Listeners do hear sounds, not tongues[a),b)]

999

Carol A. Fowler
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511-6695

(Received 30 August 1994; revised 10 February 1995; accepted 3 March 1995)

The paper first distinguishes the two perceptual theories, the motor theory and the theory of direct perception, that nearly agree in the claim that listeners to speech perceive vocal tract gestures. Next it justifies the claim of the direct realist theory that listeners perceive gestures and consider some experimental evidence in its favor. Finally it addresses evidence and arguments judged by Ohala to disconfirm the theory. The argument is made that most of the evidence put forward by Ohala is irrelevant to a distinction between theories that we perceive acoustic signals and theories that we perceive gestures. The arguments are inaccurate or highly selective in the data upon which they draw. © *1996 Acoustical Society of America.*

PACS numbers: 43.71.An

INTRODUCTION

> "Why do the advocates of MT [motor theory] and DR [direct realism] think that listeners recover speech articulations? One of the motivations is the lack of obvious invariance between the assumed linguistic units of speech and their acoustic manifestation" (Ohala, 1994, p. 21).
>
> "Some researchers, abandoning the search for acoustic invariance contained in the speech-input signal, have turned to the speech-output gesture as an alternative and possible source of invariance. This theoretical position has become known as *the motor theory of speech perception* (Liberman and Mattingly, 1986 [sic]) or a direct-realist perspective on speech perception (Fowler, 1986)" (Sussman, 1989, p. 633).
>
> "There is nothing so plain boring as the constant repetition of assertions that are not true" (Austin, 1962, p. 5, quoted in Gibson, 1966).

In the literature two theories of speech perception claim that the primitives, or the smallest perceivables, of speech perception are linguistic gestures. Contrary to the sampled quotations from Ohala and Sussman above, that is almost the only significant matter on which the two theories agree, and they are not even in perfect agreement on that one. The main purposes of the present paper are to motivate the claim of my theory that we perceive linguistically significant gestures of the vocal tract and to respond to comments in this issue by Ohala (1996) against the perception of gestures. However, the motor theory and the theory of direct perception are first contrasted, in a effort to prevent the mistaken coupling of the theories represented in the pair of quotations above and elsewhere in the literature.

I. THE MOTOR THEORY AND THE DIRECT REALIST THEORY CONTRASTED

It is misleading to suggest that a major motivation for the motor theory was the apparent lack of acoustic invariants for phonological or even phonetic segments. [It is inaccurate to suggest that a lack of invariants is any kind of motivation for the theory of direct perception, which *posits* invariants or "specifiers" (Fowler, 1986, 1994a, b).] That phonetic segments are not specified by acoustic invariants was, and remains, an opinion of motor theorists (e.g., Liberman and Mattingly, 1985, p. 26); however, by itself, this opinion does not motivate a motor theory. Many theorists deny invariance but, among those who do, perhaps just two (Liberman and Mattingly) are motor theorists. So why are motor theorists motor theorists? Most crucially, it was the following coupling of negative and *positive* evidence that led to the motor theory: "there is typically a lack of correspondence between acoustic cue and perceived phoneme, *and in all these cases it appears that perception mirrors articulation more closely than sound*" (Liberman *et al.*, 1967, p. 453, italics added). Some of the findings that were most convincing to the original motor theorists have either of the structures illustrated in Fig. 1.

In synthetic two-formant syllables /di/ and /du/, for example (Liberman *et al.*, 1967), information for /d/ is a high rise in $F2$ frequency syllable initially; in /du/ it is a low fall. However, in both syllables as produced naturally, /d/ is achieved by a constriction of the tongue blade against the alveolar ridge of the palate. Gesturally, /d/ is the same in the two syllables. Because of coarticulation, acoustic consequences after release are different, but the percepts are the same, even indistinguishable.

Synthetic /pi/ and /ka/ can be constructed to illustrate a complementary case (Liberman *et al.*, 1952): An acoustic burst centered at 1440 Hz before steady-state formants for /i/ is heard as /p/; before steady-state /a/, it is heard as /k/. Here the same acoustic signal that, because of the consonant–vowel coarticulation that must occur in natural speech, had to have been the product of distinct consonantal constrictions is heard as distinct consonants in the two contexts. This kind of positive evidence (not just negative evidence regarding

[a)] We hear sounds by definition; sounds are what we hear. [For example, the first definition of the noun usage of "sound" in *Webster's Ninth Collegiate Dictionary (1990)* is "the sensation perceived by the sense of hearing."] But I especially like *American Heritage Dictionary's* (1982) fourth definition of sound ("an articulation made by the vocal tract").

[b)] This article is an expansion of the critique given by the author at the meeting session when the Ohala paper was presented.

0001-4966/96/99(3)/1730/12/$6.00

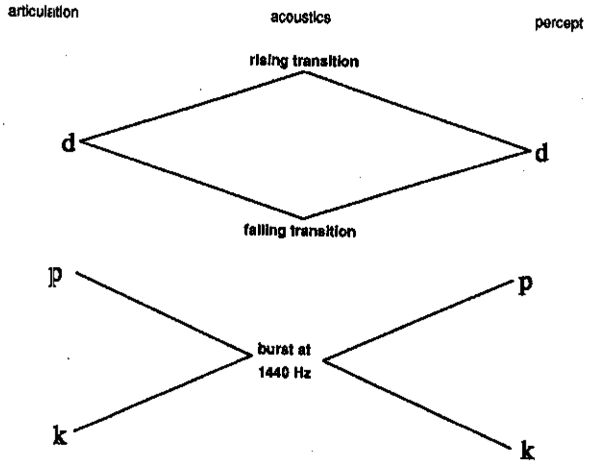

FIG. 1. Evidence from early speech research showing an apparently closer correspondence between articulation and the speech percept than between acoustics and the percept.

invariance) that percepts conform more closely to articulation than to the acoustic signal led to the motor theory.

The motor theory does not have its name only because of its claim about objects of speech perception, however. Motor theorists are in agreement with most other speech theorists (direct realists excepted) in the assumptions they make about most of auditory perception. Generally, according to Liberman and Mattingly (1989)—except for speech perception and sound localization—auditory percepts are "homomorphic" with respect to (that is, have the same form as) the acoustic signal. Extract a formant transition from a speech signal and it sounds like the pitch glide it resembles spectrographically. But the /d/s in /di/ and /du/ sound nothing like the pitch glides that signal them. Phonetic percepts are "heteromorphic" with respect to the acoustic signal (and homomorphic with respect to intended gestures of the vocal tract). In short, for motor theorists, speech perception is different from general auditory perception, and a special-to-speech account is required of how the speech percept acquires its motor character.

The account offered by motor theorists is that listeners recruit their own speech motor systems in perceiving speech. In the theory, invariant motor control structures for phonological segments give rise after coarticulation to variable vocal-tract gestures with variable acoustic consequences. The invariant control structures used by the speaker are determined by the listener with the help of his/her speech motor system (or by an "innate vocal-tract synthesizer" according to Liberman and Mattingly, 1985). The percept conforms to the recovered motor control structures, and that is why it has a motor character. Perceptual involvement of the speech motor system has two advantages in the theory. It permits perception of invariant phonological segments in coarticulated speech, and it fosters achievement of "parity." Parity is the essential ideal of communication systems that the message sent by a sending system must count as the same message for a receiving system. Yet there is an apparent difficulty with that if the sending system is a vocal tract producing gestures and the receiving system is an auditory system that works with acoustic signals. Parity is fostered when listeners' own sending systems are recruited by their receiving systems.

The theory of direct perception, to be partially motivated in the next section of the paper, disagrees in almost every respect with the motor theory. In this theory, despite coarticulation, gestures of the vocal tract themselves have invariant properties. Further, they (not their neural control structures) *are* phonological components of an utterance. That is, phonological gestures are the public actions of the vocal tract that cause structure in acoustic speech signals. By hypothesis, they will be found to cause specifiers or invariants (Fowler, 1994a, b) in the acoustic signal. Indeed, the existence of specifying acoustic properties is what allows perception of the phonological properties to be direct (that is, unmediated by processes of hypothesis testing or inference making and unmediated by mental "representations" in the literal sense of mental standins for real-world things).

In the theory, listeners perceive gestures because perceptual systems have the function universally of perceiving real-world causes of structure in media, such as light, air, and the surfaces of the body, that sense organs transduce. Accordingly, perception is generally heteromorphic with respect to structure in those media; instead, perception is not just homomorphic with, it is *of*, the real-world events that cause the structure. That is, speech perceivers, and perceivers in general, are realists (Fowler, 1987). Indeed, it is their status as perceptual realists that explains parity.

In the theory of speech perception as direct, speech perception is not special, and there is no more reason to propose a role for the speech motor system in speech perception than to propose an analogous role for the viewer's locomotor system in visual perception of walking.

With respect to most perspectives on speech perception, the theory of direct perception makes two distinctive claims. Perception is direct, and perceptual objects are phonological gestures of the vocal tract. The focus of the Acoustical Society symposium was on the latter claim, and, accordingly, that is the claim I will motivate in the next section of the paper.

II. WHY WE PERCEIVE GESTURES: THE UNIVERSAL FUNCTION OF PERCEPTION

A. Proper contexts for theory development

In my view, we cannot develop a realistic theory of speech perception unless we embed the developing theory in the context of other theories that place constraints on the form it should take. In particular, to the extent that science knows something about perception in general, that knowledge should be applied to the study of speech. Too, the form that a theory of speech perception takes can be affected fundamentally by assumptions that are made about speech production and linguistic phonology. In nature, knowers of linguistic phonology are also producers and perceivers of phonological segments. Whatever is true of a knower of linguistic phonology cannot be incompatible with what is true of a speech perceiver or producer. Accordingly, a theory of speech perception needs also to be developed in the context of theories of speech production and of linguistic phonology in such a way that the three kinds of theory are constrained to be as mutually compatible as they are, necessarily, in nature.

Following are consequences of embedding the theory of speech perception in the context of a larger theory of perception. These consequences motivate the claim that we perceive gestures of the vocal tract. [Codevelopment of compatible theories of speech production (see, e.g., Fowler, 1993, in press; Fowler and Saltzman, 1993) and of linguistic phonology (see, e.g., Fowler, 1993, in press), which largely borrow, with some adjustment, the ideas of Browman and Goldstein (1986, 1992) have served to make plausible the claim of the theory that speech perception can be direct.] Indeed, the consequences suggest that, although perception of gestures is not at all logically necessary for skilled listener/speakers, it is biologically necessary.

B. Speech perception in relation to the universal character of perception

Theories of perception, when complete, explain both public and private or covert aspects of perception. The theory of public aspects of perceiving focuses on the environment or niche of the perceiver—that is, on what an animal must perceive to survive—on the informational support for what is perceived, and on empirical investigation of the potential information that perceivers actually use in perceptual guidance of action. The theory of covert aspects of perceiving details how neural-sensory systems extract that information. The theory of perception in which I have embedded my theory of speech perception is a theory of public aspects of perceiving. It is a theory of universal perceptual function based on James Gibson's theory of direct perception (J. J. Gibson, 1966, 1979; Reed and Jones, 1982). The theory starts there, because, in my view, we lack the right tools to understand much about the neural activity supporting perception until we understand what it serves to accomplish.

Perceptual systems have a universal function. They constitute the sole means by which animals can know their niches. Moreover, they appear to serve this function in one and only one general way: They use structure in media that has been lawfully caused by events in the environment as information for the events. Even though it is the structure in media (light for vision, skin for touch, air for hearing) that sense organs transduce, it is not the structure in those media that animals perceive. Rather, essentially for their survival, they perceive the components of their niche that caused the structure.

Consider vision, where our intuitions may be clearest. We can see components of our ecological niche—objects and events—because light from a source is lawfully and causally structured by objects and events in the niche. Further, it tends to be the case that distinctive properties of objects and events structure light in distinctive ways and, consequently, there is not only a causal direction of influence (here, an "arrow") from event structure to structure in light; there is also an informational or specificational arrow that goes the other way. That is, given a certain patterning over time in light structure, the object or event that caused that structure can be known. Light imparts its structure to visual systems, and visual systems use the structure, not as something to be perceived in itself, but as the means by which perceivers can see what populates their ecological niche. Perceivers use the specificational arrow to know their world from the patterns it causes in reflected light. This is illustrated in the top panel of Fig. 2.

ecological niche	informational media	percept
visible object or event	structure in reflected light	object or event
haptic exploration of an object	skin deformations	object
sounding event (e.g., phonetic gestures)	acoustic signal	event (phonetic gestures)

FIG. 2. A schematic illustration of the universal character of perceptual function.

Now consider haptic perception (middle panel of the figure). As we explore an object haptically, it causally deforms the skin (among other consequences). It tends to be the case that distinct properties of haptically explored objects deform the skin in distinctive ways. (Imagine, for example, using a hand to explore a held pencil versus a held soda can versus a woolen blanket.) Consequently, there is not only a causal "arrow" from event structure to the patterning of skin deformations over time; there is also a specificational arrow that goes the other way. Given a patterning of skin deformations over time, characteristics of the exploratory event that caused it can be known. Patterns of skin deformations are imparted to sensory receptors, and the haptic perceptual system uses the imparted patterns to perceive their causes in the world. Humans feel a long rigid, cylindrical object in the hand (a pencil), or a larger, colder, cylindrical, more flexible metal object (a soda can) or a soft, fuzzy deformable surface (a blanket). That is, we feel the components of our ecological niche via the specificational structure they cause on the surface of the body. That must be an essential function of haptic perception; its essential function cannot be to feel skin deformations any more than the function of visual perception can be to see structure in light.

These perceptual systems were shaped by natural selection to serve the function of acquainting perceivers with components of their niches. Auditory perception can only have been selected for the same function. There is no survival advantage to hearing structured air, but there is an advantage, for example, to locating a large lumbering animal out of view and to detecting which way, in respect to one's self, it is lumbering.

In auditory perception, events causally structure air (Fig. 2, bottom). It tends to be the case (see, e.g., Gaver, 1993) that distinctive events structure air in distinct ways. Accordingly, acoustic signals caused by car engines are different from those produced by water glasses being filled or broken, by slamming doors and by speech. Therefore, there is not only a causal arrow from properties of events to patterns of structure in air; there is also an informational or specificational arrow going the other way. That is, given a patterning of structure in the air over time, properties of the event that caused it can be known.

As noted earlier, evidence that listeners perceive speech gestures has been taken by motor theorists to imply that speech perception is different in character from general auditory perception. In the view of motor theorists (e.g., Liberman and Mattingly, 1989), in auditory perception generally, but not in speech perception, the percept is "homomorphic" with respect to the acoustic signal. However, from my theoretical perspective, evidence that listeners perceive gestures[1] is evidence that speech perception is not special in this way. Indeed, to make the claim that listeners hear acoustic speech signals is to propose that speech perception is special, auditory-system based theories of speech perception to the contrary (e.g., Kluender, 1994). Perception of gestures is not logically necessary for skilled speaker/listeners, but it is biologically necessary.

III. EVIDENCE THAT GESTURES ARE PERCEIVED

Ohala (this issue) argues that no experimental findings unequivocally prove perception of gestures and that all findings given that interpretation have alternative interpretations. That opinion may explain his failing to address any of the evidence purporting to show that listeners perceive gestures. It is not scientifically defensible, however. There are very few crucial experiments in science, and speech science is not special in this regard. Despite that, experimentation in speech science is central to development of an understanding of speaking and listening. A potent tool in experimental research is the use of converging evidence (Garner *et al.*, 1956).[2] Imagine that there are, say, five classes of findings that a direct realist or a motor theorist would argue demonstrates perception of gestures. These theories may provide just one account of the five classes of findings: Gestures are perceived. If "acoustic" theorists have alternative accounts, the merit of the challenges they mount can be evaluated by counting the accounts they provide and determining the scope of each one. If just one alternative account explains all five sets of findings, then the two classes of theory are, on those grounds, on equal footing. However, if acoustic theories have five different accounts for the five kinds of findings or no account of some, then the gestural account is better supported by the evidence. Following, I will describe findings that I conclude demonstrate perception of gestures. Some have clear alternative interpretations; some do not. I will suggest that the only consistent account of the findings is that listeners to speech hear gestures.

A. The McGurk effect

When a speaker mouths, say, the syllable /da/, synchronized with the acoustic signal for /ma/, perceivers are most likely to report *hearing* /na/—a percept that integrates the visibly perceived place of articulation with the auditorily perceived manner and voicing (e.g., Dekle *et al.*, 1992; MacDonald and McGurk, 1978; McGurk and MacDonald, 1976; Massaro, 1987; Summerfield, 1987); for example, listeners reported /na/ on 92% of optical /da/–acoustic /ma/ pairings in the study of MacDonald and McGurk (1978). This finding, called the "McGurk effect" after one of its discoverers, is phenomenally very striking. It is not that the listener says, in effect, "I thought I heard /ma/, but I saw that the lips did not close; therefore, it must have been /na/." Rather, looking at the talker, the listener experiences hearing /na/; with eyes closed, he or she hears /ma/.

One interpretation for this finding is that listeners perceive gestures, and some gestures are specified optically as well as acoustically. An alternative interpretation, however (Massaro, 1987, and perhaps Diehl and Kluender, 1989), is that listeners have past experience both seeing and hearing people speak so that memories of optical as well as acoustic cues for a phoneme (or a syllable in Massaro's account) are associated with the mental concept of the phoneme (or syllable), and those associations explain the lip reading that we all do. These are the only accounts of which I am aware of the McGurk effect, and I believe that a colleague and I have disconfirmed the second (Fowler and Dekle, 1991).

Literate perceivers have ample experience both seeing printed words and hearing the words spoken. Further, research shows (e.g., Seidenberg and Tanenhaus, 1979; Tanenhaus *et al.*, 1980) that, in experiments involving only auditory presentation of words, associated spellings come to subjects' minds unbidden. Accordingly, memory includes lexical knowledge in which sound and print are associated. If the McGurk effect arises from analogous associations, then an effect of seeing a printed word on an experience of hearing spoken words should occur that is analogous to the McGurk effect.

In contrast, people have very little experience associating the haptic feel on the hand of a face producing speech with the corresponding acoustic speech signal. Feeling the face of a speaker with one's hand is considered rude in most contexts, and most of us have done so considerably less often than we have seen the face of a speaker whom we hear talking or than we have seen print and heard it read. If the McGurk effect arises from associations in memory between the sight and sound of a speaker, then an analogous effect of the haptic feel of a speaker talking on a listener's experience of hearing speech should be very weak or absent.

My theory makes different predictions about the relative size of the cross-modal influences in these two conditions. The McGurk effect occurs when the experimenter tricks the perceiver into experiencing one event of a speaker talking. Information from the sight of the speaker then must be about some of the same gestures as information from the acoustic signal, and integration of cross-modal information occurs for that reason. However, in an experiment in which print substitutes for the face, sight (the printed word or syllable) and sound (a spoken word or syllable) are no longer conjoint consequences of one event of speaking. Accordingly, no perceptual effect of the one on the other should occur. In contrast, if an experimenter can trick a perceiver into experiencing as consequences of one event, the haptic, manual, feel of a face producing a word or syllable and an acoustic signal of an appropriately chosen similar word or syllable, a McGurk-like effect should occur.

Dekle and I (Fowler and Dekle, 1991) synthesized a continuum of syllables from /ba/ to /ga/. For one group of subjects, on each trial, a syllable, either BA or GA was printed on a computer terminal screen synchronized with acoustic presentation of one of the continuum members. The subjects' task was first to identify the syllable they heard and then to identify the one they saw. For a second group of subjects, the haptic, manual, feel of a face mouthing /ba/ or /ga/ was substituted for the printed syllable. (See Fowler and Dekle, 1991, for details of the procedure.) The task of these subjects was first to identify the syllable they heard and then to identify the syllable they felt. We told the subjects in both groups (accurately) that we had independently and randomly paired the printed (felt) and acoustic syllables and therefore that they should not let themselves be influenced in their judgments of the heard syllable based on what they saw printed (or felt spoken), and vice versa for the judgments of printed (mouthed) syllables. This kind of instruction does not affect the original McGurk effect appreciably. We used it in an effort to reduce effects of response bias, whereby, when subjects hear an ambiguous syllable, they might choose to report what they had less ambiguously experienced from the other modality.

When printed words were masked to bring identification performance down to the level at which felt syllables were identified (about 78% correct), results were very striking. There was no effect of printed syllables on reports of heard syllables, but there was a large effect of felt syllables on heard syllables (and a highly significant reverse effect of heard syllables on judged mouthed syllables).^{3} This is incompatible with an associationist account of the McGurk effect, but it is exactly as predicted by the theory of direct perception.

B. Shadowing response times

In most circumstances, there is a marked difference in "simple" as contrasted with "choice" reaction times in appropriately matched tasks, with the latter being appreciably longer than the former [on the order of 150 ms longer according to Porter, 1978; see also Luce's (1986) estimate of a 100- to 150-ms difference]. In a simple reaction time procedure, the task is to respond—for example, to hit a button—whenever any stimulus (say, any tone) occurs. In a choice reaction time procedure, the task is to respond differently depending on the stimulus. For example, a subject might be instructed to hit one button if the stimulus is a high tone and a second button if it is a low tone. The difference in response times for simple and choice response tasks presumably reflects the decisions that must be made in the latter, but not the former, case (e.g., which tone occurred and therefore which response button should be pushed).

There is an exception or, sometimes, a near exception, to this generalization, however. If a subject's task is to shadow (that is, repeat after) an utterance, response times can be as fast or nearly as fast as simple response times (Kozhevnikov and Chistovich, 1965; Porter, 1978; Porter and Castellanos, 1980; Porter and Lubker, 1980), even though the task is a choice task because the vocal response varies with the stimulus.

For example, Porter and Lubker (1980) had subjects shadow a synthetic vowel–vowel sequence in one condition (choice reaction time). All stimulus sequences began with /a/, and subjects were to begin producing /a/ as soon as they heard it begin. A variable interval after /a/ onset, the vowel changed to one of three. Response time was measured as the time to begin the second vowel defined either electromyographically or acoustically relative to the acoustic onset of the second stimulus vowel. In another condition, stimuli were the same, but now subjects produced the same vowel–vowel sequence (/ao/) regardless of the synthetic sequence they heard (simple reaction time). In exactly comparable trials in which the stimulus was the sequence /ao/ and the response was /ao/, choice and simple response times, measured acoustically, differed by a nonsignificant 12 ms. Acoustic response times averaged 180 ms in the choice reaction time task and 168 ms in the simple task. In a task in which listeners either shadowed VCVs (with all V's /a/ and five different consonants), or they provided a single response to stimulus VCVs, response times were about 50 ms slower in the choice (shadowing) as compared to the simple response time condition (Porter and Castellanos, 1980). Kozhevnikov and Chistovich (1965) found that shadowing response times were considerably shorter (by 100–200 ms) than times to initiate a written identification of spoken syllables.

If listeners perceive gestures, these findings are no mystery. Perceiving how the speaker produced a syllable or word makes replicating the perceived action—that is, imitating it—easy. Presumably, an accelerated version of Simon Says would yield similar findings for other bodily actions. If an actor's task is to imitate the actions of a model, and if actors perceive the model's actions, response latencies will be closer to those of simple than to typical choice tasks, precisely because perception of what the model did in effect constitutes instructions for the response.

If, instead of perceiving phonological gestures, we perceive acoustic signals (or, as in most acoustic theories, acoustic signals that we assign to mental phonological categories), how are the findings to be explained? I do not know. As far as I know, no interpretation has been offered of these data from other theoretical perspectives. If we perceive acoustic signals that we then assign to abstract phonological categories, the shadowing task is still a choice reaction time task. The listener has to choose which vocal control structures should implement a response if the perceived category is /ba/ and which if it is /ga/, for example.

Perhaps I have overlooked a viable account of these findings from the perspective of an acoustic theory. Even so, I am confident that the account will not be the same as its (failed) account of the McGurk effect. The two findings must have different explanations unless it is supposed that listeners perceive gestures.

C. Synthetic /di/ and /du/

The synthetic syllables, /di/ and /du/, of Fig. 1 are heard as having identical syllable–initial consonants. In alphabetic writing systems crosslinguistically, those initial consonants have the same spelling, and that is precisely because they sound identical. Why do they? For direct realists and motor theorists, the reason is that the acoustic signal informs the listener that the same gesture (for motor theorists, intended gesture) of the vocal tract formed the consonant constriction in both syllables. Because of coarticulation, the acoustic signals after release (and the only acoustic signal offered in these synthesized tokens) are different. But they inform about the same consonantal constriction and so are heard as the same.

What other explanations are there for invariant percepts corresponding to invariant articulations? Massaro (1987) appears to have no account. For him, perceived phonological categories are syllables, not phonemes. Accordingly, it is no problem that acoustic signals are different for /di/ and /du/, but it must be a mystery why the syllables have the same "first names."

For Sussman (1989), the explanation is that, across CVs with the same initial consonant, $F2$ at syllable onset plotted against $F2$ at vowel midpoint gives a straight line, the slope of which serves as invariant information for place of articulation. A difficulty with this interpretation (among others, see Fowler, 1994b) is that listeners to /di/, for example, only get a pair of (x,y) coordinates; they do not get a line or its slope. How accurate could perceivers be if they identify consonants by determining the locus equation line to which the coordinates fall closest? Not as accurately as they can perceive the consonants. A discriminant analysis that classifies tokens into the categories /b/, /d/, and /g/ based on the (x,y) coordinates succeeded on about 65%–70% of attempts in my research and slightly more accurately in Sussman's (Sussman *et al.*, 1991). In contrast, listeners can classify such tokens (syllables produced in isolation) correctly on approximately 95% of occasions (Brancazio and Mitra, 1994).

For Stevens and Blumstein (e.g., Stevens and Blumstein, 1981), there is invariant information in the spectrum at stop release for /d/ in the context of different vowels. This information is not present in synthetic two-formant /di/ and /du/. However, Stevens and Blumstein propose that, in experience, we learn to associate context-sensitive "secondary" cues for consonant place, such as $F2$ formant transitions, with invariant primary cues, and eventually, the secondary cues can stand in for the primary ones. This account runs into difficulty. First, there is no evidence that giving audibly distinct acoustic signals the same name (e.g., calling audibly distinct high and low tones both "dee" or calling the $F2$ transitions of both /di/ and /du/ both /d/) leads them eventually to sound the same, and to sound the same as their shared name. Second, research suggests that listeners, even children, depend more on secondary formant transitions for place than on information in the ostensibly invariant spectra at stop release (Blumstein *et al.*, 1982 Walley and Carrell, 1983). That is, there is no evidence, as yet, that secondary cues are secondary for any listeners.

Those difficulties aside, again notice that both of these explanations from acoustic theories for perception of the alveolar place of the initial consonants of /di/ and /du/ differ from the explanations for the McGurk effect and for the simple reaction time latencies of shadowing performances. So far, we need three accounts for these outcomes if we deny that gestures are perceived and just one if we accept that gestures are perceived.

D. Parsing of the acoustic speech signal

To me, the most striking evidence that listeners perceive gestures derives from the way that they "parse" the acoustic speech signal. Perception of fundamental frequency (f_o) provides a good example. The f_o of a voiced portion of a speech utterance is a coherent characteristic of the acoustic signal. But listeners do not treat it as such; that is, there is no coherent percept (such as variation in voice pitch) that corresponds to the variation in f_o during a speech utterance.

Different gestures affect f_o. Laryngeal maneuvers implement the intonation contour of an utterance, and those maneuvers have consequences for f_o that are heard as variation in voice pitch. However, variation in vowel height also affects f_o (e.g., Silverman, 1987, and the summary of the literature therein; Sapir, 1989); so does production of voicelessness of an obstruent preceding a vowel (Hombert, 1978); so does exhaling during production of an utterance within a single inspiration (i.e., declination occurs, e.g., Gelfer, 1987). Remarkably, listeners, "parse" f_o along gestural lines. They only hear as intonational pitch, variation in f_o that talkers produced to implement the intonation contour. They parse effects of vowel height (Silverman, 1987) and declination (Pierrehumbert, 1979) from f_o and from their perception of the intonation contour. They hear f_o perturbations due to variation in vowel height, not as variation in pitch, but as variation in vowel height (Reinholt-Peterson, 1986). Compatibly, they hear perturbations due to consonant voicelessness as voicelessness (Silverman, 1986, 1987).

The findings on "parsing" are not unique to f_o. Compatible findings are obtained on perception of duration (Whalen, 1989) and of coarticulated speech (Fowler, 1981, 1984; Martin and Bunnell, 1981; Whalen, 1984). Why do listeners parse coherent acoustic dimensions, such as f_o, duration, and $F1$ and $F2$ (in perception of coarticulated speech)? My answer is that listeners perceive the coordinated actions of the vocal tract that constitute phonological properties of a speech message. Generally, gestures that occur in overlapping time frames have converging influences on f_o, duration, and the speech spectrum, and listeners, extracting information for gestures, parse the acoustic signal along gestural lines.

Is there an alternative account from the perspective of an acoustic theory? To my knowledge, none has been proposed.

Notice that, looked at in a complementary way, listeners' attention to the acoustic signal has another remarkable characteristic. Not only do listeners parse single acoustic dimensions along gestural lines. In addition, in the appropriate setting, they hear the two phonological segments as the same despite their being signaled by given different acoustic consequences of the same phonological gesture. For example, either a lowering in $F1$ or a raising of f_o can lead to a shift in identification from a lower to a higher vowel (Reinholt-Peterson, 1986). Research described in the next section examines this phenomenon more closely and shows that, not only can such different acoustic signals be classified as the same phonological segment, they may indistinguishable one from the other—even though the acoustic fragments that distinguish them are perfectly discriminable in other contexts.

E. Trading relations: Perceived cohesion of disparate acoustic consequences of the same gesture

Gestures tend to have constellations of diverse acoustic consequences. Producing /p/ in "split," for example, causes "split" to differ acoustically from "slit" in multiple ways. Fitch and her colleagues (Fitch *et al.*, 1980) explored perception of two of those consequences. A stop consonant after /s/ is associated with a silent closure interval, and its release causes labial transitions before "lit." Fitch *et al.* first showed that the two "cues"[4] for /p/—silence after /s/ frication and labial transitions "trade" in signaling the presence of /p/. That is, shorter silence is required for listeners to hear "split" rather than "slit" in the presence versus the absence of labial transitions before "lit." Accordingly, transitions traded for silence in signaling the presence of /p/. Next Fitch *et al.* investigated discrimination of pairs of syllables that differed either in just one of these cues (the presence or absence of labial transitions) or in two, and, if they differed in two cues, the experimenters varied whether the cues "cooperated" or "conflicted." In a cooperating cues condition, a relatively long silence (32 ms) and labial transitions occurred in one syllable of a pair, and a lesser silence (8 ms) and no transitions occurred in the other. In this case, the pair of cues cooperated in signaling "split" in the first pair member described and "slit" in the other. In a conflicting cues condition, the short silence was paired with labial transitions and the long silence with no transitions. Here, in each syllable, one cue signaled "split" and the other "slit." Notice that in both two-cue conditions, cooperating and conflicting, members of a pair differed in two ways acoustically. Further, they differed in the same two ways, in the duration of a silent interval after /s/ frication and in the presence and absence of labial transitions. In the one-cue condition, they differed by just the spectral cue.

In an oddity discrimination task, results were as expected if listeners perceive gestures. Syllables differing in two cues were discriminated better than those differing in just one *only* if the two cues cooperated in signaling the presence in one member of the pair and the absence in the other of a labial gesture. Remarkably, discrimination was even worse in the conflicting cues condition than in the one-cue condition despite the fact that, in the former condition, members of a discriminated pair differed in two ways and, in the latter, they differed in just one of those two ways. Clearly discrimination depended on how the members of a pair were identified. If one was heard as "split" and one as "slit," as in the cooperating cues condition, discrimination was easy. If they were heard as repetitions of the same word, discrimination was hard.

Massaro's theory offers another account of these findings. In this theory, listeners extract features from the acoustic speech signal. Features in the signal are compared with features associated with syllable templates in memory to determine which syllable has associated attributes most compatible with the featural input. In the theory, soon after a speech signal is perceived, memory for the signal itself decays. All that is left in memory is information about what syllables were identified based on the input. Massaro's theory can explain the findings of Fitch *et al.* if it is supposed that listeners were discriminating syllable names rather than the acoustic features that signal the names or categories. Listeners may do so, for example, because the discrimination task separated members of a to-be-discriminated pair by enough time (1 s in the experiment) that the acoustic input for the first syllable had decayed by the time the second syllable was identified. If the names were the same, as would typically be the case in the conflicting cues condition, discrimination would be impossible. It would be easy in the cooperating cues condition in which the perceived names were typically different within a pair.

Massaro's theory predicts that, if syllables were to be presented in conditions permitting and promoting discrimination based on the acoustic cues rather than the syllable templates, then discrimination in the conflicting and cooperating cues conditions should be good and both better than in the one cue condition. When the syllables differ in two ways, they must be easier to discriminate than when they differ in just one of those two ways. To my knowledge, this prediction has not been tested against that of a gestural view that the response pattern should be unchanged under these conditions.

F. Conclusion

A single interpretation of the findings above is that we perceive gestures. That is why we integrate auditory, visual, and haptic information for a speech event, why shadowing reaction times can approach simple reaction times, why synthetic /di/ and /du/ have identical-sounding initial consonants, why listeners parse the acoustic speech signal in the way they do, and why some pairs of acoustic signals that differ in two ways can be nearly impossible to discriminate whereas other pairs of acoustic signals with one of those differences swapped are easy to discriminate. No other theory that I know of can explain all of the findings, and the explanations that they can provide must be different for each distinct phenomenon.

IV. OHALA (1996): EVIDENCE AGAINST PERCEIVING GESTURES

A. Phonological inventories and sound change

Ohala (1996) argues that phonological inventories of languages develop so as to maintain sufficient perceptual distinctiveness of inventory members (see also Lindblom, 1989) and that it is acoustic, not gestural, distinctiveness that is maintained. In his view, this is contrary to expectations if gestures are perceived. Consider these examples: (1) Languages prefer obstruents, which cause acoustic pops and hisses, to more sonorant consonants, which do not. (2) In languages that have a gap in their inventory of voiceless obstruents, it is typically /p/ that is missing. There is nothing particularly unsalient about a voiceless–labial gesture pair, but the consequent acoustic signal has a weak burst. (3) CV syllable structures provide better acoustic information than do VCs for the C, and CV structures are more popular in languages than are VCs. (4) Jakobson *et al.* (1963) proposed a feature [grave] that groups labial and velar segments as [+grave] and segments with intermediate places as

[−grave]. In Ohala's example, English lacks sequences in which the diphthong [au] is followed by [+grave] consonants, but it tolerates [au] followed by [−grave] consonants.

Although I might raise some quibbles about some of these pieces of evidence (e.g., CV syllables are not only more perceivable than are VC syllables, they are also articulatorily more stable, e.g., Kelso *et al.*, 1986; Tuller and Kelso, 1990; Stetson, 1951), there is no need to for the most part, because, for the most part, they are not relevant to the disagreement between theories that we hear acoustic signals or gestures. They are not relevant, because in my theory, as in any viable theory, the acoustic signal plays a pivotal role in speech perception (see Fig. 2). The acoustic signal is, after all, what the ear transduces; ears do not transduce articulations. The theories do not disagree on this point; they disagree on what the acoustic signal *counts as* for the perceiver. For acoustic theorists, it counts as a perceptual object; for me it counts as a specifier of speech events.

In the theory of direct perception, auditory perception in general, and speech perception in particular, can be only as successful as the specifying information provided by the acoustic signal. If two similar gestures structure the air in very distinctive ways, then listeners will have no difficulty knowing which was produced. If a gesture does not structure the air in noticeable ways, it is likely to go unnoticed. Accordingly, the reason why languages prefer obstruents to sonorants may be because obstruents have more salient acoustic consequences than sonorants; they also provide a better acoustic contrast with vowels than do sonorants (cf. Bondarko, 1969) The reason why /p/ is more often omitted from language inventories than other voiceless consonants may be exactly because the information for it in the signal is weak. Compatibly, a reason (but perhaps not the only one) why languages prefer CV syllables over VCs may well be that postvocalic consonants are not well signaled acoustically.

As for [grave], Jakobson *et al.* (1963) did propose a unified articulatory characteristic of [+grave] consonants ("a large and less comparted mouth cavity," p. 30). Even so, I do find [grave] a surprising feature. But, then, it does not appear to have stood the test of time even in nongestural phonologies (see, e.g., Kenstowicz and Kisseberth, 1979; Rocca, 1994—less and more recent texts on "generative phonology" in which the feature does not appear), nor is it required to explain the absence of [au]–labial or –velar sequences. Ohala's own account (provided in the talk, but not in the paper) may be on the right track. Coarticulation of a vowel that originally was monophthongal ([u]) with consonants that were signaled acoustically by $F2$'s far from that of [u] created a noticeable offglide. Listeners misparsed the offglide as an intentional part of the vowel, which, therefore, became diphthongized. Labial and velar consonants, with low $F2$'s, did not cause an offglide and so, in those contexts, /u/ was not diphthongized. There seem to me to be two compelling reasons not to augment this explanation by invoking a featural description.

First, regardless of whether Ohala's account is right, should we assign a common feature value to consonants because they collectively fail to cause anything to happen in some context? Other evidence for [grave] appears to have this same flavor. Ohala cites a paper by Hyman (1973) that provides evidence in favor of the feature [grave]. The evidence that Hyman finds most convincing comes from the language FeʔFeʔ in which reduplication occurs. Following are examples of reduplication in FeʔFeʔ:

za→zɯza

to→tɯto

sii→sisii

pée→pɯpée

tee→titee

kée→kɯkée

Hyman proposes that /ɯ/ (a high, back, unrounded vowel) is the underlying form for the vowel of reduplicated syllables. This underlying vowel becomes /i/ in either of two contexts: in the context of the stem vowel /i/ or the stem vowel /e/ and a [−grave] stem consonant. This appears to be a rule like Ohala's in which [+grave] consonants all fail to cause anything to happen. Rather, something happens in the context of [−grave]. But the defense of [grave] as a feature cannot be based on justifying [−grave] consonants as sharing a feature value. It is not difficult to justify shared features values for consonants with adjacent places of articulation; it is [+grave], the feature value for discontinuous places, that needs justification.

This leaves the example of [w] and other consonants having labio–velar double articulations. These paired gestures apparently cause good, informative acoustic signals, and perhaps that is why they are preferred over other double articulations. But that does not constitute grounds for assigning labial and velar gestures the same value of a phonetic feature.

Consider now evidence from sound change generally. If I understand Ohala's argument here, it is that sound changes frequently occur in which a new gesture occurs (or replaces one that may have been quite different from it) leaving the acoustic signal changed rather less than the gestural complex used to produce the word. Just one of his examples illustrates his argument. Breton and Hindi both exhibit "spontaneous nasalization"—the occurrence of nasalization in vowels that, historically, never occurred in the context of nasal consonants. The conditioning environment for spontaneous nasalization is the occurrence near the vowel of a consonant associated with high airflow. Ohala speculates that coarticulation of the wide glottal opening for these consonants with the vowel created acoustic consequences during the vowel that are similar to consequences of a lowered velum. Listeners misperceived the coarticulatory effects as effects of nasalization and produced the words with nasalization in the vowel.

This account is plausible. But what has it to do with the issue at hand, whether listeners perceive vocal-tract gestures or acoustic signals? It appears to me either irrelevant to the issue or even slightly biased in favor of perceiving gestures. An essential part of the account has to be that the listeners consider the vowels to have been produced with velums lowered. Otherwise there is no explanation for their own velums lowering when they produced the words. So either they perceived lowered velums, or they assigned that interpretation after perceiving the acoustic signal. My account is simpler.

Whether or not the evidence from phonological inventories and sound change points either to acoustic signals or to gestures as perceptual objects, does the evidence imply that "listeners are able to differentiate the elements of speech on the basis of their sound" (i.e., their acoustic signal; Ohala, 1996)? It does. But this is not "in opposition to the view of speech-as-gestures." From my perspective, it is only to acknowledge that information for gestures is acoustic.

Evidence both from the nature of phonological systems of language and from sound change suggests to me no particular primacy for acoustic aspects of speech. Ohala has provided a highly selective look at the data. The discussion will be be brief here, in part because I believe that Ohala's view on this is idiosyncratic and because the evidence does not bear on the nature of perceptual objects.

As for phonological systems, Lindblom (e.g., 1989) has shown, it is true, that he can successfully predict the vowel inventories of languages having inventories of different sizes based only on a criterion (for predicting inclusion of a vowel in an inventory) of perceptual distinctiveness, defined acoustically. However, it is also the case that the acoustic criterion by itself does not predict consonant inventories well. It must be augmented by a criterion of articulatory cost.

A second example of evidence for articulatory shaping of phonological systems is provided by Clements (1985). He pointed out that there are marked differences in the relative independence of different phonetic features with respect to their joint participation in phonological processes cross linguistically. He created a feature tree in which features separate early on in the structure if they tend to be mutually independent crosslinguistically and separate farther down to the extent that they tend not to be. The hierarchy that the evidence yielded looks very much like the structure of the vocal tract (see Clements' Fig. 13). This should not be surprising. Structures that are far apart in the vocal tract (for example, the larynx and the lips) will, in general, be more independently controllable than structures closer together, and these varying degrees of independence will be respected in phonological processes of languages.

As for sound change, changes can occur that appear to be motivated by the talker-based disposition to reduce spoken words where possible (see, e.g., Mowrey and Pagliuca, 1987; Pagliuca, 1982). For example, Mowrey and Pagliuca point out that instances of vowel and consonant weakening or decay considerably outnumber instances of strengthening. For consonants this means moving from those consonants such as voiceless stops that are produced by several gestures or by high-amplitude gestures toward consonants with fewer or weaker gestures. In general, Mowrey and Pagliuca report, "liquids, glides and vowel qualities...descend from weakly-stopped or fricative configurations, which in turn descend from more fully stopped configurations" (p. 37).

As an example of voiceless stops becoming fricatives, Pagliuca (1982) points out that in the High German Consonant Shift, /k/ became /x/, /t/ became /s/, and /p/ became /f/ intervocalically and finally after a vowel. A quick look at Miller and Nicely's (1995) confusion matrices does not suggest that this pattern reflects asymmetries in the likelihood of mishearing. For example, in 17 listening conditions in which signal-to-noise ratios and the frequencies of filter cutoffs were varied, weaker /f/ was reported for stronger spoken /p/ predominantly in five conditions; the reverse error pattern predominated in 11 conditions. (In one condition, there was no difference.) Two hundred eight and 248 total errors of the two kinds occurred, respectively. In the case of /t/–/s/ confusions, "strengthening" errors (/t/ reported for /s/) predominated in 11 conditions and weakening in 4. One hundred eight strengthening errors and 76 weakening errors occurred.

B. Infants, quail, bugles, and automobiles

Ohala (1996) acknowledges that language learning infants must recover gestures from acoustic speech signals. However, his experience in training phonetics students suggests to him that gesture recovery is hard (and, therefore, perhaps, that people only recover gestures when they have to). He infers, in claiming that listeners routinely recover gestures, that motor theorists and direct realists are exhibiting "projection" in ascribing their own perceptual peculiarities to the rest of the world.

In rebuttal, I point out that gesture recovery is not hard and that evidence from phonetic students is not relevant. Infants perceive gestures well before they produce speech—indeed, before they babble. Kuhl and Meltzoff's 1982, 1984 and 1988) youngest subjects (4-month-old infants) selectively watched the one of two TV screens that showed a speaker producing the acoustic speech signal the infants were hearing. As for phonetic students, their inability to reproduce heard gestures does not imply that that they did not perceive gestures (any more than the typical person's inability to perform a triple axel implies that he or she cannot see them). Nor, however, does students' inability to spell what they hear mean that they are not hearing gestures. Perceptual learning does occur, and, over learning, perceivers get better and better at information pickup (E. J. Gibson, 1969).

Ohala doubts that chinchillas, budgies, quail, and other animals can be supposed to perceive actions of the human vocal tract from acoustic speech signals. I do not share his doubt. The perceptual systems of these other animals, like ours, were shaped by natural selection to recover real-world causes of structure in media to which they are sensitive. Their survival, like ours, has always depended on their knowing about the world in which they live; perceiving sounding events specified by acoustic signals promotes survival more than does hearing structured air itself.

Take the quail and first consider what it *sees*. Imagine turning a quail away from its response key and toward its human harasser. Will the quail not see a human face? Intuition suggests a positive answer, but how is such a percept possible? It is possible, because human faces causally structure light in ways distinctive to themselves and the quail's visual system evolved to recover real-world causes of specifying structure. Notice, too, that the quail must see the cage it is in and the response key it must peck even though these are human artifacts that evolution cannot have anticipated their seeing. Perceiving the world based on structure in media is an evolved function, but it cannot depend on built-in mappings of structures to events.

The reason why quail can perceive gestures, then, is that gestures of the human vocal tract causally structure the air in ways distinctive to themselves, and consequently gestures of the vocal tract are what the information in the acoustic speech signal is about. Evolution shaped the quail's auditory system to recover sounding events from the specifying structure they cause.

As for buglers, Ohala (1996) suggests that he is "unable to recover exactly what the bugler does to produce...notes." I suspect that he is right, but that does not refute a theory of direct perception? The theory of direct perception is not a theory about magic. Listeners can, at most, hear those properties of sounding events that causally structure the air in ways specific to themselves. Unless the notes of a bugle specify their causal sources exactly, listeners cannot hear them exactly.

There are harder questions to pose about music, however. Does our aesthetic appreciation for music have anything to do with recovering sounding events? Perhaps not. I believe that we do recover those properties of sounding events that are well specified in musical acoustic signals. However, that may not be what we appreciate about music. My own view (which, admittedly, is not well developed on this topic) is that informational media vary in "transparency." Reflected light is wholly transparent, in the sense that, no matter how hard we try, we cannot experience it under ordinary conditions of seeing. I think that we can feel skin deformations, however, and perhaps we can experience acoustic signals. What we cannot do is short-circuit using the structured media to recover real-world events to the extent that the media specify them. That is, as we explore a hand-held object haptically, even if we concentrate on the pressure it exerts on the skin, we cannot fail to feel the object in our hand.

As for automobiles, Ohala (1996) suggests that "recovery of the mechanisms of the sound source is unlikely or impossible." Again, my theory does not propose that we perceive everything about car engines from the acoustic signals that car engines cause. Here, perhaps, is what we can perceive:

> "In the case of the automobile, some proportion of the energy produced by burning gasoline causes vibrations in the material of the car itself.... Things tap, scrape, slosh, rub, roll and flutter. Because each of these events is determined by the physical attributes of its source, the entire pattern of the car's vibrations is meaningfully structured by its components. These mechanical vibrations, in turn, produce waves of alternating high and low pressure on the air surrounding the car.... These spreading pressure waves, then, may serve as information about the vibrations that cause them and thus about the event itself" (Gaver, 1993, p. 7).

V. ANALOGIES AND EXAMPLES

In two places in his manuscript, Ohala (1996) suggests that I use analogies with vision and haptics to support my claim that we perceive gestures. In his view, had I only focused on olfaction, instead, my conclusions about speech would be different. On the contrary, however, my claim is not based on analogies with vision and haptics, and my only reason for not bringing olfaction into the story was that human intuitions about smell are poor.

My claim that we perceive gestures rests on a more fundamental claim that perceptual systems have a universal function, one of acquainting perceivers with their ecological niche. I use vision and haptics, not as analogies, but as examples that illustrate the universal function. Our intuitions are clearer in those domains than they are about audition (or smell). Smell also serves that universal function (e.g., J. J. Gibson, 1966), albeit less effectively for humans than for bloodhounds.

VI. EXTRAVAGANCE IN THEORIZING: THE CHICKEN LITTLE CRITERION

Ohala's (1996) Chicken Little criterion is that less extravagant theories are to be preferred over more extravagant ones. Ohala argues that the motor theory and the theory of direct perception are extravagant. Here is why they are extravagant in his view:

> "[L]et us ask, why do the advocates of MT and DR think that listeners recover speech articulations? One of the motivations is the lack of obvious invariance between the assumed linguistic units of speech and their acoustic manifestation. Invariance, they claim, is to be found in the speech articulations (or, if not there, then further 'upstream'; see Lisker *et al.*, 1962). The reasoning here is of the sort 'I can't find it here, so it must be over there.' This is not a particularly compelling motivation... . MT and DR are like the theory that the sky is falling, extravagant hypotheses that leave unexplored a host of other less costly explanations" (Ohala, 1996).

Excerpts such as this justify my opinion that speech scientists do far more writing than they do reading. As noted earlier, the motor theory was developed when scientists found *positive* evidence for a close correspondence between percepts and articulation in stimuli in which they judged there to be a lack of correspondence between the percepts and the acoustic signal. It is fine to dispute these data if they appear disputable, but it is not accurate or fair to the theory's developers to write as if the data do not exist. Nor is it right to concoct imaginary, uncompelling motivations for the theory when the real motivations are readily available in the best known of the motor theory papers. As for the direct realist theory, the reasoning Ohala ascribed to it could hardly be accurate. The theory could not propose that perception is direct if it subscribed to a view that specifying information is absent from the acoustic signal. But the *direct* realist theory does propose that speech perception is *direct* (Fowler, 1994a).

Are the theories extravagant? In my view, the motor theory has one extravagant feature. It was never necessary to propose that the perception of gestures requires recruitment of the listener's own speech motor system. Perception of gestures occurs because the acoustic speech signal serves listeners as information for its source. For its part, the theory of direct perception is the least extravagant theory out there. It is one theory that takes seriously the idea that speech perception is wholly unspecial. It is just like perceiving everything else. What could be less extravagant? Other theories to which Ohala alludes and which he considers less costly are, in fact, much more costly. If animals were to subscribe to the theories of auditory perception that these theories imply, it would cost them their lives.

ACKNOWLEDGMENTS

Preparation of this manuscript was supported by NICHD Grant No. HD-01994 to Haskins Laboratories. I thank James Jenkins and Michael Studdert-Kennedy for their comments on the first draft.

[1] Notice, by the way, that a claim that we perceive phonological gestures is not a claim that we do not also, for example, perceive the words that gestures compose. Perceivables are constrained to be things that causally structure a medium to which some perceptual system is sensitive. But, at least in Gibson's theory, that encompasses a lot, including such a thing as the *function* (or "affordance") of mailboxes (J. J. Gibson, 1979, p. 139; see also pp. 253–255) alluded to by Beckman (1994) in her comments on Ohala's presentation. That is, the events to which we are sensitive are ecologically specific and can span very long time periods in Gibson's theory.

[2] I thank James Jenkins for referring me to this article.

[3] Without masking, we obtained a very small, marginally significant, McGurk-like effect with print (see also Massaro *et al.*, 1988). Even here, where the print had an advantage in identifiability over felt syllables, the effect was considerably (and significantly) smaller than we had obtained in the haptic condition.

[4] The term "cues" is used here, as by Fitch *et al.*, to refer to acoustic fragments that, independently manipulated by experimenters, each affect the listener's identification of a phonological segment. The term is not meant to imply that the fragments serve as distinct cues for the listener.

American Heritage Dictionary (**1982**). (Houghton Mifflin, Boston).

Austin, J. (**1962**). *Sense and Sensibilia* (Oxford U.P., New York).

Beckman, M. (**1994**). "Critique," paper presented at the 127th Meeting of the Acoustical Society, 7 June 1994, Cambridge, MA [J. Acoust. Soc. Am. **95**, 2849 (1994)].

Blumstein, S., Isaacs, E., and Mertus, J. (**1982**). "The role of gross spectral shape as a perceptual cue to place of articulation," J. Acoust. Soc. Am. **72**, 43–50.

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

Brancazio, L., and Mitra, J. (**1994**). "A reconsideration of locus equations as invariants for place of articulation of stop consonants," J. Acoust. Soc. Am. **95**, 2934(A).

Browman, C., and Goldstein, L. (**1986**). "Towards an articulatory phonology," Phonol. Yearb. **3**, 219–252.

Browman, C., and Goldstein, L. (**1992**). "Articulatory phonology: An overview," Phonetica **49**, 155–180.

Clements, G. N. (**1985**). "The geometry of phonological features," Phonol. Yearb. **2**, 225–252.

Dekle, D., Fowler, C. A., and Funnell, M. (**1992**). "Audio-visual integration in perception of real words," Percept. Psychophys. **51**, 355–362.

Diehl, R., and Kluender, K. (**1989**). "On the objects of speech perception," Ecol. Psychol. **1**, 121–144.

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

Fowler, C. (**1981**). "Production and perception of coarticulation among stressed and unstressed vowels," J. Speech Hear. Res. **46**, 127–139.

Fowler, C. (**1984**). "Segmentation of coarticulated speech in perception," Percept. Psychophys. **36**, 359–368.

Fowler, C. (**1986**). "An event approach to the study of speech perception from a direct-realist perspective," J. Phon. **14**, 3–28.

Fowler, C. (**1987**). "Listeners as realists, talkers too: Commentary on the papers by Strange, Diehl *et al.*, and Rakerd and Verbrugge," J. Mem. Language **26**, 574–587.

Fowler, C. (**1993**). "Phonological and articulatory characteristics of spoken language," in *Linguistic Disorders and Pathologies: An International Handbook*, edited by G. Blanken, J. Dittman, H. Grimm, J. Marshall, and C.-W. Wallesch (de Gruyter, Berlin), pp. 34–46.

Fowler, C. (**1994a**). "Speech perception: Direct realist theory," *Encyclopedia of Language and Linguistics* (Pergamon, Oxford), Vol. 8, pp. 4199–4203.

Fowler, C. (**1994b**). "Invariants, specifiers cues: An investigation of locus equations as information for place of articulation," Percept. Psychophys. **55**, 597–610.

Fowler, C. (**in press**). "Speaking," in *Handbook of Motor Skills*, edited by H. Heuer and S. Keele (Verlag fuer Psychologie Dr. C. J. Hogrefe, Goettingen).

Fowler, C. A., and Dekle, D. J. (**1991**). "Listening with eye and hand: Crossmodal contributions to speech perception," J. Exp. Psychol. Hum. Percept. Performance **17**, 816–828.

Fowler, C. A., and Saltzman, E. (**1993**). "Coordination and coarticulation in speech production," Language Speech **36**, 171–195.

Garner, W. R., Hake, H. W., and Eriksen, C. W. (**1956**). "Operationism and the concept of perception," Psychol. Rev. **63**, 149–161.

Gaver, W. (**1993**). "What in the world do we hear?: An ecological approach to auditory event perception," Ecol. Psychol. **5**, 1–29.

Gelfer, C. (**1987**). "A simultaneous physiological and acoustic study of fundamental frequency declination," Ph.D. dissertation, City University of New York.

Gibson, E. J. (**1969**). *Principles of Perceptual Learning and Development* (Appleton–Century–Crofts, New York).

Gibson, J. J. (**1966**). *The Senses Considered as Perceptual Systems* (Houghton Mifflin, Boston).

Gibson, J. J. (**1979**). *The Ecological Approach to Visual Perception* (Houghton Mifflin, Boston).

Hombert, J. (**1978**). "Consonant types, vowel quality and tone," in *Tone: A Linguistic Survey*, edited by V. Fromkin (Academy, New York), pp. 77–112.

Hyman, L. (**1973**). "The feature [Grave] in phonological theory," J. Phon. **1**, 329–337.

Jakobson, R., Fant, G., and Halle, M. (**1963**). *Preliminaries to Speech Analysis* (MIT, Cambridge, MA).

Kelso, J. A. S., Saltzman, E., and Tuller, B. (**1986**). "The dynamical perspective on speech production: Data and theory," J. Phon. **14**, 29–59.

Kenstowicz, M., and Kisseberth, C. (**1979**). *Generative Phonology* (Academic, New York).

Kluender, K. (**1994**). "Speech perception as a tractable problem in cognitive science," in *Handbook of Psycholinguistics*, edited by M. A. Gernsbacher (Academic, San Diego), pp. 173–214.

Kozhevnikov, V., and Chistovich, L. (**1965**). *Speech: Articulation and Perception* (Joint Publications Research Service, Washington, DC),

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

Kuhl, P., and Meltzoff, A. (**1984**). "The intermodal representation of speech in infants," Infant Behav. Dev. **7**, 361–381.

Kuhl, P., and Meltzoff, A. (**1988**). "Speech as an intermodal object of perception," in *Perceptual Development in Infancy: The Minnesota Symposia on Child Psychology*, edited by A. Yonas (Erlbaum, Hillsdale, NJ), pp. 235–266.

Liberman, A., and Mattingly, I. (**1985**). "The motor theory revised," Cognition **21**, 1–36.

Liberman, A., and Mattingly, I. (**1989**). "A specialization for speech perception," Science **243**, 489–494.

Liberman, A., Delattre, P., and Cooper, F. (**1952**). "The role of selected stimulus variables in the perception of the unvoiced-stop consonants," Am. J. Psychol. **65**, 497–516.

Liberman, A., Cooper, F., Shankweiler, D., and Studdert-Kennedy, M. (**1967**). "Perception of the speech code," Psychol. Rev. **74**, 431–461.

Lindblom, B. (**1989**). "Some remarks on the origin of the phonetic code," in

Brain and Reading: Proceedings of the Seventh International Rodin Remediation Conference, edited by C. van Euler, J. Linberg, and G. Lennerstrand (Macmillan, London), pp. 27–44.

Lisker, L., Cooper, F. S., and Liberman, A. (**1962**). "The uses of experiments in language description," Word **18**, 82–106.

Luce, R. D. (**1986**). *Response Times* (Oxford U.P., Oxford).

MacDonald, J., and McGurk, H. (**1978**). "Visual influences on speech perception," Percept. Psychophys. **24**, 253–257.

Martin, J., and Bunnell, H. T. (**1981**). "Perception of anticipatory coarticulation effects," J. Acoust. Soc. Am. **69**, 559–567.

Massaro, D. (**1987**). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry* (Erlbaum, Hillsdale, NJ).

Massaro, D., Cohen, M., and Thompson, L. (**1988**). "Visible language in speech perception," Visible Language **22**, 8–31.

McGurk, H., and MacDonald, J. (**1976**). "Hearing lips and seeing voices," Nature **264**, 746–748.

Miller, G., and Nicely, P. (**1955**). "An analysis of perceptual confusions among some English consonant," J. Acoust. Soc. Am. **27**, 338–352.

Mowrey, R., and Pagliuca, W. (**1987**). *The Reductive Character of Phonetic Evolution*, unpublished manuscript, State University of New York at Buffalo.

Ohala, J. (**1994**). "Speech perception is perceiving sounds not tongues," paper presented at the 127th Meeting of the Acoustical Society, 7 June 1994, Cambridge, MA [J. Acoust. Soc. Am. **95**, 2849(A)].

Ohala, J. J. (**1996**). "Speech perception is hearing sounds, not tongues," J. Acoust. Soc. Am. **99**, 1718–1725.

Pagliuca, W. (**1982**). "Prolegomena to a theory of articulatory evolution", Ph.D. dissertation, State University of New York at Buffalo.

Pierrehumbert, J. (**1979**). "The perception of fundamental frequency," J. Acoust. Soc. Am. **66**, 363–369.

Porter, R. (**1978**). "Rapid shadowing of syllables: Evidence for symmetry of speech perceptual and motor systems," paper presented at the Psychonomic Society Meeting, November 1978, San Antonio.

Porter, R., and Castellanos, F. X. (**1980**). "Speech production measures of speech perception: Rapid shadowing of VCV syllables," J. Acoust. Soc. Am. **67**, 1349–1356.

Porter, R., and Lubker, J. (**1980**). "Rapid reproduction of vowel–vowel sequences: Evidence for a fast and direct acoustic-motoric linkage in speech," J. Speech Hear. Res. **23**, 593–602.

Reed, E., and Jones, R. (**1982**). *Reasons for Realism: Selected Essays of James J. Gibson* (Erlbaum, Hillsdale, NJ).

Reinholt-Peterson, N. (**1986**). "Perceptual compensation for segmentally-conditioned fundamental-frequency perturbations," Phonetica **43**, 31–42.

Rocca, I. (**1994**). *Generative Phonology* (Routledge, London).

Sapir, S. (**1989**). "The intrinsic pitch of vowels: Theoretical, physiological and clinical observations," J. Voice **3**, 44–51.

Seidenberg, M., and Tanenhaus, M. (**1979**). "Orthographic effects on rhyme monitoring," J. Exp. Psychol. Hum. Learning Mem. **5**, 546–554.

Silverman, K. (**1986**). "F_o cues depend on intonation: The case of the rise after voiced stops," Phonetica **43**, 76–92.

Silverman, K. (**1987**). "The structure and processing of fundamental frequency contours," Ph.D. dissertation, Cambridge University.

Stetson, R. (**1951**). *Motor Phonetics* (North-Holland, Amsterdam).

Stevens, K., and Blumstein, S. (**1981**). "The search for invariant correlates of phonetic features," in *Perspectives of the Study of Speech*, edited by P. Eimas and J. Miller (Erlbaum, Hillsdale, NJ), pp. 1–38.

Summerfield, A. Q. (**1987**). "Some preliminaries to comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lip Reading*, edited by B. Dodd and R. Campbell (Erlbaum, Hillsdale, NJ), pp. 3–51.

Sussman, H. (**1989**). "Neural coding of relational invariance in speech: Human language analogs to the barn owl," Psychol. Rev. **96**, 631–642.

Sussman, H., McCaffrey, H., and Matthews, S. (**1991**). "An investigation of locus equations as a source of relational invariance for stop place categorization," J. Acoust. Soc. Am. **90**, 1309–1325.

Tanenhaus, M., Flanigan, H., and Seidenberg, M. (**1980**). "Orthographic and phonological activation in auditory and visual word recognition," Mem. Cognit. **8**, 513–520.

Tuller, B., and Kelso, J. A. S. (**1990**). "Phase transitions in speech production and their perceptual consequences," in *Attention and Performance XIII*, edited by M. Jeannerod (Erlbaum, Hillsdale, NJ), pp. 429–452.

Walley, A., and Carrell, T. (**1983**). "Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants," J. Acoust. Soc. Am. **73**, 1011–1022.

Webster's Ninth Collegiate Dictionary (**1990**). (Merriam Webster, Springfield, MA).

Whalen, D. (**1984**). "Subcategorical mismatches slow phonetic judgments," Percept. Psychophys. **35**, 49–64.

Whalen, D. (**1989**). "Vowel and consonant judgments are not independent when cued by the same information," Percept. Psychopys. **46**, 284–292.