Reply to Commentators

Randy L. Diehl
University of Texas

Keith R. Kluender
University of Wisconsin

The commentaries have helped us to consider ways in which our various arguments may be clarified and strengthened, and they have also raised certain general problems that any adequate theory of speech percepts must address. However, we do not believe that the criticisms pose a fundamental challenge to our theoretical framework or to our interpretation of key experimental results.

Because there was relatively little overlap among the three commentaries, they will be addressed separately.

CAROL FOWLER

Fowler suggests that some of our criticisms of the direct-realist account of speech perception are based on our misunderstanding of her use of the notion of “phonetic gesture.” Moreover, she argues that key examples used to support our position are either factually incorrect or subject to counterreinterpretation, and that, in general, we have tended to ignore production-oriented accounts of the phenomena that we discuss.

Perceptual Explanation in Phonetics and Phonology

Although Fowler agrees with us that “perceptual constraints guide the development of sound inventories and of phonological processes in languages,” she
equivocates somewhat as to the nature of these perceptual constraints. In her
discussion of point vowels, she seems to endorse an articulatory version of
Lindblom's perceptual dispersion principle: "Back vowels have large front
cavities, and the tendency for languages to round those vowels enhances that
distinctive property." But elsewhere she states: "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" (her italics). Instead of perceptual
distinctiveness and least effort as guiding forces in the development of sound
inventories, Fowler appeals to misperception ("misparing by listeners") as a
basis for sound change in languages (Ohala, 1981). She points out that,
although some sound changes may enhance intelligibility, that is not true in
general.

We suggest that Fowler has not distinguished two quite different sets of
linguistic phenomena: (a) persisting and more-or-less universal properties of
sound systems on the one hand and (b) language-specific sound changes on the
other. Lindblom's principles of auditory dispersion and least effort predict the
former class of properties (e.g., that almost all languages use the point vowels /i/,
/a/, and /u/ and the stop consonants /b/, /d/, and /g/) with reasonably high
accuracy (Lindblom, MacNeilage, & Studdert-Kennedy, 1988), but these prin-
ciples have little to say about language-specific sound change (except to stipulate
that, whatever sound change occurs in a language, sufficient perceptual distinc-
tiveness and ease of articulation must be maintained). By contrast, Ohala's
(1981) theory of sound change (resulting from auditory misperception) can
account for such phenomena as the trading of a voicing distinction for a tonal
distinction in Punjabi (Fowler's example), but it offers no basis for expecting
universal structural tendencies such as the prevalence of point vowels and the
consonants /b/, /d/, and /g/. The two types of theory are designed to explain
mostly nonoverlapping classes of phonetic/phonological facts.

The auditory enhancement hypothesis is an elaboration of Lindblom's dis-
persion principle, and, not surprisingly, most of the phonetic/phonological facts
that form the basis of our argument are of the persisting and universal variety.
Therefore, we do not think that they can generally be subsumed under Ohala's
theory of sound change. As for language-specific sound changes, we think
Ohala's account is largely correct. But we note that Ohala is no direct realist (see
Ohala, 1981, 1986). Like us, he views the objects of speech perception as
acoustic/auditory events, not as phonetic gestures. So the applicability of his
theory of sound change to a subdomain of phonetic/phonological facts is by no
means inconsistent with our overall approach.

On the other hand, we think it may be difficult to square Ohala's notion of
systematic misperception with a strong version of direct realism. Direct realists
do not deny that illusions occur, but they typically attribute them to artificial or
impoverished stimulus information or to specific deficiencies in perceptual
organs caused by abnormal stimulation.¹ We do not believe that these types of conditions can be plausibly invoked as a cause of systematic misperception (by a significant portion of a language community) that could give rise to sound change.

Phonetic Gestures and the Many-to-One Problem

We pointed out in the target article that the mapping between articulatory events (whether defined in terms of articulatory gestures or vocal-tract area functions) and acoustic signals is many-to-one. The same acoustic/auditory event may be realized by a variety of articulatory means, and this is a principal reason why we view acoustic/auditory events, and not articulatory events, as the objects of speech perception. However, Fowler suggests that the many-to-one problem is mitigated if the articulatory events are described in the proper way, namely, as phonetic gestures. A defining property of a phonetic gesture is equifinality; although the phonetic end in question may be achieved by various specific means, all such means are instances of the same phonetic gesture to the extent that they are implemented by the same "coordinative structure" or "synergy." For example, a variety of different positionings of upper lip, lower lip, and jaw, all implementations of the same coordinative structure, can achieve the same phonetic end of closing the lips, and thus they are all instances of the same phonetic gesture. On this view of phonetic gestures as equivalence classes, the many-to-one problem seems to disappear.

We would find none of this objectionable if phonetic ends were defined acoustically. But for Fowler, phonetic ends are articulatory (e.g., closing the lips). To see whether Fowler's account provides a fully adequate solution to the many-to-one problem, let us consider two examples: the production of the vowel

¹For example, Gibson (1966) offered the following tentative list of reasons for misperceptions and illusions in the case of visual perception:

First, the available stimulus information for perception can be inadequate. The energy may be inadequate. The energy may be minimal, or the structure of an array may be blurred, or it can be masked, or the information in structure may be contradictory. The interval during which energy is available may be cut short, or the angular size of the optic array may be narrowed down. The spatial and temporal structure of light or sound can be displaced, or biased, or distorted. Mirrors, prisms, and lenses can alter the structure of light and modern electronic gadgetry can alter the structure of sound.

Second, the physiological process of information pickup can be deficient. Even normal perceptual organs fail to work at high intensities of energy. Their capacity can be overstrained. They get out of calibration after abnormal stimulation and the recovery process then yields aftersensations. The physiological action of the receptors as such may be so obtrusive as to distract the observer. The action of a whole system may be subject to perceptual adaptation followed by after-effects. And finally there are errors of attention due to false expectations. (p. 318)
/ʃ/ and the production of certain other vowels when the jaw position is fixed by a bite block. As Diehl and Kluender (1989) noted, /ʃ/ is normally produced with constrictions at the lips, the midpalate, and the lower pharynx (Lindau, 1985; Ohala, 1985). These constrictions occur at just the locations where antinodes exist in the standing volume–velocity waveform corresponding to the third formant, and, accordingly, each independently contributes to the distinctive frequency lowering of that formant (Fant, 1960). (The many-to-one mapping between articulatory events and acoustic events is revealed here by the fact that the same degree of third-formant lowering can be achieved by varying amounts of constriction at the three vocal-tract locations.) Now, by the auditory enhancement hypothesis, the particular combination of constriction locations used for /ʃ/ is naturally explained: Each contributes to the same distinctive acoustic property.

In the model of articulatory phonology developed by Browman and Goldstein (1986, in press), after Fowler (1977), there are separate synergies or coordinative structures for each phonetic gesture such as lip closing, alveolar constriction, and pharyngeal constriction. Although the model attempts to specify how coordination of articulators is achieved within a phonetic gesture, presently it says nothing about how such coordination is achieved across different gestures. In particular, the model offers no account of the coordination of the three phonetic gestures—lip constriction, palatal constriction, and pharyngeal constriction—that comprise the production of /ʃ/. (We note that most of the examples that we cited in the target article to illustrate the strategy of auditory enhancement involve similar cross-gestural coordination.)

Moreover, neither the model of Browman and Goldstein (1986, in press) nor the related task-dynamic model of Saltzman (1986) provides any rationale for the selection of what Fowler calls "phonetic ends," whether these are construed as individual phonetic gestures or as the ensembles of phonetic gestures that make up vowels and consonants. Where do phonetic ends come from? When this question was posed by Lindblom and MacNeilage (1986), some action theorists (Kelso, Saltzman, & Tuller, 1986) replied: "This is like asking a cosmologist: Where does the universe come from? The answer is that no one knows, and the level of abstraction necessary to even approach the problem is difficult to imagine" (p. 187). However, it is precisely the aim of the auditory enhancement hypothesis and of Lindblom's theory of adaptive dispersion (Lindblom, 1986; Lindblom et al., 1988) to provide a principled account of universal tendencies in the selection of segment inventories and of gestural ensembles that comprise favored segments.

It would be possible to stipulate that the individual phonetic gestures (and their corresponding coordinative structures) that make up the vowel /ʃ/ are nested with a larger synergy, namely, the one that achieves the phonetic end of lip constriction + palatal constriction + pharyngeal constriction. But this fails to explain the particular combination of nested gestures.
Our second example concerns the production of vowels when jaw position is fixed by a bite block. As we (1989) noted, the high vowels /i/ and /u/ are normally produced with the jaw in an almost closed position, whereas low vowels, such as /a/, are produced with a more open jaw. When a bite block prevents jaw closure in the case of /i/, the talker can compensate by elevating the tongue body and blade so as to approximate the normal area function (at least in the region of oral constriction) and produce an acoustically acceptable version of the vowel (Lindblom, Lubker, & Gay, 1979). Drawing an analogy to perturbation studies of lip closure, Fowler (1989) suggests that such compensations can be explained in terms of coordinative structures that implement equifinal phonetic gestures.

But what, then, is the phonetic end in the case of /i/ produced under both normal and bite-block conditions? It is insufficient to say that the end is to produce an /i/-like ensemble of phonetic gestures; one must be able to state on independent theoretical grounds what that ensemble of gestures consists of and how it might be realized under bite-block conditions.

A bite-block study by Gay, Lindblom, and Lubker (1981) sheds a good deal of light on the issue of whether phonetic ends are articulatory or acoustic. They found that the area functions of the vowels /i/, /a/, /u/, and /o/ produced with bite blocks approximate those of normally produced vowels in the region of the major oral constriction. Elsewhere, the area functions tend to differ markedly. The results of a companion vocal-tract modeling experiment help to explain this pattern of area-function similarities and differences. Perturbations in the area function were found to cause the greatest acoustic changes in regions where the area function has the smallest values, that is, in regions of the major constriction. Thus, in approximating a normal oral constriction at the expense of significant discrepancies from the normal cavity configuration on either side of the constriction, talkers with bite blocks are able to achieve the least possible deviation from normal vowel formant patterns.

We can imagine explaining this pattern of results by assuming that it is the oral constriction per se that constitutes the phonetic end in vowel production. An /i/-like palatal constriction does, after all, constitute a phonetic gesture in the model of Browman and Goldstein (in press). However, this would disregard other gestural components of /i/ such as lip spreading. In addition, such an account would be at odds with Fowler’s (1989) emphasis on cavity size (e.g., anterior to the constriction) as a main distinctive property of vowels. In sum, we do not believe that the many-to-one mapping problem can be eliminated by appealing to the notions of “phonetic gesture” and “coordinative structure.” Saltzman’s (1986) model of task dynamics and the Browman–Goldstein (Browman & Goldstein, 1986, in press) model of articulatory organization offer attractive solutions to the problem of how individual phonetic gestures can be articulatorily realized by varying means. But neither model addresses the issues that are the principal concern of Lindblom’s theory of adaptive dispersion and
of the auditory enhancement hypothesis: What governs the selection of phonetic gestures in the first place and what constrains their covariation in the production of phonetic segments?

Articulatory Versus Auditory Constraints on Sound Production

Let us first restate and clarify our general position with respect to possible articulatory constraints that shape sound inventories. We agree with Lindblom (1986, Lindblom et al., 1988) that universal tendencies in the structure of segment inventories and in the detailed characteristic of on-line speech production result from the joint operation of a principle of auditory dispersion and a principle of least articulatory effort (see Diehl, in press, Diehl & Kolodzey, 1981). Some phonetic tendencies (e.g., the preference for peripheral over central vowels) are clearly a reflection of the dispersion principle, but other tendencies (e.g., the infrequent use of click sounds) are more likely due to the least-effort constraint. In the target article, we focused almost entirely on the dispersion principle (in the elaborated form of the auditory enhancement hypothesis) because the question we were addressing was not "What are all the constraints on the formation of segment inventories?" but rather "What are the objects of speech perception?" Our acceptance of the principle of least effort (as formulated by Lindblom) simply has no bearing on the issue posed by the latter question, so we relegated our discussion of the principle to a brief footnote. In any case, we do believe that most universals of gestural covariation are largely the result of a strategy of auditory enhancement.

However, to the extent that other possible articulatory constraints compete with the auditory enhancement hypothesis as explanations of a particular gestural/acoustic covariation universal, we are obligated to examine them very carefully. We have made a consistent effort to be as knowledgeable as possible about hypothesized production constraints and to accept the auditory enhancement account only when we are reasonably satisfied that alternative production-oriented explanations are incompatible with the data.

Next, we review the various instances of phonetic covariation for which Fowler offers explanations different from our own. We apologize to the reader in advance for having to concentrate on some technical details.

Tongue position and lip configuration in vowel production. In the target article, we suggested that the covariation between tongue position and degree of lip rounding could be explained in terms of auditory enhancement, because retracting the tongue and rounding the lips both contribute to a distinctive lowering of F2 frequency. We admitted, however, that the same gross facts could also be explained in terms of articulatory distinctiveness, because the two gestures jointly lengthen the front cavity. Our follow-up argument was that
when one examines the fine structure of this covariation as well as other gestural covariates, such as palatalization and larynx lowering, an articulatory version of the dispersion principle cannot account for the facts.

We noted that the lip rounding gesture actually consists of two components, lip protrusion (which lengthens the front cavity) and constriction of the labial aperture, and that both of these components contribute independently to F2 frequency lowering. Fowler replies that "these gestural components are articulatorily coupled, because the orbicularis oris muscle surrounding the lips both 'closes the mouth and puckers the lips' (Zemlin, 1968, p. 258)." It appears, however, that one can easily protrude the lips with a much more open labial aperture than is characteristic of, say, the back vowel /u/. Quite a large number of facial muscles other than the orbicularis oris are known to affect lip configuration (Dickson & Maue-Dickson, 1982), and the consequences of contracting the orbicularis oris can be modified by the diverse actions of these other muscles. The potential separability of lip protrusion and lip constriction is proven by the case of Swedish rounded vowels. One subclass of these vowels is referred to as inrounded and shows a high degree of lip constriction with very little lip protrusion, whereas the subclass of outrounded vowels shows considerable lip protrusion with relatively little constriction of the labial aperture (Harris, Hirose, & Hadding, 1975).

With regard to lip spreading in the case of /i/, our point was that the outward flaring of the vocal-tract opening enhances the F2 frequency raising effects of tongue fronting and palatalization (Ohala, 1980). Thus, just as lip protrusion and constriction conspire to lower the frequency of F2, lip spreading raises it by both shortening the front cavity and by enlarging the labial aperture. And, just as talkers can easily protrude the lips without forming an /u/-like constriction, they can also shorten the vocal tract (by placing the lips close to the teeth) without flaring the vocal-tract opening. The fact that language communities typically do not select these alternatives appears to demand an acoustic, not an articulatory, rationale.

**Vowel height and pitch variation.** In the target article, we noted that perceived vowel height is signaled not merely by the frequency of F1 but by the difference (in Bark units, an auditory transformation of frequency) between F1 and the fundamental frequency (F0) (Syrdal, 1985; Syrdal & Gopal, 1986; Traunmüller, 1981). We hypothesized that the well-known positive correlation between vowel height and pitch was attributable to the talker actively raising F0 for high vowels so as to reduce their F1 – F0 Bark difference and thus make them more distinct perceptually from low vowels, which have a large F1 – F0 Bark difference. In partial support of this hypothesis, we pointed out that two different accounts of vowel pitch differences based on alleged physical or physiological constraints had been ruled out experimentally.

Against our hypothesis, Fowler cites a study by Honda (1981) as supporting a
passive mechanical account of the pitch correlation with vowel height. Honda found correlations between posterior genioglossus activity, forward position of the hyoid bone, and vowel height. He suggested that contraction of the genioglossus during the production of high vowels pulls the hyoid bone forward, which, in turn, rotates the thyroid cartilage relative to the cricoid cartilage, placing greater tension on the vocal folds and raising F0.

Although Honda's data may be explained in terms of the mechanical coupling that he proposed, the data appear to be equally compatible with our claim that the vowel height/pitch correlation is under the talker's active control. Rather than pursuing this argument here, we refer to two additional pieces of evidence that we take to be incompatible with the purely passive interpretation of vowel pitch differences endorsed by Fowler. First, Gandour and Weinberg (1980) studied esophageal speech of laryngectomized patients and found pitch differences between high and low vowels similar in magnitude to those observed in normal subjects. Such patients use the esophagus as an accessory lung and the esophageal sphincter as a voicing source. Because esophageal speakers obviously lack laryngeal cartilages (as well as a hyoid bone), the observed pitch variation cannot be attributed to the anatomical couplings that Honda posited to explain vowel pitch differences in normal talkers.

Gandour and Weinberg (1980) appealed to possible aerodynamic factors to account for the vowel height/pitch correlation in both normal and esophageal talkers. Specifically, they hypothesized that the greater oral impedance of high vowels prompted talkers to exert greater respiratory force in producing them. This greater force would cause a heightened subglottal pressure and hence a higher F0. Unfortunately, the required difference in average subglottal pressure between high and low vowels simply has not been observed (see, e.g., Ladefoged, 1967). Alternatively, we suggest that esophageal speakers, like normal speakers, actively regulate pitch in order to enhance perceptually the vowel height distinction. Whereas normal talkers regulate pitch by varying the tension of the vocal folds, esophageal speakers do it by varying the tension of the esophageal sphincter.

A second piece of evidence against Honda's purely passive account of vowel pitch differences comes from a recent study by N. J. Dyhr at the Institute of Phonetics of the University of Copenhagen (Niels Reinholt Petersen, personal communication, August 10, 1988). If our interpretation of vowel pitch differences is correct, then we predict that high vowels will be produced with greater activation of the cricothyroid muscle (the primary muscle involved in active control of pitch). Dyhr obtained exactly this result. As compared to low vowels, high vowels were associated with earlier and greater levels of cricothyroid activity. We view this as reasonably strong support for our hypothesis.

Vowel length differences before voiced and voiceless consonants. We proposed that the nearly universal tendency for vowels to be longer before voiced
than before voiceless consonants reflects a strategy of speech communities to enhance perceptually the closure-duration cue for the voicing distinction. A longer preceding vowel makes a short closure interval appear even shorter (hence, more voiced), whereas a shorter vowel makes a long closure interval seem even longer (hence, more voiceless). In support of this durational-contrast hypothesis, Kluender, Diehl, and Wright (1988) showed that, for /aba/-/apa/ stimuli varying in medial closure duration and for square-wave stimuli that temporally mimic these speech stimuli, a longer initial segment causes a reliable shift in subject's two category labeling boundaries toward greater medial gap durations.

Fowler criticizes our account on several grounds. First, based on data from Chen (1970), she claims that the voicing-conditioned vowel-length effect (VLE) is considerably larger in English than in other languages (presumably ruling it out as a deliberate perceptual enhancement effect in those other languages). We quote from Kluender et al. (1988):

In fact, however, the degree to which English is exceptional in respect to the VLE has almost certainly been exaggerated. The problem with the evidence cited by Chen (1970) is that the cross-language variable was virtually always confounded with other variables known to influence the size of the VLE. The English materials (Chen, 1970; House & Fairbanks, 1953; Peterson & Lehiste, 1960; Zimmerman & Sapon, 1958) consisted almost exclusively of monosyllabic words, whereas the non-English materials (Chen, 1970; Fintoft, 1961; Zimmerman & Sapon, 1958) consisted largely, and in most cases entirely, of disyllabic words, with the critical voicing contrast occurring at the end of the first syllable. Klatt (1973) has shown that, in English utterances, the VLE (defined as the ratio of vowel lengths before voiced and voiceless consonants) is reduced from a value of about 1.5 in monosyllables to a value of about 1.27 in disyllables. The latter value is comparable to the disyllabic VLE ratios reported in various other languages [e.g., 1.30 in Korean, 1.22 in Russian (Chen, 1970), 1.22 in Norwegian (Fintoft, 1961)].

As for whether such vowel-length differences are detectable, Fujisaki, Nakamura, & Imoto (1975) found that the just-noticeable difference for vowel length in disyllables was 10 ms, given a segment duration of 100 ms. Thus, even for a language such as French, which has relatively low VLE ratios of about 1.15 in disyllables (Chen, 1970) and 1.35 in monosyllables (Mack, 1982), the vowel-length differences should be detectable. (pp. 162–163)

Nonetheless, Fowler is persuaded that the VLE is exaggerated in English partly because, among the Russian materials, there were some monosyllables that could be compared to the English monosyllables. However, such a comparison is not informative because, in utterance-final position, the Russian voicing distinction is phonologically neutralized (Lunt, 1968). That is to say, despite some residual phonetic differences (e.g., a small vowel-length difference), the surface realizations of phonologically voiced and voiceless consonants in
utterance-final position are quite similar, both being phonetically voiceless.2 The point is that comparing the vowel-length effect in English and Russian monosyllables involves an even more serious confound than comparing English monosyllables with non-English disyllables. Thus, the claim that English is highly exceptional in respect to the vowel-length effect does not appear to be warranted by the facts.

Fowler's second point is that Chen (1970) showed that the difference in vowel duration in non-English languages is probably due to a more forceful and faster closing gesture in the case of voiceless consonants, which is required because of the greater intraoral air pressure. Kluender et al. (1988) critically evaluated four different production-oriented accounts of the vowel-length effect (as well as one other perception-oriented account). Drawing on earlier criticisms by Lisker (1974) and Javkin (1976), we (1988) made the following comments about Chen's (1970) hypothesis:

Although Chen (1970) reported velocity measurements that were broadly consistent with this account, a later study by Sussman, MacNeilage, & Hanson (1973) provided equivocal support at best. Following /ɛ/ and /æ/, the net velocity of the closing gesture was slightly (about 3%-4%) greater for /p/ than for /b/. However, following /i/, the closing velocity was actually greater for /b/ than for /p/. This last outcome is troublesome for Chen's account, since /i/ appears to be no less susceptible to the VLE than other vowels (Peterson & Lehiste, 1960; Sharf, 1962).

Furthermore, as Javkin (1976) pointed out, Chen's hypothesis predicts that the VLE should be smaller for high vowels such as /i/ than for low vowels such as /a/, because a closure rate difference should have less effect with shorter articulatory trajectories. In fact, however, the mean ratio of vowel duration before voiceless and voiced stops is almost exactly the same for /i/ (.65) as for /a/ (.67) (Peterson & Lehiste, 1960; Sharf, 1962).

Javkin (1976) also suggested that Chen (1970) may have confused cause and effect. If closing gestures are faster before voiceless consonants, this may simply reflect a requirement of shorter vowel durations in that environment. (p. 155)

Fowler next suggests that the voicing-conditioned vowel-length effect must be "viewed in a broader context, this one of manifold 'compensatory shortenings.'" Compensatory shortening is said to occur, for example, when the measured duration of the vowel in the word pipe (/paip/) is reduced with the addition of a consonant as in piped (/paip/) Kluender et al. (1988) reviewed the literature on such compensatory shortening effects (Chen, 1970; Fowler & Tassinary,

---

2 An experiment by Fourakis and Iverson (1984) suggested that the incomplete neutralization of the Russian final voicing distinction evidenced in Chen's (1970) study may have been artificial. They found that a similar incomplete neutralization in German occurred only under conditions that favored hypercorrect spelling pronunciation (Labov, 1966). Under more natural speaking conditions (in which the subject's attention was drawn away from the orthographic structure of the word list), word-final voicing neutralization was essentially complete.
1981; Lindblom, Lyberg, & Holmgren, 1981) and concluded that the vowel-length effect (as well as related contrastive phonological phenomena involving inverse durational variation) cannot be viewed simply as instances of compensatory shortening. In all the studies of compensatory shortening just cited, the addition of a consonant or consonant cluster to either end of a consonant-vowel-consonant syllable reduced vowel duration by only a small amount relative to the increase in total consonant duration. The ratio of vowel shortening to consonant cluster lengthening was typically less than 0.1. By contrast, in the voicing-conditioned vowel-length effect, the ratio of vowel shortening to consonant lengthening is typically greater than 1.0 (Port, 1981). That is, the two effects differ in size by at least an order of magnitude. Thus, inverse durational variation in cases where duration is a phonologically contrastive property bears little similarity to such variation in cases where duration is not contrastive (e.g., in cases of compensatory shortening).

Fowler’s final arguments on this subject concern the interpretation of the close parallel between the speech and nonspeech results of Klunder et al. (1988). She suggests that the nonspeech boundary shift (shown in Figure 1 of the target article) might be explained by the fact that there are many events that show negative durational covariation (e.g., time-to-bounce and bounce durations of a rubber ball hit with a paddle at varying force levels), and subjects could presumably have such a model of durational patterning in mind when they label the square-wave patterns. However, because the subjects do not hear the square-wave stimuli either as speech or as bouncing balls but rather as “computer sounds,” “video game noises,” and the like, it is difficult to understand why they would appeal to models of inverse durational patterning appropriate to a restricted subset of mechanical phenomena. We believe that a large and varied sample of events in the world would yield patterns of durational correlation with coefficients spanning the full range between −1 and +1. The point is that citing one particular kind of durational covariation to provide a post hoc account of our square-wave labeling results does not amount to a convincing explanation. Moreover, it is not obvious that the model of durational patterning to which Fowler appeals actually predicts the labeling results in question. Whether it does so depends on how the model is applied by listeners in making their category judgments, and this has been left unspecified. In the end, Fowler concludes that comparisons of perceptual performance on speech and acoustically analogous nonspeech stimuli are uninterpretable. We would agree that the observed parallels in performance are not clearly interpretable from a direct-realist perspective. However, they are interpretable from our theoretical viewpoint; indeed, our theory predicts them.

Quail

Fowler states that it does not make sense for us to write that “nonhumans obviously lack specific adaptations for perceiving speech, and it is particularly
difficult to conceive of them recovering the underlying articulatory gesture." The quoted passage was directed against both motor-theoretic (first clause) and direct-realist (second clause) accounts of speech perception, and it makes sense (i.e., has intelligible content) just to the extent that motor-theoretical statements like the following make sense: "perception of gestures occurs in a specialized mode, different in important ways from the auditory mode, responsible also for the production of phonetic structures, and part of a larger specialization for language" (Liberman & Mattingly, 1985, p. 3).

Next, Fowler suggests that studies of speech perception in nonhuman subjects are not informative because performance parallels between humans and nonhumans do not imply that humans lack a specialization for speech perception. However, we did not claim that such parallels logically exclude a specialized mode of speech perception in humans. Rather, we claimed that such parallels remove much of the theoretical motivation for positing such a mode. For example, Liberman and Mattingly (1985) noted that they were led to the motor theory by the finding that humans are able to categorize speech sounds despite context-conditioned acoustic variability. But the finding that quail can do the same (Kluender, Diehl, & Killeen, 1987) eliminates noninvariance as a sufficient reason to appeal to a motor-theoretic explanation.

The most recent evidence offered to support a specialized mode of speech perception is a phenomenon referred to as "duplex perception" (Liberman & Mattingly, 1985; Whalen & Liberman, 1987). Fowler and Rosenblum (in press) found that many of the relevant aspects of duplex perception could be demonstrated using stimuli derived from a nonspeech event (viz., the sound of a door closing), and they concluded that duplex perception of speech sounds need not be attributed to a specialized mode of speech perception. We agree with this conclusion, and we note that the logical form of the argument is similar to that of our argument concerning the quail results. (Of course, quite different inferences are drawn about the objects of speech perception in the two cases).

Parallels between human and nonhuman perception of speech sounds are not only problematic for the motor theory, they are also difficult to square with direct realism. We guess that many ecological psychologists would be reluctant to accept Fowler's argument that quail, no less than humans, are capable of recovering human phonetic gestures from the speech signal. Speech has no natural ecological significance for quail, and ecological significance has always been cited by direct realists as a principal criterion for what counts as an object of perception (Gibson, 1966, 1979). But even if one dispenses with the criterion of ecological significance, there are other, more compelling reasons to reject the possibility that nonhumans are capable of directly perceiving human articulatory events.

In the target article, we suggested that a direct-realist account of visual detection of surface layout is plausible, but that such an account applied to speech perception is not. Although we applaud Fowler's attempt to promote a
"universal theory of perception," visual detection of surface layout is different in crucial respects from auditory detection of sound sources, and we think these differences pose a significant challenge to Fowler's universal version of direct realism. For direct realism to work, the mapping between informational media and distal events must be (virtually) unique and unambiguous. In the case of optical specification of surface layout, this essential condition is typically satisfied (assuming certain general constraints such as rigidity). An observer free to explore his or her environment can, in principle, acquire unambiguous information about the layout and motion of environmental surfaces. However, in the case of sound sources, it is simply not true that the acoustic signal uniquely specifies (even in principle) the kinematics and dynamics of the sound-producing event. Given an acoustic signal, there are, in general, an infinite set of ways that it could be produced. This is true even of the simplest kinds of sounds. The same frequencies produced by, say, a vibrating string can be generated by an infinite set of different combinations of string tension and length. And the ambiguity only gets worse when the sounds become more complicated.

We have already noted the ambiguity of acoustic specification of sound sources in the case of speech. But, in our discussion of the many-to-one mapping problem, we limited ourselves to natural and actualizable vocal-tract shapes and gestural types. If this restriction is eliminated and we consider also artificial and nonactualizable vocal-tract shapes and gestures (i.e., arbitrary specification of area functions), then the many-to-one problem becomes even less tractable (see, e.g., Ladefoged, Harshman, Goldstein, & Rice, 1977). Now, this enormous expansion of the many-to-one problem for speech sounds may conceivably be avoided by humans because they "know" (e.g., via somesthesia) which vocal-tract shapes and gestures are actualizable. But nonhumans do not have this recourse. Thus, the view that quail can directly perceive human articulatory events is even more difficult to sustain than the view that humans can do so. This is one reason why we must reject the claim that performance parallels between humans and nonhumans are uninformative.

ROBERT REMEZ

Remez prefaces his critique with an historical overview of the tendency of linguists to neglect auditory/acoustic explanations of phonetic facts and to favor instead articulatory explanations. We, too, initially held a strong articulatory bias (Diehhl & Kluender, 1987), but this yielded in the face of the kinds of experimental findings that were reviewed in the target article. Remez is unconvinced by these findings and offers several challenges to our claim that listeners are sensitive to the auditory properties of phonetic segments.
Linguistic Phonetics and the Objects of Speech Perception

Remez is correct to point out that theories of linguistic phonetics and theories of phonetic perception address rather different kinds of questions. Ideally, a theory of linguistic phonetics will address, among other matters, the question of what physical, motoric, and auditory factors constrain the selection of segment inventories. A theory of phonetic perception will be concerned with the question of what constitutes the perceptual objects and the question of how these objects are recognized. Although the two types of theory address somewhat different questions, it will generally be true (and it is particularly true in our case) that the choice of phonetic theory constrains the choice of perceptual theory and vice versa. For example, we have argued that an adequate phonetic theory must include Lindblom’s dispersion principle as a predictor of the structure of segment inventories and that this principle must apply to a phonetic space defined auditorily rather than articulatorily. A corollary of these phonetic claims is that the objects of speech perception are essentially auditory. In the other direction, independently motivated principles of auditory functioning are required to help define the auditory–phonetic space within which the dispersion principle applies (Lindblom, 1986). Remez refers to this kind of theoretical interdependency as “vaguely circular,” but we would reject any suggestion that it is viciously circular, because there are independent empirical grounds for both the phonetic and the perceptual claims. For example, the dispersion principle was initially confirmed for a purely acoustic representation of vowel sounds, a representation that did not presuppose any auditory model or auditorily defined distance metric (Liljencrants & Lindblom, 1972). On the other hand, the auditory models and distance metrics that are now typically used to define the auditory–phonetic space have been generated independently of any facts about speech perception or phonetic segment inventories (see, e.g., Bladon & Lindblom, 1981).

Because we argued that the objects of speech perception are auditory and that an auditorily based dispersion principle plays a significant role in the selection of segment inventories, Remez mistakenly concludes that we favor an exclusively auditory phonetic theory. As we noted in our reply to Fowler, we acknowledge the important role of nonauditory constraints (e.g., a least-effort principle) on the selection of segment inventories and on the detailed characteristics of segment production. To repeat: We focused on the auditory enhancement hypothesis because the question posed by the target article was “What are the objects of speech perception?” not “What are all the constraints on the selection of segment inventories?”

Remez points out that, like motor theorists and direct realists, we accept the common view that listeners perceive phonetic segments. (We differ from motor theorists and direct realists in that we think that, for the listener, phonetic
segments are largely auditory events.) Although we did not offer any justification for our segment-based approach, others (e.g., Lindblom, MacNeilage, & Studdert-Kennedy, 1983; Studdert-Kennedy, 1987) have presented what we think are reasonably convincing arguments for the reality of segments. In any event, it is important to point out that the validity of the auditory enhancement hypothesis does not rest on the particular choice of phonetic units (i.e., whether the units are construed as segments, diphones, syllables, or even something larger).

Three "Empirical Proofs"

Remez presents three claims ("empirical proofs") that he argues: (a) must be true if our theory is true and (b) are false. However, as we will try to show, none of the three claims follows from our theory, and so Remez's argument here does not amount to a falsification of the theory. Remez's first extrapolation from our theory is that "the equation of phonetic segments and sounds holds if listeners are able to identify auditory qualities of their phonetic experience." After listing examples of auditory properties (e.g., pitch, loudness, apparent duration, timbre), Remez concludes that "the meager evidence on the ability of listeners to experience elements of a speech signal as nonphonetic auditory impressions is not encouraging." First, it is important to understand that there is no such thing as a definitive list of auditory properties, and that the list offered by Remez (borrowed from Hirsch, 1988) can only be viewed as partial and illustrative. Although some of the qualities on this list are potentially relevant to phonetic perception, there are many other phonetically relevant auditory properties that are not included here. Of particular importance are the auditory properties corresponding to the time-varying spectral and excitation characteristics of speech sounds. In our own preliminary attempts to model computationally the auditory representation of speech, these latter properties assume a central role. In any case, it is fair to pose this question: Given that auditory properties such as "oboe-like" or "machine-like" are included in Hirsch's (1988) list, what could motivate the exclusion of properties such as "speech-like," or "vowel-like"?

Second we have not claimed that listeners generally experience speech sounds as "nonphonetic auditory impressions." That speech sounds are usually perceived phonetically has never been at issue. The issue is whether the phonetic percept is essentially articulatory, auditory, or something else.

Third, Remez seems to assume that the critical test of our theory is phenomenological (i.e., subjects' being able to identify their auditory impressions of speech sounds). But nowadays almost no psychological theory stands or falls on the basis of phenomenal description. Certainly, neither motor-theoretic or direct-realist accounts of the objects of speech perception could pass such a test (Remez, 1986). In the case of auditory percepts, the descriptive vocabulary available to naive subjects is much too limited to yield anything approaching a theoretically
adequate phenomenal analysis of speech sounds. However, it is nonetheless legitimate to ask: Are listeners aware, or can they be made aware, of auditory properties of speech sounds? The answer to this seems clearly to be yes, at least in certain cases. For example, Bladon and Lindblom (1981) found that a computational model of peripheral auditory processing (together with an auditory distance metric) predicted listeners judgments of the auditory similarity of vowels with a relatively high degree of accuracy. The same auditory distance metric was found earlier by Plomp (1975) to yield equally good prediction of auditory similarity judgments of nonspeech timbres.

Our reference to auditory qualities may be partly responsible for Remez's emphasis on phenomenal evidence as the proper basis for judging our theory. The more neutral term properties (which we also used) is perhaps less misleading. By auditory properties we mean aspects of the auditory representation of speech sounds that may influence speech discrimination and categorization performance, whether or not the listener is consciously aware of them or can describe them meaningfully to an experimenter. The notion of auditory properties can be made reasonably precise by referring to the output representations of speech sounds in a computational model of auditory processing (see, e.g., Bladon & Lindblom, 1981; Cooke, 1986).

Remez's second "empirical proof" is that the equation of phonetic segments and sounds holds "if each definitive auditory quality occurs whenever the phonetic element to which it corresponds occurs." This misrepresents our theory in two ways. First, recall that we explicitly rejected the notion that natural categories (including phonetic categories) typically have single necessary and sufficient (i.e., definitive) properties. Phonetic categories, like most other natural categories, are "polymorphous," a fact that appears to have little adverse effect on category acquisition in either humans or quail. Second, in his rejection of the "coincidence of auditory and phonetic impressions," Remez again restricts the domain of possible auditory properties of speech to simple psychoacoustic properties. In our view, phonetic percepts depend on the full range of auditory properties of speech, many of them quite complex. Consider, for example, our discussion in the target article of medial voicing distinctions. We noted that these distinctions are signaled by a variety of cues, many of which are mutually enhancing auditorily. Preceding vowel length and closure pulsing both enhance the closure-duration cue; closure pulsing enhances the vowel-length cue; pitch-contour differences enhance the closure-pulsing cue. Nowhere did we assert (as Remez claims we did) that "voicing is equivalent to psychoacoustic duration." Nor would we claim that "a 60 ms silence in speech should appear to be the same as a 60 ms silence in a tone-complex." (Perceived duration is known to be affected by stimulus properties other than physical duration.) What we claimed was this: If durational contrast accounts for the fact that a longer preceding vowel shifts the voiced/voiceless boundary toward longer closure durations, then we should observe an analogous shift for nonspeech stimuli with the same temporal
properties. The experiment that we performed was a fair and strong test of the durational-contrast hypothesis.

Remez's third "empirical proof" is that equation of phonetic segments and sounds holds "if speech perception fails when the auditory quality of a signal departs from likeness to the natural acoustic products of vocalization." His rejection of this follows from the fact that some listeners are able to extract the phonetic message from sine-wave analogs of speech (i.e., stimuli in which the amplitudes and center-frequency trajectories of the first three formants are modeled by time-varying sine waves). However, to the extent that listeners can recover the phonetic message from sine-wave speech, they must be doing so on the basis of its acoustic/auditory resemblance to real speech. Sine-wave speech preserves what most phoneticians would agree are the crucially informative acoustic properties of real speech, viz., its formant trajectories. Because the sine waves are time-varying in frequency and amplitude, their bandwidths are nonnegligible, and the filter characteristics of the auditory periphery expand their auditory bandwidths even further in the direction of real formant bandwidths. Nevertheless, sine-wave speech remains acoustically and auditorily unnatural. But, then, in the study by Remez, Rubin, Pisoni, and Carroll (1981), fewer than half of the subjects heard these stimuli as carrying a phonetic message when they were not coached to do so, and only 2 of 18 listeners correctly deciphered the message.

Remez believes that the ability to perceive sine-wave speech phonetically "could not depend on the specific likeness of the sound of the tones and the typical sounds of consonants and vowels." But, on the other hand, he admits that "the coherent variation of the sinusoids imitates that coarse-grain pattern of variation in the energy maxima [i.e., the formants] of the speech signal." So there is obviously some degree of acoustic resemblance (however limited), and there must be a corresponding degree of auditory resemblance, between sine-wave speech and the real speech patterns on which they are modeled. What basis other than this acoustic/auditory resemblance could support phonetic perception of sine-wave speech?

The fact that most listeners fail to perceive sine-wave speech phonetically unless coached to do so is interpreted by Remez as showing that "auditory impressions alone are inadequate to produce phonetic perception." Although the auditory support for phonetic perception may (typically) be inadequate in the highly artificial case of sine-wave speech, it clearly does not follow that such support is inadequate in the case of real speech. Moreover, our claim that phonetic segments are essentially auditory events for the perceiver in no way excludes the possibility of alternative categorizations of highly degraded stimuli. Being told that sine-wave patterns are "speech" rather computer bleeps makes it more likely that a listener will attend to phonetically relevant auditory information and, therefore, be able to extract the phonetic message. As we noted in the target article, it is necessary to distinguish between the notion of a space in
which auditory transforms of acoustic signals are represented and the notion of a functional partitioning of that space into categories. Presumably, in the case of degraded stimuli such as sine-wave speech, there are multiple partitionings of the corresponding regions of the auditory space.

Although the sine-wave speech results are not at odds with our theory, they do pose a challenge to both motor-theoretic and direct-realist accounts of speech perception. Whether listeners hear sine-wave speech phonetically or not, most are quite certain that it did not originate from a human vocal tract. The narrowness of the “formant” bandwidths serves to inform the listener that the acoustic source is not sufficiently damped to be flesh and bone. So the objects of perception here could not plausibly be human phonetic gestures or their underlying control structures.

Speech Perception by Nonhumans

Remez is right to express an attitude of caution regarding claims of behavioral homology between humans and nonhumans. Not every similarity in behavior warrants a presumption of common psychological processes: One would be loathe to suggest, for example, that a carnival chicken playing the piano shares much with Duke Ellington playing “Solitude.” Nevertheless, one cannot hold a theory of evolution to be true and, at the same time, reject the possibility that behavioral homologies exist. (By homology, we mean that the performances are both functionally similar and that they have common evolutionary origins.) We acknowledge that, in order for a claim of homology to be justified, rather extensive parallels should be demonstrated, and we ask the reader to consider the following examples.

1. Syllable-initial voicing. Budgerigars (Dooling, Okanoya, & Brown, 1988), chinchillas (Kuhl, 1981; Kuhl & Miller, 1978), and macaques (Kuhl & Padden, 1982) all evidence voicing boundaries at voice-onset-time values very near those found for human listeners (Lisker & Abramson, 1970) for each of the three places of articulation—labial, alveolar, and velar. As is the case for humans, discrimination for all three species is better for stimulus pairs that straddle a phonetic boundary than for pairs that do not. Recent data (Kluender, 1988) suggest that such similarities are far from superficial. The shift in human voice-onset-time boundaries at different places of articulation is apparently due to variation in F1-onset frequency (Summerfield & Haggard, 1977), and differences in F1-onset frequency result in analogous boundary shifts for Japanese quail.

2. Place of articulation. Macaques (Kuhl & Padden, 1983) discriminate stimuli along a synthetic series ranging between /ba/-/da/-/ga/ in a manner very similar to human performance (Mattingly, Liberman, Syrdal, & Halwes, 1971). Japanese quail (Kluender et al., 1987) learn to categorize natural consonant-
vowel syllables beginning with /d/ (vs. syllables beginning with /b/ or /g/), and this learning generalizes to syllables having different vowels (and, hence, acoustic shapes) than the training stimuli. Memorization of individual tokens (Greene, 1983) cannot easily explain this generalization performance.

3. Manner of articulation. Macaques (Stevens, Kuhl, & Padden, 1988) are not only able to discriminate /ba/ from /wa/, but, also, their discrimination performance is influenced by the duration of the following vowel in a manner very similar to what has been observed for human subjects (Miller, 1980; Miller & Liberman, 1979).

Additionally, in the target article, we described an experiment (Kluender & Diehl, 1987) in which quail use properties such as gender of talker and vowel quality together with voicing or place information to form categories. We noted that the internal category structure revealed by quail performance is remarkably like that reported in studies of human categorization. Quail seem to form prototypes (Posner & Keele, 1968; Rosch, 1978), and they show some preference for exemplars with which they have had experience (Medin & Schaffer, 1978).

In view of this list of striking similarities between human and nonhuman performance, we find the claim for behavioral homology to be scientifically conservative. To extend Remez's metaphor, if animal behavior is but a wax-like form of human behavior, it seems capable of withstanding a good deal of heat.

Remez believes that detailed similarity of performance, even across many different speech dimensions, does not warrant an inference of common perceptual mechanisms between humans and nonhumans. Of course, we do not claim that such an inference is logically necessary, only that it is fully in keeping with the principle of parsimony. It would be theoretically extravagant to hold that all of the observed parallels are strictly accidental. Notice that the Tinbergen example cited by Remez (a male stickleback confusing a red postal van for rival) provides no basis for behavioral similarity across species. A compelling counterargument to our inference of common auditory mechanisms must be able to rationalize the behavioral similarities that were reviewed here.

Remez cites the occurrence of evolutionary convergences (e.g., the chambered eye of octopus and vertebrates), as sufficient reason to doubt that common auditory mechanisms underlie parallels in speech perception performance between humans and nonhumans. When cross-species similarities arise from separate evolutionary sources, they are referred to as analogies. The term homology is reserved for anatomical or behavioral parallels that have a common evolutionary source. It is important to note that even in the case of analogies, the similarities across animal classes are almost never accidental. Rather, they reflect the fact that species from quite diverse evolutionary lineages sometimes come to occupy similar ecological niches and to have similar "ways of life" (Dawkins, 1987).

What is the likelihood that performance similarities among humans,
macaques, budgerigars, and quail reflect convergent evolution? If these diverse species can be said to share a common way of life, at least in respects relevant to possible convergent auditory capabilities, the commonalities must be very general indeed. The species certainly share an acoustic environment in which most of the energy occurs at frequencies below 6 or 7 kHz and in which the sound is primarily airborne. But these same very general features of the acoustic environment would have been present at virtually all points along the evolutionary trajectories of these species, presumably as far back as their common ancestor. In other words, there is no reason to suppose that the parallel selection pressures that might conceivably have resulted in convergent auditory evolution of, say, humans and quail would ever have been absent in the first place. Thus, the scenario of initial auditory divergence and subsequent reconvergence is far less plausible (and also less parsimonious) than one in which birds and mammals, including humans, retained many of the general properties of a common ancestral auditory system. This is why we believe that parallels in speech perception between humans and nonhumans constitute behavioral homologies rather mere analogies.

Notice that even if we are wrong and the parallels turn out to be analogies, there remains a strong likelihood that structural and functional commonalities underlie similarities in behavior. Despite diversity of evolutionary origins, most analogies exhibit a rather high degree of relevant structural and functional correspondence (Dawkins, 1987). Moreover, in the event that the parallels in question are analogies rather than homologies, we must presume that the convergent evolution was driven by similar selection pressures. This effectively excludes factors unique to human speech communication from playing a significant role in the selection process (at least to the extent that the selection process is convergent across species).

Distinctiveness, Simplicity, and Representational Reductionism

Remez concludes his critique with this characterization of our approach: "the auditory enhancement theory favors representational reductionism, parsimony, and distinctiveness as its perceptual predispositions." He believes that distinctiveness is a poor choice because, in fluent speech, talkers often depart from maximal distinctiveness. But, as we have acknowledged repeatedly, distinctiveness is not the only constraint governing segment inventories and speech production. Talkers willingly trade some perceptual distinctiveness for greater ease of articulation provided there is enough redundancy and signal-independent information in the communication situation to ensure reception of the message (Lindblom & Engstrand, in press). We think that phonologies are designed to be effective signaling systems even under unfavorable conditions of communication (e.g., under conditions of high background noise and low
redundancy), and this is why distinctiveness figures prominently as a constraint on segment inventories. However, under favorable conditions of communication, maximal distinctiveness is simply not required, and talkers opt instead for sufficient auditory contrast.

We confess that we do not know exactly what Remez means when he asserts that “parsimony” is a perceptual predisposition in our theory. Nowhere did we argue that phonetic segments are perceptually simple (whatever that means). As for the fact that there tend to be multiple allophonic variants of a phoneme class, nothing in the auditory enhancement hypothesis or Lindblom's theory of adaptive dispersion precludes this. In fact, allophonic variation has been shown to be a natural consequence of segment selection based on the principles of auditory distinctiveness and least effort (Lindblom et al., 1988).

In using the term *representational reductionism*, Remez (personal communication, October 5, 1988) means that we assume listeners are attuned to “sensory forms” rather than “object attributes” when they hear speech. If by *sensory forms* he means “auditory properties of phonetic segments,” and if by *object properties* he means “gestures or vocal-tract properties,” then he is correct in his characterization. However, we note that his usage of *representational reductionism* is somewhat different from that of Bever (1984), who defined it as the claim that “an animal [human or nonhuman] organizes each newly acquired behavior with the most concrete mechanisms available” (p. 62). For our part, we do not view auditory properties as any more or any less “concrete” than articulatory properties.

MICHAEL STUDDERT-KENNEDY

Although he applauds our “endeavor to redress the balance between auditory and articulatory accounts of the origin of speech sound patterns,” Studdert-Kennedy believes that we hold “too narrow a view of speech function.” He claims, in particular, that our theory fails to address the important question of how children learn to speak through imitation.

Amodal Objects of Speech Perception

According to Studdert-Kennedy, the speech percept is “an abstract, amodal, perceptuomotor structure” consisting of an “isomorphism” between the neural patterns activated by speech when it is perceived and the neural patterns that control the production of speech. He further asserts that “this isomorphism is clearly necessary, because imitation of an act requires that the perceptual process induce in the imitator a neuromotor control structure isomorphic with that which produced it.”

We would distinguish between weak and strong version of such a perceptuomotor isomorphism. The weak version implies simply that, because
there is a fairly direct correspondence between articulatory variation and the resulting acoustic/auditory variation, the neural codings of these two types of variation will also stand in fairly direct correspondence. The strong version implies, in addition, that this mapping between acoustic/auditory and articulatory representations is perceptual, such that articulatory events are detected via the acoustic/auditory representations. We accept the weak version of the isomorphism, while rejecting the strong version. If we read him correctly, Studdert-Kennedy favors the strong version, partially because he believes that it provides the most natural account of the ability of children to learn to imitate speech.

Without denying that human children appear to be predisposed to imitate speech sounds, we do not think that the strong version of the perceptuomotor isomorphism is required in order to account for speech imitation. The prebabbling and babbling phases of vocalization can be viewed as a period in which the infant receives intensive training in the acoustic/auditory consequences of a wide range of vocal-tract gestures (see Fry, 1966). This training is almost continuous, the feedback is immediate, and the infant is free to explore the relation between gesture and sound with little outside interference. A situation more conducive to rapid learning is difficult to imagine. We suggest that when a child imitates the vocalizations of others, he or she does so by selecting gestures that experience has taught are most similar auditorily to the target utterance. Notice that, although this account accepts the weak version of the perceptuomotor isomorphism, it does not presuppose that speech sounds map perceptually onto the vocal-tract events that produced them. We also emphasize that learning the acoustic/auditory consequences of speech gestures is not equivalent to learning a set of arbitrary associations. Rather, the child is mastering vocal-tract physics (both its laws and boundary conditions) in such a way that the acoustic consequences of certain novel articulations may be partially predictable.

It appears that none of the evidence reviewed by Studdert-Kennedy is incompatible with our alternative account of speech imitation. Moreover, there is at least one compelling reason to prefer our account. Children are not only adept at imitating the speech sounds of others, they are also quite good at vocally modeling various nonspeech sounds such as musical melodies. The ability to whistle or hum "Dixie" after hearing it played on, say, a xylophone cannot easily be attributed to an isomorphism between the neuromotor control structure that produced the notes on the xylophone and the one that produced the corresponding notes vocally. The imitating child in this case does not attempt to duplicate the neuromuscular and mechanical activities of the musician/instrument system; he or she tries rather to model the musically relevant auditory properties of pitch and rhythm using an entirely different neuromuscular control structure and sound-producing system. Studdert-
Kennedy might reply that the isomorphism he has in mind is sufficiently abstract to encompass this kind of example. But if so, then it would seem to be too abstract to give the child much help in the specific domain of speech imitation. There would simply be too many activities, most of them inappropriate, subsumed under such an isomorphism.

In the target article, we criticized Studdert-Kennedy's (1987) notion that the speech signal is mapped onto perceptuomotor objects that specify "a range of functionally equivalent articulatory actions" (p. 71). We argued that this formulation is unparsimonious because the equivalence class it refers to is virtually unbounded in size and that it obscures the real (auditory) basis of the functional equivalence. Studdert-Kennedy replies that "an equivalence class of perceptuomotor objects is exactly as 'unbounded' as the 'equivalence classes of acoustically diverse . . . speech tokens'" (Diehl & Kluender, 1989) that we taught our quail. But there is a crucial difference between our referring to a phonetic category as an equivalence class and Studdert-Kennedy's claim that the percept corresponding to an individual speech token is an equivalence class. We do not suppose that a speech token is heard (either by humans or quail) as anything other than an individual instance of a phonetic equivalence class. We interpreted Studdert-Kennedy to mean that the speech signal is perceived not as an instance of the phonetic equivalence class but rather as an array of functionally equivalent actions. By such an account, the percept itself is an unbounded set, and this is what we claimed is unparsimonious.

In response to this point, Studdert-Kennedy (personal communication, October 10, 1988) said: "I nowhere claimed that the percept was an equivalence class. I stated that the percept specified an equivalence class. Specify means 'is specific to or peculiar to.' The percept is, in my view, an index or an instance of a class, exactly as it is for you." However, if the objects of speech perception are sets of "direct mappings between sound and gesture" and if the articulatory side of each such mapping is an equivalence class, then it would seem that the percept itself includes an articulatory equivalence class, that is, a set of different articulatory events and not just an instance of this set. Alternatively, if the articulatory side of this perceptual mapping is just an "index" or symbolic signifier of the articulatory equivalence class, then it is hard to see how it could account for the child's ability to imitate the utterances of others. What, in other words, is the operational or algorithmic significance of the index?

As for the functional basis of this articulatory equivalence class, Studdert-Kennedy agrees that it is "perceptual" although he disagrees that it is "acoustic/auditory." Because he views speech perception as involving a direct mapping between sound and gesture, this reply seems to beg the question. If a class of articulatory events is "functionally equivalent" on perceptual grounds, but those perceptual grounds reduce to the claim that all members of the class implement the same articulatorily defined end (e.g., lip closure), then we are right back
where we started. As we argued in our reply to Fowler, the many-to-one problem
cannot generally be eliminated by appealing to articulatory notions of func-
tional equivalence such as "phonetic gesture."

Medial Voicing Distinctions and the Auditory
Enhancement Hypothesis

In his discussion of our treatment of medial voice cues, Studdert-Kennedy
correctly notes that Lisker (1978) and Abramson (1977) took a quite different
approach to explaining the diverse acoustic correlates of the voicing distinction.
According to Lisker (1978), the voicing correlates "can all be referred back to a
single crucial articulatory difference in the management of the larynx" (p. 132).

Although we would not be surprised if some voicing correlates (other than the
four that we discussed) derived from physical or physiological constraints on the
production of voicing distinctions, this remains to be convincingly demon-
strated. Our reasons for focusing on the four voicing correlates—closure dura-
tion, preceding vowel duration, presence or absence of glottal vibration during
closure, and F0 contour following closure release—were that these have been
shown to play a major perceptual role in signaling the medial voiced/voiceless
contrast, and they have also been shown to be very common across languages.
There appears to be broad consensus among experimental phoneticians that
these four correlates are among the most important medial voicing cues. Neither
Lisker (1978) nor Abramson (1977) presented any evidence or arguments to
support the claim that the four voicing correlates on which we focused merely
"fall out" of the relative timing of gestures. Instead, there are cautious statements
such as the following: "Whether this feature [closure duration] is independent or
somehow has a dependency relationship with laryngeal timing is not known at
this time" (Abramson, 1977, p. 300) and "[s]ince . . . voicing distinctions in final
position are likely to be characterized by differences in laryngeal timing, namely,
voice offset time, the question arises as to whether the concomitant difference in
vowel duration is completely independent of laryngeal timing" (Abramson,
1977, p. 301). We admire the important work of Abramson and Lisker on the
voicing distinction, but we think that the theoretical caution expressed in these
quoted passages is fully warranted.

Lip Reading and Other Evidence

After reviewing various experimental findings that he believes tend to support
his notion of amodal objects of speech perception, Studdert-Kennedy acknowl-
dges that the evidence may not be conclusive. We would go somewhat further
and suggest that the evidence he cites is highly equivocal with respect to the
objects-of-perception issue.

Studdert-Kennedy refers to studies of short-term memory interference
(Campbell & Dodd, 1980; Crowder, 1983; Crowder & Morton, 1969; Spoehr & Corin, 1978) in support of his theory. Just as an auditorily presented suffix can reduce the recency effect for a list of auditorily presented words, a similar suffix effect occurs when, for example, the word list is auditorily presented and the suffix is lip read. We do not view such cross-modality effects as incompatible with our own account of speech percepts. Although we claim that phonetic segments are primarily auditory events for the listener, we do not deny that facial cues can also provide segment information. If the interfering effects of the suffix occur at the level of segmental representation (or, alternatively, if the perception of lip reading is mediated by what Crowder (1983) calls “phantom audition”), then it is unnecessary to assume that auditory and visual cues specify the same gestural event to the perceiver. Of course, the hypothesis that suffix interference occurs at a level of segmental representation is at odds with Crowder’s claim that the effect is precategorical. However, because the suffix effect in the crossmodal condition extends over the last three or four items in the word list (Crowder, 1983), we think it is unlikely that the interference occurs only before phonetic categorization takes place.

If Studdert-Kennedy wishes to identify his amodal speech percepts with the contents of the kind of short-term memory investigated by Crowder and his colleagues, he needs to show how his theory is compatible with several well-established features of this memory. First, the suffix and modality effects that are hallmarks of the memory are much greater for stimulus lists with varying vowel quality than for lists with varying consonant place of articulation (Crowder, 1971). That is, vowel information seems to be more strongly represented in this memory than consonant information. Second, the suffix effect is highly sensitive to voice quality of the talker, being much reduced when the word list and the suffix are spoken by different talkers with dissimilar vocal characteristics (Morton, Crowder, & Prussin, 1971). This suggests that the speech representation at this level may be too speaker-dependent to serve one of the principal functions that Studdert-Kennedy assigns to it, namely, to provide a basis for the child’s imitation of adult speech.

As for the other evidence cited by Studdert-Kennedy in support of amodal speech percepts, we have the following comments. First, the fact that 4- to 6-month-old infants appear to be sensitive to the correlated optic and acoustic patterns of speech is by no means inconsistent with the kind of perceptual learning account that we offered for the McGurk-MacDonald effect in the target article. Second, the fact that much younger infants (12 to 21 days old) can imitate arbitrary mouth and hand movements provides little evidence about how they later come to imitate vocalization, although we agree with Studdert-Kennedy that it does indicate the early presence of at least one kind of perceptuo-motor link. Third, evidence for left-hemispheric lateralization of linguistically relevant perception of both sounds and lip cues is quite consistent with our own theory. Contrary to what Studdert-Kennedy asserts, we never
claimed in the target article that phonetic perception reduces to general audition. Rather, we emphasized that “while the auditory–phonetic space may be largely given, its functional partitioning is not.” Phonetic categorization requires that the child learn what aspects of the speech signal (and of accompanying facial gestures) are linguistically relevant. We have no reason to doubt that phonological partitioning of the auditory–phonetic space and the use of linguistically relevant facial cues are typically left-hemisphere activities.

Quail

Studdert-Kennedy has just a few brief comments about our quail studies. His first point is that “species’ differences in rate of learning the arbitrary response and in hemispheric specialization suggest that showing a particular discriminative task to be within the psychophysical competences of two different species is not necessarily to show that the two species’ percepts are equivalent.” This is correct, but the following is also true: Species’ differences in rate of learning and in hemispheric specialization do not necessarily show that two species’ percepts are different. Studdert-Kennedy’s second point is that teaching quail phonetic categories is only interesting if such categories play a functional role in speech perception, and he doubts that they do play such a role. Because Studdert-Kennedy (1987) elsewhere equated phonemes with his hypothesized amodal perceptuomotor structures, and because phonemes must be viewed as a type of phonetic category, we find his remark difficult to interpret. In any case, he offers no argument to support his suggestion that phonetic categories are epiphenomenal. Third, he asserts that “we have no reason to believe that they [phonetic categories] are polymorphous, have no single invariant property, or are learned by conditioning.” Our belief that phonetic categories are polymorphous is grounded in the fact that, despite many serious attempts over the past 30 years, speech researchers have consistently failed to discover single invariant properties of such categories. In our first quail study (Kluender et al., 1987), a full battery of acoustic analyses failed to reveal any single invariant and distinctive property of our test stimuli that could support the observed categorization performance. As for the issue of phonetic acquisition, no one disputes that phonetic categories are at least partly learned by humans. Studdert-Kennedy’s point here is that human learning of phonetic categories does not occur by means of the kind of conditioning procedures used with nonhuman subjects, and we certainly agree. We would also agree that phonetic categories have a different functional significance for humans than they do for the quail in our experiments. However, neither of these points alters our conclusions that phonetic categories are auditorily natural groupings, and that similarities between human and nonhuman phonetic categorization are due to cross-species commonalities in mechanisms of general audition and perceptual learning.
CONCLUSION

The comments of Fowler, Remez, and Studdert-Kennedy have been very helpful in prompting us to consider various ways that our theoretical claims must be clarified, refined, and given sturdier justification. On close examination, however, the criticisms do not appear to pose an insurmountable challenge to our overall theoretical framework or to our interpretation of relevant experimental results. The auditory enhancement hypothesis (with its corollary that the objects of speech perception are acoustic/auditory, not articulatory, events) provides a principled account of a diverse set of phonetic/phonological facts, including universal tendencies in the structure of segment inventories and in the covariation of gestures. An auditory interpretation of phonetic percepts can also explain parallels between the perception of speech and nonspeech signals, certain cross-language similarities in, for example, the perception of phonemic tones, and clear behavioral homologies between humans and nonhumans in the perception of speech. The evidence suggests to us that this is an approach worth pursuing.

REFERENCES

Little, Brown.
REPLY TO COMMENTATORS


Sussman, H. M., MacNeilage, P. F., & Hanson, R. (1973). Labial and mandibular movement dynamics during the production of bilabial stop consonants: Preliminary observations. *Journal of
Speech and Hearing Research, 16, 397–420.