669

ECOLOGICAL PSYCHOLOGY, 1(2), 145–160


# Real Objects of Speech Perception: A Commentary on Diehl and Kluender

Carol A. Fowler
*Dartmouth College and*
*Haskins Laboratories*

I agree with Diehl and Kluender (1989) that perceptual constraints guide the development of sound inventories and of phonological processes in languages. I disagree that these constraints are primary in comparison with other influences on sound inventories, such as articulatory ones. More important, I disagree that any of the evidence that Diehl and Kluender cite, indicates that objects of perception are auditory or acoustic rather than phonetic gestural. None of the evidence is persuasive; all of it is consistent with a view that perceptual objects are gestural. Viewed in a larger context—of a universal theory of perception—a theory that perceptual objects are gestural, whereas acoustic structure serves as information for gestures, is strongly promoted.

Diehl and Kluender propose that factors affecting the development of phonological-segment inventories of languages largely promote perceivability of spoken messages; talker-centered factors, such as ease of articulation, are less important. Examination of the evidence leads them also to conclude that the objects of speech perception—and the things whose perceivability is promoted in the development of sound systems of languages—are "acoustic or auditory events" but are not the vocal-tract gestures that cause acoustic events.

Although I disagree with the relative emphases that Diehl and Kluender place of perceivability constraints as compared to articulatory ones, it is easy to agree with their general claim (following Lindblom, e.g., 1986) that perceivability constrains the development of phonological-segment inventories of languages. Indeed, it could hardly fail to do so, at least in a very general way. In order for any two phonetic segments to count in a language community as members of distinct phonological categories, talkers must be able to use the segments

Requests for reprints should be sent to Carol A. Fowler, Haskins Laboratories, 270 Crown Street, New Haven, CT 06511–6695.

distinctively, and they cannot do that unless they hear the differences between the segments. Diehl and Kluender (again, following Lindblom) propose that the constraints are stronger and more particular than this. I have no reason to disagree.

I do disagree with the second claim that examination of factors that promote perceivability reveals that objects of perception are "acoustic or auditory events." That disagreement will be the major focus of my commentary, discussed later in the section, "Objects of Speech Perception." First, however, I turn to two smaller disagreements with Diehl and Kluender's discussion of the shaping of sound inventories of languages.

## PERCEPTUAL CONSTRAINTS ON THE DEVELOPMENT OF PHONOLOGICAL SYSTEMS

One disagreement concerns the relative weights of perceptual and articulatory sources of constraint. Diehl and Kluender propose that "within certain broad limits imposed by physics and physiology, talkers can, in principle, exert independent control over most of the component structures involved in speech production." They suggest that, given all that freedom, language communities "select" properties of speech to covary "largely because" they have mutually enhancing auditory effects, thereby increasing the auditory distinctiveness of members of phonological inventories of languages.

A statement that talkers can exert independent control over most vocal tract structures is misleading. They cannot do so all at once; nor would they want to. As Weiss (1941) and Pattee (1970) pointed out in different ways, reductions in degrees of freedom via coordination is the means by which larger scale entities are formed in complex systems composed of multiple small parts. In speech, coordination is the means by which phonetic segments are produced.

Perhaps Diehl and Kluender mean instead that, when talkers reduce degrees of freedom and produce phonetic segments by means of coordination, there are many ways in which they could do so, and, generally they choose ways that will create covariations that enhance auditory distinctiveness among phonological segments of a language. Even so, I think that there is an error of emphasis here. The perspective on talking that Diehl and Kluender offer is analogous to the following one on perceiving: The human auditory system is capable of handling frequencies between A Hz at X dB and B Hz at Y dB; accordingly, the only auditory constraint on sound inventories is that the sounds of language must fall within that range. Diehl and Kluender will see immediately that this perspective is misleading. As I just noted, they consider the perceptual constraints on sound inventories much stronger and more particular than that. So are the articulatory constraints on sound inventories. Why should they be weaker and less particular than constraints on perceivability?

Diehl and Kluender sometimes overlook articulatory sources of covariations.

To give just one example, they suggest that lip rounding gestures prototypically include two components: protrusion and constriction. They propose that rounding exhibits both components because both independently lower F2, and so the two components of the gesture work together to enhance that distinctive acoustic property of rounded vowels. However, the literature on lip rounding identifies the orbicularis oris muscle as the prime mover for the rounding gesture (see, e.g., Bell-Berti & Harris, 1979; Harris, Lysaught, & Schvey, 1965), and as Zemlin (1968, p. 258) pointed out, the orbicularis oris both "closes the mouth and puckers the lips." That is, the coupled movements are coupled by articulatory disposition. Unless talkers use other muscles to undo the coupling, rounding gestures will show both protrusion and constriction. When talkers use the oris to produce rounding, protrusion and constriction constitute a single gesture, not two "covarying" gestures.

A second disagreement concerns how constraints, whatever they may be, have their effects on the development of sound inventories. I do not think that it is correct to propose that languages *select* properties to covary *because* that covariation makes a phonetic feature or gesture easy to hear, easy to distinguish from others, or even because it makes it easy to say. That is, members of language communities do not decide by vote what covariations will be efficacious; rather, sound inventories evolve in the course of communicative exchanges among language-community members. Segments that are frequently misheard will not persist unmodified in an inventory.

From that perspective, it is a fascinating fact about sound change that it frequently involves elevation, often with stylization and exaggeration, of a

TABLE 1
**Sample of Analogous Dispositional Phonetic Regularities Common to Many Languages and Phonological Rules Common to Just a Few**

| *Dispositional Regularities* | *Phonological Rules* |
|---|---|
| 1. Final devoicing | Final devoicing rule (e.g., German) |
| 2. Vowel shortening before voiceless consonants | Phonologically long vowels before voiced consonants (Middle High German; Comrie, 1980) |
| 3. Vowel shortening before consonants in coda | Vowel length variation as a function of number of following consonants (Old English, Kiparsky, 1968; Yawelmani, Kenstowicz, & Kisseberth, 1979) |
| 4. Vowel shortening before one or more unstressed syllables | Trisyllabic shortening (Old English, e.g., Kiparsky, 1968), prohibition of long vowels before antepenult in a word in Chimwi:ni: (Kenstowicz & Kisseberth, 1979) |
| 5. Tonal consequences of consonant voicing | Tonogenesis (e.g., in Punjabi; Ohala, 1981) |
| 6. Declination | Tonal downstep (e.g., in Igbo; Hyman, 1973) |
| 7. Vowel-to-vowel coarticulation | Vowel harmony (e.g., Turkish), fronting of vowels before /i/ (e.g., Old High German; Kiparsky, 1968) |

dispositional behavior common to many languages into the phonologies of some languages. Table 1 lists some examples.

The dispositional behaviors are not parts of the phonologies of languages—that is, they do not create essential differences among distinct phonological segments, and they are not used to signal lexical distinctions. I claim that they are covariations or other systematicities that are easier to exhibit than not to. Consider the vowel duration difference before voiced and voiceless consonants. This is present in most languages as a 20–35 ms difference, owing probably to the more forceful (Öhman, 1967) and, hence, faster (Chen, 1970; Summers, 1987) closing gesture for voiceless consonants. (That, in turn, is associated with a tighter seal during voiceless consonants required to withstand the higher intraoral air pressure when the vocal folds are open.) At one time in the history of German, this difference was phonologized so that there was a rule selecting phonologically long vowels in the context of following voiced consonants and short vowels before voiceless consonants (see Comrie, 1980). In English, the difference may be in the process of becoming phonologized (see, e.g., de Chene, 1985); talkers actively lengthen English vowels before voiced consonants (Raphael, 1975) and, accordingly, the duration difference in the two consonantal contexts is considerably larger than in languages that have shorter vowels before voiceless consonants (Chen, 1970) and have shorter ones only because that consonantal closing gesture is fast.[1]

A plausible source of these parallels between dispositional regularities of many languages and phonological regularities of just a few, is, as Ohala (1974, 1981) suggested, misparsing by listeners. One example Ohala offered is development of tones in some tone languages. In utterances of all languages with a consonantal voicing distinction, vowels following a voiceless consonant will have a high falling tone over its first 100 ms or so; vowels following a voiced consonant will have a lower tone (e.g., Ohde, 1984). When talkers abduct the vocal folds for a voiceless consonant, they tense the cricothyroid muscle to stiffen the vocal folds against closing (Baer, Löfqvist, McGarr, & Story, in press). When the cords are adducted for the next vowel, the carryover tensing of the cricothyroid increases the rate of vocal fold opening and closing at the beginning of the vowel. In some languages in which a consonantal voicing distinction was lost, words formerly distinguished by a voicing difference subsequently were distinguished by the

[1]I am aware that Randy (personal communication, September 22, 1988) doubts that English is special in this way. However, I am persuaded by the evidence. For example, Chen (1970) provided measurements of vowels before voiced and voiceless consonants in words of several languages. On comparable word types (CVCs), English shows up as markedly deviant. For example, vowels before unvoiced consonants in English CVCs averaged 161 ms; in Russian, they averaged 153 ms. However, vowels in English CVCs exhibited 156 ms average lengthening before voiced consonants; vowels in Russian CVCs showed 34 ms of lengthening. The French CVCs were considerably longer than those in English and Russian; vowels average 440 ms in CVCs with voiceless final consonants. Despite their long duration, vowels exhibited just 64 ms of lengthening before voiced consonants.

tones on a following vowel. A vowel preceded earlier by a voiceless consonant now was associated with a high tone; one preceded earlier by a voiced consonant was produced with a low tone (e.g., Ohala, 1974). Possibly, as the voicing distinction weakened and became difficult to hear, listeners failed to recognize the consonantal source of the tones and mistakenly identified them as intentional impositions on vowels by talkers. When talkers produced the words themselves, they failed to reproduce the voicing distinction and produced a tone on the vowel instead.

There are many of these analogies between dispositional regularities common to most languages and phonological regularities common to many fewer languages. Elevation of a regularity into the phonology may well have an origin in systematic perceptual failures by listeners (encouraged generally by weakening of a distinction in productions by talkers). Some of these elevations may improve intelligibility of words. Some of the examples offered by Diehl and Kluender may be apt in this respect. However, it does not always happen that way. When the Punjabi language developed a tonal distinction, it lost one distinction (voicing) and gained another. I do not see any covariation here that enhances perceptibility. In any case, its relative, Hindi, maintained the voicing distinction and did not develop tones. Why? My guess is that these elevations do not occur in order to improve intelligibility of words. Perhaps they occur due to systematic mishearings by listeners. A multitude of factors, including perceptual ones, articulatory ones, and the existing phonological system of the language, will determine which elevations become popular and survive.

## OBJECTS OF SPEECH PERCEPTION

Diehl and Kluender argue that perceptual objects in speech perception are auditory, not articulatory. This proposal marks a major and central theoretical disagreement between us, and so it is worth careful analysis. Here is my account and its theoretical foundations.

My proposal concerning the objects of speech perception derives from Gibson's direct-realist perspective (e.g., Gibson, 1966, 1979). I propose (Fowler, 1986, in press; Fowler & Rosenblum, in press), following Gibson, that all of perception works in just one general way. Organisms always perceive properties of real-world events by recovering information about them in media such as light and air. In these media, information takes the form of structure that has been caused by properties of real-world events; the structure specifies those properties both because the properties caused it and because the structure is largely specific to its source. In perception, these structured media serve not as objects of perception themselves, but as information for their sources. We see objects and events in the environment, not reflected light. By the same token, we hear properties of real-world events that lawfully structure the air, and we feel

properties of real-world events that lawfully deform the skin. This way of perceiving, and only this way, ensures that, via information transduced by the different perceptual modalities, perceivers come to know just one world, the same world across all the modalities, the world out there.

If perception has this universal character, then how should perception of speech be characterized? What are the real-world events that are perceived, and what is the medium that carries information about them to perceivers?

The real-world events of speech that structure a medium are, most locally, activities of the vocal tract. I discuss this matter later. The informational medium causally structured by those activities is the acoustic signal. In the theory, then, activities taking place in the vocal tract are objects of perception. Properties of acoustic signals are not perceptual objects, and "auditory" properties cannot be perceptual objects for the reason given in the preceding section. In any case, to accept a proposal that objects of perception are acoustic or auditory and are not activities of the vocal tract or the larger units of speech that they compose is to give up the claim that perception universally recovers real-world events from information about them in media. In my opinion, that is too high a price to pay, and I suggest later that the evidence ostensibly in its favor has alternative interpretations that are not associated with the same cost.

First, however, I consider in more detail, the nature of the objects of speech perception from a direct-realist perspective. I have referred to them so far as *vocal-tract activities*. However, if they were all that listeners recover, then direct perception of vocal-tract activities would not yield direct perception of the talker's phonetic message. However, to the extent that phonetic segments are, in their physical aspect, activities of the vocal tract, they can be directly perceived, and that, indeed, is what I proposed. In my view, phonetic segments are essentially physical activities of the vocal tract that have linguistic (and, hence, psychological) significance in virtue of the uses to which language communities put them (Fowler, 1986, in press). Understanding how phonetic segments may be considered physical events taking place in the vocal tract requires understanding two concepts not yet introduced here. One is *coordinative structure* (or *synergy*); the other is *phonetic gesture*.

Evidence suggests that, during speech, talkers organize their vocal tracts into coordinative structures, each responsible for realizing a fundamental phonetic property of a consonant or vowel. The concept of coordinative structure or synergy is familiar in the literature on motor control, but is less familiar in the speech literature. A coordinative structure is a temporary organization, here of articulators of the vocal tract, that ensures achievement of some goal even in the presence of perturbations to individual movement effectors. For example, in speech, they ensure achievement of phonetic ends despite perturbations to articulators caused by competing demands of coarticulating phonetic segments on common articulators (e.g., Saltzman, 1986). A well-worn example is the coordinative structure that achieves bilabial closure for consonants such as /b/,

/p/ and /m/. Evidence that bilabial closure is achieved by a temporarily marshalled organization among the jaw, the upper lip, and the lower lip derives from experiments in which an articulator is perturbed during speech. When the jaw is braked unexpectedly during bilabial closure (Kelso, Tuller, Vatikiotis-Bateson, & Fowler, 1984), within 20–30 ms of the perturbation, extra activity in the orbicularis oris superior muscle of the upper lip is observed, which ensures achievement of closure despite the unusually open position of the jaw caused by the perturbation. (See, e.g., Abbs & Gracco, 1984, for similar findings when the lower lip is perturbed.) The same perturbation to the jaw applied during alveolar closure does not lead to an extra downward excursion of the lip—a nonfunctional response in this instance—but rather to extra activation of a muscle of the tongue that ensures alveolar closure despite the perturbation to the jaw (Kelso et al., 1984). The very short-latency response to perturbation signals a low-level organizational relationship among the articulators; the specificity of the compensatory response to the phonetic goals of the talker implies that these relationships are not hard-wired, but are established temporarily for the task at hand. These temporary relationships among articulators are called *coordinative structures* or *synergies*. The coordinative structures identified in the experiments just cited most likely function in speech outside the laboratory to ensure achievement of bilabial closure and alveolar closure despite perturbations to the jaw caused by coarticulating vowels (e.g., Keating, in press; Sussman, MacNeilage, & Hanson, 1973).

In speech, the coordinative structures that have been identified do not realize whole phonetic segments, but they do realize components of them that, following Browman and Goldstein (e.g., 1986), I call *phonetic gestures*. For example, the coordinative structure that realizes bilabial closure for /p/ does not also realize its laryngeal devoicing gesture, although there is evidence that the oral and laryngeal gestures themselves are coupled (e.g., Löfqvist & Yoshioka, 1984). In short, a phonetic segment is realized in the vocal tract as a set of coupled phonetic gestures.

The phonetic gestures achieved by coordinative structures are close in scope to phonetic features of linear phonologies (e.g., Chomsky & Halle, 1968), but they have the advantage over those abstract features of mapping directly onto talkers' vocal behaviors. In light of that, it is exciting to learn from the recent literature on nonlinear phonologies (e.g., Browman & Goldstein, 1986, in press; Clements, 1985; Sagey, 1986) that fundamental phonetic properties of utterances that participate as individuals in phonological rules of languages may be gestural in nature. That is, phonetic segments of languages, as linguistic objects, may be composed of phonetic gestures of the vocal tract. If so, then gestures are minimal linguistic components of an utterance and, at the same time, they are products of the minimal organizational structures of speech production. By hypothesis, too, they are the minimal phonetic objects of perception.

Notice that there is an important distinction between vocal-tract gestures and

phonetic gestures that Diehl and Kluender miss in their target article. They write that there is a determinate relation between vocal-tract area functions and resonances of the vocal tract for vowels and, hence, between vocal-tract area functions and vowel quality, but there is a many-to-one correspondence between vocal tract gestures and area functions. Because listeners will hear the same vowel when the resonances are the same, different vocal-tract gestures may underlie production of vowels heard as indistinguishable.

This is correct, but there is not the same many-to-one relationship between phonetic gestures and area functions as there is between movements of the articulators and area functions. The data Diehl and Kluender cite as demonstrating a many-to-one relationship between vocal-tract gestures and area functions are provided in a study by Lindblom, Lubker, and Gay (1979). In this research, talkers produce vowels with bite blocks clenched between their teeth. The bite blocks prevent jaw movement and fix the jaw position in an unusually open or closed position. Despite this disruption to normal ways of producing vowels, talkers produce near-normal vowels, both acoustically and perceptually, from the first pitch pulse of the first-produced tokens. Hence, different vocal-tract gestures achieve perceptually very similar vowels. I interpret data from bite block experiments (Fowler & Turvey, 1980) in the same way as data from the perturbation studies (Abbs & Gracco, 1984; Kelso et al., 1984) cited earlier. They reveal the coordinative structures that implement phonetic gestures. For vowels, these synergies will include organized relations among jaw, tongue, and sometimes the lips, that allow equifinal realizations of the vowels' height and front- or backness. Accordingly, despite the bite block, synergies allow talkers to produce the same phonetic gestures as they produce when the jaw is free to move; they produce a given phonetic gesture with different members of the family of equifinal vocal-tract gestures allowed by the synergy.

If talkers literally produce the phonetic gestures that compose phonetic segments, then phonetic segments are not private things; they are the activities of the vocal tract that causally structure the acoustic signal. Accordingly, the acoustic speech signal directly structured by phonetic segments themselves can specify those segments directly to perceivers to the extent that the signals are specific to their phonetic gestural sources.

Elsewhere (Fowler, in press; Fowler & Rosenblum, in press) I reviewed evidence that listeners do recover activities of vocal-tract synergies from acoustic signals. I do not repeat the summary here, but instead I offer just one example. A consequence of activating the genioglossus muscle of the tongue to produce a high vowel apparently is a stretching of the vocal folds and, therefore, raising of the fundamental frequency of the voice (Honda, 1981). (The genioglossus muscle has attachments to the hyoid bone; Honda suggested that when the genioglossus contracts, it pulls the hyoid bone forward, rotates the thyroid cartilage, and stretches the vocal folds.) Listeners do not hear those tonal

consequences of vowel height as part of the changing pitch melody (intonation contour) of an utterance, however (Silverman, 1987). Indeed, apparently they do not hear them as tones at all, but rather as vowel height (Reinholt Petersen, 1986). The intonation contour is implemented largely by tensing muscles of the larynx; tonal effects of vowel height are products of a synergy for producing the vowel. Listeners hear the two sources of influence on vocal fold opening and closing as distinct. Apparently listeners use the constellation of acoustic products of gestural synergies to recover the separate synergies themselves. Perceptual objects are not acoustic; they are not articulator movements. They are, minimally, gestures realized by coordinative structures.

## COVARIATION AND DISTINCTIVENESS: SOME EXAMPLES REINTERPRETED

If listeners do recover phonetic gestures, what about the evidence Diehl and Kluender cite suggesting that languages choose clusters of gestures for segments that will enhance the segments' auditory distinctiveness? Consider what auditory enhancement means from a direct-realist perspective by considering an analogy from visual perception. Generally, visible events look different to the extent that they are different kinds of events. Sometimes it does not work out quite like that, however. Heat rising from the highway is not much like a puddle, but sometimes it looks like a puddle. By the same token, generally, audible events will sound different to the extent that they are different events. However, sometimes it may not work out quite like that. Sometimes different vocal tract behaviors may give rise to similar acoustic signals. Perhaps one perceptual constraint on development of phonological-segment inventories, then, may be either to avoid those rare sets of distinct gestures that structure the air in similar ways or else to avoid counting them as distinct (e.g., Bell-Berti, Raphael, Pisoni, & Sawusch, 1979). More positively, a constraint may be to employ behaviors that not only are distinct, but that effectively broadcast their differences (cf. Warren, 1985) and to do so in a way that human auditory systems can readily detect.

Diehl and Kluender propose, however, that perceptibility is enhanced, not by selecting behaviors that are distinct and that broadcast their distinctness, but by selecting behaviors that maximize audible acoustic differences per se, however they are produced. More than that, they propose that sometimes behaviors are selected because they maximize some distinction-enhancing auditory illusion—in their example, auditory contrast. I disagree with both proposals. Next, I reinterpret one of their examples of maximization of acoustic differences and one of maximizing an auditory illusion. Finally, I very briefly consider their research on speech perception by quail.

## Point Vowels

Diehl and Kluender acknowledge that, in general, the popular point vowels are maximally distinctive among possible vowels, both articulatorily and acoustically. However, they find some evidence suggesting that acoustic, not articulatory distinctiveness, is primary. For example, there is a tendency across languages for back vowels to be rounded and front vowels unrounded. For a back vowel, such as /u/, rounding the lips lowers the already low F2. In contrast, in English, /i/ is produced with lip spreading and with a deformation of the tongue relative to the jaw that pushes the tongue blade up near the palate relative to its position for /I/ (Wood, 1982). Both maneuvers increase the already high upper formants for /i/. Diehl and Kluender do not find any articulatory motivation for these maneuvers that accompanies the enhancement of acoustic distinctiveness. I think I can, however.

Lip rounding both increases the length of the vocal tract and, especially, lengthens the already-large front cavity. Back vowels have large front cavities, and the tendency for languages to round those vowels enhances that distinctive property.[2] If listeners recover the physical events that structure the air in speech, they will be helped to the extent that those events are distinct. Vowel gestures that approach or achieve markedly different sizes and shapes of cavities on either side of the tongue constriction should be perceptually distinct.

Lip spreading for a front vowel shortens the tract and, especially, the front cavity, which is already distinctively short for front vowels. The tongue gesture relative to the palate may have a similar consequence if it increases the longitudinal extent of constriction between tongue and palate. Once again, this enhances the produced distinctiveness of front vowels, not just its acoustic distinctiveness.

## Vowel Lengthening and Consonantal Closure

Diehl and Kluender propose that languages show covariation between vowel duration before voiced and voiceless consonants (longer before voiced consonants) and stop-closure duration (shorter for voiced consonants), because longer vowels will induce an auditory illusion—a contrast effect that will enhance the

[2]Riordan's (1977) study, to which Diehl and Kluender refer, shows compensation by the larynx for an inability to round the lips for back vowels. There is at least one failure to replicate this finding (Tuller & Fitch, 1979; see also, Wood, 1982, who found no tendency for talkers to compensate for token-to-token variability in larynx height by changes in constriction location for vowels). I would not be surprised if it did not replicate, because although larynx lowering compensates for the overall change in vocal-tract length effected by rounding, it lengthens the wrong cavity—the one behind the constriction, not the one in front.

contrast in closure duration between voiced and voiceless consonants for listeners. They cite evidence that there is such a contrast effect both for speech signals and for nonspeech analogues as support for the hypothesis.

I cannot believe that language communities will ever work to maximize illusory effects of stimulation. In any case, this example is not persuasive.

Earlier I mentioned that the vowel duration difference shows up in two places in language—in the phonologies of some languages and in the phonetics of all (or most; see Keating, 1985). As a phonological lengthening of a vowel, the effect takes its place with other elevations of dispositional behaviors into phonologies. Many of these elevations have no apparent motivation in increased auditory distinctiveness. (See Table 1). As a phonetic systematicity, the lengthening or shortening also needs to be viewed in a broader context, this one of manifold "compensatory shortenings" (Lindblom & Rapp, 1973; see also Lehiste, 1971) in speech. Vowels are shorter before voiceless consonants than voiced consonants, in closed syllables than in open syllables, in syllables ending in two consonants than in syllables ending in one consonant; stressed vowels are shorter before one unstressed syllable than before none and before two unstressed syllables than before one. Also, the shortenings are audible to listeners who expect them (e.g., Nooteboom, 1973). Is it plausible that we shorten a stressed syllable before two unstressed ones to make the two unstressed syllables sound longer than they are? This is unlikely; indeed, here auditory contrast would seem to reduce the distinctiveness of stressed and unstressed syllables along the dimension of perceived duration.

As I argued earlier, the small amount of durational shortening of vowels before voiceless consonants that is common to languages may well reflect the more forceful closing gesture for the voiceless consonant. If that is the reason, then similar covariation might be found in some other physical systems.

Consider the case of a hollow rubber ball bounced on a carpeted surface. When the ball is hit by a paddle at a fixed distance from the floor, time for the ball to reach the floor is shorter when the ball is hit more forcefully than when it is hit gently. The acoustic duration of the bounce sound is longer when the

TABLE 2
Time to Bounce and Bounce-Sound Duration for Gently and Hard-Hit Racquet Balls

| | *Gently Hit* | | *Hard-Hit* | |
|---|---|---|---|---|
| | *Time-to-Bounce* | *Bounce* | *Time-to-Bounce* | *Bounce* |
| | 237 | 40 | 153 | 64 |
| | 267 | 39 | 153 | 52 |
| | 240 | 39 | 176 | 78 |
| | 240 | 41 | 197 | 51 |
| | 286 | 40 | 185 | 58 |
| M | 254 | 40 | 173 | 61 |

ball is hit hard than when it is hit softly. Data from five trials of each type are presented in Table 2. Why is there negative covariation between time to hit the floor and bounce-sound duration? Certainly it is not there to induce an auditory contrast effect in listeners so that they will hear the forceful bounce as longer than it really is. Presumably it is there largely because there is more reverberation the more forceful the impact.[3]

There may be large classes of events that show negative durational covariation when one duration reflects a forceful or less forceful gesture of some kind. Indeed, I suspect that this is why the "nonspeech" stimuli used by Diehl and Kluender show results similar to those obtained using speech syllables. It is difficult to guess what kind of event the square-wave stimuli signal to a listener. However, if a principle holds generally that similar events tend to structure air similarly, then the square-wave stimuli most likely specify an event similar to speech in the respects that have given rise to this negative covariation.

I have a final comment about this speech and nonspeech comparison. Although this kind of comparison is made frequently in the literature, I do not think that it provides interpretable findings. There is no natural category of "nonspeech" events in the way that there is a category "speech," and the particular sounds that researchers use in these experiments are very special. They are special in that they are not consequences of a natural sound-producing source and, therefore, researchers have no way to know what kind of physical event is signaled by them or even to know whether a coherent event is signaled at all. When subjects' responses to speech and "nonspeech" signals look similar, researchers infer that the same perceptual processes, presumably common to speech and nonspeech, are applied to each kind of signal. But they could just as well infer that the physical events specified by the signals are similar (in respect to the property tested; they need not otherwise be similar). That is, they could just as well draw an inference about the nature of sound-producing events in the world as about processes inside the heads of perceivers. Were subjects to respond differently to speech and "nonspeech" signals, researchers would infer that different perceptual processes are applied to the signals, one set ostensibly special to speech and one general to nonspeech. But they could just as well infer that the sound-producing events specified by the signals differ in respect to the property under test (even if they are otherwise quite similar). What justifies the inferences that are made in terms of perceptual processes?

[3]Another possibility that superficial analysis of the acoustic signals for the bounces leads me to reject (in consultation with Richard McGowan, August 2, 1988, whom I thank for his advice) is that the longer duration bounce reflects the greater deformation of the ball on impact when it is hit hard as compared to gently. Although this is the case, presumably, the harder hit ball is also deformed faster, and evidence in the signals suggests that these differences offset each other almost exactly.

## Quail

I'll confine myself to two comments on this research line. First, Diehl and Kluender write that "Nonhumans obviously lack specific adaptations for perceiving speech, and it is particularly difficult to conceive of them recovering the underlying articulatory gestures. . . ." This does not make sense, however, as the following substitutions may illustrate: *Humans obviously lack specific adaptations for perceiving bouncing balls and it is particularly difficult to conceive of them recovering their underlying trajectories.* Perceivers do not need specific adaptations for perceiving most things; they only need to be able to pick up structure in media that has been caused by "to-be-perceived" properties of those things. And phonetic gestures no more "underlie" the acoustic signal than bouncing balls underlie the reflected light. Quail may detect gestures taking place in the vocal tract, because that is what the acoustic information in a speech signal is providing information about.

The other comment is that I do not find this line of research sufficiently informative about speech perception and its neurological support to justify the sacrifice to science that the quail are required to make. If animals can do the classification task, what have we learned? We have learned that the information supporting performance must be available in the signal (but we can learn that in other ways), and we have learned that some particular species of nonhuman has an auditory system that is sensitive to the kind of acoustic structure that phonetic gestures produce. We have not learned from it that humans have no specialization for speech perception. That inference would only be justified if we knew that neurological specializations for perceiving something invariably make its perception possible when it would be impossible without the specialization. I do not think we know that about neurological specializations. The research shows only that quail pick up information necessary to do the classification task. This obviously is not the same as detecting the linguistic significance of the phonetic place distinction. Accordingly, we do not know that quail recover the kind of information that, ostensibly, is the domain of any specialization for speech. On the other hand, the outcome of this line of research is uninterpretable if the animal does not succeed in the classification task. Possibly, the information is not in the signal; possibly it is, but the animal's auditory system is not sensitive to it; perhaps the animals could not be motivated to learn the classification.

## CONCLUDING REMARKS

In my view, Diehl and Kluender are correct to argue that perceptual constraints guide development of segment inventories in languages, but they are mistaken in

their view that talker-centered constraints are considerably weaker than perceptual ones. I disagree that investigation of the factors that shape sound inventories of languages reveals that the things being perceived are acoustic or even auditory things, not behavioral ones. Finally, in my view, understanding the objects of speech perception will not be furthered by comparing speech perception to perception of signals with no well-understood perceived distal source or by asking whether nonhumans can led to exhibit human-like classifications.

## ACKNOWLEDGMENTS

Preparation of this article was supported by a fellowship from the John Simon Guggenheim Foundation and by National Institutes of Health–National Institute of Child Health and Human Development Grant No. HD-01994 and National Institute of Neurological and Communicative Disorders and Stroke Grant No. NS-13617 to Haskins Laboratories. I thank Patricia Kuhl, William Mace, Elliott Saltzman, and George Wolford for their comments on an earlier draft of this article.

## REFERENCES

Abbs, J., & Gracco, V. (1984). Control of complex motor gestures: Orofacial muscle responses to load perturbations of the lip during speech. *Journal of Neurophysiology, 51,* 705–723.

Baer, T., Löfqvist, A., McGarr, N., & Story, R. (in press). The cicrothyroid muscle in voicing contrasts. *Journal of the Acoustical Society of America.*

Bell-Berti, F., & Harris, K. (1979). Temporal patterns of coarticulation: Some implications from a study of lip rounding. *Journal of the Acoustical Society of America, 71,* 449–454.

Bell-Berti, F., Raphael, L., Pisoni, D., & Sawusch, J. (1979). Some relationships between speech production and speech perception. *Phonetica, 36,* 373–383.

Browman, C., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook, 3,* 219–254.

Browman, C., & Goldstein, L. (in press). Tiers in articulatory phonology with some implications for casual speech. In J. Kingston & M. Beckman (Eds.), *Papers in laboratory phonology, I: Between the grammar and the physics of speech.* Cambridge, UK: Cambridge University Press.

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

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York: Harper & Row.

Clements, G. N. (1985). The geometry of phonological features. *Phonology Yearbook, 2,* 225–252.

Comrie, B. (1980). Phonology: A critical review. In B. Butterworth (Ed.), *Language production, I: Speech and talk* (pp. 297–334). London: Academic.

de Chene, B. (1985). *The historical phonology of vowel length.* New York: Garland.

Diehl, R., & Kluender, K. (1989). On the objects of speech perception. *Ecological Psychology,* this issue.

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

Fowler, C. (in press). Listener-talker attunements in speech. In T. Tighe, B. Moore, & J. Santrock (Eds.), *Human development and communication sciences.* Hillsdale, NJ: Lawrence Erlbaum Associ-

ates, Inc.

Fowler, C. A., & Rosenblum, L. D. (in press). The perception of phonetic gestures. In I. G. Mattingly & M. Studdert-Kennedy (Ed.), *Modularity and the motor theory of speech perception.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Fowler, C., & Turvey, M. T. (1980). Immediate compensation for bite-block speech. *Phonetica, 37,* 3-7-326.

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

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

Harris, K., Lysaught, G., & Schvey, M. (1965). Some aspects of the production of oral and nasal labial stops. *Language and Speech, 8,* 135-147.

Honda, K. (1981). Relationship between pitch control and vowel articulation. In D. Bless & J. Abbs (Eds.), *Vocal-fold physiology* (pp. 286-297). San Diego: College-Hill.

Hyman, L. (1973). The role of consonant types in natural tonal assimilations. In L. Hyman (Ed.), *Consonant types and tone* (pp. 152-179). Los Angeles: University of Southern California.

Keating, P. (1985). Universal phonetics and the organization of grammars. In V. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 115-132). Orlando, FL: Academic.

Keating, P. (in press). Mechanisms of coarticulation: Articulatory evidence. In J. Kingston & M. Beckman (Eds.), *Papers in laboratory phonology, I: Between the grammar and the physics of speech.* Cambridge, UK: Cambridge University Press.

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

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

Kiparsky, P. (1968). Linguistic universals and linguistic change. In E. Bach & R. Harms (Eds.), *Universals in linguistic theory* (pp. 170-202). New York: Holt, Rinehart, & Winston.

Lehiste, I. (1971). Temporal compensation in a quantity language. *Seventh international congress of phonetic sciences* (pp. 929-937). The Hague, The Netherlands: Mouton.

Lindblom, B. (1986). On the origin and purpose of discreteness and invariance in sound patterns. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 493-510). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lindblom, B., Lubker, J., & Gay, T. (1979). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics, 7,* 147-161.

Lindblom, B., & Rapp, K. (1973). Some regularities of spoken Swedish. *Papers in Linguistics from the University of Stockholm, 21,* 1-59.

Löfqvist, A., & Yoshioka, H. (1984). Intrasegmental timing: Laryngeal-oral coordination in voiceless consonant production. *Speech Communication, 3,* 279-289.

Nooteboom, S. (1973). The psychological reality of some prosodic durations. *Journal of Phonetics, 1,* 25-45.

Ohala, J. J. (1974). Experimental historical phonology. In J. Anderson & C. Jones (Eds.), *Historical linguistics, II: Theory and description in phonology* (pp. 353-389). Amsterdam, The Netherlands: North Holland.

Ohala, J. J. (1981). The listener as a source of sound change. In C. S. Masek, R. A. Hendrick, & M. F. Miller (Eds.), *Papers from the parasession on language and behavior* (pp. 178-203). Chicago: Chicago Linguistics Society.

Ohde, R. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. *Journal of the Acoustical Society of America, 75,* 224-240.

Öhman, S. (1967). Peripheral motor commands in labial coarticulation. *Speech Transmission Laboratory: Quarterly Progress and Status Report, 4,* 30-63.

Pattee, H. H. (1970). The problem of biological hierarchy. In C. H. Waddington (Ed.), *Towards a theoretical biology* (pp. 117-135). Chicago: Aldine.

Raphael, L. (1975). The physiological control of durational differences between vowels preceding

voiced and voiceless consonants in English. *Journal of Phonetics, 3,* 25–33.
Reinholt Petersen, N. (1986). Perceptual compensation for segmentally-conditioned fundamental-frequency perturbations. *Phonetica, 43,* 31–42.
Riordan, C. (1977). Control of vocal-tract length in speech. *Journal of the Acoustical Society of America, 62,* 998–1002.
Sagey, E. (1986). *The representation of features and relations in non-linear phonology.* Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.
Saltzman, E. (1986). Task-dynamic coordination of the articulators. In H. Heuer & C. Fromm (Eds.), *Generation and modulation of action patterns* (Experimental Brain Research Series 15, pp. 129–144). New York: Springer-Verlag.
Silverman, K. (1987). *The structure and processing of fundamental frequency contours.* Unpublished doctoral thesis, Cambridge University, Cambridge, UK.
Summers, W. V. (1987). Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. *Journal of the Acoustical Society of America, 82,* 847–863.
Sussman, H., MacNeilage, P., & Hanson, R. (1973). Labial and mandibular dynamics during the production of bilabial consonants: Preliminary observations. *Journal of Speech and Hearing Research, 16,* 397–419.
Tuller, B., & Fitch, H. (1979). Preservation of vocal-tract length in speech: A negative finding. *Haskins Laboratories Status Reports on Speech Research, SR-59/60,* 237–244.
Warren, W. (1985, June). *Environmental design as the design of affordances.* Paper presented at the Third International Conference on Event Perception, Uppsala, Sweden.
Weiss, P. (1941). Self-differentiation of the basic pattern of coordination. *Comparative Psychology Monographs, 17,* 21–96.
Wood, S. (1982). X-ray and model studies of vowel articulation. *Lund University Working Papers, 23,* 1–50.
Zemlin, W. R. (1968). *Speech and hearing science: Anatomy and physiology.* Englewood Cliffs, NJ: Prentice-Hall.