, feints or sculpts, the actions which we witness are themselves the things which he is intelligently doing . . . He is bodily active and mentally active, but he is not being synchronously active in two different "places", or with two different "engines". There is one activity, but it is susceptible of and requiring more than one kind of explanatory description. (pp. 50–51)

This way of characterizing intelligent action does not eliminate the requirement that linguistic utterances must be planned: rather it eliminates the idea that covert processes are privileged in being mental or psychological, whereas overt actions are not. Instead, we may think of the talker's intended message as it is planned, uttered, specified acoustically, and perceived as being replicated intact across different physical media from the body of the talker to that of the listener.

An event theory of speech *production* must aim to characterize articulation of phonetic segments as overlapping sets of coordinated gestures, where each set of coordinated gestures conforms to a phonetic segment. By hypothesis, the organization of the vocal tract to produce a phonetic segment is invariant over variation in segmental and suprasegmental contexts. The segment may be realized somewhat differently in different contexts (for example, the relative contributions of the jaw and lips may vary over different bilabial closures (Sussman, MacNeilage & Hanson, 1973)), because of competing demands on the articulators made by phonetic segments realized in an overlapping time frame. To the extent that a description of speech production along these lines can be worked out, the possibility remains that phonetic segments are literally uttered and therefore are available to be directly perceived if the acoustic signal is sufficiently informative. Research on a "task dynamic" model of speech production (e.g. Saltzman, in press; Saltzman & Kelso, 1983; KST, this issue) may provide, at the very least, an existence proof that systems capable of realizing overlapping phonetic segments nondestructively can be devised.

2.3. *The third barrier: the acoustic signal does not specify phonetic segments*

Putting aside the question whether phonetic segments are realized non-destructively in articulation, there remains the problem that the acoustic signal does not seem to reflect the phonetic segmental structure of a linguistic communication. It need not, even if phonetic segments are uttered intact. Although gestures of the vocal tract cause disturbances in the air, it need not follow that the disturbance specify their causes. For many researchers, they do not. Figure 2 (from Fant & Lindblom, 1961, and Cutting & Pisoni, 1978) displays the problem.

A spectrographic display of a speech utterance invites segmentation into "acoustic segments" (Fant, 1973). Visibly defined, these are relatively homogeneous intervals in the display. Segmentation lines are drawn where abrupt changes are noticeable. The difficulty with this segmentation is the relation it bears to the component phonetic

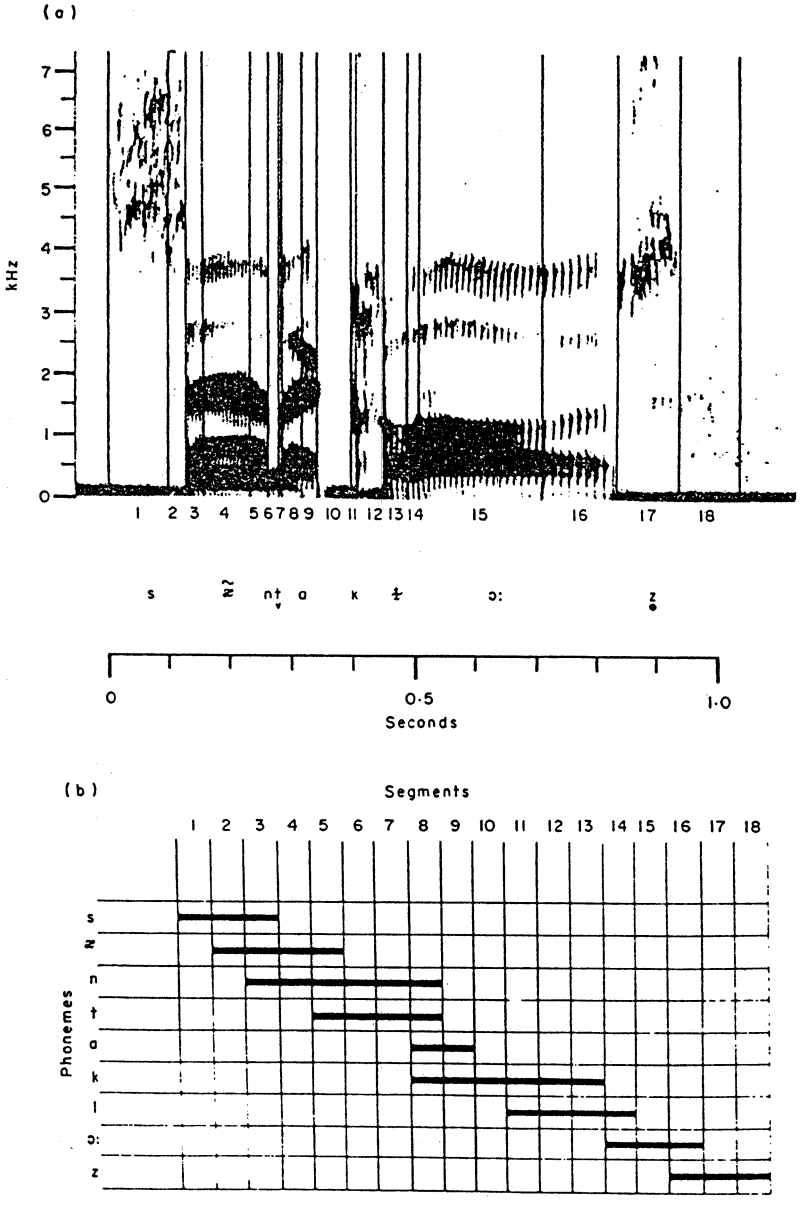

Figure 2. (a) Spectrographic display of "Santa Claus". (b) Schematic display of the relationship between acoustic and phonetic segments (reprinted with permission from Cutting & Pisoni (1978). and Fant & Lindblom (1961)).

segments of the linguistic utterance. In the display, the utterance is the name, "Santa Claus", which is composed of nine phonetic segments, but 18 acoustic segments. The relation of phonetic segments to acoustic segments is not simple, as the bottom of Figure 2 reveals. Phonetic segments may be composed of any number of acoustic segments, from two to six in the figure, and most acoustic segments reflect properties of more than one phonetic segment.

How do listeners recover phonetic structure from such a signal? One thing is clear; the functional parsing of the acoustic signal for the perceiver is not one into acoustic segments. Does it follow that perceivers impose their own parsing on the signal? There must be a "no" answer to this question for an event theory devised from a direct–realist perspective to be viable. The perceived parsing must be in the signal; the special role of the perceptual system is not to create it, but only to select it.

Notably, there is more than one physical description of the acoustic speech signal. A spectrographic display suggest a parsing into acoustic segments, but other displays suggest other parsings of the signal. For example, Kewley-Port (1983) points out that in a spectrographic display the release burst of a syllable-initial stop consonant looks quite distinct from the formant transitions that follow it (for example, see the partitioning of /k/ in "Claus" in Figure 2). Indeed, research using the spectrographic display as a guide has manipulated burst and transition to study their relative salience as information for stop place (Dorman, Studdert-Kennedy & Raphael, 1977). However, Kewley-Port's "running spectra" for stops (overlapping spectra from 20 ms windows taken at successive 5 ms intervals following stop release) reveal continuity between burst and transitions in changes in the location of spectral peaks from burst to transition.

It does not follow, then, from the mismatch between acoustic segment and phonetic segment, that there is a mismatch between the information in the acoustic signal and the phonetic segments in the talker's message. Possibly, in a manner as yet undiscovered by researchers but accessed by perceivers, the signal is transparent to phonetic segments.

If it is, two research strategies should provide converging evidence concerning the psychologically relevant description of the acoustic signal. The first seeks a description of the articulatory event itself—that is, of sequences of phonetic segments as articulated —and then investigates the acoustic consequences of the essential articulatory components of phonetic segments. A second examines the parsing of the acoustic signal that listeners detect.

The research that comes closest to this characterization is that of Stevens & Blumstein (1978, 1981; also Blumstein & Stevens, 1979, 1981). They begin with a characterization of phonetic segments and, based on the acoustic theory of speech production (Fant, 1960), develop hypotheses concerning invariant acoustic consequences of essential articulatory properties of the segments. They then test whether the consequences are, in fact, invariant over talkers and phonetic-segmental contexts. Finally, they ask whether these consequences are used by perceivers.

Unfortunately for the purpose of an event approach, perhaps, they begin with a characterization of phonetic segments as bundles of distinctive features. This characterization differs in significant ways from one that will be developed from a perspective on phonetic segments as coordinated articulatory gestures. One important difference is that the features tend to be static; accordingly, the acoustic consequences first sought in the research program were static also. A related difference is that the characterization deals with coarticulation by presuming that the listener gets around it by focusing his or her attention on the least coarticulated parts of the signal. As I will suggest shortly, that does not conform with the evidence; nor would it be desirable, because acoustic consequences of coarticulated speech are quite informative (cf. Elman & McClelland, 1983).

To date, Stevens and Blumstein have focused most of their attention on invariant information for consonantal place of articulation. Their hypotheses concerning possible invariants are based on predictions derived from the acoustic theory of speech

production concerning acoustic correlates of constrictions in various parts of the vocal tract. When articulators adopt a configuration, the vocal tract forms cavities that have natural resonances, the formants. Formants create spectral peaks in an acoustic signal; that is, a range of frequencies higher in intensity than their neighbors. With the exception of posterior places of articulation, e.g. pharynx, a consonantal constriction in the vocal tract lowers F_1 (the lowest formant) relative to its frequency for a corresponding vowel and it affects the frequencies and intensities of higher formants. Stevens & Blumstein (1978) argue that stop consonants with different places of articulation have characteristic burst spectra independent of the vowel following the consonant, and independent of the size of the vocal tract producing the constriction.

Blumstein & Stevens (1979) created "template" spectra for the stop consonants, /b/, /d/, and /g/, and then attempted to use them to classify the stops in 1800 CV and VC syllables in which the consonants were produced by different talkers in the context of various vowels. Overall, they were successful in classifying syllable-initial stops, but less successful with final stops, particularly if the stops were unreleased. Blumstein & Stevens (1981) also showed that listeners could classify stops by place better than chance when they were given only the first 10–46 ms of CV syllables.

However, two investigations have shown that the shape of the spectrum at stop release is not an important source of information for stop place. These studies (Blumstein, Isaacs & Mertus, 1982; Walley & Carrell, 1983) pitted place information contributed by the shape of the spectrum at stop release in CVs, against the (context-dependent) information for place contributed by the formant frequencies themselves. In both studies, the formants overrode the effect of spectral shape in listeners' judgments of place.

Recently, Lahiri, Gewirth & Blumstein (1984) have found in any case that spectral shape does not properly classify labial, dental and alveolar stops produced by speakers of three different languages. In search of new invariants and following the lead of Kewley-Port (1983), they examined the information in running spectra. They found that they could classify stops according to place by examining relative shifts in energy at high and low frequencies from burst to voicing onset. Importantly, pitting the appropriate running spectral patterns against formant frequencies for 10 CV syllables in a perceptual study, Lahiri *et al.* found that the running spectral patterns were overriding. The investigators identify their proposed invariants as "dynamic", because they are revealed over time during stop release, and relational because they are based on relative changes in the distribution of energy at high and low frequencies in the vicinity of stop release.

Lahiri *et al.* are cautious whether their proposed invariants will withstand further test —and properly so, because the invariants are somewhat contrived in their precise specification. I suspect that major advances in the discovery of invariant acoustic information for phonetic segments will follow advances in understanding how phonetic segments are articulated. However, the proposals of Lahiri *et al.* (see also Kewley-Port, 1983) constitute an advance over the concept of spectral shape in beginning to characterize invariant acoustic information for gestures rather than for static configurations.

2.4. *The fourth barrier: perception demonstrably involves "top down" processes and perceivers do make mistakes*

Listeners may "restore" missing phonetic segments in words (Samuel, 1981; Warren, 1970), and talkers shadowing someone else's speech may "fluently restore" mispronounced

words to their correct forms (e.g. Marslen-Wilson & Welsh, 1978). Even grosser departures of perceptual experience from stimulation may be observed in some mishearings (for example "popping really slow" heard as "prodigal son" (Browman, 1980) or "mow his own lawn" heard as "blow his own horn" (Garnes & Bond, 1980)).

These kinds of findings are often described as evidence for an interaction of "bottom-up" and "top-down" processes in perception (e.g. Klatt, 1980). Bottom-up processes analyze stimulation as it comes in. Top-down processes draw inferences concerning stimulation based both on the fragmentary results of the continuing bottom-up processes and on stored knowledge of likely inputs. Top-down processes can restore missing phonemes or correct erroneous ones in real words by comparing results of bottom-up processes against lexical entries. As for mishearings, Garnes & Bond (1980) argue that "active hypothesizing on the part of the listener concerning the intended message is certainly part of the speech perception process. No other explanation is possible for misperceptions which quite radically restructure the message . . ." (p. 238).

In my view (but not necessarily in the view of other event theorists), these data do offer a strong challenge to an event theory. It is not that an event theory of speech perception has nothing to say about perceptual learning (for example, Gibson, 1966; Johnston & Pietrewicz, 1985). However, what is said is not yet well enough worked out to specify how, for example, lexical knowledge can be brought to bear on speech input from an direct–realist, event perspective.

With regard to mishearings, there is also a point of view (Shaw, Turvey & Mace, 1982) that when reports of environmental events are in error, the reporter cannot be said to have perceived the events, because the word "perception" is reserved for just those occasions when acquisition of information from stimulation is direct and, therefore, successful. The disagreement with theories of perception as indirect and constructive, then, may reduce to a disagreement concerning how frequently bottom-up processes complete their work in the absence of top-down influence.

I prefer a similar approach to that of Shaw *et al.* that makes a distinction between what perceivers *can do* and what they may do in particular settings. As Shaw *et al.* argue, there is a need for the informational support for activity to *be able to be* directly extracted from an informational medium and for perception to be nothing other than direct extraction of information from proximal stimulation. However, in familiar environments, actors may generally guide their activities based not only on what they perceive, but also on what the environment routinely affords. In his presentation at the first event conference, Jenkins (1985) reviews evidence that the bat's guidance of flying sometimes takes this form. Placed in a room with barriers that must be negotiated to reach a food source, the bat soon learns the route (Griffin, 1958). After some time in which the room layout remains unchanged, a barrier is placed in the bat's usual flight path. Under these novel conditions, the bat is likely to collide with the barrier. Although it could have detected the barrier, it did not. By the same token, as a rule, we humans do not test a sidewalk to ensure that it will bear our weight before entrusting our weight to it. Nor do we walk through (apparent) apertures with our arms outstretched just in the case the aperture does not really afford passage because someone has erected a difficult-to-see plate-glass barrier. In short, although the affordances that guide action *can be* directly perceived, often they are not wholly. We perceive enough to narrow down the possible environments to one likely environment that affords our intended activity, and other remotely likely ones that may not.

Perceptual restorations and mishearings imply the same perceptual pragmatism among perceivers of speech. It is also implied, I think, by talkers' tendencies to adjust the formality of their speaking style to their audience (e.g. Labov, 1972). Audiences with whom the talker shares substantial past experiences may require less information to get the message than listeners who share less. Knowing that, talkers conserve effort by providing less where possible.

It may be important to emphasize that the foregoing attempt to surmount the fourth barrier is intended to do more than translate a description of top-down and bottom-up processes into a terminology more palatable to event theorists. In addition, I am attempting to allow a role for information not currently in stimulation to guide activity, while preserving the ideas that perception itself must be direct and hence, errorless, and that activity *can be* (but often is not) guided exclusively by perceived affordances.

As to the latter idea, the occurrence of mishearings that depart substantially from the spoken utterance should not deflect our attention from the observation that perceivers *can* hew the talker's articulatory line very closely if encouraged to do so. One example from my own research is provided by investigations of listeners' perceived segmentation of speech. Figure 1 above, already described, displays coarticulation of the primary articulators for vowels and consonants produced in a disyllable. This overlap has two general consequences in the acoustic signal (one generally acknowledged as a consequence, the other not). First, within a time frame that a conventional acoustic description would identify with one phonetic segment (because the segment's acoustic consequences are dominant), the acoustic signal is affected by the segment's preceding and following neighbors. Secondly, because the articulatory trajectories for consonants overlap part of the trajectory of a neighboring vowel (cf. Carney & Moll, 1971; Öhman, 1966), the extent of time in the acoustic signal during which the vowel predominates in its effects—and hence the vowel's measured duration—decreases in the context of many consonants or of long consonants as compared to its extent in the context of few or short consonants (Fowler, 1983; Fowler & Tassinary, 1981; Lindblom & Rapp, 1973).

Listeners can exhibit sensitivity to the information for the overlapping phonetic segments that talkers produce in certain experimental tasks. In these tasks, the listeners use acoustic information for a vowel within a domain identified with a preceding consonant (for example, within a stop burst or within frication for a fricative consonant) as information for the vowel (Fowler, 1984; Whalen, 1984). Moreover, listeners do not integrate the overlapping information for vowel and consonant. Rather, they hear the consonant as if the vowel information had been factored out of it (Fowler, 1984) and they hear the vowel as longer than its measured extent by an amount correlated with the extent to which a preceding consonant should have shortened it by overlapping its leading edge (Fowler & Tassinary, 1981).

These studies indicate that listeners can track the talker's vocal-tract activities very closely and, more specifically, that they extract a segmentation of the signal into the overlapping phonetic segments that talkers produce, not into discrete approximations to phonetic segments and not into acoustic segments. Of course, this is as it must be among young perceivers if they are to learn to talk based on hearing the speech of others. But whether or not a skilled listener will track articulation this closely in any given circumstance may depend on the extent to which listener estimates that he or she needs to in order to recover the talker's linguistic message.

3. Perception of speech events in an expanded time frame: sound change

Two remarkable facts about the bottom tier of dually structured language are that its structure undergoes systematic change over time, and that the sound inventories and phonological processes of language reflect the articulatory dispositions of the vocal tract and perceptual dispositions of the ear (Donegan & Stampe, 1979; Lindblom *et al.*, 1983; Locke, 1983; Ohala, 1981). There are many phonological processes special to individual languages that have analogues in articulatory—phonetic processes general to languages. For example, most languages have shorter vowels before voiceless than voiced stops (e.g. Chen, 1970). However, in addition, among languages with a phonological length distinction, in some (for example, German; see Comrie, 1980), synchronic or diachronic processes allow phonologically long vowels only before voiced consonants. Similarly, I have already described a general articulatory tendency for consonants to overlap vowels in production, so that vowels are measured to shorten before clusters or long consonants more than before singleton consonants or short consonants. Compatibly, in stressed syllables, Swedish short vowels appear only before long consonants or multiple consonants; long vowels appear before a short consonant or no consonant at all. In Yawelmani (see Kenstowiscz & Kisseberth, 1979), a long vowel is made short before a cluster. Stressed vowels also are measured to shorten in the context of following unstressed syllables in many languages (Fowler, 1981; Lehiste, 1972; Lindblom & Rapp, 1973; Nooteboom & Cohen, 1975). Compatibly, in Chimwi:ni (Kenstowiscz & Kisseberth, 1979), a long vowel may not generally occur before the antepenultimate syllable of a word.

These are just a few examples involving duration that I have gathered, but similar examples abound, as do examples of other phonological tendencies. We can ask: how do linguistic–phonological processes that resemble articulatory dispositions enter language?

An interesting answer that Ohala (1981) offers to cover some cases is that they enter language via sound changes induced by systematic misperception by listeners. One example he provides is that of tonal development in "tone languages", including Chinese, Thai, and others. Tonal development on vowels may have been triggered by loss of a voicing distinction in preceding consonants. A consequence of consonant voicing is a rising tone on the following vowel (e.g. Hombert, 1979). Following a voiceless consonant, the tone is high and falling. Historical development of tones in Chinese may be explained as the listeners' systematic failure to ascribe the tone to consonant voicing —perhaps because the voicing distinction was weakening—and to hear it instead as an intentionally produced characteristic of the vowel.

This explanation is intriguing because, in relation to the perspective on perceived segmentation just outlined, it implies that listeners may sometimes recover a segmentation of speech that is not identical to the one articulated by the talker. In particular, it suggests that listeners may not always recognize coarticulatory encroachments as such and may instead integrate the coarticulatory influences with a phonetic segment with which they overlap in time. This may be especially likely when information for the occurrence of the coarticulating neighbor (or its relevant properties, as in the case of voicing in Chinese) is weakening. However, Ohala describes some examples where coarticulatory information has been misparsed despite maintenance of the conditioning segment itself. Failures to recover the talker's segmental parsing may lead to sound change when listeners themselves begin producing the phonological segment as they recovered it rather than as the talkers produced it.

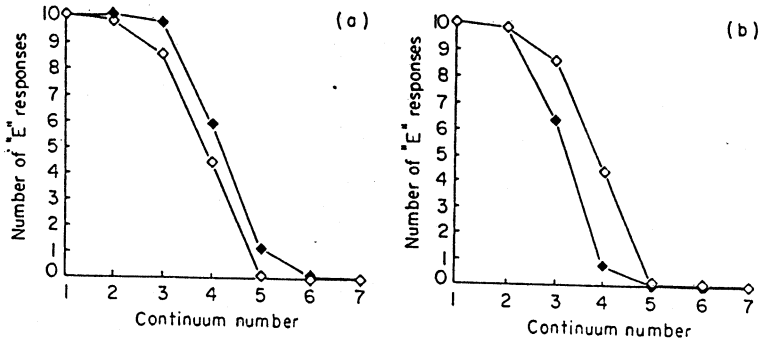

Figure 3. Identification of vowels in the experiment of Krakow, Beddor, Goldstein & Fowler (1985): see text for explanation. (a) ◆, Original [bẽnd]; ◇, original [bɛd]: (b) ◆, original [bẽd]; ◇, original [bɛd].

Recent findings by Krakow, Beddor, Goldstein & Fowler (1985) suggest that something like this may underlie an on-going vowel shift in English. In English, the vowel /æ/ is raising in certain phonetic contexts (e.g. Labov, 1981). One context is before a nasal consonant. Indeed, for many speakers of English, the /æ/ in "can", for example, is a noticeably higher vowel than that in "cad".

One hypothesis to explain the vowel shift in the context of nasal consonants is that listeners fail to parse the signal so that all of the influences of the nasalization on the vowel are identified with the coarticulatory influence of the nasal consonant. As Wright (1980) observes, the nasal formant in a nasalized vowel is lower in frequency than F_1 of /æ/. Integrated with F_1 of /æ/ or mistakenly identified as F_1, the nasal formant is characteristic of a higher vowel (with a lower F_1) than F_1 of /æ/ itself.

Krakow *et al.* examined this idea by synthesizing two kinds of continua using an articulatory synthesizer (Rubin, Baer & Mermelstein, 1979). One continuum was a /bɛd/ to /bæd/ series (henceforth, the bed–bad series) created by gradually lowering and backing the height of the synthesizer's model tongue in seven steps. A second, /bɛnd/ to /bænd/, continuum (henceforth bend–band) was created in similar fashion, but with a lowered velum during the vowel and throughout part of the following alveolar occlusion. (In fact, several bend–band continua were synthesized with different degrees of velar lowering. I will report results on just one representative continuum.) Listeners identified the vowel in each series as spelled with "E" or "A". Figure 3(a) compares the responses to members of the bed–bad continuum with responses to a representative bend–band series. As expected, we found a tendency for subjects to report more "E"s in the bend–band series.

We reasoned that if this were due to a failure of listeners to parse the signal so that all of the acoustic consequences of nasality were ascribed to the nasal consonant, then by removing the nasal consonant itself, we would see as much or even more raising than in the context of a nasal consonant. Accordingly, we altered the original bed–bad series by lowering the model velum throughout the vowel. (I will call the new /bɛd/–bæd/ continuum the bed(N)–bad(N) series. Again, different degrees of nasality were used over different continua. I will report data from a representative series.) Figure 3(b) shows the results of this manipulation. Rather than experiencing increased raising, as expected, the listeners experienced significant lowering of the vowel in the bed(N)–bad(N) series. Although this outcome can be rationalized in terms of spectral changes to the oral

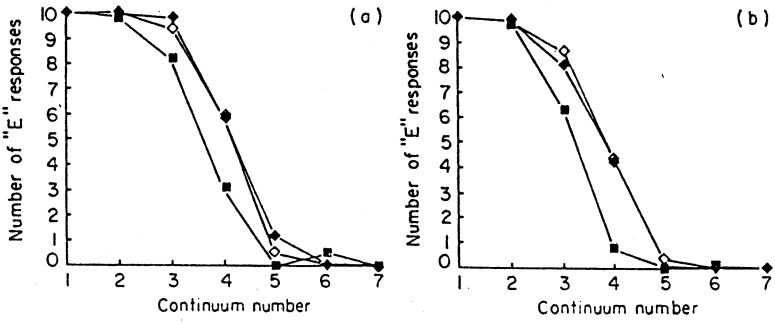

Figure 4. Identification of vowels in continua having vowels matched in measured duration (data from Krakow *et al.*, 1985). (a) Short: ◇, [bɛd]; ■, [bɛ̃d]; ◆, [bɛ̃nd]. (b) Long: ◇, [bɛd]; ■, [bɛ̃d]; ◆, [bɛ̃nd].

formants of the vowel due to the influence of the nasal resonance on them, it does not elucidate the original of the raising observed in the first study.

A difference between our bend–band and bed–bad series was in the measured duration of the vowels. Following measurements of natural productions, we had synthesized syllables with shorter measured vowels in the bend–band series than in the bed–bad series. We next considered the possibility that this explained the raising we had found in the first experiment. /ɛ/ is an "inherently" shorter vowel than /æ/ (e.g. Peterson & Lehiste, 1960). It seemed possible that raising in the bend–band series was not due to misparsing of nasality, but to misparsing of the vowel's articulated extent from that of the overlapping nasal consonant. In particular, the vowels in the bend–band continua might have been perceived as inherently shorter (rather than as more extensively overlapped by the syllable coda) than vowels in the bed–bad series, and hence as more /ɛ/-like.

To test that idea, we synthesized a new bend–band series with longer measured durations of vowels, matching those in the original bed–bad (and bed(N)–bad(N)) series, and new bed–bad and bed(N)–bad(N) series with vowels shortened to match the measured duration of those in the original bend–band series. Figures 4(a) and (b) show the outcome for the short and long series respectively. Identification functions for bed–bad and bend–band are now identical. Listeners ascribe all of the nasality in the vowel to the consonant, and when vowels are matched in measured duration, there is no raising. Stimuli in the bed(N)–bad(N) series show lowering in both Figures 4(a) and (b).

These results are of interest in several respects. For the present discussion, they are interesting in suggesting limitations in the extent to which these listeners could track articulation. Although listeners do parse speech along its coarticulatory lines in this study, ascribing the nasality during the vowel to the nasal consonant, they are not infinitely sensitive to parts of a vowel overlaid by a consonant. The difficulty they have detecting the trailing edges of a vowel may be particularly severe when the following consonants are nasals as in the present example, because, during a nasal, the oral cavity is sealed off and the acoustic signal mainly reflects passage of air through the nasal cavity. Consequently, information for the vowel is poor. (There is vowel information in nasal consonants, however, as Fujimura (1962) has shown.)

In a study mentioned earlier, Fowler & Tassinary found that in a vowel-duration continuum in which voicing of a final alveolar stop was cued by vowel duration

(cf. Raphael, 1972), the "voiceless" percept was resisted more for vowels preceded by consonants that, in natural productions, shorten the measured extents of the vowels substantially than by consonants that shorten them less. In the study, however, the effect on the voicing boundary was less than the shortening effect of the preceding consonant would predict. Together, this study and that by Krakow *et al.* suggest that although listeners do parse the speech signal along coarticulatory lines, they do not always hear the vowels as extending throughout their whole coarticulatory extent.[4]

As Ohala has suggested (1981), these perceptual failures may provoke sound change. Thereby they may promote introduction into the phonologies of languages, processes that resemble articulatory dispositions.

What are the implications of this way of characterizing perception and sound change for the theory of perception of speech events? In the account, perceivers clearly are extracting affordances from the acoustic signal. That is, they are extracting information relevant to the guidance of their own articulatory activities. (See the following section for some other affordances perceived by listeners.) However, just as clearly, the distal event they reported in our experiment and that they reproduce in natural settings is not the one in the environment. The problem here may or may not be the same as that discussed as the "fourth barrier" above. In the present case, the problem concerns the salience of the information provided to the listener in relation to the listener's own sensitivity to it. Information for vowels where consonants overlap them presumably is subtle and therefore difficult (but not impossible, see Fowler, 1984; Whalen, 1984) to detect. One way to handle the outcome of the experiment by Krakow *et al.* within a direct–realist event theory is to suppose that listeners extract less information from the signal than they need to report their percept in an experiment or to reproduce it themselves, and they fill in the rest of the information from experience at the time of report or reproduction. An alternative is that listeners are insensitive to the vowel information in the nasal consonant and use that lack of information as information for the vowel's absence there. Presumably it is just the cases where important articulatory information is difficult to detect that undergo the perceptually driven sound changes in languages (cf. Lindblom, 1972).

4. How perception guides action

4.1. *Some affordances of phonetically-structured speech*

For those of us engaged in research on phonetic perception, it is easy to lose sight of the fact that, outside of the laboratory, the object of perceiving is not the achievement of a percept, but rather the acquisition of information relevant to guidance of activity. I will next consider how perception of phonetically structured vocal activity may guide the listener's behavior. This is not, of course, where most of the action is to be found in speech perception. More salient is the way the perception of the linguistic message guides the listener's behavior. This is a very rich topic, but not one that I can cover here.

Possibly, the most straightforward activity for listeners just having extracted information about how a talker controlled his or her articulators (but not, in general, the most appropriate activity), is to control their own articulators in the same way—that is,

[4]Javkin (1976) has provided evidence for the opposite kind of error. In his research, listeners heard vowels as longer before voiced than voiceless consonants, perhaps because of the continuation of voicing during the consonant.

to imitate. Indeed, research suggests that listeners can shadow speech with very short latencies (Chistovich, Klaas & Kuzmin, 1962; Porter, 1977) and that their latencies are shorter to respond with the same syllable or one that shares gestures with it than with one that does not (Meyer & Gordon, 1984).

Although this has been interpreted as relevant to an evaluation of the motor theory of speech perception (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967) it may also, or instead, reflect a more general disposition for listeners to mimic talkers (or perhaps to entrain to them). Research shows that individuals engaging in conversation move toward one another in speech rate (defined as the number of syllables per unit time excluding pauses; Webb, 1972) in loudness (Black, 1949) and in average duration of pauses (Jaffe, 1964), although the temporal parameters of speaking also show substantial stability among individual talkers across a variety of conversational settings (Jaffe & Feldstein, 1970). In addition, Condon & Ogston (1971; see also Condon, 1976, for a review), report that listeners (including infants aged 1–4 days; Condon & Sander, 1974) move in synchrony with a talker's speech rhythms.

Although it is possible that this disposition for "interactional synchrony" (Condon, 1976) has a function, for example, in signaling understanding, empathy, or interest on the listener's part (cf. Matarazzo, 1965), the observations that some of the visible synchronies have been observed when the conversational partners cannot see one another, and some have been observed in infants, may suggest a more primitive origin. Condon (1976) suggests that interactional synchrony is a form of entrainment.

The disposition to imitate among adults may be a carry-over from infancy, when presumably it does have an important function (Studdert-Kennedy, 1983). Infants must extract information about phonetically structured articulations from the acoustic speech signals of mature talkers in order to learn to regulate their own articulators to produce speech. Although it seems essential that infants do this, very little research does more than hint that infants have the capacity to imitate vocal productions.

Infants do recognize the correspondence between visible articulation of others and an acoustic speech signal. They will look preferentially to the one of two video displays on which a talker mouths a disyllable matching an accompanying acoustic signal (MacKain, Studdert-Kennedy, Spieker & Stern, 1983). Moreover, infants recognize the equivalence of their own facial gestures to those of someone else. That is, they imitate facial gestures, such as lip or tongue protrusion (Meltzoff & Moore, 1985) even though, as Meltzoff & Moore point out, such imitation is "intermodal", because the infants cannot see their own gestures. Together, these findings suggest that infants should be capable of vocal imitation.

However, relatively few studies have examined infants' imitation of adult vocalizations. Infants are responsive to mothers' vocalizations, and indeed, vocalize simultaneously with them to a greater-than-chance extent (Stern, Jaffe, Beebe & Bennett, 1975). There are a few positive reports of vocal imitation (e.g. Kessen, Levine & Wendrick, 1979; Kuhl & Meltzoff, 1982; Tuaycharoen, 1978; Uzgiris, 1973). However, few of them have been conducted with the controls now recognized as required to distinguish chance correspondences from true imitations.

Of course, imitative responses are not the only activities afforded by speech, even speech considered only as phonetically structured activity of the vocal tract. A very exciting area of research in linguistics is on natural variation in speaking (e.g. Labov, 1966/1982, 1972, 1980). The research examines talkers in something close to the natural environments in which talking generally takes place. It is exciting because it reveals

a remarkable sensitivity and responsiveness of language users to linguistically, psychologically and socially relevant aspects of conversational settings. Most of these aspects must be quite outside of the language users' awareness much of the time; yet they guide the talker's speech in quite subtle but observable ways.

Labov and his colleagues find that an individual's speaking style varies with the conversational setting in response, among other things, to characteristics of the conversational partner, including, presumably, the partner's own speaking style. Accordingly, adjustments to speaking style are afforded by the speech of the conversational partner.

An example of research done on dialectal affordances of the speech of other social groups is provided by Labov's early study of the dialects of Martha's Vineyard (1963). Martha's Vineyard is a small island off the coast of New England that is part of the state of Massachusetts. Whereas residents were traditionally farmers and fishermen, in recent decades the island has become a popular summer resort. The addition of some 40 000 summer residents to the year-round population of 5000–6000 has, of course, had profound consequences for the island's economy.

Labov chose to study production of two diphthongs, [ai] and [au], both of which had lowered historically from the forms [əi] and [əu]. These historical changes were not concurrent; [au] had lowered well before the settlement of Martha's Vineyard by English speakers in 1642; [ai] lowered somewhat after its settlement.

Labov found a systematically increasing tendency to *centralize* the first vowel of the diphthongs—that is, to reverse the direction of sound change just described—in younger native residents when he compared speakers ranging in age from 30 years upwards. The tendency to centralize the vowels was strongest amongst people such as farmers, whose livelihoods had been most threatened by the summer residents. (The summer residents have driven up land prices as well as the costs of transporting supplies to the island and products to the mainland.) In addition, the tendency to centralize was correlated with the speaker's tendency to express resistance to the increasing encroachment of summer residents on the island. Among the youngest group studied, 15-year-olds, the tendency to centralize the vowels depended strongly on the individual's future plans. Those intending to stay on the island showed a markedly stronger tendency to centralize the diphthongs than those intending to leave the island to make a living on the mainland. Labov interpreted these trends as a disposition among many native islands to distinguish themselves as a group from the summer residents.

I find these data and others collected by Labov and his colleagues quite remarkable in the evidence they provide for listeners' responsiveness to phonetic variables they detect in conversation. In natural conversational settings, talkers use phonetic variation to psychological and social ends; and, necessarily given that, listeners are sensitive to those uses.

4.2. *What enables phonetically structured vocal-tract activity to do linguistic work and how is that work apprehended?*

Confronted with language perception and use, an event theory faces powerful challenges. Gibson's theory of perception (1966, 1979) depends on a necessary relation between structure in informational media and properties of events. The physical law relating vocal-tract activities to acoustic consequences may satisfy that requirement. But how is the relation between word and referent and, therefore, between acoustic signal and

referent, to be handled? These relations are not universal; that is, different languages use different words to convey similar concepts. Accordingly, in one sense, they are not necessary and not, apparently, governed by physical law.

I have very little to offer concerning an event perspective on linguistic events (but see Verbrugge, 1985), and what I do have to say, I consider very tentative indeed. However, I would like to address two issues concerning the relation of speech to language. Stated as a question, the first issue is: What allows phonetically structured vocal-tract activity to serve as a meaningful message? The second asks: Can speech *qua* linguistic message be directly perceived?

As to the first question, Fodor (1974) observes that there are two types of answer that can be provided to questions of the form: "What makes X a Y?". He calls one type of answer the "causal story" and the other the "conceptual story". To use Fodor's example, in answer to the question: "What makes Wheaties the Breakfast of Champions?", one can invoke causal properties of the breakfast cereal, Wheaties, that turn non-champions who eat Wheaties into champions. Alternatively, one can make the observation that disproportionate numbers of champions eat Wheaties. As Fodor points out, these explanations are distinct and not necessarily competing.

In reference to the question, what makes phonetically structured vocal-tract activity phonological (that is, what makes it serve a linguistic function), one can refer to the private linguistic competences of speakers and hearers that allow them to control their vocal tracts so as to produce gestures having linguistic significance. Alternatively, one can refer to properties of the language user's "ecological niche" that support lingusitic communication. Vocal-tract activity can only constitute a linguistic message in a setting in which, historically, appropriately constrained vocal-tract activity has done linguistic work. A listener's ability to extract a linguistic message from vocal-tract activity may be given a "conceptual" (I would say "functional") account along lines such as the following: listeners apprehend the linguistic work that the phonetically structured vocal-tract activity is doing by virtue of their sensitivity to the historical and social context of constraint in which the activity is performed.

According to Fodor (p. 9):

> Psychologists are typically in the business of supplying theories about the events that causally mediate the production of behavior . . . and cognitive psychologists are typically in the business of supplying theories about the events that causally mediate intelligent behavior.

He is correct; yet there is a functional story to be told, and I think that it is an account that event theorists will want to develop.

As to the second question, whether a linguistic message can be said to be perceived in a theory of perception from a direct–realist perspective, (direct) perception depends on a necessary relation between structure in informational media and its distal source. But as previously noted, this does not appear to apply to the relation between sign and significance.

Gibson suggests that linguistic communications, and symbols generally, are perceived (rather than being apprehended by cognitive processes), but indirectly. His use of the qualifier "indirect" requires careful attention (1976/1982, p. 412):

> Now consider perception at second hand, or vicarious perception; perception mediated by communications and dependent on the "medium" of communication, like speech sound, painting, writing or sculpture. The perception is indirect since the information has been

> presented by the speaker, painter, writer or sculptor, and has been *selected* by him from the unlimited realm of available information. This kind of apprehension is complicated by the fact that direct perception of sounds or surfaces occurs along with the indirect perception. The sign is often noticed along with what is signified. Nevertheless, however complicated, the outcome is that one man can metaphorically see through the eyes of another.

By indirect, then, Gibson does not mean requiring cognitive mediation, but rather, perceiving information about events that have been packaged in a tiered fashion, where the upper tiers are structured by another perceiver/actor.

What is the difference for the perception of events that have a level of indirect as well as of direct specification? I do not see any fundamental difference in the *manner* in which perception occurs, although *what* is perceived is different. (That is, when I look at a table, I see it; when I hear a linguistic communication about a table, I perceive selected *information about tables*, not tables themselves.)

When an event is perceived directly, it is perceived by extraction of information for the event from informational media. When a linguistic communication is indirectly perceived, information for the talker's vocal-tract activities is extracted from an acoustic signal. The vocal-tract activity (by hypothesis) *constitutes* phonetically structured words organized into grammatical sequences, and thereby indirectly specifies whatever the utterance is about.

It is worth emphasizing that the relation between an utterance (uttered in an appropriate setting) and what it signifies *is* necessary in an important sense. The necessity is not due to physical law directly, but to cultural constraints having evolved over generations of language use. These constraints are necessary in that anyone participating in the culture who communicates linguistically with members of the speech community must abide by them to provide information to listeners and must be sensitive to them to understand the speech of others.

Indeed, in view of this necessity, it seems possible that the distinction between direct and indirect perception could be dispensed with in this connection. Both the phonetically structured vocal-tract activity and the linguistic information (i.e. the information that the talker is discussing tables, for example) are directly perceived (by hypothesis) by the extraction of invariant information from the acoustic signal, although the origin of the information is, in a sense, different. That for phonetic structure is provided by coordinated relations among articulators; that for the linguistic message is provided by constraints on those relations reflecting the cultural context of constraint mentioned earlier. What is "indirect" is apprehension of the table itself—which is not directly experienced; rather, the talker's perspective on it is perceived. Therefore, it seems, nothing is indirectly *perceived.*

I have attempted to minimize the differences between direct and "indirect" perception. However, there is a difference in the reliability with which information is conveyed. It seems that this must have to do with another sort of mediation involved in linguistic communications. As already noted, in linguistic communications the information is packaged into its grammatically structured form by a *talker* and not by a lawful relation between an event and an informational medium. And as noted much earlier, talkers make choices concerning what the listener already knows and what he or she needs to be told explicitly. Talkers may guess wrongly. Alternatively, they may not know exactly what they are trying to say and therefore may not provide useful information. For their

part, listeners, knowing that talkers are not entirely to be trusted to tell them what they need to know, may depend relatively more on extraperceptual guesses.

Preparation of this paper was supported by NICHD Grant HD 16591 to Haskins Laboratories. I thank Ignatius Mattingly for his comments on an earlier draft of the manuscript.

References

Beattie, G. (1983). *Talk: an analysis of speech and non-verbal behaviour in conversation.* Milton Keynes, England: Open University Press.

Black, J. W. (1949). *Loudness of speaking. I. The effect of heard stimuli on spoken responses.* Joint Project No 2 Contract N 7 Nmr-411 T. O. I., Project No NM 001 053 US Naval School of Aviation, Medicine and Research. Pensacola, Florida and Kenyon College, Gambier, Ohio (cited in Webb, 1972).

Blumstein, S., Isaacs, E. & Mertus, J. (1982). The role of the gross spectral shape as a perceptual cue to place of articulation in initial stop consonants, *Journal of the Acoustical Society of America,* **72**, 43–50.

Blumstein, S. & Stevens, K. (1979). Acoustic invariance in speech production: Evidence from measurement of the spectral characteristics of stop consonants, *Journal of the Acoustical Society of America,* **66**, 1001–1017.

Blumstein, S. & Stevens, K. (1981). Phonetic features and acoustic invariance in speech, *Cognition,* **10**, 25–32.

Browman, C. (1980). Perceptual processing: evidence from slips of the ear. In: V. Fromkin (ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand.* New York: Academic Press.

Carney, P. & Moll, K. (1971). A cinefluorographic investigation of fricative-consonant vowel coarticulation, *Phonetica,* **23**, 193–201.

Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment, *Phonetica,* **22**, 129–159.

Chistovich, L., Klaas, I. & Kuzmin, I. (1962). The process of speech sound discrimination, [translated from] *Voprosy Psikhologii,* **6**, 26–39.

Comrie, B. (1980). Phonology: a critical review. In: B. Butterworth (ed.), *Language production, I.* London: Academic Press.

Condon, W. (1976). An analysis of behavioral organization, *Sign Language Studies,* **13**, 285–318.

Condon, W. & Ogston, W. (1971). Speech and body motion synchrony of the speaker-hearer. In: D. Horton & J. Jenkins (eds.), *The perception of language.* Columbus, Ohio: Charles C. Merrill.

Condon, W. & Sander, W. (1974). Neonate movement is synchronous with adult speech: interactional participation and language acquisition, *Science,* **183**, 99–101.

Cooper, A., Whalen, D. & Fowler, C. A. (1984). Stress centers are not perceived categorically. Paper presented to the 108th meeting of the Acoustical Society of America, Minneapolis, Minnesota.

Cutting, J. E. & Pisoni, D. B. (1978). An information-processing approach to speech perception. In: J. F. Kavanagh & W. Strange (eds.), *Speech and language in the laboratory school, and clinic.* Cambridge, Mass.: MIT Press.

Donegan, P. & Stampe, D. (1979). The study of natural phonology. In: D. Dinnsen (ed.), *Current approaches to phonological theory.* Bloomington, Indiana: Indiana University Press.

Dorman, M., Studdert-Kennedy, M. & Raphael, L. (1977). Stop-consonant recognition: release bursts and formant transitions as functionally-equivalent context-dependent cues, *Perception and Psychophysics,* **22**, 109–122.

Elman, J. & McClelland, J. (1983). Speech perception as a cognitive process: The interactive activation model. ICS Report No 8302. San Diego: University of California, Institute of Cognitive Science.

Fant, G. (1960). *Acoustic theory of speech production.* The Hague: Mouton.

Fant, G. (1973). *Speech sounds and features.* Cambridge, Mass.: MIT Press.

Fant, G. & Lindblom, B. (1961). Studies of minimal speech and sound units, *Speech Transmission Laboratory: Quarterly Progress Report,* **2/1961**, 1–11.

Fitch, H., Halwes, T., Erickson, D. & Liberman, A. (1980). Perceptual equivalence of two acoustic cues for stop consonant manner, *Perception and Psychophysics,* **27**, 343–350.

Fodor, J. (1974). *The language of thought.* New York: Thomas Y. Crowell.

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

Fowler, C. A. (1981). A relationship between coarticulation and compensatory shortening. *Phonetica,* **38**, 35–50.

Fowler, C. A. (1983). Converging sources of evidence for spoken and perceived rhythms of speech: cyclic production of vowels in sequences of monosyllabic stress feet, *Journal of Experimental Psychology: General,* **112**, 386–412.

Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception and Psychophysics,* **36**, 359–368.

Fowler, C. A. & Smith, M. R. (1986). Speech perception as "vector analysis": an approach to the problems of segmentation and invariance. In: J. Perkell & D. Klatt (eds.), *Invariance and variability of speech processes*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Fowler, C. A. & Tassinary, L. (1981). Natural measurement criteria for speech: the anisochrony illusion. In: J. Long & A. Baddeley (eds.), *Attention and performance IX*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Fujimura, O. (1962). Analysis of nasal consonants, *Journal of the Acoustical Society of America*, **34**, 1865–1875. Reprinted in I. Lehiste (ed.), *Readings in acoustic phonetics*. Cambridge, Mass.: MIT Press, 1967.

Garnes, S. & Bond, Z. (1980). A slip of the ear: A snip of the ear? A slip of the year? In: V. Fromkin (ed.), *Errors in linguistic performance: slips of the tongue, ear, pen, and hand*. New York: Academic Press.

Gibson, J. J. (1966). The problem of temporal order in stimulation and perception, *Journal of Psychology*, **62**, 141–129. Reprinted in E. Reed & R. Jones (eds.), *Reasons for realism*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1982.

Gibson, J. J. (1971). A preliminary description and classification of affordances. In: E. Reed & R. Jones (eds.), *Reasons for realism*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1982.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston, Mass.: Houghton-Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, Mass.: Houghton-Mifflin.

Griffin, D. R. (1958). *Listening in the dark: the acoustic orientation of bats and men*. New Haven: Yale University Press.

Hammarberg, R. (1976). The metaphysics of coarticulation, *Journal of Phonetics*, **4**, 353–363.

Hammarberg, R. (1982). On redefining coarticulation, *Journal of Phonetics*, **10**, 123–137.

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

Hockett, C. (1955). *Manual of phonology*. Publications in Anthropology and Linguistics, No 11. Bloomington, Indiana: Indiana University Press.

Hockett, C. (1960). The origin of speech, *Scientific American*, **203**, 89–96.

Hombert, J.-M. (1979). Consonant types, vowel quality and tone. In: V. Fromkin (ed.), *Tone: a linguistic survey*. New York: Academic Press.

Hornbostel, E. M. von (1927). The unity of the senses, *Psyche*, **7**, 83–89.

Jaffe, J. (1964). Computer analyses of verbal behavior in psychiatric interviews. In: D. Rioch (ed.), *Disorders of communication: Proceedings of the Association for Research in Nervous and Mental Diseases*. Vol. 42. Baltimore: Williams and Wilkins.

Jaffe, J. & Feldstein, S. (1970). *Rhythms of dialogue*. New York: Academic Press.

Javkin, H. (1976). The perceptual bases of vowel-duration differences associated with the voiced/voiceless distinction, *Reports of the Phonology Laboratory* (Berkeley), **1**, 78–89.

Jenkins, J. (1985). Acoustic information for objects, places and events. In: W. Warren & R. Shaw (eds.), *Persistence and change*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Johnston, T. & Pietrewicz, A. (eds.), *Issues in the ecological study of learning*. Hillsdale: N.J.: Lawrence Erlbaum Associates.

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E. & Fowler, C. A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures, *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 812–832.

Kenstowiscz, M. & Kisseberth, C. (1979). *Generative phonology: Description and theory*. New York: Academic Press.

Kessen, W., Levine, J. & Wendrick, K. (1979). The imitation of pitch in infants. *Infant Behavior and Development*, **2**, 93–100.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants, *Journal of the Acoustical Society of America*, **73**, 322–335.

Klatt, D. (1980). Speech perception: A model of acoustic–phonetic analysis and lexical access. In: R. Cole (ed.), *Perception and production of fluent speech*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Krakow, R., Beddor, P., Goldstein, L. & Fowler, C. A. (1985). Effects of contextual and noncontextual nasalization on perceived vowel height. Paper presented at the 109th Acoustical Society of America, Austin, Texas.

Kuhl, P. & Meltzoff, A. (1982). The bimodal perception of speech in infancy, *Science*, **218**, 1138–1141.

Labov, W. (1963). The social motivation of a sound change, *Word*, **19**, 273–309.

Labov, W. (1966). *The social stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics (third printing, 1982).

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania.

Labov, W. (ed.) (1980). *Locating language in time and space*. New York: Academic Press.

Labov, W. (1981). Resolving the neogrammarian controversy, *Language*, **57**, 267–308.

Lahiri, A., Gewirth, L. & Blumstein, S. (1984). A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: evidence from a cross-language study, *Journal of the Acoustical Society of America*, **76**, 391–404.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America*, **51**, 2018–2024.
Liberman, A. M., Cooper, F. S., Shankweiler, D. & Studdert-Kennedy, M. (1967). Perception of the speech code, *Psychological Review*, **74**, 431–461.
Liberman, A. M. & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, **21**, 1–36.
Lieberman, P. (1984). *The biology and evolution of language*. Cambridge, Mass.: Harvard University Press.
Lindblom, B. (1972). Phonetics and the description of language. *Seventh International Congress of Phonetic Sciences*. The Hague: Mouton.
Lindblom, B., MacNeilage, P. & Studdert-Kennedy, M. (1983). Self-organizing processes and the explanation of phonological universals. In: B. Butterworth, B. Comrie & D. Dahl (eds.), *Explanations of linguistic universals*. The Hague: Mouton.
Lindblom, B. & Rapp, K. (1973). Some temporal regularities of spoken Swedish. *Papers in Linguistics from the University of Stockholm*, **21**, 1–59.
Lisker, L. (1978). *Rapid* vs *Rabid*: A catalogue of acoustic features that may cue the distinction. *Haskins Laboratory Status Reports on Speech Research*, **SR-54**, 127–132.
Locke, J. (1983). *Phonological acquisition and change*. New York: Academic Press.
MacKain, K., Studdert-Kennedy, M., Spieker, S. & Stern, D. (1983). Infant intermodal speech perception is a left-hemisphere function. *Science*, **219**, 1347–1349.
MacNeilage, P. & Ladefoged, P. (1976). The production of speech and language. In: E. C. Carterette & M. P. Friedman (eds.), *Handbook of perception: Language and speech*. New York: Academic Press.
Marcus, S. (1981). Acoustic determinants of perceptual center (P-center) location. *Perception and Psychophysics*, **30**, 247–256.
Marslen-Wilson, W. & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10**, 29–63.
Matarazzo, J. D. (1965). The interview. In: B. B. Wolman (ed.), *Handbook of clinical psychology*. New York: McGraw-Hill.
Mattingly, I., Liberman, A., Syrdal, A. & Halwes, T. (1971). Discrimination in speech and nonspeech modes. *Cognitive Psychology*, **2**, 131–159.
McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746–748.
McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review*, **92**, 350–371.
Meltzoff, A. & Moore, M. K. (1985). Cognitive foundations and social functions of imitation. In: J. Mehler and R. Fox (eds), *Neonate cognition*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
Meyer, D. & Gordon, P. (1984). Perceptual-motor processing of phonetic features in speech. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 153–171.
Morton, J., Marcus, S. & Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Review*, **93**, 457–465.
Neisser, U. (1967). *Cognitive psychology*. Englewood Cliffs, N.J.: Prentice-Hall.
Nooteboom, S. G. & Cohen, A. (1975). Anticipation in speech production and its implications for perception. In: A. Cohen & S. G. Nooteboom (eds.), *Structure and process in speech perception*. New York: Springer-Verlag.
Ohala, J. (1981). The listener as a source of sound change. In: C. S. Masek, R. A. Hendrick & M. F. Miller (eds.), *Papers from the parasession on language and behavior*. Chicago: Chicago Linguistics Society.
Öhman, S. (1966). Coarticulation in VCV utterances: spectrographic measurement, *Journal of the Acoustical Society of America*, **39**, 151–168.
Peterson, G. & Lehiste, I. (1960). Duration of syllabic nuclei in English. *Journal of the Acoustical Society of America*, **32**, 693–703.
Porter, R. (1977). Speech production measures of speech perception: systematic replications and extensions. Paper presented to the 93rd meeting of the Acoustical Society of America, Pennsylvania State University.
Raphael, L. (1972). Preceding vowel duration as a cue to the perception of voicing characteristic of word-final consonants in American English. *Journal of the Acoustical Society of America*, **51**, 1296–1303.
Rapp, K. (1971). A study of syllable timing. *Papers in Linguistics from the University of Stockholm*, **19**, 14–19.
Repp, B. (1981). On levels of description in speech research. *Journal of the Acoustical Society of America*, **69**, 1462–1464.
Rubin, P., Baer, T. & Mermelstein, P. (1979). An articulatory synthesizer for perceptual research. *Haskins Laboratories Status Reports on Speech Research*, **SR-57**, 1–15.
Ryle, G. (1949). *The concept of mind*. New York: Barnes and Noble.
Saltzman, E. (in press). Task dynamic coordination of the articulators: A preliminary model. *Experimental Brain Research Supplementum*.
Saltzman, E. & Kelso, J. A. S. (1983). Skilled actions: A task dynamic approach. *Haskins Laboratories Status Reports on Speech Research*, **SR-76**, 3–58.
Samuel, A. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, **110**, 474–494.

Shattuck-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production planning. In: P. MacNeilage (ed.), *The production of speech*. New York: Springer-Verlag.
Shaw, R. & Bransford, J. (1977). Introduction: psychological approaches to the problem of knowledge. In: R. Shaw & J. Bransford (eds.), *Perceiving, acting and knowing: toward an ecological psychology*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
Shaw, R., Turvey, M. T. & Mace, W. (1982). Ecological psychology: The consequences of a commitment to realism. In: W. Weimer (ed.), *Cognition and the symbolic processes, 2*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
Stern, D., Jaffe, J., Beebe, B. & Bennett, S. (1975). Vocalizing in unison and in alternation: two modes of communication within the mother–infant dyad. In: D. Aaronson and R. Reiber (eds.), *Developmental psychology and communication disorders. (Annals of the New York Academy of Sciences*, **263**, 89–100.)
Stevens, K. & Blumstein, S. (1978). Invariant cues for place of articulation in stop consonants, *Journal of the Acoustical Society of America*, **64**, 1358–1368.
Stevens, K. & Blumstein, S. (1981). The search for invariant correlates of phonetic features. In: P. Eimas & J. Miller (eds.), *Perspectives on the study of speech*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
Studdert-Kennedy, M. (1983). On learning to speak, *Human Neurobiology*, **2**, 191–195.
Sussman, H., MacNeilage, P. & Hanson, R. (1973). Labial and mandibular dynamics during the production of bilabial consonants: Preliminary observations, *Journal of Speech and Hearing Research*, **16**, 397–420.
Tuaycharoen, P. (1978). The babbling of a Thai baby: echoes and responses to the sounds made by adults. In: N. Waterson & C. Snow (eds.), *The development of communication*. Chichester: John Wiley.
Tuller, B. & Fowler, C. A. (1980). Some articulatory correlates of perceptual isochrony, *Perception and Psychophysics*, **27**, 277–283.
Uzgiris, I. (1973). Patterns of vocal and gestural imitation in infants. In: L. J. Stone, H. T. Smith & L. B. Murphy (eds.), *The competent infant*. New York: Basic Books.
Verbrugge, R. (1985). Language and event perception: Steps toward a synthesis. In: W. Warren & R. Shaw (eds.), *Persistence and change*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
Walley, A. & Carrell, T. (1983). Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants, *Journal of the Acoustical Society of America*, **73**, 1011–1022.
Warren, R. (1970). Perceptual restoration of missing speech sounds. *Science*, **167**, 392–393.
Warren, W. & Shaw, R. (1985). Events and encounters as units of analysis for ecological psychology. In: W. Warren & R. Shaw (eds.), *Persistence and change*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
Webb, J. (1972). Interview synchrony: an investigation of two speech rate measures. In: A. W. Siegman & B. Pope (eds.), *Studies in dyadic communication*. New York: Pergamon Press.
Whalen, D. (1981). Effects of vocal formant transitions and vowel quality on the English [s]–[š] boundary, *Journal of the Acoustical Society of America*, **69**, 275–282.
Whalen, D. (1984). Subcategorical mismatches slow phonetic judgments, *Perception and Psychophysics*, **35**, 49–64.
Wright, J. (1980). The behavior of nasalized vowels in the perceptual vowel space, *Report of the Phonology Laboratory* (Berkeley), **5**, 127–163.