557

6 Speech Perception as "Vector Analysis": An Approach to the Problems of Invariance and Segmentation

Carol A. Fowler and Mary R. Smith
Haskins Laboratories
and Dartmouth College

INTRODUCTION

Two central problems for a theory of speech perception are those of segmentation and invariance. The segmentation problem is to partition the acoustic signal into the phone-sized segments reported by phoneticians and (literate) listeners. The invariance problem—in the aspect that interests us here—is to explain why acoustically distinct, apparently context-sensitive, versions of a phonetic segment may sound free of contextual influences to listeners. We call this the problem of "perceptual invariance."

We suggest that the invariance problem arises in part from assumptions made to resolve the segmentation problem, and that different assumptions imply a novel solution to the problem of perceptual invariance. We approach the problem of segmentation with two hypotheses, one concerning the natural structure of the acoustic signal, and one concerning the nature of perceptual systems. These hypotheses allow us to understand why listeners credit acoustic signals with phone-sized structure, and why they can report perceptual invariance for acoustically different signals.

The Problems of Segmentation and Invariance

Language is said to have "duality of patterning" (Hockett, 1960)—a relatively large number of meaningful linguistic units composed of a relatively small number of meaningless phonological constituents. Compatibly, speakers and listeners behave as if acoustic speech signals are composed of separate and serially-ordered phonetic segments. For example, speakers misorder phonetic segments in spontaneous speech errors and speaker-hearers learn to use al-

phabetic orthographies. However, in most instances, analysis of the signals has not revealed invariant acoustic correspondents of separate and ordered phonetic segments.

Analysis does reveal acoustic segments, however (Fant, 1962). For example, in spectrographic displays, certain salient changes in the signal provide markers of the edges of acoustic segments. The difficulties are that for many signals, the acoustically defined segments outnumber the phonetic segments attributed to the signal (Fant, 1960, 1962); further, across different phonetic contexts there may be differences in the *kinds* of acoustic segments identified with a given phonetic segment, not just their number; and within the borders of an acoustic segment, typically there is information about more than one phonetic segment.

When researchers adopt a solution to the segmentation problem—for example, for purposes of measuring the durations of phonetic segments (Fowler, 1981a; Klatt, 1975, 1976; Lindblom & Rapp, 1973)—generally, they partition the signal into temporally discrete segments by drawing segmentation lines perpendicular to the axis of time. It is probably accurate to say that segmentation lines are drawn where influences of one phonetic segment cease to predominate visibly in the signal and those of the next segment take over. This manner of segmentation is illustrated in Figure 6.1a. The figure presents a schematized display of a syllable consisting of three segments with time along the horizontal axis and a provisional dimension, "prominence," along the vertical axis. The prominence of a phonetic segment refers to the extent to which the acoustic signal takes its character from properties of that phonetic segment. For example, an interval of frication is identified with a fricative consonant and not with a coarticulated vowel even though the frication may bear spectral evidence of the vowel. Therefore, during a period of frication, a fricative consonant has more prominence than a coarticulating vowel. As illustrated in Figure 6.1b (i), a consequence of segmentation along prominence lines is that the acoustic interval identified with a phonetic segment is context-sensitive. A perceptual theory is required to explain why listeners treat distinct acoustic signals as tokens of the same phonetic-segment type, and, why intrinsic allophones of a phoneme *sound* free of contextual influences to listeners. This is the problem of perceptual invariance.

Proposed resolutions to both the segmentation and invariance problems have come either from reexamining the acoustic basis on which perception rests (Stevens, 1981b), or from invoking special perceptual mechanisms and strategies in the listener (Oden & Massaro, 1978). We take a tactic here that requires us to look *both* at how the signal is structured and at how the listener may accomplish the task of perception.

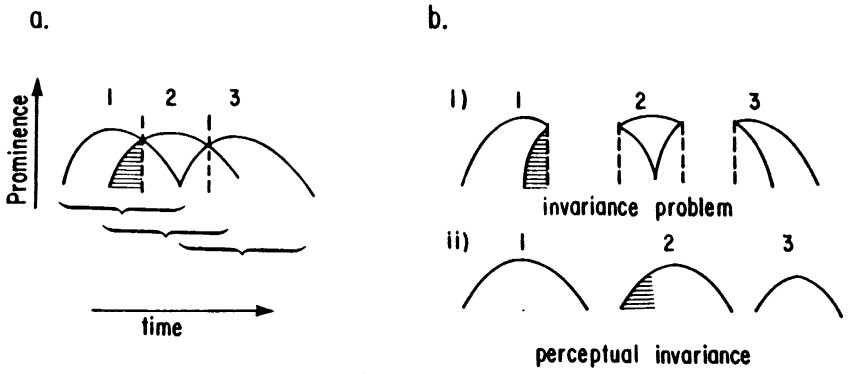

FIG. 6.1. A schematic display of coarticulated speech and two ways that it may be segmented.

The Structure of the Signal and Two Strategies for its Perception

Talkers coarticulate neighboring phonetic segments in speech—that is, their productions of neighboring phonetic segments overlap, as illustrated in Figure 6.1a. In Figure 6.1b, we illustrate two possible perceptual strategies for partitioning such a signal into phonetic segments: strategy (i) was described above as conventional for researchers measuring durations of phonetic segments in speech; strategy (ii) follows the natural structure of the speech as produced by the talker. In (i), the signal is partitioned into segments with boundaries perpendicular to the axis of time according to the dashed lines in Figure 6.1a. The resulting segments are discrete and context-sensitive. In (ii), the segmentation procedure follows natural coarticulatory lines (indicated by the braces in Figure 6.1a). Intervals of overlap among neighboring phonetic segments have, as it were, been factored from one another; therefore, the resulting segments are separate and free of contextual influences. Our research contrasts these two perceptual strategies.

We favor the hypothesis that listeners use a strategy similar to (ii) on two grounds. First, the strategy yields perceptual invariance; potentially, therefore, it can explain why, for example, the [d]s in [di] and [du] sound alike to listeners despite substantial acoustic differences between them. Second, it yields a "realistic" percept in recovering the structure of the segments as produced by a talker. Next, we provide an elaboration of this second perceptual strategy.

Perceptual Vector Analysis

Our proposal that listeners segment speech along coarticulatory lines implies that they will not always group together acoustic events that cooccur in time. Similarly, where appropriate, they will hear temporally successive events as coherent. In short, the hypothesis implies that information for segmenta-

tion and its complement, coherence, is not solely temporal succession and simultaneity.

Researchers in other domains confirm that perceivers' parsings of complex events rest on information other than coincidence and temporal or spatial separation. Johansson (1974) and Bregman (1978) show this clearly. We see our proposal as analogous to "perceptual vector analysis" as described by Johansson, and to the process of "auditory stream segregation" as described by Bregman.

Johansson (1974) finds that viewers use common motions of components of an event as important sources of information for coherence of spatially separated components of the event and use distinctive motion as information for segmentation. An attractive aspect of this approach is that it tends to yield a realistic percept. That is, when viewers perceive a vector analysis of motions of point lights filmed on the joints of a walking person or of a dancing couple, they perceive, respectively, a walking person or a dancing couple.

Bregman (1978) offers an analogous analysis of complex acoustic patterns. He describes a number of auditory displays for which listeners report separate "streams" in the signal. As Bregman points out, the principles whereby components are grouped are the same principles whereby components are dissociated or segmented. In Johansson's terms (and the terms we have adopted here), vector analysis captures both the notion of finding the common fate of components and that of segregating the remaining material relative to the unified material. Thus, both coherence and separation among components of a complex event emerge perceptually from a vector-analysis on the stimulus.

Predictions of a Vector Analysis Hypothesis for Speech

Two complementary predictions can be derived from a hypothesis that segmentation in speech perception occurs along natural coarticulatory lines. One is that acoustic consequences of coarticulation will be ascribed to the *influencing* segment. A second is that they will not contribute to the listener's perceptual experience of the *influenced* segment.

Figure 6.1 shows why these are two major consequences of a perceptual vector analysis of speech and why at least the second consequence is not expected if segmentation of speech creates temporally *discrete* segments. In Figure 6.1a, acoustic influences of segment 2 to the left of the dashed line—during a time frame in which segment 1 predominates in the signal—are identified as "anticipatory coarticulation." If listeners were to segment the signal perceptually along the dashed lines, anticipatory coarticulatory influences of segment 2 should not be ascribed to 2 itself, but rather, integrated with influences of 1, should contribute to perception of 1 as a context-sensitive phonological segment as illustrated in Figure 6.1b (i).

Alternatively, if listeners segment the signal along natural coarticulatory lines, anticipatory influences of 2 should be ascribed to 2. For its part, segment 1 should sound invariant to listeners over influences of different neighboring 2s because, by hypothesis, those influences are "factored" from 1. This alternative, labeled "perceptual invariance," is illustrated in Figure 6.1b (ii).

The first prediction has been confirmed in recent research by Whalen (1982) and Martin and Bunnell (1982). Whalen cross-spliced frication noises from CV syllables across different vocalic contexts. Subjects classified the vowels in a choice reaction-time study. They were slower and less accurate when fricative noises or transitions provided misleading information about the vowels than when the information was accurate. Martin and Bunnell obtained a similar outcome when stimuli were VCVs in which the initial vowel had been cross-spliced across different final-vocalic contexts and subjects classified the final vowels.

In themselves, these outcomes are compatible with speech segmentation by the listener into either discrete (i) or overlapping (ii) segments. By strategy (i), the percept of the fricative in Whalen's study and of the initial vowel in Martin and Bunnell's work is context-sensitive, and the nature of the contextual influence can be used by a listener to predict the identity of the following vowel. By strategy (ii), the utterance-final vowel has its onset during the frication in Whalen's study and during the utterance-initial vowel in Martin and Bunnell's experiment. That onset, no less than the later-coming information for the vowel, contributes to identification.

Strategy (ii) would be favored by its convergence with tests of the next prediction—that anticipatory information for a segment does not contribute to the perceptual experience of the segment with which it cooccurs. Rather, it is perceptually "factored" from cooccurring information for an earlier segment. This can be tested using a discrimination paradigm (Fowler, 1981b).

If listeners factor anticipatory and carryover effects of neighboring segments from the acoustic domain of a phonetic segment, then two consequences are expected. First, a given phonetic token should sound different from itself in different contexts because the contexts will cause different information to be factored from the token. Second, versions of a given phonetic type produced in different coarticulatory contexts should sound alike (free of contextual influences) as long as each is presented in its original context so that contextual influences can be factored out. These predictions are tested in the present experiments.

The studies we report are the initial ones in a series that pairs the choice reaction time procedure of Whalen (1982) and Martin and Bunnell (1982) with a discrimination paradigm first used for this purpose by Fowler (1981b). The reaction-time procedure determines whether anticipatory coarticulatory information for a forthcoming segment is ascribed to that segment. The

discrimination procedure determines whether it is factored from a phonetic segment with which it cooccurs in time.

THE EXPERIMENTS

To date, we have applied the reaction-time and discrimination procedures to two sets of stimuli, both involving coarticulatory influences of a stressed vowel on unstressed schwa. One stimulus set includes the disyllables [bəbi] and [bəba]; the other includes trisyllables [ibəbi] and [abəba] and, as filler items in the choice reaction-time study, [ibəba] and [abəbi]. In the first stimulus set, schwa receives contextual influences from a following stressed vowel (anticipatory coarticulation); in the second set, contextual influences on schwa are both anticipatory and perseverative.

Materials

Disyllables. Two tokens of the disyllables [bəbi] and [bəba] produced by a female talker were digitized at a 20 kHz sampling rate and filtered at 10 kHz. The stimuli were electronically divided into syllables at the onsets of closure for the second [b] in each disyllable. This created two unstressed [bə$_i$] syllables from the two tokens of [bəbi] (hereafter, a subscript preceding or following [bə] indicates the context in which it originated), two unstressed [bə$_a$]s, and two tokens each of the stressed syllables [bi] and [ba]. Three types of disyllables were constructed from these unstressed and stressed syllables: "original" productions in which an unstressed syllable was appended to the stressed syllable with which it had been produced originally, "spliced" productions in which an unstressed syllable was appended to a different token of the same phonetic type of stressed syllable with which it had been produced originally, and "cross-spliced" productions in which unstressed syllables were appended to stressed syllables different in phonetic type from their original neighbors. These stimulus types are illustrated in Figure 6.2.

Trisyllables. Two tokens each of [ibəbi] and [abəba] and one each of [abəbi] and [ibəba] produced by a male talker were digitized at a 20 kHz sampling rate and filtered at 10 kHz. They were divided into syllables at the onsets of closure for each of the [b]s. The tokens of [$_i$bə$_i$] from [ibəbi] and of [$_a$bə$_a$] from [abəba] were used to create "original", "spliced" and "cross-spliced" stimuli as illustrated in Figure 6.2. The fillers [ibəba] and [abəbi] appeared only as "original" productions in the choice reaction time study to eliminate the redundancy between the initial and final vowels.

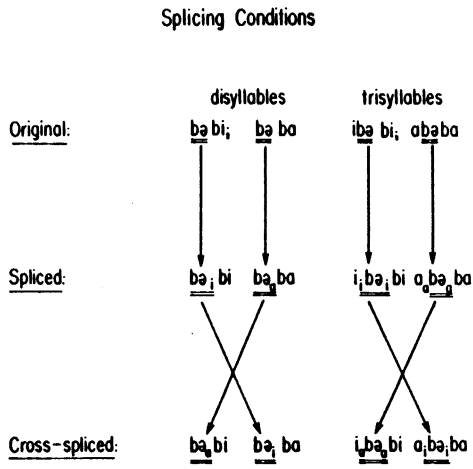

FIG. 6.2. The splicing conditions of the reaction-time and 4IAX studies.

Test Orders and Procedures

Choice reaction time. Subjects received four blocks of trials, the first serving as a practice block. There were 48 trials per block in the disyllable study and 64 per block, consisting of 32 fillers and 32 test items, in the trisyllable study. Each block included one-quarter original productions, one-quarter spliced productions, and one-half cross-spliced productions of the test stimuli, giving nine responses per subject per test item in the disyllable study and six in the trisyllable study.

Stimuli were presented on-line to listeners over headphones. In both studies, listeners were instructed to identify the final vowel of a stimulus as [i] or [a] by pressing a labeled key on the computer terminal.

If listeners use anticipatory coarticulatory information for the final vowel, choice reaction times should be slowed and accuracy reduced in the cross-spliced condition as compared to the other two conditions.

Discrimination. The paradigm was a 4IAX discrimination procedure. Only spliced and cross-spliced productions served as stimuli. On each trial, subjects received four stimuli, grouped temporally into pairs. Their task was to decide which pair had members that sound more alike one to the other. Two sample trials from the disyllable version of the study are given below:

A: bə$_i$bi---bə$_i$bi---------bə$_i$bi---bə$_a$bi

B: bə$_i$bi---bə$_a$ba---------bə$_i$bi---bə$_i$ba

In trials of type A, within and across pairs, the stressed vowels are the same. In addition, one pair of the two has identical members, whereas the second has members that differ. When the members differ, one member has

a spliced and one a cross-spliced unstressed vowel. If subjects are sensitive to the different contextual influences on these two versions of [bə], they should pick the identical pair members (in the example, the first pair) as more similar than the different members.

Trials of type B provide the critical test of our segmentation hypothesis. In these trials, stressed vowels within a pair are different, but, as in trials of type A, stressed vowels do not differentiate the two pairs of stimuli. In addition, in a B trial, one pair has identical [bə] syllables, one spliced and one cross-spliced (in the example, the second pair); the other has two different spliced [bə]s.

In the sample B trials, if listeners segment the speech signal as in Figure 6.1b(i), they should pick the members of the second pair as more similar than the members of the first pair. Alternatively, if listeners segment the signal along natural coarticulatory lines as indicated by Figure 6.1b(ii), then they should pick the members of the first pair as more similar than the members of the second. This should occur because vector analysis of spliced [bə] syllables (as in the first three disyllables in the example above) should leave the same, perceptually-invariant schwa vowel; vector analysis of a cross-spliced [bə] should leave a different residual. In the example, schwa in the last disyllable should sound high because effects of the low vowel [a] will be factored from an already raised schwa vowel.

In the two studies, the test order consisted of three blocks of 64 trials; the first block served as practice. A trials and B trials appeared equally often. Similarly, [bə$_i$] and [bə$_a$] occurred equally often as the more frequent unstressed syllable within a trial. (In the example, [bə$_i$] is the more frequent unstressed syllable.) Finally, if the four stimuli in the sample A and B trials above are given the numbers 1, 2, 3, and 4, then the stimuli within a trial appeared equally often in the orderings: 12–34, 21–43, 34–12, and 43–21.

Subjects were instructed to provide first, a "1" or a "2" signifying respectively that the members of the first or the second pair sounded more alike and, second, a confidence judgment (1: guess; 2: uncertain; 3: certain). Neither response was timed.

Subjects. Subjects were undergraduates at Dartmouth College, who received course credit for their participation. Nineteen students participated both in the choice reaction time study involving disyllables, and, in the same session, in the corresponding 4IAX discrimination study. Ten students participated in the choice reaction-time study involving trisyllables and a different group of 18 students performed the corresponding 4IAX discrimination.

Results

Choice Reaction Time. Results of the two choice reaction-time studies are shown in Table 6.1 collapsed over identity of the stressed vowel. In the

TABLE 6.1

Response Times (in msecs) and Proportion Correct in the Choice of Reaction-time Studies

Disyllables

	Original	*Spliced*	*Cross-spliced*
RT	473	464	512
S	46	43	41
Prop. correct	.97	.97	.91

Trisyllables

	Original	*Spliced*	*Cross-spliced*
RT	563	568	604
S	39	33	49
Prop. correct	.95	.94	.92

disyllable study, an analysis of variance performed on the response times revealed a significant main effect of splicing condition ($F[2,36] = 59.71$, $p < .0001$), but no effect of vowel and no interaction (both $Fs < 1$). Analysis of accuracy provided a similar outcome (splicing condition: $F[2,36] = 21.34$, $p < .0001$; vowel: $F[1,18] = 2.82$, $p = .11$; vowel by splicing condition: $F < 1$). In both analyses, the significant effect of splicing condition was due to the difference between the cross-spliced and the other two conditions, which did not differ.

The results were essentially the same in the study involving trisyllables. The effect of splicing condition on response time was highly significant ($F[2,18] = 25.03$, $p < .001$). Neither the effect of vowel nor the interaction reached significance. In the analysis of accuracy, the effect of splicing condition did not reach significance ($F[2,18] = 2.74$, $p = .09$). The main effect of vowel and the interaction were nonsignificant (both $Fs < 1$).

These results are predicted by our hypothesis that listeners segment speech along natural coarticulatory lines, but this hypothesis is not unique in making the prediction. The special prediction of our hypothesis—that anticipatory information for a forthcoming segment is factored from the phonetic segment with which it cooccurs—is tested by the discrimination studies.

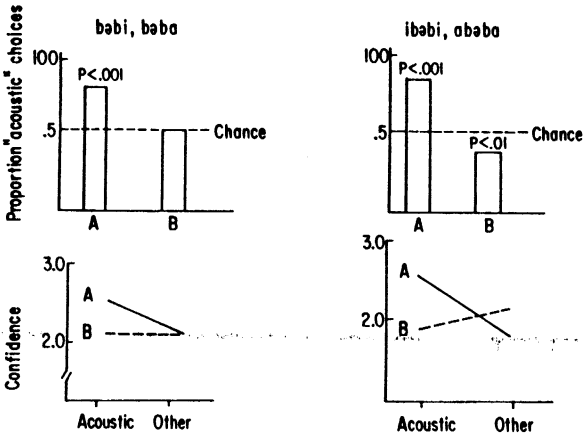

FIG. 6.3. Discrimination (top) and confidence (bottom) judgments in the disyllable (left) and trisyllable (right) studies.

Discrimination test. Figure 6.3 displays the results for the two tasks. The top half of the Figure 6.3 presents the proportions of A and B trials on which listeners select acoustically identical schwas as more similar than acoustically different schwas. They are predicted to exceed the chance or no-preference value of .5 on A trials, but to fall below .5 on B trials. The bottom half of the figure presents subjects' confidence judgments. In view of the observation that subjects tend to be more confident of their correct responses than of their erroneous responses, we predicted that they would be more confident of their selections of acoustically identical schwas on A trials than of their selections of different schwas; on B trials, their pattern of confidence judgments should reverse.

In the disyllable study, displayed on the left side of the figure, subjects selected the acoustically identical schwa vowels as more similar than the different schwas on .75 of the A trials, significantly more frequently than the chance value of .5 ($t[18] = 10.97$, $p < .001$). Listeners' confidence judgments also followed the expected pattern.

Subjects were unable to make consistent choices on B trials, however, selecting acoustically identical schwa vowels on exactly half of the trials. Their confidence judgments on B trials verify that listeners were equally unconfident when they made judgments based on acoustic identity as when they made judgments based on vector analysis.

We know both from the A trials of the present study and the reaction time study reported above that listeners are sensitive to the anticipatory coarticulatory information about [i] or [a] in the schwa vowels of the disyllables used there. We ascribe the subjects' chance performance on B trials in this study to three factors. First, in comparison to A trials, B trials are

difficult because stressed vowels are different within a pair. In A trials, schwa vowels may be compared against an invariant backdrop. Next, as compared to the reaction-time procedure, the 4IAX procedure is relatively insensitive[1] in placing memory demands on listeners not imposed by the reaction-time studies. In addition, the 4IAX procedure is difficult in requiring that listeners make explicit judgments of similarity. In contrast, the reaction-time procedure requires a very easy judgment (classifying the stressed vowel) and looks for subtle influences of the schwa vowels on classification times and errors.

The difficulties with the 4IAX paradigm are less evident using the trisyllabic stimuli in which coarticulatory influences on schwa are bidirectional.[2] Results using trisyllables (a partial replication of Fowler [1981b]) are shown on the right side of Figure 6.3. As in the disyllable study, subjects consistently picked acoustically identical schwas as more similar one to the other than different schwas on A trials. Subjects made that selection on .82 of the trials, differing significantly from the chance value of .5 ($t[17] = 16.58$, $p < .0001$). On B trials, as predicted, subjects selected acoustically identical schwas (in different coarticulatory contexts) as more similar than acoustically different schwas (each in its proper coarticulatory context) less than half (.37) of the time ($t[17] = 3.46$, $p = .003$).

Subjects' confidence judgments mirrored their discrimination selections. On A trials, subjects were more confident when they chose acoustically identical schwas than when they did not; on B trials they showed the opposite pattern. An analysis of variance on the confidence judgments shows significant main effects of trial type. Subjects were more confident overall on the relatively easy A than B trials ($F[1,17] = 10.97$, $p = .004$)—a main effect of choice type. They were more confident of their choices of acoustically identical than different schwas ($F[1,17] = 14.28$, $p = .002$), and, most importantly, a significant interaction occurred ($F[1,17] = 32.21$, $p < .001$). The interaction occurs because subjects are in fact only more confident on A than B trials when choices of acoustically identical schwas are made (A: 2.54, B: 1.89); they are less confident on A than B trials when the opposite selection is made (A: 1.86, B: 2.13). Similarly, the main effect of choice type holds only for A trials; on B trials, confidence is higher when a coarticulation-based

---

[1] By 'insensitive' we mean that the procedure places demands on subjects that may preclude their exhibiting discriminations that they may in fact make.

[2] A reviewer suggested an alternative reason why the results were positive with the trisyllables and not with the disyllables. Possibly perceivers do not work backward to adjust earlier perceptual identifications based on later ones. Therefore, only perseverative coarticulatory influences are subject to factoring. This account, however, is disconfirmed on two grounds. First, if subjects did no factoring in the disyllables, performance should have patterned identically on A and B trials. Second, in a recent study we have found factoring of anticipatory coarticulatory information for a vowel using the 4IAX procedure (Fowler, in press).

(vector analytical) choice is made than when acoustically identical schwas are judged the more similar.

CONCLUSIONS

Our findings suggest that in respect to the stimuli we selected for study, listeners segmented the acoustic signals along natural coarticulatory lines. They used anticipatory coarticulatory information for a phonetic segment as information for that segment. Moreover, they behaved as if they had "factored" those anticipatory influences from the segment with which they cooccured, hearing the influenced segment as free of contextual influences.

We are currently testing the generality of our findings across a broader range of contexts. We turn now to issues that will be important and relevant if our extensions are successful.

Finding Acoustic Support for Segmentation Along Coarticulatory Lines

We have not yet suggested what acoustic support there might be for the segmentation strategy we have observed in our listeners' behaviors. We can only suggest an approach that may be productive.

In effect, the segmentation strategy we have observed indicates (along with other evidence—for example, Fitch, Halwes, Erickson & Liberman, 1980; Fowler, 1979) that listeners use the acoustic speech signal as information about articulation. We see this role of the acoustic signal as analogous to that of reflected light in vision. In visual perception, light reflected from distal objects and events serves as a proximal stimulus providing information about the objects and events. It can provide information about them because, in being reflected from objects and events, it takes on structure specific to them. In Figure 6.4, we show the analogy we recognize between the role of reflected light in vision and that of the acoustic signal in speech perception.

The acoustic signal, structured as it passes through the moving vocal tract, can provide information about the changing shape of the vocal tract and about the articulatory gestures taking place. According to the analogy, the acoustic signal is not the object of perception itself, but rather, like reflected light in vision, is *proximal* stimulation. The "object" of speech perception is the distal source of the structure in the acoustic signal, the moving vocal tract (cf. Gibson, 1966; Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967).[3] If this analogy is apt, the place to start looking

[3] In fact, of course, the moving vocal tract is only the most peripheral of the perceptual objects. The perceptual object of which the listener is most aware is the talker's message.

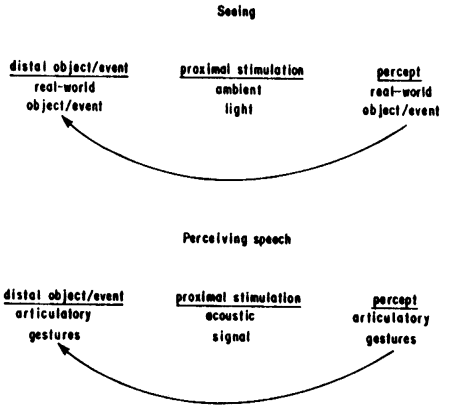

FIG. 6.4. Schematic representation of the analogous roles of reflected light and acoustic speech signals in perception.

for support for the listeners' perceptual reports is in articulation. Then, having discovered the perceived articulatory structure of speech we can look for its reflections in the acoustic signal.

Examination of coarticulated speech may reveal articulatory analogues of the "common" and "relative" motions studied by Johansson (1974). Stressed vowels—produced both during neighboring unstressed vowels, as in our studies, and during neighboring consonants—reveal themselves as relatively slow and continuous gestures of the tongue and jaw that together effect a global change in the shape of the vocal tract. Consonant gestures, produced either relatively independently of vowel gestures (as in bilabial consonants) or superimposed on vowel gestures (as in velar consonants), are rapid, local occluding gestures. Possibly, then, gestures for overlapping segments can be identified as coherent or, alternatively, as separate based on their common or distinctive rates and kinds of movements.

Other Applications of "Vector Analysis" to Speech

Listeners correctly classify phonetic tokens of a type over variations in speaking rate and over variation in vocal-tract size. These "normalizations" may be understood as factoring of invariant information for rate or for vocal-tract size from invariant information for phonetic identity.

A different application of the idea of vector analysis derives from Ohala's proposals (1981) concerning the origins of some sound changes in language. Ohala suggests that some sound changes occur when listeners fail to detect coarticulatory influences of (or, in Ohala's words, "distortions" by) neighboring segments. For example, some languages may have developed tones

for linguistic uses when listeners failed to ascribe the tonal influences of consonants on a following vowel to the consonants themselves and interpreted them instead as tones intentionally imposed on the vowels by talkers. We interpret this hypothesis as one that listeners fail to recover the natural segmentation of the produced signal. The raising of low vowels in English particularly before nasal consonants (Labov, 1981) may have a similar interpretation (Wright, 1980). This phenomenon, like the development of tones, may be seen as an inaccurate parsing of the acoustic speech signal along its natural coarticulatory lines.

ACKNOWLEDGMENTS

This research was supported by NSF Grant BNS-8111470 and NICHD Grant 16951-01 to Haskins Laboratories.

Philip Lieberman: Comment

The data of Fowler and Smith are consistent with the theory that there is a syllabic level in the perception of speech that precedes the identification of segmental phonetic elements. The responses of their listeners indicate that they can sort out sequences that are mismatched at a syllabic level. The syllable being in this case a unit of speech in which articulatory preprogramming takes place. Sven Öhman showed in 1966 that the vowels in a CV_1CV_2 sequence were coded together (Öhman, 1966). The formant frequencies of V_1 were affected by V_2. Fowler and Smith demonstrate that listeners at some level of perception are aware of this effect. Their data however, do not appear to address the theoretical issues that were raised.

Carol A. Fowler and Mary R. Smith: Response

We will briefly respond to Lieberman's published comments and then discuss other comments that were raised following the oral presentation of our paper.

Lieberman finds our data consistent with a theory that there is a syllabic level of information extraction prior to extraction of phonetic segments in speech perception. The findings he cites are that "listeners can sort out sequences that are mismatched at a syllabic level." In view of the fact that the same sequences are mismatched at the phonetic segmental level (and, for that matter, at the articu-

latory level), however, the findings do not particularly address the question of a syllabic level of information extraction; nor, accordingly, do they address the question *when* syllables are extracted, if they are, in relation to extraction of phonetic segments.

Lieberman does not explain why, in his view, our data do not address the issues of segmentation and perceptual invariance. In our view, defended earlier and in a different way below, they do.

Much of the discussion following the presentation of our paper focused on a fundamental question whether our research is misguided from the outset in assuming that phonetic or phonological segments are recovered in speech perception. Although several other interesting and useful comments were made on other topics, we will restrict our response to this very fundamental question.

Stevens pointed out that there is no strong evidence that phone-sized segments are recovered by listeners, and that, in any case, recovery of lexically-specified segments would be difficult from casual-speech productions. (His example was a casual-speech production of the word "international" realized as something like [Inɝnaešn̩ḷ].) Perkell argued that whereas production of speech must be more-or-less exhaustive ("serial" to use his wording), putting into the signal everything that a listener should recover, perception may involve short cuts when information is redundant. In particular, recovery of phonological or phonetic segments may be short circuited. (Stevens' example of "international" suggests that talkers take short cuts too; however, there must be somewhat stronger constraints on the talker's short cuts than on the listener's.) Finally, Lisker suggested that phonemes[1] are constructs derived from words.

The disagreement between at least the first two of these views and the premises of our research is not, we think, as substantial as it may have appeared to Stevens and Perkell. It is not a disagreement between two views of *processing* in speech perception, one (ours) that listeners go through a phonetic or phonological stage of processing in speech perception and another (theirs) that they do not. The premises of our research do not concern *processes* in speech perception, nor is the research designed to uncover any covert psychological processes by which vector analysis may be achieved, if it is. Instead, the premises concern a set of *conditions* that, in our view, must obtain if the essential duality of patterning in language (Hockett, 1960) is to be perpetuated. The conditions define a perceptual and articulatory "realism" (cf. Fowler, 1983) whereby talkers actually produce and listeners actually perceive the units constituting part of their "linguistic competence." Our research supports the view that conditions like these do obtain.

---

[1] We are not always certain what range of phone-sized segments the commentator intended. For example, it is a more radical claim that phonetic segments are constructions than that phonemes are. However, we would disagree with either claim. In the view of articulatory and perceptual "realism" we are working from here (see also, Fowler, 1983 and Fowler, Rubin, Remez and Turvey, 1980), spoken utterances have phone-sized structure despite coarticulatory overlap; the phonetic segments, in addition, are phonological in their deployment. Therefore, units at both levels of abstraction are in the linguistic behavior of the talker and, realized acoustically, are available for the listener to recover.

Relevant premises of our research can be characterized as follows:

1. Talkers realize linguistic units of various types in articulation (see footnote 1). In casual speech, the units they realize may not be precisely the ones they use in formal speech, but in either case, spoken words have a phone-sized segmental structure as evidenced by the occurrence of phonetic segment errors in speech production.
2. The acoustic signal provides information about those linguistic units and serves the role of message carrier about the units to listeners.
3. Listeners extract information about produced linguistic units from the acoustic signal. Among those extracted units can be phonetic units (or, even, we would argue, phonological ones [again, see footnote 1]).

Our research examines a situation in which listeners extract phonetic segments from speech and asks: *when* this occurs, what kind of segmentation do they achieve, and can that manner of segmentation explain the phenomenon of perceptual invariance?

Perkell's comment that listeners ordinarily need not, and do not, bother to recover the segmental structure of speech[2] may be correct. If it is, processing in speech perception is an example among many processes under investigation in cognitive psychology currently that can be characterized as involving an interactive relationship between "top-down" and "bottom-up" processes. Top-down processes may occur in speech, for example, when the listener knows what is forthcoming in the talker's message, and, knowing that, shortcircuits the bottom-up recovery of individual phonetic segments.

We have no particular reason to dispute that, but we find it important that the bottom up recovery of segmental structure in speech is the sine qua non for perpetuation of languages' duality of structure—which, by most accounts, is a universal, special and essential property of the language. Top-down short cuts are overlaid processes that develop based on experience with the language and with the world. As Studdert-Kennedy pointed out, speech errors provide strong evidence that talkers produce phonetic (and/or phonological) segments, our premise (1). For a child to become a talker, information for the existence of segmental structure in words must be available in the acoustic signal, our premise (2), because the acoustic signal is the only (or at least the major) source of evidence the child has. Finally, the would-be talker has to extract the produced segmental structure from the signal, our premise (3).

However, there are apparent barriers to the views that phonetic, and especially phonological, segments are in articulated speech, that they are specified in acoustic speech signals, and, hence, that they can be recovered from the signals by listeners. Some of those barriers are philosophical considerations that these segments are

[2] There is an assumption here that extraction of phone-sized segments requires more work than not extracting them when the message is redundant. This may or not be the case. It probably is the case if extraction of the segments requires a stage of processing prior to lexical identification. It need not be, however, if listeners extract words directly that have phonetic and phonological structure.

mental kinds of things and, hence, can exist in the mind, but not in the mouth (see, for example, Hammarberg, 1982; Repp, 1981). One of us has discussed these apparent barriers elsewhere (Fowler, 1983), and we will not repeat the discussion here. Two remaining barriers are the problems bof segmentation and perceptual invariance.

If, as we believe, the child must extract segments from the acoustic speech signal, there must be solutions to these problems. Our research examines one solution (viz., perceptual vector analysis) that we believe extends to both problems. We conduct our research with adult subjects rather than with children in part because it is more easier to do so but also because we believe that the capacity to perceive the phonological structure in speech is fundamental to perception at all levels of development. However, our choice of subjects does require us to prevent their normal use of top-down short cuts. We do that by using isolated nonsense words and by giving listeners tasks that coerce them into using phonetic information.

In summary, then, we agree that in conversation listeners may use short cuts and, indeed, that talkers, who say such things as [Inɝ$^{\vee}$naeš n̩ l̩], count on their doing so. However, in our view, bottom-up extraction of segmental structure is a prior and fundamental capacity. In a world without top-down short circuiting of phonetic or phonological perception, talkers would have to use formal rather than casual speech all of the time, and, perhaps, would have to slow down significantly, but communication would not be prevented. In contrast, in a world limited to top-down processes, there would be no communication, only hallucination.

ACKNOWLEDGMENTS

Preparation of this manuscript was supported by NSF Grant BNS 8111470 and NICHD Grant HD 16591-01 to Haskins Laboratories.