JOURNAL OF MEMORY AND LANGUAGE 26, 574–587 (1987)

597

Perceivers as Realists, Talkers Too: Commentary on Papers by Strange, Diehl et al., and Rakerd and Verbrugge

CAROL A. FOWLER

Dartmouth College and Haskins Laboratories

The preceding papers emphasize the importance of information in the speech signal that reflects movement. In addition, R. Diehl et al. (*Journal of Memory and Language* 26, 564–573) suggest that perception is responsive to the coarticulatory overlap of segments. These findings are concordant with a view that the components of phonology as perceived and produced are essentially gestural rather than postural. Each paper is reviewed in the light of what can be added from studies of the production of speech and the perception of other speechlike signals. The conclusion is reached that a viable approach to understanding speech perception is one that views listeners as realists, extracting information from the acoustic speech signal that specifies what talkers do. © 1987 Academic Press, Inc.

The studies by Strange, Diehl et al., and Rakerd and Verbrugge (this issue) are exciting for the unified perspective they offer on perceivers of speech—a perspective quite distinct from that offered by target theory as described and criticized by Jenkins. This new perspective is all the more promising in its compatibility with recent progress in understanding how talkers produce speech.[1]

Preparation of this manuscript was supported by NSF Grant BNS 8111470 and by NINCDS Grant NS 13617 and NIH Grant HD-01994 to Haskins Laboratories. I thank Robert Fox and an anonymous reviewer for their comments on an earlier draft of the manuscript. Requests for reprints should be sent to Dr. Carol A. Fowler, Department of Psychology, Dartmouth College, Hanover, NH 03755.

[1] *Caveat lector.* It will be apparent to readers that my commentary on these papers is both biased and incomplete. I am going to suggest that the experimental findings that the set of manuscripts describe are remarkable in their consistency with evidence on how talkers produce speech, and also with theories of perception such as a direct-realist theory and a motor theory (Liberman & Mattingly, 1985) in which listeners are presumed to recover information about speech gestures from acoustic speech signals. However, I will not attempt to defend an argument that these accounts of speech perception are the only ones that can handle the data reported here. Nor will I consider how alternative accounts might handle the data; I leave that as an exercise for the interested reader.

As for perception, research by Strange and by Rakerd and Verbrugge underscores the importance to perceivers of information in the speech signal that reflects movement or change. Research by Diehl et al. suggest, in addition, that listeners perceive phonological segments in speech in a way that is responsive to the segments' coarticulatory overlap.

These two observations—that listeners extract information from change specified by acoustic speech signals and that their manner of identifying phonological segments reflects coarticulatory overlap—are, in fact, closely related from the perspective of recent theoretical claims that components of the phonology are essentially gestural (Browman & Goldstein, 1986; see also Griffen; 1985; Pagliuca, 1982), rather than being either postural as target theory proposes or than being even further abstracted from events at the articulatory surface as some theorists have asserted (Hammarberg, 1982; Parker & Walsh, 1985; Repp, 1981). From such a view of phonological segments, both observations suggest that listeners are perceptual "realists"—that is, they perceive phonological segments as they are produced by talkers. Listeners focus on acoustic change, by hy-

0749-596X/87 $3.00

pothesis, because changing regions of the spectrum best reveal the gestural constituency of the talker's utterances. That their identification of consonants reflects coarticulatory overlap likewise follows from their attention to segments as talkers produce them.

I elaborate on these ideas below as I discuss the findings from each paper in more detail.

Phonetic Identification and Articulatory and Acoustic Change

As outlined by Jenkins, target theory proposes that vowels, as talkers intend to produce them, are static postures of the vocal tract with characteristic formant frequencies. According to target theory, the requirement that speech segments be produced in sequence and at high rates frequently prevents the vocal tract from achieving the target postures for vowels; moreover, even when it does achieve them, it cannot move from canonical posture to canonical posture without traversing intermediate states of the vocal that are not part of the talker's originally intended utterance. Worse yet, inertia ensures some additional smearing and distortion of articulatory realizations of phonological segments caused by the physical limitations of the vocal tract. Hockett (1965) likened these effects of coarticulation on planned phonological segments to the passage of an array of colored, raw Easter eggs through a wringer. Discrete planned phonological segments are, according to this analogy, irrevocably smeared in their translation from mind to mouth.

From this characterization, we should expect the best information for vowels to be provided by the least coarticulated parts of the acoustic signal—that is, by the spectral information in the temporal center of the vowel. However, research by Strange and her colleagues (e.g., Strange, Jenkins, & Johnson, 1983; Strange, Verbrugge, Shankweiler, & Edman, 1976; Verbrugge, Strange, Shankweiler, & Edman, 1976) shows that information anywhere in the vowel, accompanied by information for the vowel's "inherent duration," generally yields accurate identification. Moreover, where there is a difference in identification performance associated with information from the vowel's center or from its margins, the margins yield superior performance. Indeed, in studies reported here, Strange shows that vowel identification performance may be very little impaired over that for intact syllables when the middle 50–60% of the vowel is removed (see also Parker & Diehl, 1984; Rakerd & Verbrugge, this issue, who perform even more radical surgery) leaving the remainder split between the syllable margins. Thus, the least coarticulated part of the signal does not provide the best information for vowel identification.

It does not for consonant identification either. In several studies, Stevens and Blumstein (1978, 1981; Blumstein & Stevens, 1979, 1981) identified ostensibly invariant information for stop-consonant identification in the short-time static spectra of the consonants at their releases. Applying spectral templates of the stop's spectral properties at release to 1800 spoken CVs and VCs produced by six talkers, Stevens and Blumstein (1979) were able to classify the CV syllables with about 85% accuracy and the VCs with a somewhat lower accuracy. However, two independent studies, pitting the spectral information at stop release against information provided by the vowel-dependent consonant transitions, showed that listeners nearly always identified the stops based on the transitions rather than the spectra at release (Blumstein, Isaacs, & Mertus, 1982; Walley & Carrell, 1983). Once again, listeners are found to learn more from the changing, most coarticulated parts of the signal (cf. Kewley-Port & Luce, 1984) and less from static, albeit possibly invariant (but see Lahiri, Gewirth, & Blumstein, 1984), spectral cross-sections.

The findings of Strange and her col-

leagues on vowels are unexpected in a target theory, and, along with findings just described on consonants, they are also unexpected from Hockett's compatible view of coarticulation as a destructive coalescing of intended phonological segments in the vocal tract. They are easier to understand from a different way of characterizing phonological properties and their articulatory realizations.

As noted, a recent development in linguistics is to view essential phonological properties as gestural rather than static-featural in nature (Browman & Goldstein, 1986). To the extent that this approach succeeds, it clears the way for a phonology in which phonological segments have only properties that vocal tracts are capable of realizing nondestructively. That is, if the approach succeeds, the Easter egg analogy can be rejected as inaccurate; the vocal tract is not like a wringer. By the same token, it fosters the idea that articulation is, literally, the production of phonological segments (Fowler, 1983a) rather than the talkers' provision of incomplete articulatory and acoustic cues to ineffable phonological intentions (cf. Hammarberg, 1982; Parker & Walsh, 1985).[2]

If phonological segments are gestural, then how should consonants and vowels be characterized? In their physical aspects, they must be organizations among articulators ("coordinative structures"; see, for example, Fowler, Rubin, Remez, & Turvey, 1980; Turvey, 1977) that get the vocal tract moving so as to realize their component gestures. Indeed, recent research reveals some of these organizational relations among articulators during speech.

Vowels produced with a bite block clenched between the teeth to prevent jaw movement are acoustically normal or nearly normal from the first pitch pulse (e.g., Gay & Turvey, 1979; Lindblom & Sundberg, 1971; Lindblom, Lubker, & Gay, 1979)—even when response times are constrained so as to prevent articulatory exploration before the first pitch pulse is uttered (Fowler & Turvey, 1980). How do talkers do it?

One likely possibility is that they produce vowels by setting up organizational relations among jaw, tongue, and other articulators that cause movement toward a characteristic tongue-palate relationship. When jaw movement is eliminated, the tongue moves by virtue of its coordinative relations with the jaw, so that the intended tongue-palate relationship is approached insofar as the bite block permits.

Analogous relations among articulators have been studied somewhat more elegantly in investigations of jaw-lip and jaw-tongue coordination in production of consonants (e.g., Folkins & Abbs, 1975; Folkins & Zimmerman, 1982; Kelso, Tuller, Vatikiotis-Bateson, & Fowler, 1984). In research by Kelso et al., an experimentally induced sudden braking of the jaw during closure for final /b/ in /bæb/ led to a short latency response by the upper lip that achieved bilabial closure despite an unusually open position of the jaw. Compensatory movement by the upper lip was not present when the jaw closing gesture during /z/ of /bæz/ was perturbed in the same way. However, when closing for /z/ was perturbed, a short latency response was observed in the genioglossus muscle of the tongue, while no such response was seen during perturbations of closing during of final /b/ of /bæb/.

This evidence of "equifinality" in ges-

[2] Theorists who take the line that phonological segments are not physical in nature because they are mental or psychological kinds of things are, in my view, confusing two aspects of linguistic elements. As Ryle (1951) pointed out, many things are physical in nature, but are also psychological in virtue of their function. So, for example, a phonological segment can be articulatory in nature, because that is how it makes its public appearances, but it is also mental or psychological—not because it exists only in the mind (Hammarberg, 1982) and because it is the product of complex perceptual and cognitive processes in the brain (Repp, 1981)—but because it does linguistic work in a spoken communication.

tures of phonological segments[3] reveals coordinative relations among the articulators responsible for realizing phonological segments during speech. Research on jaw movement during unperturbed speech shows, too, that these findings of equifinality in articulatory systems for vowels and consonants are not special to conditions of experimenter-contrived perturbations. Sussman, MacNeilage, and Hanson (1973) report that the jaw position adopted during closure for a bilabial consonant varies with the height of a preceding or following vowel. This means that, for bilabial stops produced in the context of low vowels, the jaw contributes less to bilabial closure and the lips contribute more than when the same stop is produced in the context of high vowels.

This research also reveals just how elegant coarticulation can be. Far from being a destructive smearing of phonological segments, here it may be seen as a way of coordinating the sometimes competing demands of more than one phonological segment on the vocal tract and even, sometimes, on the same articulators.

Presumably, had Sussman et al. used a wider time window, they would have seen the jaw gradually marshaled increasingly by a following vowel and less by the preceding bilabial stop. Elsewhere, I (Fowler, 1984) have referred to the gradual waxing and waning influence of phonological-segmental demands on articulators as overlapping "prominence" waves. Each wave reflects the strength of influence over time of the coordinated relations among articulators for a given segment; the overlap among the waves reflects coarticulation, now seen as an elegant way of coordinating and ordering phonological segments in speech.

From this perspective, why is the best perceptual information for vowels found in the changing, most coarticulated part of the acoustic signal? One answer may be that listeners extract better information from acoustic consequences of the gestures that realize a segment than from the acoustic consequences of achievement of the gestures' target endpoints because segments are essentially gestural, not static, in nature. Another possibility, at least as interesting, is that the changing, most coarticulated parts of the signal are most informative because they best reveal coordinations of two important kinds: those among the articulators responsible for producing a segment and those among the different articulatory systems responsible for producing overlapping segments.

To see why this may be so, consider by way of analogy, a finding by Johansson (e.g., 1973) on visual perception of biological motion. Johansson placed point lights on the joints of actors' limbs and filmed the actors in the dark so that only the point lights were visible (Maas, Johansson, Jansson, & Runeson, 1970, 1971). Actors engaged in such activities as walking, running, dancing, and so on. These activities were immediately recognizable to vieweres of the film. However, they are not recognizable under two other display conditions. The actions are not identifiable from stopped frames or from an almost spectrogramlike display with time represented on the horizontal axis and the different point lights from head to ankle arrayed vertically (e.g., Johansson, 1980). In this display, movements appear as point-light tracks across the page. Possibly, people could be taught to read these displays just as they can be taught, with great difficulty, to read spectrograms. However, the coordinations among the limbs so obvious in the film are not any more evident in the display than are the coordinations of the articulators, so obvious, it seems, to perceivers of the acoustical signal, evident in a spectrographic display.

In the film, point lights exhibit complex

[3] Equifinality is the tendency that certain kinds of systems have to reach a common goal from a variety of initial conditions and by a variety of routes (e.g., von Bertalanffy, 1968).

motion trajectories consisting of vectors common to several joints ("common" motions) and vectors relative to them that are specific to just one or two joints ("relative" motions). For example, the whole body, and hence all of the joints, move along the path of locomotion. In addition, however, rotation of a leg about the hip will introduce another motion vector common to the knee and ankle, but not to the joints of the arm on the same side of the body. Finally, if the knee bends, the ankle will have a motion vector not shared by the knee. Common motions reveal coupling among the joints; relative motions reveal a degree of independence.

By analogy, vocal tract movement and its acoustic reflections may reveal the coordinations among the articulators—the tight coupling presumably characteristic of articulators jointly realizing a phonological segment and the looser couplings characteristic of articulators for different, coarticulating segments. For those articulators shared by coarticulating segments, movement over time may also allow detection of the consonantal and vocalic systems as separate by revealing the waning influence of the one segment and the waxing influence of the other.

In short, then, data on perception of vowels reported by Strange are compatible, in broad outline, with a hypothesis that listeners are sensitive to information in the acoustic speech signal about vowel production and about coarticulation of vowels and consonants. However, it is not yet evident that this hypothesis can account for some of the interesting details of subjects' response patterns that Strange reports. Most interesting, perhaps, is the asymmetry between tense and lax vowels in the importance of duration information to their identifiability. Identification of lax vowels is impaired when durations are manipulated; identification of tense vowels is not. It is not obvious from anything we know about the production of these vowels why the asymmetry should occur. However, if the hypothesis is correct that acoustic information specifies its gestural source to listeners then the perceptual asymmetry should be explainable by differences in the articulatory realizations of tense and lax vowels. That is, the articulatory basis for laxness should be such that the duration (or perhaps the extent) of the opening gestures for a lax vowel is constrained. The basis for tenseness should be less constraining of opening duration. This perceptual finding, indeed, should promote design of experiments examining articulatory foundations of tenseness and laxness of vowels as they might bear on the perceptual asymmetry.

Phonetic Identification and Coarticulation

As Diehl et al. point out, the findings they report, although compatible with more than one account of coarticulation, fit the idea of coarticulation as coproduction very neatly. Many researchers have found articulatory evidence of more or less continuous vowel production in VCV productions (Barry & Kuenzel, 1975; Butcher & Weiher, 1976; Carney & Moll, 1971; Ohman, 1966) so that the tongue body moves smoothly from its posture for the first vowel toward that for the second during a constriction for a medial consonant—even if the consonantal constriction involves the tongue in some way (Barry & Kuenzel, 1975; Carney & Moll, 1971). It is as if a consonant interposed between two vowels occludes part of the vowels' trajectories. Indeed, as I have suggested elsewhere (Fowler, 1983b), this may explain the finding that vowels are measured to shorten (in their acoustic manifestations, not necessarily in their articulatory realizations) as consonants are added to the end, or to a lesser extent, the beginning of a syllable. Perhaps they are not shortened in fact, but rather are occluded by overlapping consonants (however, see below for a qualification to the occlusion metaphor).

From this perspective, a finding that time to identify a syllable-initial consonant de-

pends on the duration of a following vowel leads to an inference that the dependence may have its roots in the coarticulatory relation of vowel to consonant. Indeed, Diehl gives the findings exactly the interpretation I would like to put on them. That the vocalic gestures span the syllable may give them the status of "common" motions, somewhat analogous to the vectors of motion shared by several joints of a locomoting body. During production of any syllable-initial consonants, the movements of the articulators include the common vectors of motion contributed by the vowel and local movements of articulators relative to them contributed by the consonant. To identify a motion as "relative" to others, the accompanying common motions must be identified. Therefore, by hypothesis, talkers adopt a vowel-first identification strategy; or at least they adopt a strategy of identifying the consonant only after a certain substantial proportion of the vowel's trajectory has taken place.

This interpretation is attractive; however, it is also premature. There are reasons for caution in accepting the interpretation, and several follow-up investigations are needed to test and develop it further.

One cautionary consideration is that consonants can be identified with very little vocalic information, and, in according to at least one report, can be identified when following vowels cannot. Studdert-Kennedy, Kewley-Port, and Pisoni (1984) report that when CV syllables (/b,d,g/ × /i,e,a,o,u/) were truncated 10, 20, 30, 50, or 70 ms after burst onset, identification of consonants was generally accurate at all syllable durations (with occasional exceptions); however, identification of vowels typically was poor at the two shortest durations.

Possibly the findings of that study are misleading. Listeners had to choose among three consonants, but five vowels. This not only lowered the guessing probability for vowels, but it may also have made the discrimination among vowels more difficult than among consonants. We should, however, take note of the conclusion of the investigators, in contrast to that of Deihl et al., that "while consonant and vowel information is carried on in parallel over the initial portions of a voiced stop-vowel syllable, the distribution of information is not temporally uniform so that phonetic decisions on the two segments may be ordered and independent" (p. 523).

As for studies to pursue the findings, as noted, we need to ask whether it is really vowel duration that is crucial. Possibly, duration *is* crucial, but anything that lengthens the word, including adding a consonant or even an unstressed syllable after the stressed vowel, may lengthen time to identify a syllable-initial consonant. To address this question, targeting times for syllable-initial consonants might be compared for syllables differing in the size of a syllable-final consonant cluster (for example, /dIs/ versus /dIst/) or for syllables followed within the same word by zero or one unstressed syllables (for example, /dIs/ versus /dIsiy/). This would make an interesting test because the extra consonant or syllable would shorten the vowel acoustically (e.g., Fowler, 1981, 1983b; Lindblom & Rapp, 1973) while lengthening the whole word.

Alternatively, possibly it is specifically the duration of the vowel that is effective in the correlation, but then we must ask what it is about vowel duration that explains the significant correlation with monitoring time for a syllable-initial consonant. Is it that listeners identify the vowel first, and a vowel's duration affects a listener's time to identify it? If so, then anything that affects vowel identification time should affect time to identify a preceding consonant. For example, reducing the number of vowels from 10 as in the experiment reported by Diehl et al. to just 2 acoustically quite distinct ones should reduce time to identify the syllable-initial consonant. Alternatively, is it that listeners can only identify the conso-

nant once a significant proportion of the vowel's trajectory beyond the end of the consonant's own trajectory has gone by? If so, then reducing the vowel inventory should not affect consonant identification time.

A final suggestion is that it might be useful to look for properties of the target syllables that may be confounded with duration and that might be responsible for the effects on targeting time. For example, in view of pervasive "compensatory" durational relations between vowels and consonants (e.g., Lindblom & Rapp, 1973), it may be the case that acoustically longer vowels are accompanied by shorter consonants, and the shortening may make the consonant difficult to identify.

In a recent study, Munhall, Fowler, Saltzman, and Hawkins (1987; see also Fowler, Munhall, Saltzman, & Hawkins, 1987) examined articulatory correlates of compensatory shortening of vowels by consonants in the syllable rhyme. An unanticipated finding in the study was that two of three talkers we examined showed shorter closure durations (measured acoustically) for syllable-initial bilabial stops followed by long vowels (/æ/) than followed by shorter vowels (/ɛ/). That is, the consonants were shortened "compensatorily" by the following vowel. For both talkers, the shorter acoustic duration of closure was accompanied by a shorter duration closing gesture of the jaw for the stop. A similar effect on the closure duration and the closing gesture of the jaw for a syllable-initial stop was seen when the syllable was lengthened by addition of a syllable-final consonant rather than by a long vowel. Obviously, these data are limited both in number of talkers and in number of phonetic contexts for the syllable-initial consonants; however, one possibility that the findings raise is that the acoustic signal for a stop consonant followed by a long vowel or a long syllable rhyme more generally may be shorter and hence less salient perceptually than one for a stop consonant followed by a shorter vowel or syllable rhyme. By the same token, the distinction that Diehl et al. make between relevant the irrelevant (because it is remote) variation in vowel length may be true only approximately.

Support for the finding that syllable-initial consonants are shortened when syllables are lengthened in other ways may be provided by two studies (Miller & Liberman, 1979; Summerfield, 1981). These studies show that listeners expect shorter syllable-initial consonants in syllables of different durations when the duration difference does not necessarily signal a rate change. Miller and Liberman report that the category boundary along a /ba/-/wa/ continuum varying in transition duration is shifted toward /ba/ when the syllables are closed by addition of syllable-final transitions for /d/ that increased duration of the syllable. Compatibly, Summerfield found that the boundary along a /bi/-/pi/ continuum varying in VOT shifted toward /b/ when the syllable was closed by addition of syllable-final frication for /s/ or /z/. Possibly listeners expect syllable-initial consonants to be shorter in these contexts because they are shorter, in fact, both articulatorily and acoustically.[4]

I offer these comments not in criticism, but rather in the hope that this finding by Diehl and his colleagues will be pursued.

[4] Compatibly with findings of Lisker (1974) cited by Diehl et al., and with the findings of Diehl et al. themselves (this issue), in the study by Munhall et al., the shortening of the consonant in the context of /æ/ as compared to /ɛ/ was restricted to the closure interval of the stop; it did not extend to events following consonantal release. However, compatible with the perceptual findings of Miller and Liberman (1979) and of Summerfield (1981), shortening due to the addition of a syllable-final consonant was accompanied by shortening of closure, VOT, and of the vowel itself. In general, we find that factors that shorten the vowel (such as addition of a consonant to the rhyme) also shorten VOT; longer vowels (/æ/ versus /ɛ/) are associated with longer VOTs.

The dependence of consonant identification on characteristics of the following vowel offers the possibility that listeners are sensitive to the coarticulatory overlap of consonantal and vocalic segments. This is just as it should be if listeners are "perceptual realists," hearing phonological segments as talkers produce them. By the same token, one of the great problems remaining to be worked out for perceivers is to understand how they achieve a segmentation of the speech signal into the ordered consonants and vowels intended by the talker. The findings of Diehl et al. (see also Fowler, 1984; Fowler & Smith, 1986) may be showing us that the listener does so by tracking the coarticulatory overlap of the talker's articulations.

Phonetic Identification and Normalization

If vowels were, essentially, target postures of a canonical vocal tract specified by a target formant pattern, a problem for listeners, as Jenkins points out, would be one of normalizing for the many different sizes of vocal tracts and the many different formant patterns they produce for the same vowel.

Normalization does not seem to pose much of a problem for perceivers, however (Assman, Nearey, & Hogan, 1982; Peterson & Barney, 1952; Shankweiler, Strange, & Verbrugge, 1977; Verbrugge, Strange, & Shankweiler, 1976), and the research by Rakerd and Verbrugge (see also Verbrugge & Rakerd, 1986) offers some insight as to why.

In their research, vowels in "hybrid" silent-center syllables—that is, silent-center syllables in which vocal tracts responsible for the starting and ending transitions differ in sex and no doubt in size as well—are identified as well as silent-center vowels from a single vocal tract. Rakerd and Verbrugge show just how distinct the "target" formant patterns are for the two syllable margins in the hybrid silent-center vowels in their first figure. The two syllable margins point to markedly different target frequencies for the same vowel.

What is a vowel such that it survives a change in vocal tract in midutterance? And, by the same token, how do listeners go about identifying vowels from an acoustic speech signal so that even changes in vocal tract during the middle of a (silent-center) vowel are not disruptive?

Evidently, vowels—and presumably phonetic segments more generally—must be, as Rakerd and Verbrugge suggest, patterns of articulatory change that can be realized in any vocal tract or vocal-tract-like system. Just as the visible transformation, "rolling," for example, can be realized by any object of the appropriate sort in the appropriate environmental setting—regardless of its size and other properties—so, for example, /a/ may be seen as an articulatory transformation applicable to any vocal tract—indeed, to any physical system with appropriate physical characteristics—independently of its size.

As for how listeners identify phonetic segments from an acoustic speech signal, research by Rakerd and Verbrugge suggests very strongly that they look for evidence of the appropriate transformation, and when they find evidence for it, are remarkably insensitive to competing evidence that the transformation could not, after all, have taken place, because it would have had to span two vocal tracts.

I must confess that I find the results of these experiments strongly counterintuitive in their suggestion that listeners hear vowels that span two quite distinct vocal tracts. The finding appears inconsistent with an idea that listeners literally track the articulators of a talker. (It may not be, however; see below.) If that were precisely what listeners do, then, it would seem information for a shift in vocal tracts should be recognized as such and listeners should not use information from two distinct vocal tracts to identify a common phonetic seg-

ment. Indeed, because the research runs counter to this view of perception, I would look hard for an alternative interpretation were it not for confirming evidence from two independent sources in the speech literature and for compatible evidence from elsewhere.[5]

Remez and Rubin and their colleagues (1983; Remez, Rubin, Pisoni, & Carrell, 1981) have found good listener performance identifying words in "sine-wave" speech. Sine-wave speech is created by eliminating most of the acoustic signal and replacing the rest, most notably the center frequencies of formants, by sine waves. The frequency and amplitude variations of the formants are preserved. However, the signal provides strong evidence that it was not produced by a vocal tract. There are no glottal vibrations corresponding to voicing nor any harmonic relations among the formant-tracking tones to allow a fundamental frequency to be extracted. There are no resonances at all—only sine waves tracking the center frequences of formants in real speech. As Remez and Rubin (1983) point out, these signals lack the acoustic "cues" traditionally considered to underlie listeners' ability to identify phonetic segments. All that remain in the signal are time variations of tones, in frequency and amplitude, which follow those of the formants of the original signal. These tones, therefore, simultaneously provide information *for* the phonetic segments and words of the message but, paradoxically, also provide evidence *against* the idea that the source was a vocal tract.

Left to themselves, listeners may or may not identify such sequences as composed of words. However, if listeners are told that the sequence is a sentence in English and are asked to identify the words of the sentence, they often do quite well, picking out about 70% of the words. Compatibly with the research by Rakerd and Verbrugge, these findings help to reveal both what in an acoustic speech signal provides crucial phonetic information—namely, information for gestural transformations of various sorts—and what listeners attend to when they extract phonetic information. They focus on the transformations and can somehow leave aside compelling information that the source of the signal was not a vocal tract at all.

A second piece of evidence pointing in the same direction is provided by the phenomenon of "duplex" perception (e.g., Liberman, Isenberg, & Rakerd, 1981). Here, a "base" syllable, consisting, for example, of /s/ frication followed by a short silent period and a vowel with steady state formants, is presented to one ear of a listener, while second- and third-formant transitions appropriate to /p/ or /t/ are presented to the other. Heard in isolation, the base is identified as "sa" or "s-a" or sometimes "sta." Presented binaurally with the transitions attached, the syllable is identified as "spa" or "sta" as appropriate for the particular transitions presented. When the base and transitions are presented dichotically, listeners report hearing "spa" or "sta" in the ear receiving the base and "chirps"—the isolated transitions—in the other ear. Thus, although listeners receive a strong indication that the transitions could not be part of the base because they occur in the wrong location in space, they receive other strong indications in the spectral and temporal continuity of base and transition that the two fragments form a single cohesive syllabic vocal-tract event—and they hear the fragments integrated. Remarkably, they also hear the transitions not integrated with the syllable, in their ear of origin; hence the percept is duplex.

[5] One route to take in seeking another interpretation might be to point out that, in the second experiment described by Rakerd and Verbrugge, listeners generally report hearing one source, not two. Therefore, from the listeners' perspectives, they are not identifying vowels that span two vocal tracts. This finding in itself requires explanation, however. In view of the enormous spectral mismatch between the syllable margins, why do listeners hear only one source? Perhaps it is a consequence of their hearing a cohesive syllable.

Before pursuing an interpretation of these findings, it may be worth pointing out that analogous findings of dominance of information for change can be found in other perceptual domains. Observers experience apparent motion not only when the same object is presented successively in two locations, but also when different-shaped objects are presented in the two locations. Reviewing these findings, Hochberg and Brooks (1978) conclude that "shape preservation thus does not appear to be an essential feature in stroboscopic motion" (p. 267). Compatible findings have been obtained by researchers studying the slower transformation of growth and aging. A cardioidal transformation mimics craniofacial and certain other forms of growth (Mark, Shapiro, & Shaw, 1986). Some objects are seen to grow and age when they are transformed in this way that do not grow and age in the real world (for example, Volkswagen bugs; Mark, Todd, & Shaw, 1981; Pittenger, Shaw, & Mark, 1979; Shaw & Pittenger, 1977). Thus, under some conditions, information provided by dynamic change is more compelling than structural information indicating that the dynamic change could not, in fact, be taking place.

It is some comfort to know that there are limits to this, however. Not all objects can be seen as aging when they undergo the cardioidal transformation. Work is underway (see Mark et al., 1986) to establish the priniciples for classifying objects that will and will not be seen as aging under the transformation.

Returning to speech and the implications of findings of the dominance of change, we should take note of an interpretation offered by Liberman and Mattingly (1985). In their revised motor theory, duplex perception is one among other indications that phonetic perception is "modular." A modular perceptual system is highly specialized in the information it accepts for processing. Generally, its role is to recover properties of a distal event from proximal stimulation that specifies the properties somewhat opaquely. An example is the neural system responsible for perceiving location in space based on time of arrival and intensity differences of an acoustic signal at the two ears. Proximal time and intensity differences at the ears are opaque with respect to location in space in that they do not inherently signify location in space. They signify location in space only to an organism having two appropriately spaced ears. A second example is the neural system that uses binocular disparity as information for depth.

For Liberman and Mattingly, coarticulation renders the speech signal somewhat opaque with respect to the phonological segments that the talker intended to produce. Phonetic segments are perceived, according to the revised motor theory, by a special neural system that recovers intended phonetic gestures from the acoustic speech signal.

The phonetic-perception system proposed by Liberman and Mattingly extracts information only for phonetic events and only from acoustic signals that specify appropriate linguistic-gestural characteristics. In virtue of its selectivity, it is insensitive to other, sometimes important, information such as the locations in space of different parts of the signal, perhaps changes in identity of the vocal tract producing the signal, or even information that the signal could not, in fact, have been produced by a vocal tract.

I would like to offer an alternative interpretation, and one intended also to salvage the idea that listeners are realists, perceiving what talkers do. Certainly, they do not perceive what talkers do when they identify vowels specified by acoustic signals that span two vocal tracts, when they perceive words in sine-wave sentences or when they hear "spa" based on acoustic signals from two spatial locations.

I do not think that these data warrant or even promote jettisoning the important idea of perceptual realism (cf. Fowler, 1986; Shaw & Bransford, 1977); however,

FOWLER, C. (1984). Segmenttion of coarticulated speech in perception. *Perception & Psychophysics,* **36,** 359–368.

FOWLER, C., MUNHALL, K., SALTZMAN, E., & HAWKINS, S. (1986). *Acoustic and articulatory evidence for consonant-vowel interactions.* Paper presented as the 112th meeting of the Acoustical Society of America, Fall.

FOWLER, C., RUBIN, P., REMEZ, R., & TURVEY, M. (1980). Implications for speech production of a general theory of action. In B. Butterworth (Ed.), *Language production. I.* (pp. 373–420). Orlando/ London: Academic Press. 373–420.

FOWLER, C., & SMITH, M. (1986). Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In J. Perkell & D. Klatt (Eds.), *Invariance and variability of speech processes.* (pp. 123–136). Hillsdale, NJ: Erlbaum.

FOWLER, C., & TURVEY, M. (1980). Immediate compensation for bite block speech. *Phonetica,* **37,** 306–326.

GAY, T., & TURVEY, M. (1979). Effects of efferent and afferent interference on speech production: Implications for a generative theory of motor control. *Proceedings of the Ninth International Congress of Phonetic Sciences.* Copenhagen: University of Copenhagen.

GRIFFEN, T. (1985). *Aspects of dynamic phonology.* Amsterdam: Benjamins.

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

HOCHBERG, J., & BROOKS, V. (1978). The perception of motion pictures. In E. Carterette & M. Friedman (Eds.) *Handbook of perception. X: Perceptual ecology* (pp. 259–306). New York: Academic Press.

HOCKETT, G. (1955). *A manual of phonology. Publications in anthropology and linguistics, Memoire 11.* Bloomington: Indiana Univ. Press, 1955.

JENKINS, J. (1987). A brief history of vowel theories. *Journal of Memory and Language,* **26,** 000–000.

JOHANSSON, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics,* **14,** 201–211.

JOHANSSON, G. (1980). Event perception. *Annual Review of Psychology,* **31,** 27–63.

KELSO, J. A. S., TULLER, B., VATIKIOTIS-BATESON, E., & FOWLER, C. (1984). Functionally-specific articulatory cooperation following jaw perturbations: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance,* **10,** 812–832.

KEWLEY-PORT, D., & LUCE, P. (1984). Time-varying features of initial stop consonants in auditory running speech: A first report. *Perception & Psychophysics,* **35,** 353–360.

KRAKOW, R., BEDDOR, P., GOLDSTEIN, L., & FOWLER, C. (in press). Coarticulatory influences on the perceived height of nasal vowels. *Journal of the Acoustical Society of America.*

LAHIRI, A., GEWIRTH, L., & BLUMSTEIN, S. (1984). A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-linguistic study. *Journal of the Acoustical Society of America,* **76,** 391–404.

LIBERMAN, A., ISENBERG, D., & RAKERD, B. (1981). Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception and Psychophysics,* **30,** 133–143.

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

LIBERMAN, A., & STUDDERT-KENNEDY, M. (1978). Phonetic perception. In R. Held, H. Leibowitz, & H.-L. Teuber (Eds.), *Handbook of sensory physiology, Vol. VIII: Perception.* New York: Springer-Verlag.

LINDBLOM, B., LUBKER, J., & GAY, T. (1979). Formant frequencies of fixed-mandible vowels and a model of speech-motor programming for predictive simulation. *Journal of Phonetics,* **7,** 147–162.

LINDBLOM, B., & RAPP, K. (1973). Some temporal regularities of spoken Swedish. *Papers in Linguistics from the University of Stockholm,* **21,** 1–59.

LINDBLOM, B., & SUNDBERG, J. (1971). Acoustic consequences of lip, tongue, jaw and larynx movement. *Journal of the Acoustical Society of America,* **50,** 1166–1179.

LISKER, L. (1974). On 'explaining' vowel duration variation. *Glossa,* **8,** 223–246.

MAAS, J., JOHANSSON, G., JANSSON, G., & RUNESON, S. (1970/1971). Motion perception (Pts. 1 and 2; Film). Boston: Houghton Mifflin.

MARK, L., SHAPIRO, B. & SHAW, R. (1986). Structural support for the perception of growth. *Journal of Experimental Psychology: Human Perception and Performance,* **12,** 149–159.

MARK, L., TODD, J., & SHAW, R. (1981). Perception of growth: A geometrical analysis of how different styles of change are distinguished. *Journal of Experimental Psychology: Human Perception and Performance,* **7,** 855–868.

MILLER, J., & LIBERMAN, A. (1979). Some effects of later occurring information on the perception of stop consonants. *Perception & Psychophysics,* **25,** 457–465.

MUNHALL, K., FOWLER, C., SALTZMAN, E., & HAWKINS, S. (1987). Manuscript in preparation.

PAGLIUCA, W. (1982). *Prolegomena to a theory of articulatory evolution.* Ph.D. dissertation, SUNY, Buffalo.

PARKER, E., & DIEHL, R. (1984). Identifying vowels in CVC syllables: Effects of inserting silence and

noise. *Perception & Psychophysics,* (1985). **36,** 310–328.

PARKER, F., & WALSH, T. (1985). Mentalism versus physicalism: A comment on Hammarberg and Fowler. *Journal of Phonetics,* **13,** 147–153.

PETERSON, G., & BARNEY, H. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America,* **24,** 175–184.

PITTENGER, J., SHAW, R., & MARK, L. (1979). Perceptual information for the age-level of faces as a higher order invariant of growth. *Journal of Experimental Psychology: Human Perception and Performance,* **5,** 478–493.

RAKERD, B., & VERBRUGGE, R. (1987). Evidence that the dynamic information for vowels is talker-independent in form. *Journal of Memory and Language,* **26,** 558–563.

REMEZ, R., & RUBIN, P. (1983). The stream of speech. *Scandinavian Journal of Psychology,* **24,** 63–66.

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.

SHANKWEILER, D., STRANGE, W., & VERBRUGGE, R. (1977). Speech and the problem of perceptual constancy. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing: toward an ecological psychology,* (pp. 315–346). Hillsdale, NJ: Erlbaum.

SHAW, R., & BRANSFORD, J. (1977). Introduction. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting and knowing: Toward an ecological psychology* (pp. 1–39). Hillsdale, NJ: Erlbaum.

SHAW, R., & PITTENGER, J. (1977). Perceiving the face of change in changing faces: Implications for a theory of object perception. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting and knowing: Toward an ecological psychology* (pp. 103–134). Hillsdale, NJ: Erlbaum.

STEVENS, K., & BLUMSTEIN, S. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America* **64,** 1358–1368.

STEVENS, K., AND BLUMSTEIN, S. (1981). The search for invariant correlates of phonetic features. In P. Eimas & J. Miller (Eds.), *Perspectives on speech research* (pp. 1–38). Hillsdale, NJ: Erlbaum.

STRANGE, W. (1987). Information for vowels in formant transitions. *Journal of Memory and Language,* **26,** 550–557.

STRANGE, W., JENKINS, J., & JOHNSON, T. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America,* **74,** 694–705.

STRANGE, W., VERBRUGGE, R., SHANKWEILER, D., & EDMAN, T. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America,* **60,** 213–224.

STUDDERT-KENNEDY, M., KEWLEY-PORT, D., & PISONI, D. (1984). Independent consonant and vowel recognition in CV syllables. In M. Van den Broecke & A. Cohen (Eds.), *Proceedings of the tenth international congress of phonetic sciences* (p. 523). Dordrecht-Holland: Foris Publications.

SUMMERFIELD, A. Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance,* **7,** 1074–1095.

SUSSMAN, H., MACNEILAGE, P., & HANSON, R. (1973). Labial and mandibular dynamics during the production of bilabial consonants. *Journal of the Acoustical Society of America,* **16,** 397–420.

TURVEY, M. (1977). Preliminaries to a theory of action with reference to vision. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting and knowing: Toward an ecological psychology* (pp. 211–266). Hillsdale, NJ: Erlbaum.

VERBRUGGE, R., & RAKERD, B. (1986). Evidence of talker-independent information for vowels. *Language and Speech,* **29,** 39–57.

VERBRUGGE, R., STRANGE, W., SHANKWEILER, D., & EDMAN, T. (1976). What information enables a listener to map a talker's vocal track? *Journal of the Acoustical Society of America,* **60,** 198–212.

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.

(Received January 26, 1987)
(Revision received June 1, 1987)