Speech perception is hearing sounds, not tongues

John J. Ohala
Department of Linguistics, University of California, Berkeley, California 94720

(Received 17 September 1994; revised 28 December 1994; accepted for publication 9 January 1995)

Three types of evidence are reviewed which cast doubt on claims that recovery of the speaker's articulations is an inherent part of speech perception: (a) Phonological data (e.g., universal tendencies of languages' segment inventories, phonotactic patterns, sound changes, etc.) show unmistakably that the acoustic-auditory properties of speech sounds, not their articulations, are the primary determinant of their behavior. (b) Infants and various nonhuman species can differentiate certain sound contrasts in human speech even though it is highly unlikely that they can deduce the vocal tract movements generating the sounds. (c) Humans can differentiate many nonspeech sounds almost as complex as speech, e.g., music, machine noises, as well as bird and monkey vocalizations, where there is little or no possibility of recovering the mechanisms producing the sounds.

PACS numbers: 43.71.An

INTRODUCTION

Does speech perception by humans necessarily include their retrieving the articulatory activity which produced the heard speech? The answer to this question may have application to the task of speech recognition by machines if one believes that a good initial strategy in designing machines which mimic the behavior of humans (or other sentient creatures) is to attempt first to understand how that behavior is accomplished in nature.[1] I will argue that speech perception does *not* involve recovering the articulations that generate the acoustic signal. In opposition to this view there are, among others, the "motor theory of speech perception" (MT) (Liberman and Mattingly, 1985) and the "direct realist" theory of speech perception (DR) (Fowler, 1986).[2] According to MT, part of the auditory decoding of the speech signal includes arriving at the equivalent articulations which gave rise to the signal; speech is "recognized" by comparing this deduced articulation to a stored articulatory template. According to DR, speech articulations are perceived directly; the acoustic signal is only the medium of transmission of these gestures. There is no "encoding" of the elements of pronunciation; the directly perceived gestures are the linguistically meaningful aspects of the message.

Both MT and DR have been embedded by their proponents into a philosophical framework informed by biological and ethological considerations (Liberman and Mattingly, 1985; Fowler 1986, 1990). For Liberman (1993) this involves recognition that

> ... speech is to the human being as echolocation is to the bat or song is to the bird ... all these behaviors depend on biologically coherent faculties that were specifically adapted in evolution to function appropriately in connection with events that are of particular ecological significance to the animal.
> ...[in the case of speech] what evolved ... was the phonetic module ... The primitives of the module are gestures of the articulatory organs. These are the ultimate constituents of language.

Similarly, Fowler states that DR was inspired by J. J. Gibson's theory of perception in which

> ... perceptual systems constitute the only means by which organisms can know the environment in which they act. All perceptual systems ... take advantage of the fact that certain media—most notably, light, air, and the skin and joints of [organisms'] bodies—are lawfully, and largely distinctively, structured by the objects and events in the environment. The structure in these media in turn stimulates the sense organs ... imparting its structure to them. ... [The lawful relation between object-medium-sense organ means that the media's] structuring can serve as information for the objects and events themselves, and it is those objects and events, not the structured light, air, or skin and joints, that perceivers need to know about in order to guide their actions in the world. ...
> According to [DR], speech perception ... depends largely on the auditory system that evolved to recover acoustic-signal-producing events in the world.

It is certainly reasonable that speech perception and, indeed, all linguistic behavior must be constrained by larger biological considerations, i.e., factors applicable to any organisms' species-specific (and cross-species) behavior. (For speculations on how this might apply to a restricted range of linguistic behavior, see Ohala, 1984.) Perhaps speech production and speech perception *have* evolved in special ways and certainly the human capacity to link large numbers of symbols and meanings as well as to differentiate between different permutations of symbols are plausible candidates for this. But accepting this biological viewpoint on speech does not automatically require one to accept the specifics of MT and DR. There is no a priori biological or ethological reason why speech units couldn't be basically acoustic-auditory.

Much has been written about the Gibsonian theory of perception, and reasonable people still disagree on its appli-

 0001-4966/96/99(3)/1718/8/$6.00

cability to specific perceptual tasks or to specific sense modalities. It is not possible for me review this literature here. I will limit myself to one remark (which, for all I know, has been anticipated in prior literature). Fowler attempts to buttress her arguments about the perception of speech being direct by drawing analogies to visual and haptic perception where, she argues, we have the impression that it is the objects seen and felt that are perceived, not the reflected light or the pattern of deformation of the skin. But she does not draw analogies between hearing and another sense modality: *smell*. In this case it would be stretching a point to say that our impression is that we are perceiving the object emitting the smell as opposed to the smell itself. We can and often do detect smells without even being aware of the source of the smell—indeed, the object which emitted the smell may be absent from the immediate environment. I am not necessarily claiming that sounds are more like smells than sights or touch sensations (although a number of instructive parallels could be drawn) but simply that analogies are unsure guides. It is not a given that all sense modalities have a direct link with the objects or events which give rise to the structured medium. If commonsense analogies are to be drawn between speech perception and perception with other sense modalities, let us at least consider the full range of sense modalities.

There are unfortunately no definitive experiments or "existence proofs" (machines that successfully recognize speech well enough, in comparison to humans, that anyone would dare to hold them up as models of how humans perceive speech). The advocates of the above-mentioned theories have not so far produced an algorithm for deriving the articulations which produce any given speech signal. The arguments offered to support these theories rest purely on plausibility, part of which is based on "what-else-could-it-be?" reasoning. My counterarguments will also be based on plausibility and commonsense. To begin, I will first review some aspects of the phonological structure and behavior of speech which reveal, I claim, the primacy of acoustic-auditory factors in its design and use. After discussing the implications of this point I will then briefly allude to a body of experiments and anecdotes that pose serious difficulties for the opposing viewpoint. Finally, I will approach the question using the "Chicken Little" strategy, i.e., the epistemological tactic.

Although I will cast doubt on the proposition that human listeners recover the articulations underlying speech, it obviously is important for the *speech scientist* to attempt to work out the bidirectional mapping between articulations and sounds. This is part of the task of trying to understand how humans communicate via speech and will inevitably help us to solve the big questions in this area, both theoretical and applied.

I. EVIDENCE FROM PHONOLOGY

A desirable feature of any signaling system is the physical differentiability of the signals or ciphers it employs. Systems designed by humans, from Morse Code, semaphore, and the Grey Code, all attempt to adhere to this principle. Natural codes, such as the visual and auditory signals animals use to convey threat and nonthreat also show a maximization of the physical difference between such contrasting messages (Morton, 1977; Ohala, 1984). If speech consists of conveyed articulatory or gestural events, we should expect languages' speech sound inventories to achieve some measure of differentiability between articulations; if, on the other hand, it is *sounds* that are being conveyed, then we should find that this is the domain where differentiation occurs.

TABLE I. Hindi segment inventory (M. Ohala, 1983).

p			ʈ			ʈ		ʧ		k				
b			ɖ			ɖ		ʤ		g				
pʰ			ʈʰ			ʈʰ		ʧʰ		kʰ				
b̤			ɖ̤			ɖ̤		ʤ̤		g̤				
	f			s				ʃ				h		
				z										
m			n											
w								j						
			ɾ			ɽ								
						ɽ̤								
			l											
								i	ĩ		u	ũ		
								ɪ	ɪ̃		ʊ	ʊ̃		
								e	ẽ		o	õ		
								ɛ	ɛ̃		ɔ	ɔ̃		
										ə	ə̃			
								æ		ɑ	ɑ̃			

A. The disproportionate incidence of obstruents

When we look at the segment inventories of languages of the world, one thing we notice immediately is that consonants usually outnumber vowels and among the consonants, obstruents (stops, affricates, fricatives), usually outnumber sonorants (nasals, glides, laterals). (See also M. Ohala, 1983, p. 194.) Hindi, whose phoneme inventory is given in Table I, exemplifies this. Out of a total phoneme inventory of 54 phonemes (not counting geminate or long consonants), there are 30 obstruents (including 25 stops or affricates and 5 fricatives), 2 nasals, and 6 sonorant consonants. There is a disproportionate incidence of obstruents (56%) when one considers that the pool from which the sounds are drawn can include nasals, glides, other sonorant consonants, and vowels. Why should this be? What is "good" about obstruents?

The answer is that speech sound inventories exploit the nonlinear relation between articulation and the sound produced in order to make the elements of speech *sound* different (cf. Ohala, 1983b; Stevens, 1989). Imagine a speech "valve", e.g., tongue apex, tongue dorsum, as the sole regulator of the sound and airflow emerging from the vocal tract: in effecting a change from a large valvular opening (A_{max}) to a state of complete closure (A_{min}), it creates a disproportionate effect on the resulting sound the closer it comes to complete closure. Deviation of the vocal tract resonances from those of a uniform tube is greater at small apertures than

large apertures for an equivalent Δ area change. At some point, as A approaches A_{min}, turbulence and thus audible frication is created. Then, when A_{min} is achieved there is an abrupt attenuation of the amplitude of the acoustic energy output—whether there is an active sound generator behind the closure or not. But this is not the end of the story: after maintaining A_{min} for a while ($\geq$50 ms) the closure is released and since air pressure has risen during the closure, a "pop" occurs—the acoustically salient stop burst. Given the full range of valvular closures of the articulators, speech seems to exploit disproportionately a very narrow region near A_{min}.

This pattern of disproportionate use of articulations that make "pops" and "hisses" makes sense only if the acoustic-auditory properties of the sounds are more important than their articulation. In fact, as far as many groups of these obstruents are concerned, they have a very similar articulation that, one would suppose, would be mutually confusable if it was their articulatory shape which listeners have to recover to differentiate them. This latter point might be countered by positing that listeners' ability to resolve different articulations was nonlinear, having greater resolution at close than at more open constrictions. But one may legitimately question the wisdom of attributing to listeners such sophisticated sensory abilities when it has not even been established that they can recover articulations at all.

B. Absence of /p/

One occasionally finds revealing asymmetries in the obstruent inventories of languages. Among voiceless stops, /p/ is most often missing even among languages that have /b/, e.g., in Arabic, Tuareg, Somali, Hausa, Vietnamese, among others (Maddieson, 1984). The apparent reason for this is the following: One of the major perceptual cues for a stop is an abrupt rise in amplitude (Mack and Blumstein, 1983). In the case of voiced stops the sharp amplitude rise at the release of the stop is contributed by a burst and the simultaneous onset of the vowel which also has a relatively steep rise in amplitude. In the case of voiceless stops, at least those with some degree of voice onset lag, the abrupt amplitude rise would be given primarily by the stop burst alone; the subsequent amplitude rise at voice onset is usually much slower and thus not as reliable a cue for a stop. In the case of labial stops, [b] and [p], the amplitude of the burst may not be very great because these stops lack any resonating cavity downstream of their point of release. With a weak burst a [b] would still have the abrupt amplitude rise for the vowel as a cue for a stop but the voiceless and even slightly aspirated [p] would lose an important cue for a stop. Thus of the various voiceless stops, [p] is most likely to be confused with a nonstop (e.g., a voiceless bilabial fricative [ɸ]) and therefore be subject to a sound change the result of which would be the disappearance of this stop from a language's segment inventory.

If the preceding suggests that not all stops are created equal, there is also evidence that not all fricatives are equal either. Figure 1 is a bar-chart representation of incidence of different fricative types among the 317 languages surveyed by Maddieson (1984). The so-called sibilant fricatives [s, z, ʃ, ʒ, ʂ, ʐ] outnumber by far the nonsibilant fricatives (e.g., [f, v, θ, ð, ç, x, ɣ], etc. Why should this be? As Shadle (1990) has demonstrated, what makes sibilant fricatives special is the fact that the jet of air escaping from the tongue-palate constriction strikes the upper incisors; this creates added turbulence which makes the high-frequency noise of these fricatives relatively intense. Fricatives made at other places of articulation lack such an ideally placed barrier to their air jet.

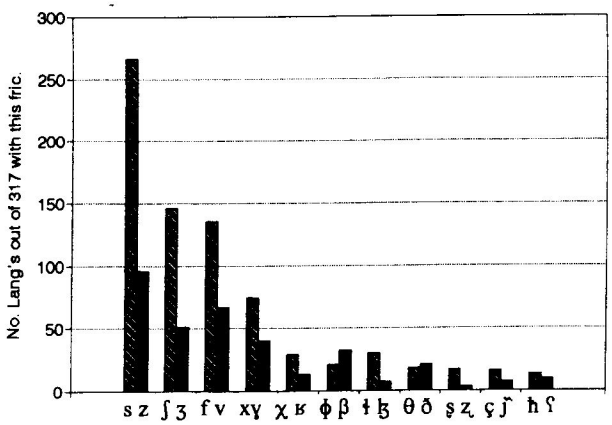

FIG. 1. Incidence of fricatives by type (from Maddieson, 1984).

C. Salient consonantal releases influence universal tendencies of syllable structure

The highly salient releases of stops and other consonants (vis-à-vis these segments' onsets) play an important role in languages' phonologies: the favored syllable structure is CV rather than V or VC, there are more phonological contrasts made in syllable-initial position than in syllable final position, and the CV junction may even be the point of synchronization for prosodic contours, etc. (Erikson and Almstermark, 1972; Erikson, 1973; Ohala, 1993).

The weak status of [p] (in comparison to other voiceless stops), the disproportionate representation of sibilant fricatives, and the asymmetrical treatment of CV and VC syllables would just be unexplained if it were articulatory configurations and gestures which are the ultimate "coin" underlying the exchange between speaker and hearer. On the other hand, they are compatible with the view that it is distinctive sounds which are aimed at in speech.

D. Jakobson's distinctive features [grave] and [flat]

Given that this special session is being held in Cambridge, MA, we should not neglect the relevance to this issue of some of the insights of the Jakobsonian distinctive feature (d.f.) system which was developed here some 42 years ago.[3] The Jakobsonian d.f.'s were defined primarily in acoustic-auditory terms because, as Jakobson noted, in the speech chain between speaker and hearer, it is the representation of the signal at the hearer's end that is the intended function of speaking. Or as he put it "We speak in order to be heard in order to be understood." Two d.f.'s are of particular interest: [±grave] and [±flat]. [+grave] sounds were those having energy predominantly in the lower end of the spectrum. They

include both labials and back velars (including uvulars), but not apical and palatal sounds—in other words, a *discontinuous* set from an articulatory point of view. But this was a precious insight: [+grave] sounds do form a natural class. In English, the diphthong [au] does not appear before [+grave] consonants in the same syllable, i.e., we find [aus], [aud], [aul], etc. as in house, loud, howl, down, and so on, but there are no words ending in [aup], [auf], [auk], etc. The Proto-Indo-European [+grave] cluster [kw] is manifested in the daughter dialects as another [+grave] sound, /p/, e.g., PIE *ekwos yields either Latin *equus* or Greek *hippos*. (For further examples, see Ohala and Lorentz, 1977; Ohala, 1993a.)

The feature [grave] is also relevant in explaining the incidence of sounds with double articulations, i.e., more or less equal simultaneous closures at two place of articulation. [w] is such a sound because it has a constriction in both the labial and velar regions. If it was articulatory distinctness that mattered, one might expect that double articulations would be made more or less randomly at any two places, e.g., labial and palatal, dental and pharyngeal, labial, and alveolar, etc. All of these are physiologically possible. This is not what we find, though. The overwhelming favorite double articulation is labial and velar: [w, ʍ, kp, gb, ŋm]. Could this have anything to do with the fact that these two places of articulation coincide with antinodes of the second resonance such that simultaneous constrictions there can effect a maximum—and thus acoustically distinct—lowering of that resonance, i.e., a maximization of [+grave]? I suggest that it does (Ohala, 1979).

There are many other examples of sound patterns that are rendered explicable by the Jakobsonian feature [±grave] (Hyman, 1973).

The feature [flat], a term borrowed from music, is defined as a sound whose higher frequencies are somewhat lower (vis-à-vis another sound which is otherwise similar except for lack of this lowering). It applies to sounds which are retroflex, as well as those with the secondary articulations of labialization, velarization (and uvularization), or pharyngealization. This is again an articulatorily discontinuous set since it does not include apicalization or palatalization. Jakobson's claim is that since all these articulatorily diverse sounds have a similar acoustic effect, one should not find any language using more than one of these manifestations of [+flat]. That is, no language should have both retroflexes and distinctively labialized sounds or both pharyngeals and labialized sounds, etc. Although it turns out that this claim is not absolutely true (the Caucasian language Abkhaz has a contrast between labialized and nonlabialized uvulars and pharyngeals), it is certainly statistically true to a high degree (see Ohala, 1985a for references).

E. Stevens' quantal theory and "ideal" segments

Another Cambridge contribution to this issue is Ken Stevens' Quantal Theory of Speech (Stevens, 1989). Although it is controversial in many of its details, some of the predictions it offers on languages' sound structure correlate well with the observed facts. To simplify, Stevens maintains that one of the characteristics of a "good" speech sound is its capacity to be detected within a short time window (approx. 40–50 ms) since the auditory system is most sensitive to changes at this rate (Stevens, 1980). Such sounds consist of a rapid modulation of various parameters in the acoustic signal, e.g., amplitude, periodicity, spectral shape. (See also Ohala, 1992a,b.) Contrasts exploiting such features as [±continuant], [±nasal], e.g., [p] versus [m] versus [f], are examples of such auditorily robust speech sounds. Not present among the list of robust sounds are those with secondary articulations such as labialization, palatalization, glottalization, and pharyngealization. These sounds are not robust because (a) they take a long time (up to 100 ms) to manifest their distinctive characteristics and (b) their secondary articulations are superimposed on and thus mask some of the distinctive aspects of the primary articulation, especially the formant transitions cueing place. One consequence of this is that all languages use the robust features and those with small consonantal inventories use them almost exclusively; the nonrobust features like labialization, palatalization, glottalization, etc. are only exploited by languages with large consonantal inventories, i.e., after they have used the robust ones.

This pattern follows if it is the acoustic-auditory aspect of speech that matters; both members of such pairs as /p/ and /p^j/, /s/, and /s^w/, etc., should be equal if articulatory gestures were primary.

TABLE II. Vowel inventory of Yakut.

i	y		ɯ	u
ɛ	œ			ɔ
		a		

F. Vowel inventories

Lindbolm (Liljencrants and Lindblom, 1972; Lindblom, 1986) has shown that at least for languages with few vowels (<c. 9) the vowels tend to distribute themselves more or less equidistantly in the acoustic-auditory vowel space. Thus given a five vowel system, [i e a o u] tend to be preferred rather than [y œ, ʌ, ɤ, ɯ] even though both sets would be approximately the same in articulatory distance. The greater acoustic contrast between, [i] and [u] can be illustrated with data from Hungarian. Magdics (1969) gives the following means from male speakers for formant 2 (the formant widely regarded as conveying the characteristic quality of a vowel) as [i:], 2300 Hz, [u:], 730, [y:], 1890. This gives $\Delta F2$ for [i:]–[u:] of 1570 Hz, but only 1160 for [y:]–[u:] and 410 for [i:]–[y:]. There would similarly be lower $\Delta F2$'s and hence less acoustic distinctiveness between any pair from the nonpreferred set and any comparable pair from the preferred set.

To be sure, some languages do have inventories which include something close to the nonpreferred set but these languages also have the more favored vowels, too, e.g., Yakut, whose vowel inventory is given in Table II. My point here is not that the vowels from the less preferred set cannot be easily detected or identified by listeners; the Yakut inventory disproves that. Rather, when vowel inventories develop

distinctiveness, it is their acoustic-auditory properties that matter, not their articulatory configuration.

G. Sound change

Sound change is the change in pronunciation from one generation to the next. It too seems to show the primacy of the acoustic-auditory aspect of speech. Considering the degrees of freedom speech has in the articulatory domain and in the acoustic-auditory domain, it is articulation which often shows more variation—some of it quite remarkable—than does the sound shape itself. I have documented this in a number of papers (Ohala, 1974, 1978, 1979, 1981, 1983a,c, 1985a,b, 1990, 1992a,c, 1993a,b,c; Ohala and Lorentz, 1977; Ohala and Ohala, 1993). I will give just a few of the more extraordinary examples here.

1. Palatalized labials are confused with apicals

It is well known in the ASR community that in English the names of the letters of the alphabet are very difficult to differentiate; in particular, "T" [t^hi] and "P" [p^hi] are highly confusable, as well as "D" and "B" although with attention to detail, they can be successful differentiated; see Fanty and Cole, 1990). It is the "carrier" vowel [_i] that creates the difficulty. What is difficult for the machine is difficult for listeners, too. The sound change of palatalized labial to apical (whether the labial is distinctively palatalized or, as in the English names for "P" and "B", phonetically palatalized) is well-attested in the historical linguistics literature. A few examples from various languages are given in (1)

(1) Examples of the sound change: palatalized labial >apical (where the form on the left exemplifies the "before" pronunciation and that on the right the "after" or "innovative" one). (For sources, see Ohala, 1978.)

a.	Standard Czech	East Bohemian	Transl.
	m^jɛstɔ	nɛstɔ	*town*
	p^jɛt	tɛt	*five*
b.	Roman Ital.	Genoese Ital.	Transl.
	pjeno	t͡ʃena	*full*
	bjaŋko	d͡ʒaŋku	*white*
c.	Proto-Bantu	Zulu	Transl.
	pia	-t͡ʃha	*new*

Reinforcing my contention that these changes arise because of the confusability of these sounds in the auditory domain is the fact that the same confusions crop up at a high rate in lab-based listening experiments, e.g., Winitz *et al.*, 1972, where there was never any opportunity for an articulatory substitution.[4]

2. Spontaneous nasalization

A rather unusual sound change is one where nasalization seems to appear in words that never had any nasal segments in it. Some examples are given in (2).

(2) Examples of spontaneous nasalization. (For sources, see Ohala and Ohala, 1993.)

a.	Sânskrit	Hindi	Translation
	sarpa	/sãp/	*snake*
	učča-	/uča/	*high*
b.	French	Breton (French loans)	Translation
	maçon	[mãson-]	*mason*
	vis	[bīs]	*screw*

It seems that one conditioning environment for this is an adjacent high-airflow segment such as a voiceless fricative, including [h], affricate, or an aspirated stop. I hypothesized that this arose from (a) the high airflow segments requiring larger-than-normal glottal opening (b) this glottal opening spreading onto the margins of the adjacent vowels (though they remained voiced), (c) this slight opening of glottis during the vowels creating acoustic conditions which mimic vowel nasalization, namely, lowered amplitude and increased bandwidth of $F1$. Then what sounds like nasalization may be reinterpreted by listeners and manifested in their pronunciations as actual (physiological) nasalization. The similarity of the acoustic consequences on vowels with nasalization and slightly open glottis has been demonstrated by Fujimura and Lindqvist (1971). Listening tests also confirm that vowel fragments excerpted from those portions of a vowel immediately adjacent to voiceless fricatives sound nasalized to listeners (Ohala, 1983a; Ohala and Ohala, 1993). (Other supporting evidence has been presented in Ohala, 1993c; Ohala *et al.*, 1992.)

3. Nasal Epenthesis Before Voiced Stops

There is evidence that in the development from Old to Modern Hindi a sequence of a *nasalized vowel+voiced stop* can change to *nasalized vowel+homorganic nasal cons +voiced stop*. But the epenthetic nasal did not appear if the stop following the nasalized vowel was voiceless. See examples in (3).

(3) Examples of Nasal Epenthesis

Old Hindi	Mod. Hindi	Transl.
ã:gana	[ãŋgən]	*courtyard*
čã:da	[ʧãnd]	*moon*
BUT:		
dã:ta	[dãt]	*tooth*

An epenthetic nasal consonant is also observable as a form of synchronic variation when a nasalized vowel and a voiced stop about across word boundaries in Modern Hindi and in French, e.g., the sequence in French ...saint bel ... reveals the cross word boundary segments [-ɛ^{m}b-]. (For additional supporting evidence, see Ohala and Ohala, 1993.) Ohala and Ohala offer the following explanation for this phenomenon:

> Among the auditory cues for a voiced stop there must be a spectral and amplitude discontinuity with respect to neighboring sonorants (if any), low amplitude voicing during its closure, and termination in a burst; these requirements are still met even with velic leakage during the first part of the stop as long as the velic valve is closed just before the release and pressure is allowed to build up behind the closure. However, voiceless stops have less tolerance for such leakage because any nasal

sound—voiced or voiceless—would undercut either their stop or their voiceless character.

This sound change, then, like the others reviewed above, seems to adhere to the constraint that articulation may vary if the auditory shape is not too greatly distorted.

H. Summary and discussion of phonological evidence

The sound patterns in language reviewed above are a few of the pieces of evidence pointing to the primacy of the acoustic-auditory aspect of speech. In those instances where a certain sound was said to be less preferred than some other sound, e.g., the set of nonsibilant fricatives as opposed to sibilant fricatives, it is not the case that these sounds cannot be detected—English, for example, employs fricatives from both sets. So if speech sounds were designed to be distinct articulatorily it would be possible for their selection and development in languages to reflect that. But this is not what we find: they are selected and they develop for their acoustic-auditory properties.

It might be maintained that speech has evolved in such a way that only those articulatory gestures have survived whose presence and distinctiveness are conveyed by the speech aerodynamics and speech acoustics. Listeners nevertheless still recover the articulations. But this would still imply that *listeners are able to differentiate the elements of speech on the basis of their sound* and this is the principal claim of those who espouse the speech-as-sounds view in opposition to the MT and DR view of speech-as-gestures. If we imagine, for the sake of argument, that the strategy of articulation is to make itself known to the listener by the agency of the acoustic-auditory signal, the result would likely be similar to the outcome of (Rostand's) Cyrano de Bergerac's courting of Roxane through the agency of Christian, namely, that Roxane [=listener] gives her affection to Christian [=sounds], not to Cyrano [=articulation]!

II. OTHER EVIDENCE

Further evidence against the plausibility of speech perception requiring the recovery of the articulations which produce the acoustic signal comes from a variety of sources both experimental and anecdotal. Over the past couple of decades there have been a number of studies demonstrating that prespeech infants and various animals (including chinchillas, macaques, budgerigars, and Japanese quail) are capable of differentiating certain speech sounds (Eimas *et al.*, 1971; Kuhl, 1986; Kluender *et al.*, 1987; Dooling *et al.*, 1987). Granted, differentiating a few isolated speech sounds—and often only after considerable training—is not the same as perceiving real speech in all of its complexity but the problems motivating MT and DR are still present in these cases: categorical perception and deriving phonemic constancy from a variable acoustic signal. We do not know how these experimental subjects solve these problems but it is unlikely to involve recovery the underlying vocal tract gestures.

It is also well known that certain species of birds, e.g., budgerigars, parrots, crows, and mynahs, can be trained to mimic human speech. Again, such mimicry is not speech in intention, but it does demonstrate that nonhumans (who presumably have no idea what is going on inside their trainer's vocal tract) can extract enough information on the relevant acoustic-auditory details to be able—literally—to parrot it. If nonhumans can do this, why not humans?

Finally, it is well attested anecdotally that humans have the capacity to perceive and differentiate a variety of complex sounds where, *contra* DR, recovery of the mechanisms of the sound source is unlikely or impossible. These include machine noises (e.g., from automobiles, computers, kitchen appliances) and, among skilled ornithologists, bird song. I think the experience of listening to musical instruments is relevant here, too. Even though I am aware that the notes of a bugle come from varying breath pressure and lip tension, I am unable to recover exactly what the bugler does to produce those notes. Again, one has to admit that machine noise, bird song, and music do not have certain key properties of speech but they are complex signals that can be analyzed without the complexities posited by MT and DR. At the very least, these cases suggest that we should not underestimate the ability of the auditory system (including the auditory cortex) to deal with complex patterns in acoustic signals as they are presented to the ear, i.e., without delving further into their origins.

As discussed above, Fowler argues for DR by drawing analogies to visual and haptic perception and appeals to commonsense: visually we see objects, not the light reflected from or emitted by them; why not the same with heard speech sounds? Even if we grant, for the sake of argument, that when we hear sounds we have the impression that what we are perceiving is doors slamming, didgeridoos, meadowlarks, or police car sirens. But, as reviewed here, it does not follow from this that we are aware of how these sounds are *produced*. Even ornithologists who study avian vocalization are not entirely sure how all birds produce their song and I doubt that very many people are familiar with the precise mechanism of sound production in sirens. Commonsense tells us that we *associate* sounds with the bodies that produce them without really having any clear idea of the precise mechanism by which the sounds are produced. This, I maintain, is true in the case of speech perception as well.

III. LANGUAGE LEARNERS HAVE TO RECOVER THE ARTICULATIONS OF HEARD SPEECH

There is one important circumstance where listeners *do* have to recover the articulations of heard speech: when the sounds of a language are being learned—more exactly, when the listeners also want to be speakers. This happens in first and second language acquisition as well as in phonetics classes and in speech therapy. In most cases—but not all—this learning is very successful. But what a lot of effort it takes! Outside formal instruction, success at imitating the heard speech seems to come about through trial-and-error. Different articulatory gestures are tried out until the feedback from native speakers and the speaker's own ears is positive. We should not impose this laboriously learned skill onto the task of speech perception. Furthermore, it is also well recognized in both first and second language learning that ability to differentiate sounds auditorily usually precedes the ability

to articulate the differences accurately—further indirect testimony that successful listening does not require recovery of the underlying articulations.

Speech scientists, of course, by virtue of their long familiarity with the vocal mechanism and its output, are very good at deducing the articulations that produce heard speech. I would speculate that this by itself may be the reason why MT and DR strikes some as plausible: it is a case of what the psychologists call "projection": attributing one's own intuitions and attitudes to others.

IV. THE "CHICKEN LITTLE" INQUIRY

Chicken Little is the principal figure in a children's story familiar to most English speakers. Chicken Little received a tremendous blow on the head one day and then set the entire barnyard into a panic with her theory that the sky was falling. The resolution of the story did not involve experiments testing whether the sky was, in fact, falling; it involved an inquiry as to why Chicken Little *thought* it was falling. The immediate evidence was a swelling on the top of her head. Where was she when the injury occurred? Under an oak tree. A large acorn was found on the very spot! The moral of the story—not stated quite so explicitly—is that before investing a lot of time and effort in costly experiments evaluating a given hypothesis, we should first look at the motivation for the hypothesis having been made. Then we should ask whether the observations prompting the hypothesis might be accounted for by other, less extravagant, hypotheses.

So, let 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 and, as noted by Lindblom in this same volume, invariance of functionally equivalent speech units has not been found at any stage in the act of speech production. MT and DR are, like the theory that the sky is falling, extravagant hypotheses that leave unexplored a host of other less costly explanations. Among these are

(1) that the invariants are present in the acoustic signal, one was just looking at the wrong features (Stevens and Blumstein, 1978);
(2) that some higher-order relations among certain acoustic parameters will show an invariant relation to linguistic units (Sussman *et al.*, 1991);
(3) that phonemes are not the units of encoding in the speech signal (though they may be in the lexicon): diphones or transemes may be a better candidate (Ohala, 1992c);
(4) that there is no single invariant feature for each unit of encoding, rather there is a collection of features and the identity of a given unit is determined if some critical number of these features appear. (Something of this sort has to underlie our recognition of faces that, inevitably, change with age, with cosmetics, with weight gain or loss, with more or less hair on various parts of the head, etc.);
(5) invariance isn't necessary: speech is like a cursor to the menu of possible messages and one can use a mouse or the arrow keys to move it (Lindblom, 1996).

V. SUMMARY

I have reviewed some phonetic and phonological as well as (briefly) experimental and anecdotal evidence that the hypothesis that speech perception by humans involves recovering articulations—as advocated in the Motor Theory and the Direct Realist Theories of speech perception—is implausible and is premature in that other, less extravagant, hypotheses have yet to be fully investigated. I am not saying that the hypothesis of recovery of articulations is untrue—no one knows that—just that it has not earned a position near the top of the list of candidate hypotheses about how speech perception works.

ACKNOWLEDGMENTS

I thank Robert Fox, Marios Fourakis, and an anonymous reviewer for their careful reading of an earlier version of this paper and their helpful editorial suggestions.

[1]Arguing that artificial intelligence need not copy nature, Raj Reddy has reminded us that "Birds fly and airplanes fly, but airplanes don't flap their wings." True, but airplanes, like birds, do have wings! Until the AI function is solved we are unsure which aspects of nature to copy and which to do in a way different from nature.

[2]See also Ladefoged and McKinney, 1963.

[3]These distinctive features were part of the lore of the Prague School in the 1920's and 1930's (Trubetzkoy, 1939/1969) but had a largely impressionistic basis. It was the collaborative effort of Roman Jakobson. Gunnar Fant, and Morris Halle which gave them an acoustic interpretation (Jakobson *et al.*, 1952).

[4]Regarding the asymmetry of confusion, i.e., that palatalized labials are confused with apicals but not, or with much lower probability, the reverse, see Ohala (1983a, 1985b).

Dooling, R. J., Soli, S. D., Kline, R. M., Park, T. J., Hue, C., and Bunnell, T. (**1987**). "Perception of synthetic speech sounds by the budgerigar (*Melopsittacus undulatus*)," Bull. Psychon. Soc. **25**, 139–142.

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

Erikson, Y. (**1973**). Preliminary evidence of syllable-locked temporal control of *F*0. Stockholm Speech Transmission Lab. STL-QPSR 2-3/1973.23–30.

Erikson, Y., and Alstermark, M. (**1972**). "Fundamental frequency correlates of the grave accent in Swedish: The effect of vowel duration," Quarterly Progress & Status Report, Speech Transmission Lab., Royal Inst. of Tech., Stockholm STL-QPSR 2-3/1972, 53-60.

Fanty, M. and Cole, R. (**1990**). "Speaker-independent English alphabet recognition: Experiments with the *E*-set," ICSLP 1990, Kobe, pp. 1361–1364.

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

Fowler, C. A. (**1990**). "Calling a mirage a mirage: direct perception of speech produced without a tongue," J. Phon. **18**, 529–541.

Fujimura, O., and Lindqvist, J. (**1971**). "Sweep-tone measurements of vocal-tract characteristics," J. Acoust. Soc. Am. **49**, 541–558.

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

Jakobson, R. Fant, C. G. M., and Halle, M. (**1952**). "Preliminaries to speech analysis. The distinctive features and their correlates," Acoustic Laboratory, MIT, Technical Report No. 13, Acoustic Laboratory, MIT, Cambridge.

Kluender, K. R., Diehl, R. L., and Killeen, P. R. (**1987**). "Japanese quail can learn phonetic categories," Science **237**, 1195–1197.
Kuhl, P. K. (**1986**). "Theoretical contributions of tests on animals to the special-mechanisms debate in speech," Exp. Biol. **45**, 233–265.
Ladefoged, P., and McKinney, N. (**1963**). "Loudness, sound pressure and sub-glottal pressure in speech," J. Acoust. Soc. Am. **35**, 454–460.
Liberman, A. M. (**1993**). "Some assumptions about speech and how they changed," Haskins Laboratories Status Report on Speech Research SR-113, 1–32.
Liberman, A. M., and Mattingly, I. G. (**1985**). "The motor theory of speech perception revised," Cognition **21**, 1–36.
Liljencrants, J., and Lindblom, B. (**1972**). "Numerical simulation of vowel quality systems. The role of perceptual contrast," Language **48**, 839–862.
Lindblom, B. (**1986**). "Phonetic universals in vowel systems," in *Experimental Phonology*, edited by J. J. Ohala and J. J. Jaeger (Academic, Orlando, FL), pp. 13–44.
Lindblom, B. (**1996**). "Role of articulation in speech perception: Clues from production," J. Acoust. Soc. Am. **99**, 1683–1692.
Lisker, L., Cooper, F. S., Liberman, A. M. (**1962**). "The uses of experiment in language description," Word **18**, 82–106.
Mack, M., and Blumstein, S. E. (**1983**). "Further evidence of acoustic invariance in speech production: The stop-glide contrast," J. Acoust. Soc. Am. **73**, 1739–1750.
Maddieson, I. (**1984**). *Patterns of Sounds* (Cambridge U.P., Cambridge).
Magdics, K. (**1966**). *Studies in the Acoustic Characteristics of Hungarian Speech Sounds* (Indiana U.P., Bloomington).
Morton, E. W. (**1977**). "On the occurrence and significance of motivation-structural rules in some bird and mammal sounds," Am. Nat. **111**, 855–869.
Ohala, J. J. (**1974**). "Phonetic explanation in phonology," in *Papers from the Parasession on Natural Phonology*, edited by A. Bruck, R. A. Fox, and M. W. LaGaly (Chicago Linguistic Soc., Chicago), pp. 251–274.
Ohala, J. J. (**1978**). "Southern Bantu vs. the world: the case of palatalization of labials," Berkeley Ling. Soc., Proc., Ann. Meeting **4**, 370–386.
Ohala, J. J. (**1979**). "Universals of labial velars and de Saussure's chess analogy," *Proc., 9th Int. Congr. of Phonetic Sciences*, Vol. 2 (Institute of Phonetics, Copenhagen), pp. 41–47.
Ohala, J. J. (**1981**). "The listener as a source of sound change," in *Papers from the Parasession on Language and Behavior*, edited by C. S. Masek, R. A. Hendrick, and M. F. Miller (Chicago Linguistic Soc., Chicago), pp. 178–203.
Ohala, J. J. (**1983a**). "The phonological end justifies any means," in *Proc. of the XIIIth Int. Congr. of Linguists, Tokyo, 29 Aug.–4 Sept. 1982*, edited by S. Hattori and K. Inoue [distributed by Sanseido Shoten, Tokyo], pp. 232–243.
Ohala, J. J. (**1983b**). "The origin of sound patterns in vocal tract constraints," in *The Production of Speech*, edited by P. F. MacNeilage (Springer-Verlag, New York), pp. 189–216.
Ohala, J. J. (**1983c**). "The direction of sound change," in *Abstracts of the Tenth Int. Congr. of Phonetic Sciences*, edited by A. Cohen and M. P. R. v. d. Broecke (Foris, Dordrecht), pp. 253–258.
Ohala, J. J. (**1984**). "An ethological perspective on common cross-language utilization of *F*0 of voice," Phonetica **41**, 1–16.
Ohala, J. J. (**1985a**). "Around flat," in *Phonetic Linguistics. Essays in Honor of Peter Ladefoged*, edited by V. Fromkin (Academic, Orlando, FL), pp. 223–241.
Ohala, J. J. (**1985b**). "Linguistics and automatic speech processing," in *New Systems and Architectures for Automatic Speech Recognition and Synthesis*, edited by R. De Mori and C.-Y. Suen [NATO ASI Series, Series F: Computer and System Sciences, Vol. 16] (Springer-Verlag, Berlin), pp. 447–475.
Ohala, J. J. (**1990**). "There is no interface between phonetics and phonology. A personal view," J. Phon. **18**, 153–171.
Ohala, J. J. (**1992a**). "Alternatives to the sonority hierarchy for explaining the shape of morphemes," *Papers from the Parasession on the Syllable* (Chicago Linguistic Society, Chicago), pp. 319–338.
Ohala, J. J. (**1992b**). "The segment: Primitive or derived?" in *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, edited by Gerard J. Docherty and D. Robert Ladd (Cambridge U.P., Cambridge), pp. 166–183.
Ohala, J. J. (**1992c**). "What's cognitive, what's not, in sound change," in *Diachrony within Synchrony: Language History and Cognition*, edited by Günter Kellermann and Michael D. Morrissey [Duisburger Arbeiten zur Sprach- und Kulturwissenschaft 14] (Peter Lang Verlag, Frankfurt/M), pp. 309–355.
Ohala, J. J. (**1993a**). "The phonetics of sound change," in *Historical Linguistics: Problems and Perspectives*, edited by C. Jones (Longman, London), pp. 237–278.
Ohala, J. J. (**1993b**). "Coarticulation and phonology," Language Speech **36**, 155–170.
Ohala, J. J. (**1993c**). "Sound change as nature's speech perception experiment," Speech Commun. **13**, 155–161.
Ohala, J. J., G. Busà, M. G., and Harrison, K. (**1992**). "Phonological and psychological evidence that listeners normalize the speech signal," *Proceedings, International Conference on Spoken Language Processing*, Banff, 12–16 Oct. 1992 (University of Alberta, Edmonton), pp. 1303–1306.
Ohala, J. J., and Lorentz, J. (**1977**). "The story of [w]: an exercise in the phonetic explanation for sound patterns," Berkeley Ling. Soc., Proc., Ann. Meeting **3**, 577–599.
Ohala, J. J., and Ohala, M. (**1993**). "The phonetics of nasal phonology: theorems and data," *Nasals, Nasalization, and the Velum*, edited by M. K. Huffman and R. A. Krakow (Phonetics and Phonology Series, Vol. 5) (Academic, San Diego, CA), pp. 225–249.
Ohala, M. (**1983**). *Aspects of Hindi Phonology* (Motilal Banarsidass, Delhi).
Ohala, M., and Ohala, J. (**1991**). "Nasal epenthesis in Hindi," Phonetica **48**, 207–220.
Shadle, C. H. (**1990**). "Articulatory-acoustic relationships in fricative consonants," in *Speech Production and Speech Modelling*, edited by W. J. Hardcastle and A. Marchal (Kluwer, Dordrecht, The Netherlands), pp. 187–209.
Stevens, K. N. (**1980**). "Discussion. (in Symposium on Phonetic Universals in phonological systems and their explanation)," *Proceedings of the 9th Int. Congr. of Phonetic Sciences, Copenhagen*, Vol. 3, pp. 185–186.
Stevens, K. N. (**1989**). "On the quantal nature of speech," J. Phon. **17**, 3–45.
Stevens, K. N., and Blumstein, S. E. (**1978**). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. **64**, 1358–1368.
Sussman, H. M., McCaffrey, H. A., and Matthews, S. A. (**1991**). "An investigation of locus equations as a source of relational invariance for stop place categorization," J. Acoust. Soc. Am. **90**, 1309–3125.
Trubetzkoy, N. (**1969**). *Principles of Phonology* (University of California Press, Berkeley) [Orig. ed. 1939, Grundzüge der Phonologie, Prague.]
Winitz, H., Scheib, M. E., and Reeds, J. A. (**1972**). "Identification of stops and vowels for the burst portion of /p,t,k/ isolated from conversational speech," J. Acoust. Soc. Am. **51**, 1309–1317.