Journal of Phonetics (1986) **14**, 149–170

Reply to commentators

Carol A. Fowler

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

I thank the commentators for their thoughtful and provocative commentaries. In response to them, I will attempt to counter those criticisms that must be countered for a direct–realist theory of the perception of speech events to be viable, and discuss comments that may serve to develop or alter the form of the theory as I proposed it.

I will begin by clarifying the goals of the target article as I saw them; next I will consider a criticism that was raised by almost every commentator in some form; finally, I will address other important topics raised by individual commentators.

1. Goals of the target article

In the target article, I attempted to work out a theoretical approach to the study of speech that is compatible with "event approaches" under development in other research domains. The effort followed an earlier one by a working group on speech and sign language convened at the first event conference (Warren & Shaw, 1985). That working group made some progress in outlining an "event approach" to the study of speech and sign, but it did not address many central issues of the theory (Fowler & Rakerd, 1985). The present effort differed from the first one in restricting itself largely to development of the theory within my particular domain of expertise, speech perception, and in trying to focus on some of the important issues that did not get aired by the working group. The target article is not a summary of the working theoretical perspective from which other ecological speech scientists and I have conducted our research. Rather, it is an attempt to outline an event approach to the study of speech that eventually can serve that role.

Accordingly, when I describe research that others and I have conducted in which listeners identify phonetic segments presented in isolated nonsense syllables, I am not necessarily advocating these experimental procedures as Remez suggests that I am. Rather, I am using the data that are out there to ground the theory as best I can. (This is not to say that I will not defend the use of these procedures for some purposes; see Section 3.6.)

Similarly, I found Massaro's comments misplaced when he objected to a "confirmation bias" I exhibited in claiming support for a direct–realist theory by data that, in his view, are equally well or better handled by other theoretical approaches, most notably his own fuzzy-logical model of perception or FLMP (e.g. Oden & Massaro, 1978). My review of experimental findings that he complains about was not meant to demonstrate the superiority of this type of theory over others (after all the review was largely borrowed from Liberman & Mattingly (1985), who provide it in support of a motor theory), but only to point out where the theory appears descriptively adequate.

0095-4470/86/010149 + 22 $03.00/0

2. Is phonetically structured articulation a perceived speech event?

Most of the commentators objected to some aspect or aspects of my suggestion that phonetically structured articulation serves as a "speech event" that listeners perceive directly. I will consider the objections under several different headings.

2.1. *Other distal events could have been proposed*

Ohala asks why, if I was going to focus on something other than the acoustic signal as an object of perception, I picked articulatory activity when there are other possibilities further "upstream," including muscle contractions, neuronal activity and mental events. Both Massaro and Studdert-Kennedy may have something similar in mind when they propose that, in my theory, articulation is a medium no less than the acoustic signal. (Remez also considers it a medium, but for a different reason that I will take up in Section 2.6.) Likewise, when Diehl brings up misarticulations, which make the acoustic signal an erroneous reflection of the talker's intended message, he also must have in mind a covert object of perception.

One answer to Ohala's question is that, in a theory of direct perception, the distal event has to structure an informational medium; otherwise it cannot be directly perceived. That rules out anything that might be truly "upstream" of articulation. However, more needs to be said because it does not rule out "mental events." In the target article, I used a quotation from Ryle to suggest that "bodily acts" need not be seen as the product of "occult" mental activies. Rather, many of them *are* mental activities themselves. This way of thinking about events at a psychological (mental, intentional, ecological) scale places the psychological aspects of the events out in the world where they can be perceived directly. So I would say, too, that grammatically structured words are in articulation. I focused on phonetic segments only because the commentators and I know them best.

This is also my response to Massaro and Studdert-Kennedy, as a first approximation anyway. It may have been misleading for me to agree with Liberman (1982) as I did (Fowler, 1983) when he proposed that phonetic "intents" are perceived, because the word "intent" may imply that something covert that may or may not have any subsequent public realization. In the target article and elsewhere I have proposed that (intended) phonological (and phonetic) segments are literally uttered and it is those uttered segments that are perceived.

Unfortunately, this may not do entirely. Diehl raises the very difficult issue of speech errors and their implications for a direct–realist approach of perception. He suggests that there is a marked difference between visual perception and speech perception just because speech errors occur. In visual perception, "the laws of optics fully determine the specificational relation between optic arrays and environmental surfaces." In contrast, whereas the laws of acoustics specify vocal-tract shapes (or gestures), if the talker has made a speech error, it does not fully specify what he meant to say, because intended parts of the utterance were not spoken.

Now, this problem of "erroneous distal events" does not, in fact, distinguish speech perception from visual perception. However, it does appear to distinguish perception of intentional activity from perception of other things. Both Norman (1981) and Reason (1979) have documented many examples of "action slips" in which an activity performed is different from the one intended. (Examples are a person stopping his car and

unbuckling his watchband instead of his seatbelt, and a woman leaving her apartment in a hurry, throwing her earrings to her dog and attempting to attach a dog biscuit to her earlobe.) When these action slips are witnessed, the optic array, like the acoustic signal, may mislead the perceiver as to the actor's intent.

In either case, auditory or visual, the problem as Diehl sets it up remains: there is a glitch in the transmission from intent to percept that may then require cognitive mediation if the perceiver is to recover the intent.

What really happens when perceivers encounter speech errors? They may perceive what the talker literally uttered (e.g. "The blueberries are ripe, and on Sunday we're going to morage in the fountains. Want to come?"). In addition, or instead, they may recover the talker's message ("Do you want to go blueberrying with us on Sunday?").

I can imagine that each of the following accounts may apply to the different possible perceptual experiences of utterances containing speech errors.

(1) In the utterance as produced, there is more information that the talker was referring to the concepts of foraging and mountains than that he was talking about moraging and fountains. (For one thing, blueberries are more plentiful on mountainsides than in fountains, and, for another, foraging is a way to find them, while moraging is not.) In short, despite the exchange of phonological segments, produced in the proper setting, the utterance conveys (and constitutes) the talker's message. If the listener perceives the message, he may, or may not, hear the utterance as it was literally said. If he does not (that is, if he reports hearing "forage in the mountains", an explanation similar to that required to handle some McGurk findings (e.g. McGurk & MacDonald, 1976) can be invoked. The message as a whole specifies the words "forage" and "mountains" while the sequence of phonetic segments specifies /mɔrəǰ/ and /fawntənz/. Some information, incompatible with the rest, is ignored.

(2) A second possibility is that the perceiver hears the utterance just as it was produced and only recovers the message via problem-solving (cognitive) tactics. This may cover instances where the errors are very salient. ("We have a laboratory in our own computer.")

(3) Finally, the utterance may be heard as it was produced and the perceiver may not recover the message at all, because the information for it was too impoverished.

None of these possibilities requires that the listener perceive something covert.[1]

2.2. *Focus on articulation or the acoustic signal is a matter of "theoretical taste"*

Diehl and Studdert-Kennedy both suggest that if acoustic information is sufficient to support direct perception, then, in Studdert-Kennedy's words, it is "a matter of theoretical taste" which is said to be perceived. However, I think that this way of setting up the issue does not make sense. To serve as information, structure in the acoustic speech signal has to be about something; in speech, it is about its articulatory source. We cannot extract the information and yet avoid perceiving the environment it informs about; extracting the information *is* perceiving the environment.

[1] Having gone through all of this, I should acknowledge that the problem Diehl raises strikes me as much more difficult once the theory of production is considered from a realist perspective. If it is crucial, as the theory of perception claims, that perceivers be able to recover real-world properties, then it must be just as important that they act felicitously with respect to them. Otherwise the advantage of errorless perception is substantially reduced.

2.3. *Acoustic signals are perceived*

Ohala offers several pieces of evidence that it is not just a matter of theoretical taste after all which we perceive. In fact, we perceive the "speech sound", not its articulatory source. (In several dictionaries I have consulted, "sound" has two definitions—just the two that make the puzzle of the tree falling in the forest unresolvable. It can refer either to the acoustic signal or to the signal's perceptual consequences. Ohala has the first definition in mind; however, if he had the second in mind, I would agree that we perceive speech sounds.) One kind of evidence is that different articulations may have indistinguishable acoustic consequences and may count as instances of the same speech sound. (Different articulations with the same acoustic consequences may also be perceived at different speech sounds (Liberman, Delattre & Cooper, 1952).) My reading of Ladefoged's recent work is that there is far less flexibility here than was previously thought (e.g. Ladefoged, 1980). Examples of the sort Ohala mentions may be largely limited to findings such as that by Sussman, MacNeilage & Hanson (1973) in which the relative contributions of the jaw and lips differ across productions of /b/ in the context of different vowels, or in which vowels are produced with different contributions to height by jaw and tongue. In cases such as these, the variants may be explained as products of just one organization of the vocal tract that has equifinality properties. In most other cases where the same acoustic segment is perceived as different phonetic segments (as in the stop burst in /pi/, /ka/ and /pu/), it is because different organizations of the vocal tract must have produced different gestures in their contexts to produce the same acoustic segment.

Not all cases can be explained in this way, however. "Compensation" by people with vocal-tract pathologies, and "talking" by parrots and mynah birds require a different account. In my view, both are examples of "mirages". [Mirages occur when one distal event structures a medium in a way that, in the short run, mimics structuring by a different distal event with greater "attensity" (that is, greater ecological salience): for example, heat rising off the road structures light, in the short run, in the way that a puddle structures it, and until we approach it closely enough, we see it as a puddle.] People with vocal-tract pathologies and birds that "talk" learn to structure the acoustic signal in ways that mimic structuring by intact human vocal tracts.

I am aware that my proposing this raises all kinds of other issues including that, having allowed that perceivers can be tricked by mirages, I have apparently abandoned the idea that perceivers are in immediate, errorless, contact with their ecological niches. In reply, I can only repeat my suggestion in the target article that perceivers *can* be in this kind of contact with the environment, but they may not always do the exploration necessary to narrow down the possible worlds out there to just one.

2.4. *Articulatory events are not perceived*

Not only is there evidence that the acoustic signal is an object of perception as brought out by Ohala, but also, according to Massaro, there is reason to discount the data I cited earlier as evidence that phonetically structured articulation is extracted by perceivers. He suggests specifically that trading relations in perception and the McGurk effect are more consistent with FLMP than with a direct–realist theory, and, in FLMP, no recourse is made to the articulatory origin of the acoustic "features" of speech.

In FLMP, perception is held to occur in three stages. First, continuous acoustic features are extracted from the acoustic speech signal; that is, acoustic features each are

assigned a value reflecting the probability that they are present in the acoustic signal. Next, that information is used to determine the degree to which speech units, represented as stored prototypes and defined as a particular collection of acoustic features, are present in the signal. Finally, the speech unit corresponding to the most likely prototype is selected.

FLMP offers an explanation of precisely how cues trade, when they do. Essentially, any feature that contributes to a prototype adds, according to the evidence for it in the signal to the probability that listeners identify the speech unit corresponding to the prototype. Therefore, if acoustic features for a given stop consonant include silence and a certain pattern of formant transitions, then an acoustic signal having lots of silence is as good as a signal with less silence but appropriate transitions.

Despite its accomplishments, FLMP is in no way in competition with the direct–realist theory (nor, for that matter, with the motor theory). In generating the precise form that cue trading takes, FLMP goes beyond what a motor theory or direct–realist theory has been worked out to do. However, FLMP has nothing explanatory or predictive to say about which acoustic cues do trade, or, indeed, why they trade at all. That is, the theory can be applied successfully to extant findings that a given phonetic category can be cued by multiple acoustic bits of information, but there is nothing in the nature of the model to explain why those multiple cues trade while others do not. Therefore, the model can be applied as comfortably to concocted data that show trading where, in fact, none occurs as it is applied to actual data. (The theory can "predict" that cues will trade whenever there is a prototype to which the cues jointly contribute, but the point is that there is nothing in the theory to explain the composition of the prototype. Moreover, it is clear that there are, as yet, no constraints on what can be a prototype; in Massaro & Cohen (1983) a prototype corresponding to /bda/, sometimes reported by subjects in a McGurk paradigm, is proposed. In the gesture experiment that Massaro describes in his commentary, arm movements can apparently contribute to prototypes for phonetic segments too.)

In contrast, a theory in which acoustic information is related back to its articulatory source can predict that only those cues will trade that are joint products of an articulatory gesture (e.g. vocal cords opening and closing), or a set of them that composes a phonetic segment. In addition, a direct–realist theory can predict that trading relations will be found whenever a cohesive sound-producing event (linguistic or not) gives rise to complex acoustic consequences.

Precisely the same line of argument applies to FLMP's handling of the McGurk effect. The interesting question is why the acoustic and visible sources of information "integrate" at all, not what the precise weights are with which they combine. They integrate because they jointly specify the same distal event.

2.5. *Articulated phonetic segments have no ecological significance*

Diehl and Remez both argue that the articulated phonetic segment fails as an ecological event. Diehl suggests two dimensions on which it fails. It has no ecological significance and it is inaccessible to a perceiver's awareness. As Remez asks, if we perceive articulation, why aren't we better at controlling an articulatory synthesizer? Perhaps, he concludes, perceived speech is only "figuratively articulatory."

Phonetic and phonological segments do have ecological significance, however, in two quite distinct respects. First, they presumably evolved in languages due to pressure from

a growing lexicon (Hockett, 1960; Lindblom, 1971; Lindblom, MacNeilage & Studdert-Kennedy, 1984). They may develop in child language under the same lexical pressures. Secondly, they have perceptual significance for perceivers, not only because talkers produce them (as speech errors reveal; e.g. Shattuck-Hufnagel, 1980), but also because social groups use them, as Labov's work, cited in my target article, indicates.

As for accessibility, that is not the only, or even the best, index as to whether articulated phonetic segments are perceived. A more telling index is that we shadow or imitate speech both well and remarkably rapidly (e.g. Porter & Castellanos, 1980; Porter & Lubker, 1980). Thus, the facts appear to be that we do perceive speech as articulated, and yet we are not easily made aware that we do. As for why we are unaware, Remez' reference to tacit knowledge may be on the right track. Language is tiered and its ecologically most significant information is provided by levels more encompassing than the phonetic level. It may be very difficult to ignore them in order to attend to their constituents.

2.6. *Articulated phonetic segments are not autonomous events*

Several commentators objected to my partitioning of a communicative event in such a way that the "speech event" as I defined it serves by itself as an event. The comment took two forms, one easy to deal with and the other not so easy.

Massaro reports evidence that he sees as showing that arm gestures contribute to phonetic perception. In his view, this disconfirms a direct–realist theory (and a motor theory at the same time) because the theory incorrectly identifies the distal event for phonetic perception as the articulating vocal tract.

Now, my partitioning of an event involving linguistic communication, vertically into the linguistic utterance on the one hand and everything else on the other, and horizontally into "linguistic events" and "speech events", was meant to be tentative, subject to discussion and test. However, I am willing to stick to my guns in the face of these data and claim that effects on identification of a spoken syllable of a talker's arm gestures are of a different order than effects of vocal-tract activity seen and heard. In particular, I make the radical claim that, whereas vocal-tract activity may be perceived as a spoken syllable, arm gestures may not. (Both vocal-tract activity and arm gestures may contribute to a *communicative* event, of course. This, by the way, is closer to McNeill's theory (1985) than is the characterization provided by Massaro.) How can this claim be maintained in light of the evidence?

Consider the model that Massaro ascribes to the direct–realist theory in order to test it against his data. The model represents identification of a syllable as occurring in two stages. First, gestural and vocal stimulation lead to separate perceptions of the referents as "ball" or "doll". (In the model, these perceptions are assigned probabilities. That is, the model identifies a spoken syllable as "doll" with some probability. This is required at least to generate the data that Massaro models, which are grouped over trials and subjects.) Secondly, a decision rule is applied to the two percepts when they differ. The rule is to go with the percept based on the vocal source of information with probability p and with that based on the gestural source with probability $1 - p$. The first stage of the model does capture the idea that vocal stimulation leads to phonetic perception without influence from the arm gesture. The decision rule has a different status, however. It is no part of the theory of perception under test. The decision rule has to be tacked on to the model because subjects in the experiment must report one referent even if they perceive two different ones.

The model cannot handle the findings of the study and so Massaro considers it (and by implication, the theory of direct perception) disconfirmed. He considers the evidence to "demand that gesture and speech information are integrated prior to phonetic identification." However, clearly it is the decision rule that is at fault in the model. As Massaro points out, subjects do not make their judgments based either on the gesture alone or else on the spoken syllable alone. (Why should they?)

That the combination of syllable and gesture is not perceptual may be demonstrated by asking subjects to discriminate what they *hear* on pairs of trials on which the identification study reveals equal probabilities of responding "ball". Interesting pairs of trials would include points to different objects (or one trial with no pointing at all) and different acoustic stimuli. I predict that listeners could easily discriminate the acoustic stimuli on such trials; this contrasts with relevant cases of trading relations such as those described by Fitch *et al.* (1979). I suspect that it would also distinguish the gesturing trials from analogous McGurk trials.

Clearly, too, there is nothing in the design of this experiment to warrant the conclusion that perceivers' use of gestural information precedes phonetic perception. Indeed, the following Gedanken experiment suggests that Massaro could replicate his findings in a setting in which integration in phonetic perception is not at issue. In the experiment, children hear the same continuum of syllables as in the study Massaro describes. However, instead of seeing a videotape of someone gesturing toward a doll or a ball, they see one of a clown (i.e. a confederate with some, but not lots of, credibility). After the spoken syllable is presented, but before the children respond, the clown says: "Boys and girls, *I* think he was talking about the Cabbage-patch Kid!" (or "Boys and girls, *I* think he was talking about the round toy!") I predict that listeners would show the same tendency to trust the confederate increasingly the more room the perception of the synthetic syllable left for influence to appear.

Finally, Massaro should consider whether his own theory should allow the arm gesture to contribute to phonetic perception. For the concept of prototype to have theoretical value, there must be constraints on what can be assumed to contribute to one. But there are countless ways of referring to/b/- and /d/-initial words.

A different type of concern also leading to the conclusion that articulated phonetic segments are not events (and hence are not "objects of perception") was raised by Ohala, Studdert-Kennedy and Remez. Ohala points out that the significance or status of /ʔ/ (a glottal stop) is different for speakers of English, who hear it optionally in front of vowel-initial words than for speakers of Arabic, for whom it is a constituent phoneme of certain words. Studdert-Kennedy suggests that a "speech event is not merely articulation, however structured, any more than a tennis serve is simply the server's swing." Finally, for Remez, "there is nothing intrinsically valuable about these movements of the anatomy of deglutition and respiration." I accept the first two comments, although I interpret them differently than Ohala and Studdert-Kennedy, but I do not accept the last (unless I am misunderstanding "valuable"). Phonetically structured articulation does constitute a natural class of activities (cf. Locke, 1983) that do not crop up elsewhere; and so they identify themselves as constituents of an utterance.

This aside, what is the import of these comments? For Remez it is that phonetically structured articulation, and, for that matter, linguistic patterning in general are not perceived events ("objects of perception"). Instead, they are informational media, and hence, not perceived at all. (It is possible that Remez and I are interpreting "object" differently in the phrase, "object of perception". I had in mind "perceived event" not

"goal of perception." Therefore, I did not mean to set up phonetic segments and grammatically ordered words as "premisses for cognitive elaboration". Rather, they are perceived constituents of a perceived speech act.)

Now the specific claim that phonetic segments are not perceived at all, in my view, is disconfirmed on the empirical grounds I consider in the target article. Moreover, the evidence Remez cites concerning phonetic perception is not relevant (at least under my construal of "object of perception"). As he points out, many studies show that identification of phonetic segments is colored by the larger units they constitute. But this is precisely as it should be, given that phonetic segments participate in these larger units of language.

The more general idea that Remez presents that the dual tiers of language structure serve as "information media" rather than as objects of perception is interesting. In trying to work out criteria by which something might be identified as informational medium or as event, I realize that the dual tiers of language structure have properties in common with both. Indeed, this may characterize a representational or communicative system.

An information medium (including air, light and the perceiver's skin in auditory, visual and haptic perception), as I have defined it, has the following properties. It takes on structure from environmental events that is specific to the properties of those events, and it makes the structure available to perceptual systems by stimulating the sense organs. These properties allow the medium to serve as a carrier of information. Additionally, perceivers are blind to the media *per se* (as experience with a Ganzfeld shows) and (accordingly) they do not interact with them, but only with events (that is, for example, tennis players swat at tennis balls, not at light, or even at expanding contours in the optic array.)

As for events, they are occurrences in the environment at the perceiver/actor's scale in which an actor participates. In serving as a focus of an actor's behavior or in being an object of perception, an event has some degree of autonomy from its context; however, except for the most encompassing ones of all, events are nested and are not fully autonomous.

The dual tiers of language structure fail the definition of information medium in not being transduceable by sense organs and in being perceived, but they pass in serving as structure that provides information about something other than themselves. They pass on all counts as events. The conclusion is, I think, that communicative systems are systems in which environmental events take on some functions of informational media. Looked at that way, there is no need to engage in contortions to show that phonetic segments and grammatically structured words are not perceived. And it is more-or-less clear why they are less reportable and memorable than the linguistic message itself. They are not usually the main goals of perception; they are not what the perceiver is looking for in the speech signal, except as Polanyi says "subsidiarily" (e.g. 1962).

How do phonetic segments take on linguistic values? *Contra* Remez, I think that they have some intrinsic linguistic value in constituting a natural class of mouth movements, reserved for participation in speech acts. (Accordingly, glossolalia—even as faked by someone on whom the gods have turned their backs—is a speech event.) However, that does not explain the different statuses of /ʔ/ in English and Arabic. That is due to "nesting". Phonetic segments are constituents of grammatically structured words produced in a setting in which conformity to phonological and grammatical constraints of some language means producing a meaningful utterance of the language. (By the same

token, in answer to Studdert-Kennedy, a tennis swing becomes a serve when nested in the context of a game.)

Because these larger events are precisely what gives utterances their communicative value, I cannot agree with Remez (or with Gibson) that indirect perception is involved in linguistic communication. (In fact, I do not see the value of a concept of indirect *perception* at all.) According to Remez, "those things to which the talker refers" are indirectly perceived. However, if I listen to a disquisition on the subject of tables, and no tables are present, I do not perceive tables at all. Via the dually structured articulations of the talker, I can directly perceive the talker's message (which conveys his perspective on the subject of tables), but I cannot perceive tables.

3. Responses to individual commentaries

3.1. *Diehl*

I will concentrate on just two additional comments here, one relating to perception and the other to speech production.

Diehl points out that as a message becomes increasingly redundant, talkers provide less phonetic information. In consequence, he concludes, information for the perceiver is less direct and more difficult to define than information in optic arrays.

As for the example of speech errors considered earlier, I doubt that this is a difference between speech perception and visual perception. Here it seems to be a difference between communicative events and others. (For example, my handwritten "the" is less readable by itself than are less overdetermined handwritten words.) In contrast to the inanimate world, actors engaging in communicative activities can provide either more or less information to listeners whom they estimate need more or less (because the information missing from the utterance is available in some other form). Talkers who utter "degraded" versions of words they said earlier can get away with it because the words' identities are specified by many other sources of information in the communicative setting. (I wish I could end by saying "It's as simple as that", but obviously it is not. It is difficult to understand how the information is preserved that makes the redundant information redundant. Clearly, the perceiver is changed, for example, by the previous productions of a given word by the talker. The trick for a theory of perception as direct is to allow for that without giving cognitive mediation a foot in the door by which it can sometimes bring inappropriate learned information to bear on stimulus information. I don't know what to say.)

On the subject of speech production, I have proposed that coarticulation be seen as (largely) coproduction of phonetic segments, rather than as assimilation of discrete segments to their contexts. This is consistent with the data (e.g. Öhman, 1966; Perkell, 1969; Carney & Moll, 1971, and others), it allows findings of compensatory shortening to be handled by the same account that handles coarticulation (see Fowler, 1981) and, apparently, it is consistent with the ways in which perceivers hear phonetically structured speech (Fowler, 1983; 1984; Fowler & Smith, 1986).

Diehl suggests that coproduction would only be possible if the coproduced segments could be independently articulated; yet clearly this is not possible for vowels that coarticulate with each other. In addition, he suggests a view of coarticulation as assimilation and a view of perception of coarticulated segments in which listeners attend to the temporal or serial ordering of information peaks for successive segments without bothering to find the boundaries between segments.

I do not think that coproduction depends on independence of the articulators for the simultaneously produced segments. It is true that, in my early papers, I referred to Perkell's idea that different muscles might be responsible for vowel and consonant gestures. However, this is not a useful idea, it seems; there are clear cases of coproduction where the same articulators are used simultaneously by two different segments (for example, the jaw in CVs; see Sussman, MacNeilage & Hanson, 1973). Moreover, in other domains, it is clear that independence is not required for perceptual "decomposition." Consider locomoting figures. Movement about the elbow is complex in combining the forward movement of the body, the swinging of the arm about the shoulder, and rotation of the forearm about the elbow itself. Even though these are all "mixed together" in one joint, they are readily decomposable. We do not see a complexly moving elbow; we see a whole figure moving forward, with arm swinging and forearm rotating about the elbow.

As for the idea that we do not bother with segment boundaries, I have to disagree if I hew to the direct–realist line and if I conclude, from speech error data, that segments are separable in some way in production. In any case, if a listener tracks the acoustic speech signal along coarticulatory lines, boundaries are simply places where one segment starts or stops influencing the acoustic speech signal (see also Elman & McClelland, 1983).

3.2. *Fujimura*

Fujimura points out that my characterization of phonetic segments from the perspective of the theory is markedly underspecified. The hypothesis I put forward that the organization of the vocal tract for a segment may be invariant over contextual variability is impossible to evaluate without that specification.

He is right that the theory is underspecified. I cannot responsibly fill in all the needed information either. I had in mind systems such as the jaw–lip system that Kelso, Tuller, Bateson & Fowler (1984) studied, for example, which allows bilabial closure over variation in the position of the jaw. I interpret the findings of Sussman, MacNeilage & Hanson (1973) that, during bilabial closure, the jaw differs in its absolute position depending on the height of the neighboring vowel, as evidence that the system revealed by perturbation studies operates during unperturbed talking as well.

Systems such as this should be invariant over intrinsic allophonic variation. I doubt that they will be invariant over extrinsic allophinic variation—that is, over differences that we can hear and that some languages treat as distinctive. However, although the articulatory organizations for extrinsic allophones cannot be identical (in order for the audibility of their differences to be explained), it is likely that they belong to a family of similar organizations.

3.3. *Massaro*

Massaro cites a study in which sentence context affects listeners' judgments as to whether a word in the sentence is "the" or "to". He argues that the study reveals top-down influences in phonetic perception. As in his gesture study, it is possible that these effects are not on perception, but on the listeners' later response selection. In studies of phonemic restoration, Samuel (1981) found that sentence context does promote

restoration. However, its effects are on **ß**, not on **d′**. (**ß** representing the listener's decision criterion, and **d′** their perceptual sensitivity).

Citing the Samuel findings is a slightly sneaky route for me to take here, however, because even if the "the"/"to" findings can be dispatched in this way, others that I cite in the target article cannot. I cannot say more than I did in the target article, however, to explain these findings.

Massaro also criticizes the idea I propose that perceivers are pragmatic in perception, sometimes being content to perceive enough to reduce the possible environments to a likely one (e.g. "that's an aperture") and other unlikely ones ("its a barrier built of low-glare plate-glass", "its a holograph") rather than eliminating the unlikely ones by further perceptual exploration. His objection is that, in a theory of perception as direct, perception does not involve effort and therefore there is none to conserve. I suspect that Massaro is confusing the information-processing concept of "central processing capacity" (something like "mental effort") with one of actual physical effort to which I refer, following Gibson. While it is true that there is no processing capacity to be conserved, in a theory of perception as direct as Gibson's writings reveal clearly (e.g. 1966; 1979), physical effort in the form of exploration is centrally involved in the perceiver's obtaining of information from the environment. Perception is not just picking up information in the environment; it is actively exploring the environment to obtain information needed to guide activity. How do we find out whether an apparent aperture is blocked by a pane of plate-glass? We slow our approach to the aperture, reach out and search for plate glass. This does involve time and effort that, ordinarily, we do not bother to expend.

3.4. *Ohala*

I will briefly consider six comments that Ohala made in his response to my target article.

(1) "Fowler identifies 'the articulating vocal tract' as the distal event which is perceived in the case of listening to speech." I do not mean to exactly. First, just any vocal tract movements do not count as distal speech events; only phonetically structured ones do. Secondly, even those movements of the vocal tract are not *the* distal event in speech perception; they are *a* distal event. As I acknowledge in the target article, they are not even primary in any sense; they are just the language-related events that the commentators and I know most about.

(2) Ohala uses the example of interpretation of Moore's "reclining figure" to suggest that what is "out there" is not always identical to what we perceive. Instead, in interpretation of art as in speech perception, what we experience perceiving can be a product of what is out there and what we interpret it to stand for.

That perceptual experience is colored in this way is not a point of contention. The question is how it is to be explained. I offer a two-part account. First, one has to be very careful in deciding "what's out there." Even in the absence of a perceiver to appreciate it, I contend that phonetically structured vocal-tract activity can stand for something. At one level of description, vocal tract noises are, perhaps, just noises (just sound-producing events) but at another, they are a linguistic message. The "interpretation" here is in the activity and its setting; it need not be imposed on the activity. (The same story might be told of a sculpture as well—at least one associated with a consistent interpretation. The figure has to support the viewers' interpretation in some way.) Secondly, the listener (or,

generally, the perceiver) may, or may not, be sensitive to all of the levels and kinds of structure in the event, and hence, may, or may not, recover the "interpretation".

(3) Ohala points out that errors in receipt of a message do occur. Although these may be ascribed to a perceiver's not doing all of the exploration necessary to extract sufficient information from the environment (as the example of the bat illustrates), if this is typical of perceivers, then the direct–realist theory may explain a rarely occurring form of information acquisition by the senses. Why reserve the word "perception" for those rare occasions?

However the word is deployed, the possibility that a perceiver can extract the information needed to support its activites is crucial to the theory and special to it. Notice that the only way that bats and people can get to a point where they can be perceptually lazy is if on some earlier occasions, they were not. In a direct–realist theory, this form of perceiving is a "bottom-line" form.

(4) Ohala cites a study by Winitz *et al.* (1972) as suggesting that the acoustic speech signal is inherently ambiguous. In the experiment, listeners are more accurate in identifying a stop given its burst alone than given the first 100 ms of the syllable, including the burst. Ohala concludes that "the greater the amount of information presented, the greater the error." He explains that research by Stevens & Blumstein (1979) has shown the burst to be a more reliable cue to stop identity than the transitions which, then, must just confuse things in the Winitz study.

Let us look at these observations carefully and then consider Ohala's interpretation. First, the study does not show that the signal is inherently ambiguous. Nor does any other. Speech researchers who interpret some findings that way should consult the literature on visual perception (for example, Gibson, 1968) where classical "inherent ambiguities" (such as world movement and self movement) have been shown not to be. Secondly, there is a missing condition in the study by Winitz *et al.* that would no doubt be informative. In which condition would subjects identify the stops better: one in which bursts alone are presented, or one in which the whole syllable is presented? Presumably, they would do better with the whole syllable. There must be no simple correlation between the proportion of a whole acoustic signal for a syllable that is presented and the amount of information that signal provides for the syllable produced. One reason why is that fragments of acoustic signals for a syllable provide misleading information. The brief signal flanked by silence suggests that a sound-producing event consisted of a very brief maneuver, not much like a syllable. Thirdly, I had the impression from Ohala's comments that the findings of higher performance on burst-alone stimuli were reliable and typical, but they were not. They were occasional outcomes that Winitz *et al.* considered anomalous. They hypothesized that they were due to the researchers' having cut the longer-duration stimuli off at 100 ms whether or not the interval included all of the transitions. That is, stimuli in which the vowel target or steady state was not reached may have provided misleading information concerning consonant identity to listeners.

Whether or not Winitz *et al.* have accurately explained these occasional anomalies in their data, the reason Ohala gives for the result is incorrect. Blumstein & Stevens (1979) do not show that a burst is a more reliable *cue* to stop place than are the transitions. They suggest that the burst spectrum contains invariant information for place, whereas transitions do not. Other studies show that which is the more relied-on cue, burst or transitions, depends on the particular consonant–vowel syllable under examination (Dorman, Studdert-Kennedy & Raphael, 1977). Moreover, in that study, the burst tended to carry less weight overall than the transitions. In addition, in two tests of

cue-hood that I describe in the target article (Blumstein, Isaacs & Mertus, 1982; Walley & Carrell, 1983), burst spectra pitted against formant transitions in syllables were not used by listeners as information for place. Further, Lahiri, Gewirth & Blumstein (1984), also cited in the target article, show that the burst spectra do not provide invariant information after all, and a better candidate is provided by running spectra including the burst and the transitions. This candidate does pass the test of perceptual salience when pitted against transitions alone. It seems likely that the findings by Winitz *et al.* on which Ohala focuses are peculiar to unnaturally fragmented speech signals.

(5) Ohala suggests that some sound changes are due to inherent ambiguity in the signal. His example, is "actual", pronounced /æktuəl/ formerly and /æktʃuəl/ now. The high vowel following /t/ leads to increased aspiration on release of the /t/ that listener/talkers may have misconstrued as a /ʃ/. He concludes that "listeners have to form a hypothesis about the function of the noise and sometimes they accept the wrong hypothesis."

I find this conclusion very improbable. First, the example is similar in nature to the example of /æ/ raising before nasals that I describe in the target article. Careful examination of the two pronunciations of "actual" (just as examination of "band" with and without a raised /æ/) would, no doubt reveal perceptible differences. After all, there was a time when the word was pronounced (and presumably perceived) as /æktuəl/. If the sound change is in fact perceptually motivated, it may be that listeners failed to detect the distinguishing information (they fail to extract the proper parsing of the word, as in the nasal example); no hypothesis testing need be involved.

Secondly, consider the plausibility of a view that phonetic-segment perception involves hypothesis testing. Consider the number of hypotheses that must be formulated and tested in speech perception if every segment is associated with one. Moreover, consider how rapidly they must be formulated and tested. Finally, consider that, at those rates, hypotheses must essentially never be considered disconfirmed by the listener (that is, they must never be evaluated by the listener in such a way that a different hypothesis must be formulated and tested); otherwise the listener runs the risk of falling hopelessly behind the talker. These "hypotheses", then, must not be real hypotheses. Rather, they must be simply the way the listener perceives the signal.

This line of argument, by the way, is essentially the one occasionally leveled at the old analysis-by-synthesis view of speech perception (e.g. Halle & Stevens, 1959). It does not matter whether the putative hypotheses are about how the segment was produced, as in the analysis-by-synthesis theory, or about what segment the talker had in mind as in Ohala's. They cannot realistically be considered hypotheses, subject to test and disconfirmation.

(6) Ohala proposes a crucial experiment, designed to test whether perceivers "induce" missing acoustic information based on what is present. The experiment is to distort a signal so that its distinguishing features are progressively obscured. The question is whether perceivers will "restore" the missing information.

Ohala predicts that if stimuli are "E"s and "F"s, distorted progressively so that the bottom horizontal segment of "E" is obscured, viewers will increasingly report "E"s, given "F". He cites a somewhat analogous study of vowels in which masking noise covers different parts of the signal. When the noise would mask a high F2, listeners misreport back vowels (with low F2s) as front vowels (with high F2s) more than with lower-frequency masking noise.

This is not a crucial experiment. The speech experiment, as described in ideal terms

by Ohala, is analogous to cases of "occlusion" in visual perception (Gibson, 1969, 1979), in which one object passes behind another one and is temporarily occluded by it. Nonetheless, its continued persistence is specified by its manner of "disappearing" behind the object and by its later reappearance. In both phonemic restoration and the analogous E/F and speech studies described by Ohala, there is information in the non-occluded parts of the signal to specify that something occurred behind the mask. (For example, in phoneme restoration, in "legi_lation" (Warren, 1970) with noise covering the gap, the articulators show no signs of slowing down to a stop at the gap. This gives evidence for fluent production through the noise.) Then there is the mask that would indeed occlude the missing parts of the signal were they present. The missing phoneme or formant need not be hallucinated, therefore, because it is at least partially specified in the signal in the ways I have just decribed.

Turning to the data, it is, in any case, not very striking. Research by Townsend (1971) shows that under conditions of distortion, "F" is a much more common response given "E" than is "E" given "F". Why do viewers resist "restoring" the invisible leg of the "E"? There appears to be another constraint—a conservative one. Viewers seem reluctant to "restore" when there is already a familiar distal object that does not require restoration.

As for the study by Pickett (1957), I find the data uninterpretable. A prediction that might be made is that the difference between u → i type errors and i → u type errors should be larger with high-frequency noise than with low-frequency noise because the high-frequency noise will promote restoration of a high F_2. However, the difference between the error percentages does not reveal a consistent effect in that direction. With low-frequency noise, the difference is 15%, while it is 10% and 23% for two S/N ratios of high-frequency noise. The most consistent effect I can see in the data is that back vowels are less frequent responses for front vowels than vice versa. This may be completely unrelated to the nature of the masks, however; it may relate, instead, for example, to the different relative frequencies of occurrence of front and back vowels in the language.

3.5. *Porter*

I have little to say about this commentary, because Porter raised no major criticisms of the approach, but, instead, helpfully suggested a type of acoustic information (amplitude and frequency modulations of the signal) that may be important in speech perception.

I will comment on one central aspect of Porter's approach as it relates to a theory of speech perception as direct. Action systems do not, naturally, adopt static postures. Accordingly, it is not likely that phonetic segments are, even ideally, static postures of the vocal tract. In consequence, too, spectral cross sections of the acoustic signal are unlikely candidates to specify crucial vocal-tract events during speech. One advantage of the modulations is that they can reflect occurrences over time. Another is that the perceptual system appears sensitive to them.

Crucial questions, however, are whether acoustic speech signals exhibit the modulations, and whether any they exhibit reflect the control systems underlying speech production. Porter is inclined to think that they may in light of evidence that the control systems have oscillatory properties.

This approach of looking in the acoustic signal for perceivable structure needs to go hand in hand with another one of looking for the crucial control systems underlying

production of phonetic-segments and then studying their ways of structuring the acoustic signal. Without the latter tactic, we are likely to be confronted with another pattern recognition problem even after we learn what perceivable acoustic structure the signal offers. That is, we will not know how to read the structure as the perceptual system does. By working from the study of control systems for articulation forward to their acoustic consequences, we may discover more than the perceptually important acoustic structure, we can also discover how it specifies what listeners recover in perceiving speech.

3.6. *Remez*

I disagree with very little that Remez argues in his commentary. Moreover, I am grateful to him for contributing to the description of an ecological approach to the study of speech in a domain where I was unprepared to make many projections—namely, the study of naturally produced linguistic communications. However, there are two aspects of his commentary that I do take issue with, and that I have not discussed elsewhere in the commentary.

(1) One is his charcterization of my theoretical claims, and the other is his view of his own approach in relation to the one I adopt.

> Fowler's program for an ecological approach to language begins by taking the phonetic segment as the event . . . The construction rests, however, on the assumption that the distal objects of speech perception are discrete perceptual-motor-linguistic units that are, curiously, hard to perceive . . . The phonetic level is an integral aspect of the linguistic system and is neither perceptually primary nor autonomous, as Fowler claims.

However, I did not establish phonetic segments as *the* distal objects for speech perception. Nor did I claim that their perception is primary and (fully) autonomous. Here is what I said:

> [T]alkers produce phonetically-structured speech, listeners perceive it as such and they use the phonetic structure they perceive to guide their subsequent behavior . . . This defense is not intended to suggest that the perception of speech events is primary or privileged in any sense. It is to defend it as one of the partitionings of an event involving linguistic communication; therefore it is an event in its own right and requires explanation by a theory of perception.

My standing by those statements does not prevent me, as I do, from applauding Remez or anyone else for studying perception of what I called linguistic and communicative events. I only assert that speech events are perceived and their perception, just as the perception of the larger events in which they participate, must be explained by a perceptual theory. Something grounds the perception of linguistic communications. What grounds it is, in part, the grounding of the linguistic message itself largely in the articulatory gestures of the talker.

To reject the idea that articulated phonetic segments are perceived because their perception is not open, generally, to introspection is to apply a criterion for the perception of phonetic segments that is not applied to perception of other things. For example, outfielders presumably perceive the "time-to-collision" (Todd, 1981; see also Lee, 1974, 1976) of the baseball and use that information to catch the ball; however, that information would be difficult to elicit from a baseball player in an interview.

In any case, phonetic perception does guide behavior, as I point out in the target article, hence, clearly it is perceived. Moreover, it cannot be all that far away from the experimental surface. After all, someone invented the alphabet and most of us can learn to use it.

(2) My second point of disagreement with Remez is with his characterization of the goal of an ecological approach to the study of speech and language and of the methods that should be used to achieve that goal. According to Remez: "Fowler's objective, shared by all, is to explain the perception of ordinary, casual speech."

That's partly true. However, as just one objective, it is too restrictive and there are other important objectives as well.

Talkers adopt a remarkable range of speaking styles. Many of those, to be sure, count as ordinary, casual speech. However, speech to strangers, foreigners, children—indeed to anyone with whom we may share few common experiences or whom we estimate will not comprehend ordinary casual speech—is formal and careful.

So, I would not agree that our objective is to understand perception of casual speech. Rather, a theory of perception will have to explain how we perceive messages provided by whatever style the talker adopts (and how we detect what the talker's style signifies about the talker and his or her estimation of our relationship).

The foregoing more broadly defined objective is central to an ecological approach to the study of language and speech. However, it pertains to the perception of linguistic events, as I defined them, a topic about which, as I explained in the target article, I do not yet feel prepared to develop a theory. There is another objective of an ecological approach about which I have more to say, however, that is excluded in Remez' characterization of *the* goal of an ecological approach. That is to understand what about perceptual systems, informational media and environment events supports perception. Whatever form that understanding ultimately takes, I am convinced that it will be essentially invariant across speaking styles. Therefore, studies using words produced in isolation (and given that, words produced in a formal, careful style) will not provide misleading findings for a theory of perception of speech events as they would were they the only studies used to develop a theory of the perception of linguistic and communicative events.

By the same token, studies using ordinary, casual speech would provide misleading findings for a theory of perception of speech events although they would not (at least in conjunction with studies of other speaking styles, and of speaking styles in general) for a theory of perception of linguistic and communicative events. As many of the commentators have pointed out, in ordinary-speech settings, less information is provided phonetically as more is available in the communicative setting. However, formal, careful speech must be the generator for those casual and rapid speech styles in which phonetic information is left out (see also Linell, 1979). That is, careful speech is the style used when no redundant, non-phonetic information is available. It is, then, the style of choice when phonetic perception, rather than linguistic-message perception is under study.

3.7. *Studdert-Kennedy*

Studdert-Kennedy argues that when I emphasize that speech errors reveal the reality of the phonological segment, I "implicitly [dismiss] 'feature' errors as unimportant." However, I should not because the coordinative structure description of a phonetic segment

suggests an internal gestural constituency to the segment. Moreover, children appear to produce words consisting of gestures before they produce words consisting of gestures grouped into segments. Therefore, it seems, I am forced to acknowledge that phonological segments must be assembled from gestures or features. However, if I agree that talkers assemble segments from gestures or features, then, to retain my view that listeners' and talkers' perspectives on speech events are interchangeable, I must then conclude that listeners assemble perceived phonological segments from auditory parts.

I did not intend to suggest that features are unimportant. There are many sources of data that show convincingly that phonological segments have an internal structure. As to whether skilled talkers assemble segments from gestures, I do not know. However, even if I accept that characterization, I do not agree thereby that my theory of perception requires listeners to assemble them from acoustic cues or the like.

Feature errors are extremely rare as compared to word and segment errors. This need not imply that features are unimportant in production, but it strongly suggests that, like the syllable, they play a different role in speaking than does the phonetic segment or the word (see Dell, 1980, and in press). In a corpus of 70 exchange errors (like "morage in the fountain") involving intended word-initial consonants differing by more than one feature (so that a feature error could be unambiguously identified as such), three were single feature errors. The remaining 67 errors were (multiple-feature) errors in which the identity of the intended segments was preserved: that is, they were segment errors. Clearly the features of a phonetic segment are strongly inclined to stick together.

Syllables and features do make themselves evident in errors. Consonant errors preserve their intended positions relative to the vowel in a syllable (and vowels only shift into other vowel slots); but, in addition, consonants tend to move to the slots of similar consonants and substitutions, when they occur, are similar to the intended consonant or vowel.

Possibly, the reason why features exchange rarely is that their participation in segments is automatized, as Studdert-Kennedy proposes. However, that may not explain why feature- and syllable-, but not segment- and word-errors, are rare. A different possibility, implemented in Dell's model of language production, is that syllable and feature errors do not occur, because they are not selected and ordered in speech planning. Substitutions occur between similar segments because segments share features; very rarely, two substitutions occur in an utterance and these are identified mistakenly as feature errors.

I do not know how to come down on this issue. However, I do know that accepting that phonetic segments are assembled in talking does not oblige me to suppose that they are also assembled in listening. Listeners recover the talker's phonetic behavior, not how the behavior got to be the way it was. (By the same token, I can see a table without reconstructing how it was built.)

I offer the same rebuttal to Studdert-Kennedy's claim that perceptual trading relations are explained, "somewhat circularly" in articulatory terms. He argues that if talkers and listeners have interchangeable perspectives on speech events, then perceived trading between, for example, closure duration and labial release transitions should imply articulatory trading in which shorter closures imply more salient labial transitions. Since this has never been checked out, the claim of an articulatory basis for trading is circular.

However, articulatory trading does not follow necessarily from perceptual trading in experimentation even if (even though) listeners do extract information for the talker's phonetically structured articulations. The reverse may be true; that is, if there is

articulatory trading that has detectable acoustic consequences, then perceptual trading should occur. However, the perceptual trading that occurs in studies such as the "slit–split" studies of Fitch, Halwes, Erickson & Liberman (1980) occurs for a different reason. Both the closure and the transitions at release are correlated acoustic consequences of stop production. Consequently, they redundantly specify the same act. Perceptual trading will occur in research whenever experimenters treat as independent "cues" parts of the acoustic signal that are not in fact independent in origin.

A different criticism that Studdert-Kennedy raises is that I should, but do not, place the articulatory and auditory systems on equal footing in theorizing about production, perception, and most importantly, the evolution of speech forms. It is true that I do not, but I do not disagree with the ideas put forward by Lindblom (1971, 1983) that perceptual distinctiveness along with articulatory dispositions shape the evolution of phonetic-segment inventories. However, I do disagree with Studdert-Kennedy's conclusion that the problems of normalization, segmentation and invariance will be significantly illuminated by studying audition.

"Normalization" usually refers to our ability to identify phonetic segments over vocal-tract size and shape variation that affects the acoustic signals vocal tracts produce. It will, then, be illuminated by studying natural variation in vocal tracts and its acoustic consequences. (By analogy, understanding how we recognize Caucasians as such [how we normalize for individuals when we identify race] will not be illuminated by studying the visual system, but it will be illuminated by studying the morphological characteristics of Caucasians and other racial types.) By the same token, the problems of segmentation and invariance will be illuminated by studying how talkers produce separate, but overlapping, context-free phonetic segments.

3.8. *Lindblom & MacNeilage*

I will respond to three comments.

(1) Lindblom & MacNeilage object to the idea suggested by KST that study of such linguistic units as phonetic segments must await their explanation (or elimination altogether) by a dynamic account of talking. I object too. When KST write that "'segments' or phonological units as typically defined by linguists may not be relevant to the speech production system", not only are they ignoring relevant data, but, worse yet, they are saved only by a hedge ("as typically defined by linguists") from jettisoning realism as a foundation of an ecological theory. Talkers are so convinced that they produce phonological segments that they misorder them occasionally when they talk. (In the sound errors of Shattuck-Hufnagel's (1980) corpus, 66% involve single phonological segments). Alphabets depend on phonological segments constituting the internal structure of a word. If talkers and literate users of the language are misled in this way, then the concept of direct realism must be incorrect. I think that KST will come around on this point.

Turning to the frame-content perspective on errors that Lindblom & MacNeilage briefly outline, I have two comments. First, I think that Lindblom & MacNeilage are hasty in suggesting that speech errors, requiring a frame-content explanation, are special and unlike errors produced by animals or by humans in non-speech activities. Consider typing errors. They do exhibit "reversals" (such as "thses" for "these"). Obviously, these reversals do not preserve V/C membership or syllable position as speech errors do. But that is because vowels, consonants and syllables are natural vocal-tract activities, produced largely as jaws open and close. Vowels, consonants and syllables are not finger

activities, however. What would count as structure-preserving errors in typing analogous to syllable-position preserving errors in talking? (My first answer is I do not know. My second is to take the following guess.) In general, they would be ones in which enduring constraints on the movements of the fingers (analogous to syllabic jaw opening and closing in the vocal tract) are preserved, while some planned variants within that constraint (analogous to selection of a particular consonant or consonants in the closing phase of the jaw cycle and a particular vowel in the opening phase) anticipate, perseverate, exchange or substitute. Consider the following as a constraint in typing. The finger movements in typing always consist of the following components: movement toward the target key (this may consist of no movement at all), a downward movement depressing the key, an upward movement releasing it and a translation movement back to the home key if needed. It is the translation movement and the selection of which finger moves that varies from key press to key press, just as consonants and vowels vary from syllable to syllable. Do errors always consist of within-component anticipations, perseverations, exchanges and substitutions? It appears likely. A typist does not mistakenly lift his finger when he should be depressing a key; nor does he try to move the finger back to the home key before he has released a pressed-down key. Rather, translation movements are anticipated, perseverated, exchanged or substituted or one finger is substituted for another. Analysis of error-producing systems from the perspective of their own dispositions and constraints may reveal more similarities to speech than an analysis that asks whether dispositions and constraints of the vocal tract are evident in the behavior of other action systems.

A second point is that it is remarkable and probably quite significant that errors occur only in some activities and not in others. We never breathe in twice before breathing out in between; nor do we take two steps with the left foot without stepping once with the right foot in between. But we do throw our earrings to the dog and try to attach its treat to our ears and we do make speech errors. Even in speaking, some errors occur much more freely (e.g. on segments and words) than others (features and syllables). Dell (1980, and in press) suggests that errors occur freely in speech where the language is open to choice (that is, on each of the dual-tiers of language). By the same token, we do not make errors in breathing because the elastic recoil forces of the lungs strongly discourage it. But no such constraint stops phoneme misorderings, or stops the earrings from being thrown to the dog.

Before concluding too quickly (based on an analogy between frames and contents in speech and bimanual activity that strikes me as highly metaphorical) that speech is special, it may be productive to analyze each error-prone action system in terms of its own dispositions and constraints and to look closely at systems that err and those that do not, to understand how each set constitutes a natural classification of action systems.

(2) Lindblom & MacNeilage argue that invariants must be mentally and not physically defined on grounds that a gesture can substitute for an utterance in a communicative event and that, indeed, speaking style, adapted to the communicative setting, ranges from hyper articulation of the segments of a word to hypoarticulation. Therefore, the only invariant possible is what the talker is referring to.

I have dealt with parts of this issue elsewhere, but I will add a few points. First, the opposition, "mental invariant", "physical invariant" is not fruitful (Fowler, 1983, 1984; also see the target article). Nor is it accurate to refer to my position as one of searching for physical invariants if that implies physical–non-mental invariants. In my exposition

of the theory, phonetic sgments and other linguistic units are explicitly intentional, psychological, ecological events that are physical in nature as well. As Ryle points out (1949), mental events are not necessarily different things in different places from physical events; they are aspects of certain physical events.

That aside, consider hypoarticulations, hyperarticulations and gestures. From the perspective of my partitioning of a communicative event anyway, Lindblom & MacNeilage are confusing invariants for "speech events", which I claim are articulatory, with invariants, if any, for a communicative event.

Consider the substitutability of gestures and speech. To take an example at random, consider a person either uttering the word "ball" in some discourse or other or else pointing to a ball. These are, in some sense, equivalent acts of referring. However, they are not equivalent at the level of speech events. Can a person perceive the difference between /bɔl/ and a gesture toward a ball? Yes. Can a person perceive the difference between the /d/s in /di/ and /du/? No. Is saying /bɔl/ equivalent to pointing to a ball in all cultures? No. Do the consonants in /di/ and /du/ always count, cross-linguistically, as the same consonant? Yes. These different answers reinforce a view that there is a realistic partitioning of communicative events that includes what I called speech events. Here, and only here, in language perception, is invariance perceived as identical-ness.

I offer the same kind of answer to the discussion of hypo- and hyperarticulations. (I agree, by the way, that this is a study of central importance to the investigation of communicative events.) The talker's adoption of a location on the HH continuum reflects his or her estimation of the extent to which information for the words s/he is uttering is specified elsewhere (in the immediate communicative setting, the listener's memory, etc.) so that it need not be supplied redundantly. Studying this, however, is not studying speech events; it is studying communicative events. I do not think that the articulatory invariants for "the", for example, *are* preserved in most naturally occurring utterances. Talkers listening to their own productions would agree that they have not produced intact "the"s. To find out what the invariants are for a word, it is necessary to elicit productions in settings where there are essentially no redundant, non-phonetic sources of information (see also Linell, 1979). The study of communicative events, then, includes the study of how these invariants are gradually lost as other sources of information are allowed into the communicative setting.

(3) "Another attractive feature of the EPACT program is the insistence of working out descriptions from bottom up both in production and perception . . . [An example is] the lack of interest in perceptual top-down processing."

Attractive or not, a search for "bottom up" descriptions is not a feature of my approach. The idea is to find the description of environmental events that is isomorphic with the perceiver's perspective on them. This will be a description in ecological, intentional terms of the physically occurring event. My objection to the concept of top-down (cognitive) intervention in perceiving is not because I want to substitute a "bottom up" alternative. It is because I want to substitute an alternative with no "vertical" movement, up or down, in it.

This is a very important point. In a "bottom up" account of perceiving speech, a listener has to reconstruct the talker's mental intent from hints that the physical acoustic signal provide. In a "top-down" account, that bottom-up effort is aided by cognitive mediation. In a top-down account of talking, talkers use a mental plan to guide physical gestures of the vocal tract. In all three accounts there is a causal process transforming inputs or outputs into or out of a mental domain from or to a physical domain. This is an impossible kind of process. Theories of speech production use a conceptual sleight of hand

known as the "motor command" to make the translation (Fowler, 1983). It is a sleight of hand because commands are clearly mental kinds of things that a mind can formulate, but the "commands" are to be obeyed by motor neurons that are physical kinds of things responsive to release of transmitter substance, not to commands. A workable theory of production and perception has to avoid translations across domains like this.

I have proposed (Fowler, 1984) that talking and listening involve replicating a linguistic message across four physical systems by means of the causal relations between the systems. The talker's plan (embodied, perhaps, in a neural medium) is realized in the vocal tract as the linguistic message by means of causal connections between the nervous system and the muscles and structures of the vocal tract. The message is reflected in the acoustic signal by means of causal effects of vocal-tract gestures on the air. It is perceived (embodied in the perceptual system of the listener) by means of causal effects that the vibrating air has on the sense organs. Although the communication is physically grounded at all times, it also always has a real level of structure at the psychological, linguistic scale. The physical and mental descriptions are different, simultaneously apt, descriptions of the same thing, not descriptions at different points in time and in different places (outside and inside the talker/listener) during a communicative event.

Preparation of this manuscript was supported by NICHD Grant HD16591 to Haskins Laboratories. I thank George Wolford for his comments on an earlier draft of the manuscript.

References

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

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

Carney, P. & Moll, K. (1971). A cinefluorographic investigation of fricative-consonant vowel coarticulation, *Phonetica,* **22**, 193–201.

Dell, G. (in press). A spreading-activation theory of retrieval in speech production, *Psychological Review*, in press.

Dell, G. & Reich, P. (1980). Toward a unified theory of slips of the tongue. In: V. Fromkin (ed.), *Errors in linguistic performance: slips of the tongue, ear, pen and hand.* New York: Academic Press.

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

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

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

Fowler, C. (1983). Realism and unrealism, *Journal of Phonetics,* **11**, 303–327.

Fowler, C. (1984). Current perspectives on language and speech production: A critical overview. In: R. Daniloff (ed.), *Speech science.* California: College-Hill Press.

Fowler, C. & Rakerd, B. (1985). Work group on speech and language. In: W. Warren & R. Shaw (eds), *Persistance and change: proceedings of the first international conference on event perception.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

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

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

Gibson, J. J. (1968). What gives rise to the perception of motion, *Psychological review,* **75**, 335–346.

Gibson, J. J. (1969). The change from visible to invisible: a study of optical transitions, *Perception and Psychophysics,* **5**, 113–116. Reprinted in E. Reed & R. Jones (eds.) (1982) *Reasons for realism.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

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

Halle, M. & Stevens, K. (1959). Analysis by synthesis. In: W. Wathen-Dunn & L. E. Woods (eds.), *Proceeding of the seminar on speech comprehension and processing.* Bedford, Mass.: Air Force Cambridge Research Laboratories.

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

Kelso, J. A. S., Tuller, B., Bateson, E. & Fowler, C. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 812–813.

Ladefoged, P. (1980). What are linguistic sounds made of? *Language*, **56**, 485–502.

Lahiri, A., Gewirth, L. & Blumstein, S. (1984). A reconsideration of acoustic invariance for place of articulation, *Journal of the Acoustical Society of America*, **76**, 391–404.

Lee, D. (1974). Visual information during locomotion. In: R. MacLeod & H. Pick (eds.), *Perception: Essays in honor of J. J. Gibson*. Ithaca: Cornell University Press.

Lee, D. (1976). A theory of visual control of braking based on information for time-to-collision, *Perception*, **5**, 437–459.

Liberman, A. (1982). On finding that speech is special, *American Psychologist*, **37**, 148–167.

Liberman, A., Delattre, P. & Cooper, F. S. (1952). The role of stimulus variables in the perception of stop consonants, *American Journal of Psychology*, **65**, 497–516.

Liberman, A. & Mattingly, I. (1985). The motor theory of speech perception revised, *Cognition*, **21**, 1–36.

Lindblom, B. (1971). Phonetics and the description of language. *Seventh International Congress of Phonetic Sciences*. The Hague: Mouton.

Lindblom, B. (1983). Economy of speech gestures. In: P. MacNeilage (ed.), *The production of speech*. New York: Springer-Verlag.

Lindblom, B., MacNeilage, P. & Studdert-Kennedy, M. (1984). Self-organizing processes and the explanation of phonological universals. In: B. Butterworth, B. Comrie & Ö. Dahl (eds.), *Explanations of linguistic universals*. The Hague: Mouton.

Linell, P. (1979). *Psychological reality in phonology*. Cambridge: Cambridge University Press.

Lisker, L. (1978). Rapid vs rabid: a catalogue of acoustic features that may cue the distinction, *Haskins Laboratories Status Reports on Speech Research*, **SR-54** 127–132.

Locke, J. (1983). *Phonological acquisition and change*. New York: Academic Press.

Massaro, D. & Cohen, M. (1983). Integration of visual and auditory information in speech perception, *Journal of Experimental Psychology: Human Perception and Performance*, **9**, 753–771.

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

McNeill, D. (1985). So you think gestures are nonverbal, *Psychological Review*, **92**, 350–371.

Norman, D. (1981). Categorization of action slips, *Psychological Review*, **88**, 1–15.

Oden, G. & Massaro, D. (1978). Integration of featural information in speech perception, *Psychological Review*, **85**, 172–191.

Öhman, S. (1966). Coarticulation in VCV utterances: spectrographic measurements, *Journal of the Acoustical Society of America*, **39**, 151–168.

Perkell, J. (1969). *Physiology of speech production: results and implications of a quantitative cineradiographic study*. Cambridge, Mass: MIT Press.

Pickett, J. (1957). Perception of vowels heard in noise of various spectra, *Journal of the Acoustical Society of America*, **27**, 613–620.

Polanyi, M. (1962). *Personal knowledge*. Chicago: University of Chicago Press.

Porter, R. & Castellanos, F. X. (1980). Speech production measures of speech perception: rapid shadowing of VCV syllables, *Journal of the Acoustical Society of America*, **67**, 1349–1356.

Porter, R. & Lubker, J. (1980). Rapid reproduction of V-V sequences: evidence for a fast and direct acoustic-motoric linkage in speech, *Journal of Speech and Hearing Research*, **23**, 593–602.

Reason, J. (1979). Actions not as planned. In: G. Underwood (ed.), *Aspects of consciousness*, Vol. 1. London: Academic Press.

Ryle, G. (1949). *The concept of mind*. New York: Barnes and Noble.

Samuel, A. (1981). Phonemic restoration: Insights from a new methodology, *Journal of Experimental Psychology: General*, **110**, 474–494.

Shattuck-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production planning. In: P. MacNeilage (ed.), *The production of speech*. New York: Springer-Verlag.

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**, 387–420.

Todd, J. (1981). Visual information about moving objects, *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 795–810.

Townsend, J. (1971). Theoretical analysis of an alphabetic confusion matrix, *Perception and Psychophysics*, **9**, 40–50.

Walley, A. & Carrell, T. (1983). Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants, *Journal of the Acoustical Society of America*, **73**, 1011–1022.

Warren, R. (1970). Restoration of missing speech sounds, *Science*, **167**, 392–393.

Warren, W. & Shaw, R. (1985). In: W. Warren & R. Shaw (eds), *Persistance and change: proceedings of the first international conference on event perception*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Winitz, H., Schieb, M. & Reeds, J. (1972). Identification of stops and vowels from the burst portion of /p,t,k/ isolated from conversational speech, *Journal of the Acoustical Society of America*, **51**, 1309–1317.