Journal of Phonetics (1990) **18**, 529–541

Calling a mirage a mirage: direct perception of speech produced without a tongue

Carol A. Fowler

Haskins Laboratories, New Haven CT 06511, U.S.A. and Dartmouth College, Hanover, NH 03755, U.S.A.

Received 7th August 1990

I address Morrish's contention that the compensatory articulations of a speaker without a tongue disconfirm a direct-realist theory of speech perception. Listeners' successful identifications of the speaker's utterances and their mistaken identifications pattern in ways that suggest to Morrish that listeners perceive acoustic signals, not their articulatory causes. I point out that the interpretation is inconsistent with other evidence in speech perception and with interpretations of analogous situations (mirages) in visual perception. Morrish argues that the better performance of listeners identifying the speaker's intended words in connected speech than in isolation reveals the central role of "top-down" processes in speech perception. I point out that she fails to show that the top-down influences are on perception, rather than on judgment, I suggest reasons whey we should hope that they are on judgment, and I point out that top-down information, acquired as it is by perceptual experience, would be itself impoverished were perception impoverished in the absence of top-down help. Acquiring useful top-down information presupposes veridical perception.

Morrish (1990) offers challenges to two claims of a direct-realist theory: its claim that perceptual objects are articulatory and its claim that perception is direct (that is, unmediated by "top-down" processes). I will consider each challenge in turn and then address some more particular criticisms that Morrish raises against the theory.

1. Objects of speech perception

The direct-realist theory of speech perception (Fowler, 1986*a*, *b*; Rosenblum, 1987; Fowler & Rosenblum, 1990) is embedded in a larger, universal theory of perception as direct (J. J. Gibson, 1966, 1979). In the universal theory, perceptual systems constitute the only means by which organisms can know the environment in which they act. All perceptual systems serve the function of perceiving the world in abstractly the same way. Individually, they take advantage of the fact that certain media—most notably, light, air, and the skin and joints of their bodies—are lawfully, and largely distinctively, structured by the objects and events in the

0095–4470/90/040529 + 13 $03.00/0

environment. The structure in these media in turn stimulates the sense organs—respectively, the eyes, ears and kinesthetic receptors of the body—imparting its structure to them. Because the media have been lawfully, and largely distinctively, structured by objects and events in the world, their structuring can serve as information for the objects and events themselves, and it is those objects and events, not the structured light, air, or skin and joints, that perceivers need to know about in order to guide their actions in the world. Accordingly, patterns in media serve perceptual systems not as perceptual objects, but as information for their causal sources in the environment. So, for example, given a rapid, symmetrical expansion of local optical structure, perceivers (from human adults to human infants to rhesus monkeys and fiddler crabs, Schiff, 1965; Schiff, Caviness & Gibson, 1962) perceive impending collision of an object; given time-of-arrival and intensity differences at the ears, perceivers hear location in space of a sounding object; given a (to-date unanalayzed; but see, e.g., Solomon & Turvey, 1988) complex pattern of skin-deformation and joint-angle changes of the fingers, perceivers handling a rod, feel a long, rigid, cylindrical object.

According to the theory of perception of speech as direct, speech perception is not special in this respect; it depends largely on the auditory system that evolved to recover acoustic-signal-producing events in the world. While there is no obvious advantage, having to do with the requirements of perceiving speech itself, to the recovery of vocal tract actions from acoustic speech signals, their recovery is necessitated by the universal function of perceptual systems to recover the world from stimulation at the sense organs. We cannot opt to hear the acoustic signal rather than vocal-tract action in speech any more than we can opt to see the reflected light rather than the printed page when we read just because it would seem to make little difference in linguistic communications whether or not distal causes of stimulation at the sense organs are recovered.

Notice that the claim that perceptual objects are environmental does not require acceptance of the second claim of a direct-realist theory, considered shortly, that perception is direct (see also, Fowler, in press *a*). Among theorists of speech perception, there are also, of course, motor theorists, who agree that perceptual objects of speech are articulatory, but deny that perception is direct (e.g., Liberman & Mattingly, 1985). Among theorists of visual perception, no one holds the view that reflected light is perceived rather than environmental events (or representations of them); yet most theorists of visual perception hold that perception is indirect.

To drive home this point and to extract an implication from it, I quote from a theorist who distances himself as far as possible from the direct-realist point of view namely Jerry Fodor (e.g., Fodor & Pylyshyn, 1981). Here are his views on the outputs of "transducers" (that is, sense organs) and of "input systems" (perceptual systems):

> Whereas transducer outputs are most naturally interpreted as specifying the distribution of stimulations at the 'surfaces' (as it were) of the organism, the input systems deliver representations that are most naturally interpreted as characterizing the arrangement of *things in the world* (Fodor, 1984, p. 42, italics in the original).

While direct realists would accept much of this,[1] they would not endorse Fodor's

[1] They disagree that the perceptual systems yield a representation of the world, rather than the world itself.

next sentence at all: "Input analyzers are thus inference-performing systems within the usual limitations of this metaphor".

He juxtaposes the same two ideas shortly thereafter (p. 45):

> The character of transducer outputs is determined, in some lawful way, by the character of impinging energy at the transducer surface; and the character of the energy at the transducer surface is itself lawfully determined by the character of the distal layout. Because there are regularities of this latter sort, it is possible to infer properties of the distal layout from corresponding properties of the transducer output. Input analyzers are devices which perform inferences of this sort.

In respect to the ideas I have quoted, Fodor's point of view is commonplace outside the field of speech perception. The implication I want to draw from this observation is that, when, *inside* the field of speech perception, Morrish and others (Ohala, 1986; Diehl & Kluender, 1989*a*) reject the direct-realist claim that perceptual objects are articulatory, they are distancing themselves not only from a direct-realist theory of speech perception, but also from most general theories of perception as indirect. Oddly, despite explicit denials (Diehl & Kluender, 1989*a*), they are thereby implying that speech perception is special after all.

Now, there is a reason why speech perception might be considered special. For some theorists and investigators (e.g., Hammarberg, 1976, 1982; Repp, 1981), linguistic objects of perception are mental categories, not real-world events. That is, objects of speech perception are not "out there" in the world to be perceived. Perhaps this idea stands behind Ohala's (1986) comment, repeated by Morrish, that other possible objects of speech perception than vocal-tract actions exist "upstream" of those actions, including perceptual appreciation of the linguistic significance of perceptual speech input.

I have addressed this idea elsewhere (Fowler, 1986*a*, *b*, 1990; in press *b*). I argue that linguistic objects of speech perception are public actions of speakers. Actions of the vocal tract have linguistic significance because of the way they are used in the linguistic community. While it is true that speakers may have mental concepts of linguistic categories, we should not confuse the concepts with the categories themselves.

Consider an analogy (consider two, simultaneously). There are chairs in the world, and there are baseball games. Those of us having previous experience with chairs and baseball games may have concepts relating to them in our minds. But even so, chairs and baseball games are undeniably in the world; what is in our heads is what we know of chairs and baseball games, they are not chairs or baseball games themselves. To count as winning a game, the Red Sox actually have to win one in the world; winning one in their manager's imagination does not count. Likewise, phonological segments are actions of the vocal tract that have linguistic significance because of the work they do in vocal communications. Members of a language community may have concepts of phonological segments in their heads, but, if they do, that is what they know of phonological segments, they are not the segments themselves. Just as in any other domain, in language, performance is primary, while competence is derived. Therefore, speech perception need not be special on account of its objects being covert; they are not.

Turning to Morrish's challenges to the theory based on the glossectomized

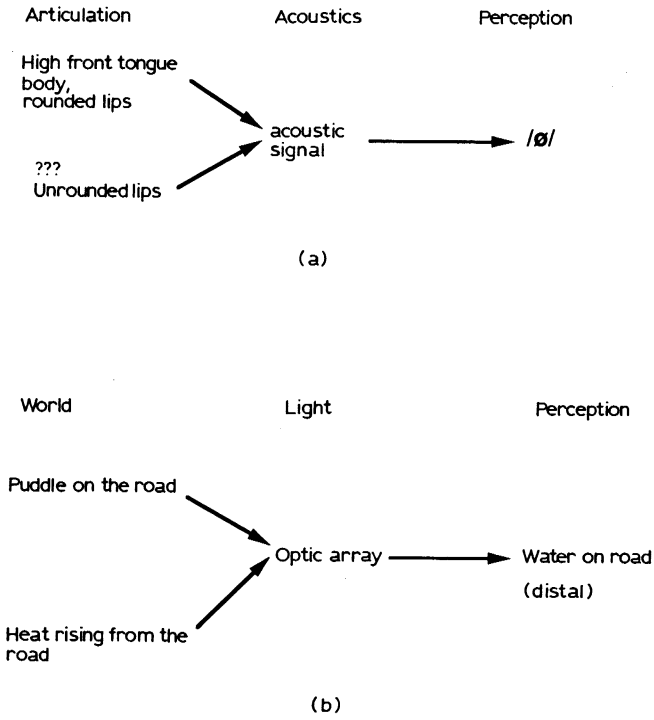

Figure 1. The structure of events in which (a) distinct articulations give rise to the same or similar acoustic signals and are perceived as the same vowel, and (b) in which distinct layouts in the world give rise to the same or similar optic arrays and are perceived as the same layout. Only in the former case do investigators draw the inference that, therefore, proximal stimuli are objects of perception.

speaker, GS, she offers two. The surface-structure of the one challenge is strong; the other is weaker.

The strong challenge has the form illustrated in Fig. 1(a): GS uses compensatory articulations that are quite distinct from normal articulations; however, the compensatory articulations sometimes yield acoustic signals like those of intact speakers. An example is GS's attempt to produce /i/ that yields an acoustic signal rather like that of a normally-produced /ø/. For its part, perception appears to conform to the acoustic signal and not to articulation in that phonetically-trained listeners transcribe GS's attempts as /ø/, when articulatorily, they are not /ø/-like. Hence, there are two articulations, (roughly) one acoustic signal and one percept; perception apparently tracks the acoustic signal.

Despite its surface persuasiveness, the argument must be wrong. First, it is precisely complementary to findings that fostered the development of the motor theory of speech perception—for example, the classic /di/–/du/ and the /pi/–/ka/–/pu/ phenomena (e.g., Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967). In the case of /di/ and /du/, the same articulation—raising of the tongue tip to the alveolar ridge of the palate—gives rise, owing to vowel-consonant coarticulation, to markedly different acoustic signals, and the percept tracks articulation, not the acoustic signal [see Fig. 2(a)]. In /pi/–/ka/–/pu/, the same acoustic signal (a stop burst centered at 1800 Hz), necessarily produced by different constrictions in the

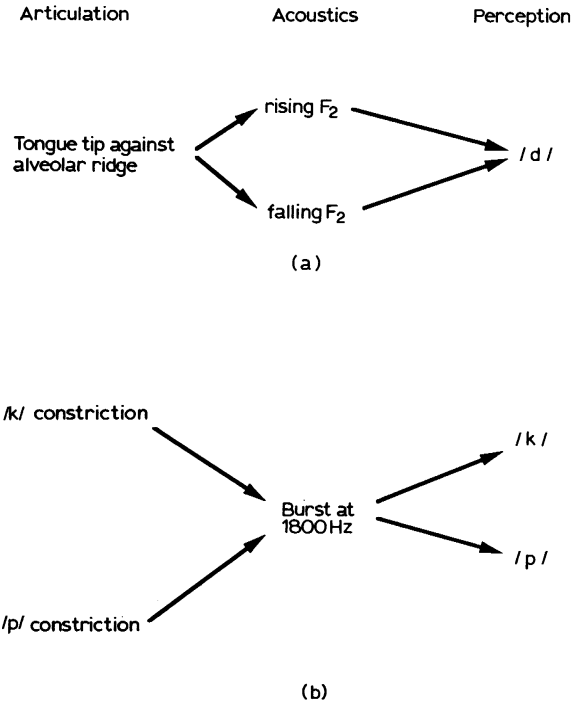

Figure 2. The structure of events in which (a) the same articulation gives rise to distinct acoustic signals and are perceived as the same consonant, and (b) distinct articulations give rise to the same acoustic signal and are perceived as distinct consonants.

vocal tract before /i/ and /u/ compared with /a/, gives rise to different consonantal percepts [Fig. 2(b)].

Obviously, we cannot have it both ways. It cannot be that perception follows acoustics and not articulation sometimes, but follows articulation and not the acoustic signal at some other times. There is an alternative solution that I have already suggested elsewhere (Fowler, 1986*b*; Fowler & Rosenblum, in press) in addressing listeners' perceptions of synthetic speech and "speech" produced by a mynah bird. It is that perception normally tracks articulation *as articulation is signaled* by the acoustic signal. Insofar as they mimic acoustic speech signals (or, in sinewave speech, insofar as it caricatures speech signals, (e.g., Remez, Rubin, Pisoni & Carrell, 1981), speech of an acoustic synthesizer and mynah bird speech can signal articulation even when the distal causes of the signals are not speech-articulation-like. For the deviant speech of the glossectomized speaker to be perceived as speech at all, the speaker must create acoustic signals that mimic those produced by normal articulations. When he succeeds, listeners hear the normal vocal-tract actions, not the compensatory articulations that, in fact, occurred. That is, they hear a mirage.

That this interpretation is viable is suggested by the example Morrish raises (after Fowler, 1986*b*) of heat rising from the road looking like a puddle. (Later I address Morrish's interpretation of this phenomenon.) This situation has the same general form as the example Morrish provides of the speaker with no tongue compared with normal speakers [compare Fig. 1(a), (b)]. That is, there are two "distal objects"—a

puddle on the road or else heat rising from the road—which give rise (approximately) to the same optic array. Regardless of the distal event, we see the same thing, namely, water on the road. Although the form of this situation is the same as that in which we perceive the same vowel regardless of its articulatory origins, notice that no one draws the inference from it, analogous to the one that Morrish draws about perception of speech, that visual perceivers see reflected light rather than the world (or a representation of the world). Our percept is just too clearly that of water on a road. Rather than draw a conclusion that is too obviously refuted by perceptual experience, we recognize that sometimes perceivers see the world as it is and sometimes (very rarely) they see the world as it is not—that is, they experience a mirage just when the light or air is structured by some event in a way that mimics its structuring by some other, presumably ecologically more salient one (see, e.g., Coss & Moore, in press, on the ecological and perceptual salience of water, and see, e.g., Whalen & Liberman, 1987, p. 171, on the "profound biological significance of speech"). Notice that we simply accept the occurrence of mirages in vision; yet under analogous circumstances in speech perception, invoking mirages counts for Morrish as "defin(ing) the problem away" and as evidence for our hearing structured air. I argue that under consistent sets of optical and auditory conditions, we need a consistent interpretation. To coin a term, I "nominate" the world as the domain of perceptual objects universally.

The weaker argument that Morrish marshals against the idea that articulation is perceived derives from the failure of speech pathologists to guess the nature of GS's deficit. As she points out, this test was done informally. My guess is that, done formally, the results would be different. How many ways are there to explain why a talker cannot produce any other than bilabial places of articulation? The speaker either has an immobile tongue or he has no tongue at all. Quite possibly, the acoustic signal does not offer evidence to distinguish those two deficits, but given just the *phonetic* evidence, the pathologists should at least be able to narrow the possibilities to those two. Of course, GS makes the task more difficult by using compensatory articulations that are aimed at creating a lingual mirage; however, according to Morrish, he does not succeed in mimicking acoustic consequences of other-than-bilabial places of articulation.[2]

2. Perception is direct

Fifty-eight percent of GS's words are identified correctly in connected speech; some speakers without tongues are even more intelligible. Moreover, GS's words are more intelligible in connected speech than in isolation, where it is apparent that only bilabial consonantal constrictions are achieved. Morrish writes "despite the fact that GS produces an invariant acoustic pattern for plosives, listeners hear both "bilabials" and "alveolars" in connected speech. We must ask what else could account for this effect other than hypothesis-driven perception" (p. 522).

Morrish fails to make a few important distinctions concerning the ways in which past experience can affect perceptual reports. One is a distinction between

[2] Critics sometimes raise another objection to the idea of articulatory perceptual objects similar to Morrish's. If vocal tract actions are perceptual objects, why do we not know it? My answer is that we generally perceive more than we are aware of. As I was careful to point out in the article to which Morrish is replying, I do not claim that linguistically-significant actions of the vocal-tract are *the* objects of speech perception; I claim that they are the smallest ones. As Polanyi (1958) points out, we generally are unaware of the perceptual primitives *from* which we attend *to* the larger objects that primitives compose.

"perceptual learning" and "hypothesis-driven" perception. Perceptual learning—that is, learning to recover increasingly subtle bits of information from stimulation—is central to direct-realist theories. Indeed, one of the classic books on the topic of perceptual learning was written by a Gibsonian direct perceptionist (in fact, by a Gibson: E. Gibson, 1969). Accordingly, a finding that experience with a language improves one's ability to recover words in that language is expected in the theory.

More relevant here, however, is a second distinction between effects of experience on perception and effects of experience on judgment. That the semantic and syntactic contexts of GS's speech improve listeners' ability to identify his words does not imply, necessarily, the they improve listeners' ability to perceive his words. The issue of where in processing "top-down" knowledge exerts its effect is a hotly contested one in psychology, largely distinguishing "interactionists" (e.g. Marslen-Wilson, 1987) from "modularists" (e.g., Forster, 1979; see Tanenhaus & Lucas, 1987, for a review). There are quite striking instances in the literature in which it is clear that perception of ambiguous words is unaffected by prior context that completely disambiguates it; disambiguation is post-perceptual (e.g. Seidenberg, Tanenhaus, Leiman, & Bienkowski, 1982).

That Morrish does not appreciate this latter distinction is apparent in her discussion of the puddle mirage. She writes: "heat rising from a road reflects light similar to a puddle. However, our experience of such matters means that we *understand* heat rising and not water collecting on the road. We bring to perception knowledge of the context in all cases" (p. 525). But clearly we do not bring that knowledge to *perception* in this case; we bring it to judgment. All the understanding in the world cannot make the road *look* dry. We see a puddle; we judge it to be heat rising (that is why it is called a mirage).

Analogously, Betty Tuller (personal communication, 27 June 1990), who listened to recordings of GS's speech, remarked that, while words having other-than-bilabial consonants are indeed identifiable in connected speech, her own transcriptions of them, made with her phonetician's hat on, correctly assigned a bilabial place of articulation to most of the consonants. (In her example, she identified the word "nuclear" in GS's speech, but transcribed it/nupwiə/.)

The passages from my article (Fowler, 1986*a*) that Morrish quotes on the subject acknowledged—honestly and modestly, I though—that I do not consider myself to understand so-called top-down effects on perceptual reports in any deep way. While many effects, such as semantic and syntactic effects on word recognition or phonemic restoration are clearly post-perceptual in origin (see, e.g., Samuel, 1981, who finds effects on ß, rather than on d' of these influences on phonemic restoration), some, such as lexical effects on phoneme identification and on compensation for coarticulation appear not to be (e.g. Samuel, 1981; Elman & McClelland, 1988). These latter I do not understand. However, I persist in adopting a conservative approach to the interpretation of top-down effects. There are reasons for expecting perceptual systems to *resist* low-level influences of top-down knowledge. To explain why, I turn once again to Fodor (1984), by way of suggesting that this idea, like the idea of perception recovering the environment, is not peculiar to a direct-realist theory; perhaps Morrish should consider it in her interpretation of listeners' understanding of GS's speech.

> If I write "I keep a giraffe in my pocket", you are able to understand me despite the fact that, on even the most inflationary

> construal of the notion of context, there is nothing in the context of the inscription that would have enabled you to predict either its form or its content. In short, feedback is [top-down influences are] effective only to the extent that *prior* to the analysis of the stimulus, the perceiver knows quite a lot about what the simulus is going to be like. Whereas the *point* of perception, is surely, that it lets us find out how the world is, even when the world is some way that we *don't* expect it to be . . . .
>
> So: The perceptual analysis of *un*anticipated stimulus layouts (in language and elsewhere) is possible only to the extent that (a) the output of the transducer is insensitive to the beliefs/expectations of the organism; and (b) the input analyzers are adequate to compute a representation of the stimulus from the information that the transducers supply. This is to say that the perception of novelty depends on bottom-to-top perceptual mechanisms (p. 67–68; italics in the original).

3. Specific comments

3.1. *The inherent ambiguity of the acoustic signal*

Morrish quotes Diehl & Kluender (1989*b*, p. 207) who assert that acoustic signals do not "even in principle" specify the kinematics and dynamics of the sound producing event. Elsewhere she refers to an "often inherently ambiguous signal" (p. 525) and cites Ohala (1986), who ostensibly provided evidence that "the only possible strategy [for perceivers] is to guess" confronted with ambiguous information for consonant place of articulation (p. 523).

Following is Molyneux's Premise (Pastore, 1971, p. 68), another claim about something being inherently unperceivable without guessing, a claim that was also made with a high degree of confidence:

> For *distance* of itself, is not to be perceived; for 'tis a line (or a length) presented to our eye with its end toward us, which must therefore be only a point, and that is invisible.

Visual perceptionists have become much better since Molyneux (a contemporary of John Locke's) at mining the optic array for the information it provides about the environment. Molyneux was wrong; so, I believe, are Diehl & Kluender, Ohala and Morrish. We speech perceptionists have to get better, too, at mining the acoustic signal for the information it provides, rather than giving up and appealing to top-down knowledge. We cannot always depend on top-down processes to fill in the missing information (infants lack it, but learn to speak), and we must be able to posit some way in which the top-down knowledge could be acquired by those who do not have it yet (Section 3.4).

For their part, Diehl & Kluender provide no reference to a demonstration that kinematics and dynamics of speech production cannot "in principle" be recovered from the signal. To date, researchers have rarely attempted that; rather they have attempted recovery of *postures* of the vocal tract from static spectra (e.g., Ladefoged, Harshman, Goldstein & Rice, 1978; Wakita, 1973).

As for Morrish's description of Ohala's comment, it is not accurate (he does not claim that /pi/–/ti/ transitions are indiscriminable); further, it is obviously not accurate to refer to burst and transitions produced by the same bilabial or alveolar gesture as "conflicting". Ohala's own comment to which she refers was mistaken in

using findings of Blumstein & Stevens (1979) to suggest that bursts provide a more "reliable cue" to place than do transitions; I pointed that out in my commentary on his remarks (Fowler, 1986*b*). Finally, the remark she ascribes to Ohala ("this is quite incompatible . . .") is indeed a quotation from Ohala, but it appeared elsewhere in the paper and "this" did not refer to the data she cites.

3.2. *GS's dialect*

Morrish writes that "one further aspect of GS's speech that forces us to question Fowler's theory is the fact that he spontaneously recovered speech" (p. 523). She does not go on to explain why spontaneous recovery of speech production skills forces her and unspecified others to question a theory of speech perception as direct, and I do not see the connection.

Morrish goes on instead to ask how a direct realist can explain GS's recovery of his Yorkshire dialect. She concludes that, because he cannot use his old manner of producing speech and he has no models of articulation that he can copy, he must be matching his new acoustic signals to his old ones or to the acoustic signals of current Yorkshire speakers.

My account is just a little different from that. The difference is that I do not believe that perceivers can hear acoustic signals any more than they can see reflected light. To sound right to himself, however, the speaker must produce signals that create a mirage—the mirage of producing speech the way Yorkshire speakers do.

3.3. *Invariance*

Morrish quotes my 1986*a* paper at length where it suggests that the search for invariance in the acoustic signal needs to be guided by a better understanding of articulatory dynamics. She concludes: "In other words, Fowler defines away the problem of invariance as an artifact of our incorrect articulatory model" (p. 526).

On my own behalf, and on behalf of the many investigators who are struggling to understand speech production, I deeply resent its cavalier dismissal as an attempt to "define away" the problem of acoustic invariance. We have actually spent rather little time on lexicography and considerable time studying speech production in relation to linguistic phonologies. That work is paving the way for an articulatorily-informed search for acoustic invariance (in contrast to the approach of Blumstein & Stevens, which I otherwise admire, that looks for invariance corresponding to nongestural phonetic features; e.g., Blumstein & Stevens, 1979; Blumstein, Isaacs, & Mertus, 1982). Here is what has been accomplished: We know that primitives of an articulatory phonology can be identified with public actions of the vocal tract (e.g. Browman & Goldstein, 1986, 1989). We know that at least some of these public actions—I believe all of them—are implemented by synergies of the vocal tract (Abbs & Gracco, 1984; Kelso *et al.*, 1984; Shaiman & Abbs, 1987; cf. Saltzman & Kelso, 1987). Synergies are collections of constraints on articulators that cause coordinated actions to achieve invariant articulatory aims. They achieve macroscopic order in action at the phonetic-gestural level and higher. On the flip side, as the elegant research by Recasens (e.g., 1984; 1987; see also Farnetani, in press) suggests, the constraints keep coarticulation in check. Strong constraints on an articulator that are implemented in production of a phonetic segment give that

segment a strong hold on the articulator in temporal regions associated with other segments as well—that is, the segment exerts a marked coarticulatory influence of its strongly-constrained articulatory actions. Compatibly, however, those same constraints prohibit or reduce influences *from* other segments *on* that articulator's actions in the temporal region of the strongly constrained segment. In short, in their temporal domain, coarticulatory influences occur freely only as long as they do not violate a synergy's pattern of constraints and hence impede achievement of the phonetic-gestural invariant that the synergy is there to realize.

As for acoustic invariance, we know that in perception generally, informative invariants in stimulation are patterns—usually over time—that are invariantly caused by an object, event or layout. Perceivers recover the causes from their distinctive effects on informational media. In speech, once the synergies of the vocal tract that implement phonological primitives have been identified, our task will be to determine the acoustic patterns that those actions invariantly and distinctively cuase.

We have not "defined" any problems "away". We are attempting to solve three problems in compatible ways: one of making phonological primitives public, rather than covert; one of understanding vocal-tract activity during speech as the implementation of these primitives; and one of understanding how acoustic signals inform about those linguistically-significant actions of the vocal tract.

3.4. *Top-down processes and depth perception*

Morrish remarks that she and many other lack the faculties for depth perception and three-dimensional vision. Yet, even so, they manage to drive, play tennis and avoid tripping over the furniture—a testimony, she believes, to the power of top-down learned information being brought to bear on the "multiplicity of [monocular] cues" that she and unfortunate others must make do with.

Morrish apparently lacks binocular vision; so do I. But I have good news. I *see* a three-dimensional world. She does too, and she could jettison her top-down processes and still do so.

As any text on visual perception will attest, perceivers who are not blind see a three-dimensional world (see, e.g., Haber & Hershenson, 1973). There is plenty of "bottom up" information available to one eye about distance (Molyneux notwithstanding) and about the third dimension. There is the change in the relative sizes of objects at different distances from ourselves that create texture gradients. (Even the lowly bacterium is able to use a chemotaxic analog of optical texture gradients, and with nowhere to put any top-down knowledge; Pittenger & Dent, 1988). There is information provided by occlusion, there is motion parallax, and there is probably much more. Perhaps these are the multiple "cues" to which Morrish refers, requiring top-down information to be interpreted. However, this kind of thinking must be misguided.

Binocular vision provides one, happily redundant, source of information about relative distance by providing two, slightly offset, perspectives on the world. But an eye or head movement also provides at least two (pre- and post-movement) perspectives on the world, now delivered over time. Why suppose that one source of information is meaningful in itself, while the other must be interpreted? (The reason cannot be that we have an innate system for handling stereopsis; we have an innate system for handling motion parallax too—remove the occipital lobe and we cannot use that information source any more.)

Suppose that other-than-binocular information does have to be interpreted to be useful. How could those of us without binocular vision have acquired the necessary top-down knowledge in our youths? I contend that it would be impossible unless at least one perceptual system provided other-than-uninterpreted-cues. Clearly the top-down knowledge could not be acquired from experience seeing; by her description, Morrish "lack(s) the faculties of depth perception and three-dimensional vision". She started out not seeing depth, and only became able to infer it when she acquired the relevant knowledge. Possibly, as Berkeley (1709) proposed, we ascribe meaning to an uninterpreted cue such as size of an object's image on the retina by correlating it with cross-modal information—his guess was the amount of effort needed to get from the observer's current location to the object whose image is on our retina. We get top-down distance information from kinesthesis. However, such a cross-modal account of the source of top-down knowledge runs into two difficulties. First, if the retinal-image size itself is a meaningless cue requiring top-down knowledge for its interpretation, then so must the kinesthetic consequences of exerting locomotory effort be meaningless until interpreted in terms of something else (which must itself be meaningless . . .). The general lesson is that there is no way to get meaning into the system in the first place if no modality is capable of direct perception (that is, of perceiving in an unmediated way what the structure in the proximal stimulation is "about"). If one modality (the haptic one for Berkeley) is capable of direct perception, why rule it out elsewhere? Second, as Turvey (1977, p. 68) points out: "the retinal size that is contiguous and thus supposedly associated with the degree of kinesthesis and effort expended in locomotion is that retinal size that occurs at the *end* of locomotion and not which was present at the outset . . . This particular formula would work only if in our early years, we spent most of our time crawling or walking backwards!" I do not recall spending my time that way.

The general point is that here, and throughout the paper, Morrish invokes top-down knowledge as savior from the inadequacies of "bottom-up" information. But it cannot be used in this way (to coin a phrase) to "define away" the difficult problem of finding stimulus information to support perception. We cannot *need* top-down knowledge for perception, because we start out life lacking it, and the only way to get it is from perceptual experience. Top-down knowledge, acquired from perceptual experience, can only be as meaningful and accurate as the experiences were by which it was acquired. But if perceptual experience is impoverished in the absence of top-down knowledge, then the top-down knowledge, itself acquired via these impoverished perceptual experiences will be impoverished itself and hence, not helpful. Getting useful top-down knowledge requires acquiring true knowlege perceptually, and only direct perception can guarantee that.

It is surely the case that speech is frequently perceived in noisy contexts (or, less frequently, that it is produced by a damaged vocal tract) so that critical acoustic information is absent. In those cases, those of us who have relevant top-down knowledge will use it—mostly, I prefer to suppose, in judgments, not in perception. However, it cannot be the case that crucial acoustic information is, in principle and necessarily, lacking so that perception is, in principle and necessarily, dependent on top-down information, because, as for visual perception, top-down knowledge could never be acquired. Acquiring top-down lexical and other linguistic information presupposes direct perception.

Preparation of the manuscript was supported by grants HD01994 and NS-13617 to Haskins Laboratories.

References

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

Berkeley, G. (1709) An essay toward a new theory of vision. In *The works of George Berkeley*, Vol. 1 (A. C. Fraser, editor). Oxford: Clarendon Press, 1871.

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.

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.

Browman, C. & Goldstein, L. (1986) Towards an articulatory phonology, *Phonology Yearbook*, **3**, 219–252.

Browman, C. & Goldstein, L. (1989) Articulatory gestures as phonological units, *Phonology*, **6**, 201–251.

Coss, R. & Moore, M. (in press) All that glistens: Connotations of water in surface finishes, *Ecological Psychology*.

Diehl, R. & Kluender, K. (1989*a*) On the objects of speech perception, *Ecological Psychology*, **1**, 121–144.

Diehl, R. & Kluender, K. (1989*b*) Reply to commentators, *Ecological Psychology*, **1**, 195–225.

Elman, J. & McClelland, J. (1988) Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes, *Journal of Memory and Language*, **27**, 143–165.

Farnetani, E. (in press) V-C-V lingual coarticulation and its spatiotemporal domain. In *Speech production and speech modeling* (W. Hardcastle and A. Marchal, editors), Dordrecht: Kluver.

Fodor, J. (1984) *The modularity of mind*. Cambridge, MA: MIT Press (third printing).

Fodor, J. & Pylyshyn, Z. (1981) How direct is visual perception? *Cognition*, **9**, 139–196.

Forster, K. (1979) Levels of processing and the structure of the language processor. In *Sentence processing: psycholingusitic studies presented to Merrill Garrett* (W. E. Cooper & E. C. T. Walker, editors), pp. 27–85. Cambridge, MA: MIT Press.

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

Fowler, C. A. (1986*b*) Reply to commentators, *Journal of Phonetics*, **14**, 149–170.

Fowler, C. A. (1990) Comments on the contributions of Pierrehumbert and Nearey, *Journal of Phonetics*, **18**,.

Fowler, C. A. (in press *a*) Sound-producing sources as objects of speech perception: Rate normalization and nonspeech perception, *Journal of the Acoustical Society of America*.

Fowler, C. A. (in press *b*) Speaking. In *Handbook of motor skills* (H. Heuer & S. Keele editors), Goettingen: Verlag fuer Psychologie Dr C. J. Hogrefe.

Fowler, C. A. & Rosenblum, L. (1990) The perception of phonetic gestures. In *Modularity and the motor theory of speech perception* (I. G. Mattingly & M. Studdert-Kennedy, editors). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gibson, E. J. (1969) *Principles of perceptual learning and development*. New York: Appleton-Century-Crofts.

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

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

Haber, R. & Hershenson, M. (1973) *The psychology of visual perception*. New York: Holt, Rinehart and Winston.

Hammarberg, R. (1976) The metaphysics of coarticulation, *Journal of Phonetics*, **4**, 353–363.

Hammarberg, R. (1982) On redefining coarticulation, *Journal of Phonetics*, **10**, 123–137.

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

Ladefoged, P., Harshman, R., Goldstein, L. & Rice, L. (1978) Generation of vocal tract shapes from formant frequencies, *Journal of the Acoustical Society of America*, **64**, 1027–1035.

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

Liberman, A., Cooper, F. S., Shankweiler, D. & Studdert-Kennedy, M. (1967) Perception of the speech code, *Psychological Review*, **74**, 431–461.

Marslen-Wilson, W. (1987) Functional parallelism in spoken word recognition. In *Spoken word recognition* (U. Freuenfelder & L. K. Tyler, editors), pp. 71–102. Cambridge, MA: MIT Press.

Morrish, E. (1990) The direct-realist theory of speech perception: counter-evidence from the analysis of the speech of a glossectomee, *Journal of Phonetics*, **18**, 519–527.
Ohala, J. (1986) Against the direct-realist view of speech perception, *Journal of Phonetics*, **14**, 75–82.
Pastore, N. (1971) *Selective history of theories of visual perception: 1650–1950*. New York: Oxford University Press.
Pittenger, J. & Dent, C. (1988) A mechanism for the direct perception of change: The example of bacterial chemotaxis, *Perception*, **17**, 119–133.
Polanyi, M. (1958) *Personal knowledge* (fifth impression, 1974). Chicago: University of Chicago Press.
Recasens, D. (1984) Vowel-to-vowel coarticulation in Catalan VCV sequences, *Journal of the Acoustical Society of America*, **76**, 1624–1635.
Recasens, D. (1987) An acoustic analysis of V-to-C and V-to-V coarticulatory effects in Catalan and Spanish VCV sequences, *Journal of Phonetics*, **15**, 299–312.
Remez, R., Rubin, P., Pisoni, D. & Carrell, T. (1981) Speech perception without traditional speech cues, *Science*, **212**, 947–950.
Repp, B. (1981) On levels of description in speech research, *Journal of the Acoustical Society of America*, **69**, 1462–1464.
Rosenblum, L. (1987) Towards an ecological alternative to the motor theory of speech perception, *PAW Review*, **2**, 25–28.
Saltzman, E. & Kelso, J. A. S. (1987) Skilled actions: a task dynamic approach, *Psychological Review*, **94**, 84–106.
Samuel, A. (1981) Phonemic restoration: Insights for a new methodology, *Journal of Experimental Psychology: General*, **110**, 474–494.
Schiff, W. (1965) Perception of impending collision: A study of visually directed avoidant behavior. *Psychological Monographs*, **79**, Whole No. 604.
Schiff, W., Caviness, J. & Gibson, J. (1962) Persistent fear responses in Rhesus monkeys to the optical stimulus of "looming", *Science*, **136**, 982–983.
Seidenberg, M., Tanenhaus, M., Leiman, J. & Bienkowski, M. (1982) Automatic access of the meanings of ambiguous words in context, *Cognitive Psychology*, **14**, 489–537.
Shaiman, S. & Abbs, J. (1987) Phonetic task-specific utilization of sensorimotor activities. Paper presented at the annual convention of the American Speech-Language-Hearing Association, University of Wisconsin.
Solomon, H. Y. & Turvey, M. T. (1988) Haptically perceiving the distances reachable with hand-held objects, *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 404–427.
Tanenhaus, M. & Lucas, M. (1987) Context effects in lexical processing. In *Spoken word recognition* (U. Freuenfelder & L. K. Tyler, editors), pp. 213–234. Cambridge, MA: MIT Press.
Turvey, M. T. (1977) Contrasting orientations to a theory of visual information processing, *Psychological Review*, **84**, 67–88.
Wakita, H. (1973) Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms, *IEEE Transactions on Audio and Electroacoustics*, **Au-21**, 417–427.
Whalen, D. & Liberman, A. (1987) Speech perception takes precedence over nonspeech perception, *Science*, **237**, 169–171.