480

Perception & Psychophysics
1984, 36 (4), 359-368

Segmentation of coarticulated speech in perception

CAROL A. FOWLER
Dartmouth College, Hanover, New Hampshire
and Haskins Laboratories, New Haven, Connecticut

The research investigates how listeners segment the acoustic speech signal into phonetic segments and explores implications that the segmentation strategy may have for their perception of the (apparently) context-sensitive allophones of a phoneme. Two manners of segmentation are contrasted. In one, listeners segment the signal into temporally discrete, context-sensitive segments. In the other, which may be consistent with the talker's production of the segments, they partition the signal into separate, but overlapping, segments freed of their contextual influences. Two complementary predictions of the second hypothesis are tested. First, listeners will use anticipatory coarticulatory information for a segment as information for the forthcoming segment. Second, subjects will not hear anticipatory coarticulatory information as part of the phonetic segment with which it co-occurs in time. The first hypothesis is supported by findings on a choice reaction time procedure; the second is supported by findings on a 4IAX discrimination test. Implications of the findings for theories of speech production, perception, and of the relation between the two are considered.

Skilled listeners sometimes behave as if they have not extracted all of the phonetic structure from an acoustic speech signal. Listeners to fluent speech recognize target syllables and words more readily than they do phonetic segments (McNeill & Lindig, 1973; Savin & Bever, 1970), and, in their perceptions of a fluently produced sequence, they are likely to "restore" phonetic segments that are overdetermined and missing (Samuel, 1981; Warren, 1970) or mispronounced (Marslen-Wilson & Welsh, 1978).

Despite this apparent inattention to the phonetic structure of speech by skilled listeners, the structure persists in languages; that is, "duality of patterning" (Hockett, 1960) is not, apparently, disappearing from them. Indeed, it is universal to languages, presumably because it is required to maintain the openness of their open lexical classes. Moreover, the phonetic structure of words *is* psychologically real, even to the skilled listeners, just described, when they talk. For example, speech errors commonly consist of phonetic-segment misorderings and substitutions (see Fromkin, 1973), and many language games (including rhyming, alliteration, and Pig Latin, among others) involve operations performed on the phonetic-(or phonological-)segmental structure of words (cf. Pisoni, in press).[1]

If the phonetic structure of words is to be perpetuated in languages, and if language learners are to become talkers who Spoonerize and play language games, the learners must be able to extract phonetic structure from an acoustic speech signal even if they will not always do so when they become skilled users of the language. This observation implies that an acoustic speech signal must provide sufficient information for extraction of the phonetic structure of the talker's intended message. Yet, the signal has provided major barriers to investigators' efforts to extract phonetic segments from it.

Two related barriers are those of segmentation and invariance. Both problems arise because speech is coarticulated—that is, because articulatory gestures for successive phonetic segments are not temporally discrete. The segmentation problem is to understand how separate phone-sized segments may be extracted from a signal in which information for the segments overlaps in time. The invariance problem is to rationalize listeners' classifications of phonetic tokens into types. It is called the "invariance" problem because the presumption has been (e.g., Stevens & Blumstein, 1981) that its solution lies in discovering acoustic invariants that exist across tokens. The search for invariance is rendered difficult by coarticulation, which ensures that the acoustic signal during a time window most closely identifiable with one phonetic segment is context-sensitive, not (wholly) invariant. Moreover, the problem of explaining listeners' classifications goes beyond the search for acoustic invariance. Certain sets of phones (e.g., the [d]s in [di], [da], and [du]) are always classified as tokens of a common phonemic type, even though acoustic information for the different tokens is largely (and, in some synthetic stimuli, entirely) context-sensitive, and even though listeners attend to the context-sensitive information (in the example, the second-formant transitions) more closely than they do any invariant information (e.g., the shape of the release-

The research was supported by NSF Grant BNS8111470 and NICHD Grant HD16591-01 to Haskins Laboratories. I thank George Wolford for his comments on the manuscript and advice on data analysis. The author's mailing address is: Department of Psychology, Dartmouth College, Hanover, NH 03755.

burst spectrum; cf. Stevens & Blumstein, 1981) that may be present when they identify the phones (Walley & Carrell, 1983).

The present research contrasts two possible ways that listeners may segment the acoustic speech signal into phone-sized segments. These strategies offer different perspectives on the problem of explaining listeners' classifications of apparently context-sensitive phonetic segments into types.

Figure 1a displays an acoustic speech signal schematically, and Figures 1b and 1c illustrate the two segmentation strategies. In Figure 1a, the horizontal axis is time and the vertical axis is a provisional dimension, "prominence." The prominence of a segment in an acoustic signal refers to the extent to which acoustic properties characteristic of that segment are salient in the signal. For example, in a syllable, /si/, /s/ is more prominent than /i/ during the frication noise, even though production of /i/ begins before or during closure for the frication (Carney & Moll, 1971) and evidence of its production is available in the signal.

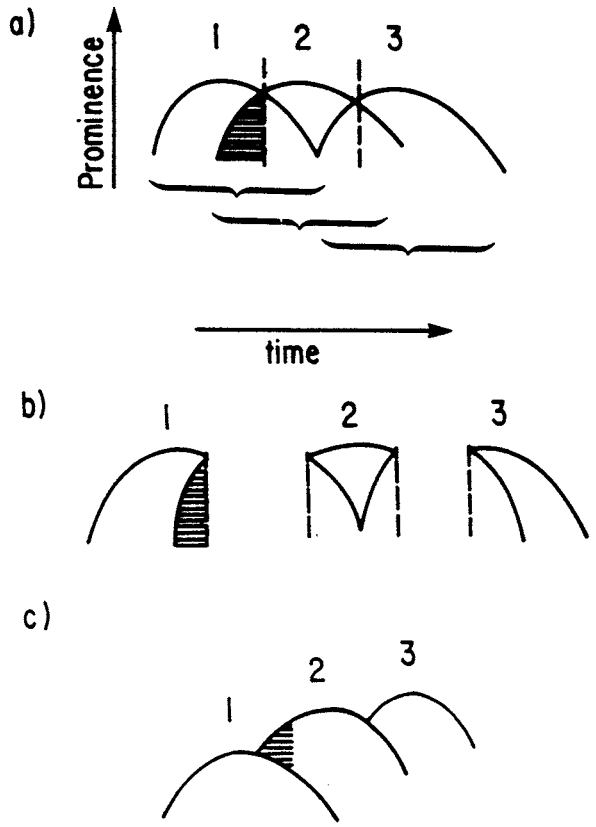

Figure 1. Schematic display of segment production. In Figure 1a, segments are produced in overlapping time frames. In Figure 1b, segmentations are made at points in time when one segment ceases to dominate in the signal and another takes over. This divides speech into discrete, context-sensitive segments. In Figure 1c, speech is segmented along coarticulatory lines into overlapping segments freed of their contextual influences.

One segmentation strategy, illustrated in Figure 1b, uses the relative prominence of successive segments to establish boundaries between them. This essentially is the procedure used in the phonetics literature when measurements are made of phonetic-segment durations, acoustically realized (e.g., Klatt, 1975; Peterson & Lehiste, 1960; and see Lisker, 1972, for other references). This procedure divides the acoustic speech signal into discrete context-sensitive phonetic segments. Disagreements concerning where to draw the segmentation lines arise when neither of two neighboring segments clearly predominates in some acoustic interval. For example, Lisker (1972) cites research in which vowel onsets sometimes include and sometimes exclude formant transitions following consonant release.

A second possible strategy is illustrated in Figure 1c. The acoustic signal is segmented along coarticulatory lines into overlapping phonetic segments, free from the contextual influences of phonetic neighbors. Thus, for example, in /si/, the onset of /i/ is identified where production of /i/ is first detectable within the /s/ frication and not where its acoustic manifestations begin to predominate in the signal.

Measurement conventions reflecting the segmentation strategy of Figure 1b are adopted in the phonetics literature to maximize reliability, not necessarily either to mimic listeners' segmentation strategies or to capture any articulatory lines of segmentation that the signal may reflect. Indeed, the literature offers a hint that the conventions do not mirror the listener's manner of segmenting the signal. The hint is provided by two independently developed, but possibly converging, lines of research.

First is evidence that listeners use anticipatory coarticulatory influences on one phonetic segment as information for the influencing segment (e.g., Alfonso & Baer, 1982; Martin & Bunnell, 1982; Ochiai & Fujimura, 1971; Whalen, 1984; but see Lehiste & Shockey, 1972). For example, Whalen (1984) cross-spliced friction noises across tokens of /sa/, /su/, /ša/, and /šu/ and asked listeners to identify the vowels of each syllable in a choice reaction time procedure. Listeners were faster and more accurate when the frication noises provided accurate anticipatory information for the vowels than when they provided misleading information.

In itself, this finding can be explained assuming the segmentation strategy of Figure 1b. Having partitioned the signal into discrete, context-sensitive allophones (Wickelgren, 1969, 1976), listeners may use the context sensitivity of the allophone that precedes a target segment to predict the target's identity.

The second line of research shows that listeners "compensate" for coarticulatory influences on phonetic segments when they identify them (Liberman, Delattre, & Cooper, 1952; Mann, 1980; Mann & Repp, 1980). For example /s/ and /š/ are distinguished acoustically in part by the relative locations of energy concentrations in their spectra, that for /s/ being higher than that for /š/. In the

context of a following /u/, however, the spectra for both consonants are lowered by anticipatory lip rounding. Compatibly, listeners accept stimuli with concentrations of energy in lower frequencies as tokens of /s/ if the frication is followed by a rounded vowel than if it is followed by an unrounded vowel (Mann & Repp, 1980). The same frication noise may be identified as /š/ before /i/, but as /s/ before /u/.

Again, in itself, this finding can be explained assuming segmentation into discrete, context-sensitive segments. The explanation is that compensation reflects an adjustment to information for an earlier segment based on knowing the effects that anticipation of the following segment should have had on it (cf. Mann & Repp, 1980).

Considered together, however, this account and the foregoing account of listeners' use of anticipatory coarticulatory information are paradoxical. For the segmentation hypothesis illustrated in Figure 1b to account for the findings from the reaction time procedure, it must be supposed that the coarticulatory effects on a segment are identifiable as such *before* the onset of the anticipated segment. Otherwise, reaction times would not be reduced when coarticulatory information is "predictive." (Nor, as Meltzer, Martin, Mills, Imhoff, & Zohar, 1976, have shown, would they be improved even further when the anticipitory information is shifted earlier in time than its natural time of occurrence.) However, for it to explain a finding of compensation, it must be supposed that the segment following a context-sensitive allophone is used to guide the identification of the contextual influence on the allophone. That is, in the one instance, the coarticulatory information facilitates *later* identification of the segment it anticipates; in the other, it can be identified as a coarticulatory influence on an allophone only *after* the segment it anticipates has itself been identified.

The alternative segmentation hypothesis under consideration satisfies both sets of findings. Listeners may segment the speech stream along its coarticulatory lines into overlapping phonetic segments (Figure 1c). There are two consequences of this segmentation strategy: Anticipatory coarticulatory information is perceived as the onset of the segment it "anticipates," and the same information therefore is not integrated with concurrent information for the preceding phonetic segment. That is, "compensation" occurs as a necessary by-product of segmentation.[2] From this perspective, compensation is symptomatic of an additional consequence of segmentation. Because sources of context sensitivity are not integrated with information for segments with which they co-occur in time, the same phonetic segment in different coarticulatory contexts is predicted to sound approximately the same to listeners. That is, listeners may perceive the tokens of a phonetic type as the same or very similar across different phonetic contexts because sources of contextual influence have not been integrated with the tokens.

The present research is designed to contrast the foregoing accounts of segmentation. In particular, the research first asks whether listeners will use coarticulatory information for a vowel within the acoustic domain of a preceding phonetic segment (in the present case, /g/) as information for the vowel (cf. Martin & Bunnell, 1982; Whalen, 1984). It next asks, if they do, whether they also show evidence of "compensation" for contextual influences of the vowel in their perceptual judgments of the preceding phonetic segment. If listeners exhibit both behaviors on the same syllables, then the segmentation strategy of Figure 1b can be ruled out on grounds previously outlined: that strategy requires that listeners identify the coarticulatory information as such before identifying the segment it anticipates in the paradigm used by Martin and his colleagues; however, it requires the reverse ordering of identification to explain apparent compensation. The segmentation strategy of Figure 1c provides a unified account of both findings.

Use of anticipatory coarticulatory information to identify a forthcoming segment will be tested using the cross-splicing, choice reaction time procedure developed by Martin and Bunnell (1982) and used by Whalen (1984). Compensation will be assessed using a 4IAX discrimination procedure on the same stimuli (see Pisoni, 1971).

Listeners are said to compensate for contextual influences if their perceptual judgments of a phonetic segment suggest that the contextual influences have been eliminated or reduced (see Mann & Repp, 1980). A 4IAX trial such as the following will allow assessment of compensation:

$$g_i i \text{-----} g_i u \text{----------} g_i i \text{-----} g_u u.$$

The trial includes four syllables temporally organized into pairs. Members of a pair have different vowels, but the vowels are the same across the pairs. "g" refers to a stop burst originally produced either in a [gi] syllable (subscripted with "i") or in a [gu] syllable (subscripted with "u"). Subjects are asked to decide which pair has members that sound more similar. If listeners make their assessments on the basis of the relative acoustic overlap between members of a pair, they should select the members of the first pair as more similar than the members of the second because the former have identical bursts. The opposite prediction is made if listeners make their judgments with contextual influences eliminated from the different consonantal segments (that is, if they "compensate" for those influences). In that case, influences of different vowels are eliminated from identical stop bursts in the first pair, yielding different residuals. In the second pair, the different influences of /i/ and /u/ are eliminated from contextually appropriate stop bursts, yielding, by hypothesis, identical context-free phonetic tokens. Thus, members of the second pair should be judged more alike than members of the first.

This research continues a series of studies reported elsewhere (Fowler, 1983b; Fowler & Smith, in press). The earlier research used the paired choice reaction time and 4IAX procedures just described to test listeners' perceptions of coarticulatory influences of stressed vowels on

preceding or following cross-spliced, unstressed schwa. Predictions based on the hypothesis that listeners segment speech along coarticulatory lines were partially supported for these stimuli. We found positive evidence for the segmentation strategy of Figure 1c (and, correspondingly, disconfirming evidence for the strategy of Figure 1b) when the coarticulatory effects under study combined both carryover and anticipatory effects of stressed vowels (as in /ibəbi/ and /abəba/). The reaction time procedure also provided positive evidence for contexts in which coarticulatory effects were only anticipatory (as in /bəbi/ and /bəba/). However, in the 4IAX task, responses were random when coarticulatory effects were anticipatory only.

We hypothesized that the chance performance in the 4IAX study in which only anticipatory coarticulation was present was due to a lack of sensitivity of the 4IAX procedure as compared with the choice reaction time procedure, and not to a restriction on the applicability of the segmentation of the segmentation hypothesis to carry over coarticulatory influences. This interpretation is plausible because anticipatory coarticulation of stressesd vowels is more limited in these contexts than is carryover coarticulation (Bell-Berti & Harris, 1976; Fowler, 1981a), and because, as compared to the choice reaction time procedure, the discrimination procedure places severe memory demands on the listener and requires a difficult judgment. However, the need to demonstrate that segmentation occurs along coarticulatory lines, whether the lines reflect anticipatory or carryover coarticulation, still remains.

The present experiment used stimuli in which anticipatory coarticulatory effects of a vowel on a preceding segment are larger than they were on the schwas of our previous study. Stimuli in the experiment are the stop-vowel syllables /gi/ and /gu/. Because the stop immediately precedes the vowel and because velar stops coarticulate extensively with vowels, I expected these stimuli to enable observation of segmentation of anticipatory coarticulatory influences of the vowel from the acoustic domain of the consonant if it occurred.

METHOD

Subjects

The subjects were 36 students at Dartmouth College. All were native speakers of English, and all reported normal hearing.

Materials

Stimuli. Stimuli were two tokens each of the monosyllables /gi/ and /gu/ produced by a female talker. They were input to a New England Digital minicomputer, sampled at 20 kHz and filtered at 10 kHz.

Based on criteria provided by Dorman, Studdert-Kennedy, and Raphael (1977), the release bursts of each utterance were identified and segmented from the remainder of the syllable. Release bursts ranged in duration from 16 to 20 msec and did not vary systematically in duration with the identity of the vowel. (These values are comparable to averages across /gid/ and /gud/ of 7.5 msec for one speaker and 22.5 msec for the second speaker, as reported by Dorman et al.) The period of aspiration following release was removed from the vocalic portion of the syllable and was replaced by an equivalent period of silence. This was done to avoid abrupt discontinuities in the spectra when bursts and vocalic segments were cross-spliced. In the test orders, stimuli were presented in low levels of white noise which improved the perceived quality of the stimuli by masking the temporal discontinuity. The intervals of aspiration ranged from 6 to 15 msec (the averaged values for talkers in Dorman et al. were 11.5 and 12.5 msec). Durations of the voiced portion of each syllable were 429 and 430 msec for tokens of /gi/ and 359 and 361 msec for tokens of /gu/.

Three types of test syllables were constructed from the syllable fragments just described. The four "original" syllables consisted of release bursts and vocalic portions that had originally been produced together. They were separated by a period of silence equivalent to the original period of aspiration for the vocalic segment. "Spliced" syllables were release bursts from one token of a syllable type attached to a silent interval and vocalic portion originally associated with production of the other token of the same syllable type. (That is, for example, a burst from one token of /gi/ was spliced onto an interval of silence and a silent interval and vocalic portion of the other token of /gi/.) There were four spliced syllables. Eight "cross-spliced" syllables were created by attaching a release burst from a token of one type onto a silent interval and vocalic portion associated with a token of the other phonemic type (that is, e.g., a burst from a token of /gi/ was spliced onto the vocalic portion of a /gu/ syllable.)

Identification test. An identification test presented release bursts, vocalic portions, and whole CVs for identification, in that order, in separate blocks of 32 trials.

The identification test was originally presented to 12 naive subjects, who had not heard the stimuli before, and to 12 subjects who had just completed the choice reaction time and 4IAX tests to be described. These 24 subjects were given an answer sheet with alternatives "bee," "dee," "gee," "boo," "doo," and "goo," arranged in three blocks of 32 rows, 1 row for each trial of the identification test. In the test, both groups of subjects exceeded chance in their ability to identify the syllables' vowels from their release bursts alone. This suggested the possibility that some or all of the subjects heard diphthongal vowels in the cross-spliced syllables. To assess that, a new group of 12 subjects took the identification test preceded by the reaction time and 4IAX procedures. The response sheet given to these subjects for the identification test allowed six new response alternatives—"bwee," "dwee," "gwee," "byoo,""dyoo," and "gyoo"—in addition to the original six.

Choice reaction time test. The choice reaction time study, modeled after the paradigm of Whalen (1984), consisted of original, spliced, and cross-spliced stimuli presented randomly one at a time in four blocks of 48 trials. Predictions were that, because the release burst would provide misleading information for the vowel in cross-spliced stimuli, reaction time and accuracy to identify the vowel in those stimuli would be inferior to the same measures taken on original and spliced stimuli.

4IAX discrimination test. The 4IAX test consisted of three blocks of 64 trials. One-half of the trials were of type A, and one half were of type B, both illustrated by example below. Stimuli in this test were either spliced or cross-spliced; no original stimuli were presented. (As before, subscripts on the /g/s indicate the vowel with which the release burst had originally been produced.)

Trials of type A were designed to test whether listeners could distinguish the different bursts in the context of a vowel; as described in the introduction, trials of type B provided the critical test of the segmentation hypotheses:

A: $g_i i - g_i i - - - - - - - - - g_i i - g_u i$

B: $g_i i - g_i u - - - - - - - - - g_i i - g_u u$

In either trial type, four stimuli were presented per trial, arranged temporally in two pairs. Members of one pair of an A trial were identical. One member of the second pair was identical to the members of the first pair. The fourth syllable differed from the others in its release burst. That syllable was always cross-spliced. Trials of type B were like trials of type A except that the vocalic segments within a pair were different. In B trials, then, the members of one pair had identical release bursts. In the other pair, one item had the same release burst as the members of the first-mentioned pair; the other had a different release burst. In pairs where bursts were identical, one member of the pair was spliced

and one was cross-spliced. For the pair with different release bursts, both members were spliced so that the bursts were in vocalic contexts compatible with those in which they had originally been produced.

In a trial, the offset-onset time was 200 msec within a pair; between pairs, it was 500 msec. If the stimuli in the sample trials above are labeled 1, 2, 3, and 4, their ordering in the sample trials is 12-34. In addition to this ordering were equal numbers of occurrences of orders 21-43, 34-12, and 43-21. In the sample trials above, release burst [g_i] occurs three times and burst [g_u] occurs just once. There were equal numbers of trials in which [g_u] was the more frequent burst in the trial.

Procedure

Group 1. The 12 subjects in this group took only the identification test. They listened to stimuli over headphones. Stimuli were presented on-line on a New England Digital minicomputer. In this test, as in the others, the stimuli were mixed with a low level of white noise.

The subjects were told that stimuli on the first third of the test were the first few milliseconds of a CV syllable and that their task was to guess the identity of the whole syllable from the fragment. The syllable types they might hear were pronounced for them. They were instructed to circle, on the answer sheet, the response choice that best represented the syllable from which the fragment had been excised. They were required to guess if necessary. In addition, they were told that they might hear all or only some of the syllables represented on the answer sheet, and that, therefore, they should circle their best guess based on what they heard and not attempt to distribute their responses evenly among the response alternatives. On the second block, they were told that the stimuli were the remainders of the CV syllables with the first few milliseconds excised. Instructions were the same as on the first block. Finally, on the third block, they were told that stimuli were the two types of syllable fragments they had just been listening to, but rejoined to make a whole CV syllable. Instructions were to identify the CV on each trial as one of the six listed on the answer sheet.

Trials were initiated individually by keypress; thus, the subjects had unlimited time to make their responses.

Groups 2 and 3. The 24 subjects in these groups took the choice reaction-time test, the 4IAX test, and the identification test, in that order. The procedures for these groups were identical; they differed only in the response sheets they received on the identification test.

In the reaction time procedure, the subjects listened over headphones to stimuli presented on-line and mixed with noise, as in the identification procedure just described. They were instructed to identify the vowel in each syllable as "ee" or "oo" by hitting the appropriate labeled key on the computer terminal's keyboard as quickly and as accurately as possible. They received response-time feedback after every trial and averaged response times and accuracy at the end of each of the four blocks of 48 trials. They were asked to keep their accuracy above .9. The first block of trials served as practice.

In the 4IAX procedure, the subjects were instructed to choose the first or second pair of stimuli on each trial as having the more similar members. They signaled their selection by typing "1" or "2" into the computer, using the calculator pad on the keyboard. They followed that selection with a confidence judgment (1, guess; 2, intermediate certainty; 3, high level of confidence). Neither response was timed.

In this test there were three blocks of 64 trials, the first block serving as practice. Trials were self-paced, and there was no feedback.

Last in the session, the subjects took the identification test.

RESULTS

Identification

Identification of bursts, vocalic portions, and original spliced and cross-spliced CVs are provided in Table 1 for all three groups of subjects. Consonant and vowel identifications are displayed separately.

In identification of consonants from isolated bursts, the subjects are close to chance in Group 1 (naive listeners). Performance for experienced subjects, particularly those in Group 3[3] is better than chance. More remarkable is performance identifying the vowel from the burst. All groups exceeded chance on this identification task.

Table 1
Identifications (Proportion of Responses) of Bursts, Vocalic Portions and Whole Syllables by Naive (Group 1) and Experienced Listeners (Group 2) with 6 Response Choices and by Experienced Listeners (Group 3) with 12 Alternatives

		b	d	g	i	wi	u	yu
BURSTS								
Group 1	gu	20	38	42	12		88	
	gi	22	42	36	89		11	
Group 2	gu	21	30	49	24		76	
	gi	13	39	48	88		12	
Group 3	gu	17	23	60	11	05	65	19
	gi	09	19	72	73	13	06	08
VOCALIC PORTIONS								
Group 1	gu	83	12	05	05		95	
	gi	25	50	25	97		03	
Group 2	gu	71	11	08	09		91	
	gi	44	29	27	95		05	
Group 3	gu	79	09	12		02	94	04
	gi	20	38	42	84	14	01	01
WHOLE SYLLABLES								
Original								
Group 1	gu			100			100	
	gi	04		96	100			
Group 2	gu			100	04		96	
	gi		02	98	98		02	
Group 3	gu		02	98			98	02
	gi	02	04	94	98	02		
Spliced								
Group 1	gu	04		96			100	
	gi	02	02	96	98		02	
Group 2	gu	02		98			100	
	gi		04	96	100			
Group 3	gu			100			98	02
	gi		06	94	96	04		
Cross-Spliced								
Group 1	gu	88	7	04	02		98	
	gi	18	35	47	100			
Group 2	gu	18	44	38	02		98	
	gi	08	18	74	88		12	
Group 3	gu	57	28	18	05	03	65	27
	gi	02	21	78	45	56		

As for the isolated vocalic portions of the syllables, the vocalic portion of /gu/ led to predominantly "b" identifications in all groups. In contrast, the vocalic portion of /gi/ evidently was more ambiguous, leading to substantial numbers of identifications in all consonantal response categories. Vowel identifications based on vocalic portions of the syllables were accurate.

Subjects in all groups were accurate in identifying the vowels and consonants of original and spliced whole syllables. As for cross-spliced syllables, "g" was the predominant identification in [g_ui], but, in two groups, "b" was the predominant consonant identification for [g_iu]—a finding also reported by Cole and Scott (1974; for a related

finding, see Liberman et al., 1952). Subjects in Group 2 gave predominantly "d" responses for the consonant in the latter syllable.

Subjects in Group 3 did report more diphthongs in the cross-spliced syllables than elsewhere, particularly in the syllable [$g_{u}i$]. These data, as well as those for subjects reporting "b"s or "d"s in cross-spliced syllables, will be used later to examine individual subject's performances in the choice reaction time and 4IAX procedures.

Overall, this test provides two pieces of information necessary to the interpretation of the next two tests. First, despite the surgery performed on the syllables, the bursts are integrated with the vocalic portions sufficiently in whole CVs for consonant identifications based on the two fragments together to be different from identifications based on the separated parts. Second, the identification test provided a finding that we had expected to uncover only in the reaction time procedure—namely, that listeners are sensitive to the information for the following vowel in the release burst. This led to the only effect that the burst appeared to have on vowel identification in the identification test. Some subjects in Group 3 identified the vowel as diphthongal.

Choice Reaction Time

Table 2 provides response times and accuracies in the choice reaction time procedure. In Table 2, means are collapsed over the subjects in Groups 2 and 3. Although subjects in Group 3 responded more rapidly (by an average of 70 msec) and more accurately (by an average of 5%), response patterns and outcomes of separate ANOVAs performed on the data from each group were the same.

Reaction times and accuracy were subjected to separate two-way analyses of variance with factors: syllable type (original, spliced, cross-spliced) and vowel (/i/, /u/). In the analysis of reaction times, the main effect of syllable type was significant [$F(2,46) = 29.40$, $p < .001$], reflecting the substantially longer response times to cross-spliced than to spliced and original syllables. In addition, the interaction of vowel and syllable type reached significance [$F(2,46) = 4.05$, $p = .02$] because the slowing caused by cross-splicing was more marked for the syllable [$g_{u}i$] than for [$g_{i}u$].

The accuracy measure provided a compatible outcome, with performance lower in cross-spliced than in original and spliced syllables [$F(2,46) = 20.87$, $p < .001$]. In this analysis, the interaction did not reach significance.

Table 2
Reaction Times (in Milliseconds) and Accuracy (Percent Correct) for Groups 2 and 3 in the Choice Reaction Time Study

	gu		gi	
	RT	Accuracy	RT	Accuracy
Original	455	93	441	98
Spliced	453	93	425	98
Cross-spliced	504	73	520	86

The identification test had revealed that some subjects heard diphthongal vowels in cross-spliced syllables, particularly in [$g_{u}i$]. This provides an alternative account of the slowing on cross-spliced stimuli. If subjects hear diphthongs, then, as predicted, they hear the vowel information in the burst; their reaction times to cross-spliced stimuli are slowed, however, because the perceived vowels include both response alternatives and subjects have to choose just one. Subjects who do not report diphthongs may also extract vowel information from the bursts in cross-spliced syllables, yet still hear the syllable vowel as monophthongal, because later vocalic information overwhelmingly contradicts information in the burst. This latter was the possibility the experiment had been designed to establish and test.

Post hoc analyses of responses by individual subjects in Group 3 were performed to determine whether subjects responded differently depending on whether they heard the vowel as monophthongal or diphthongal. For the syllable [$g_{u}i$], seven subjects consistently reported diphthongs in the identification test, three consistently reported monophthongs, and two reported some of each. [Consistency in identification was defined operationally as selection of a diphthongal (monophthongal) response on at least six of eight opportunities on the identification test.] For the syllable [$g_{i}u$], numbers of subjects falling into the three categories were 2, 10, and 0, respectively. Some subjects fell into the same category twice, because they had heard the vowel in the same way on both syllables. In those instances, their data for the two syllables were pooled. Average response times and accuracy were collapsed over syllables for the 7 subjects consistently reporting diphthongs in Group 3 and separately for the 10 subjects reporting monophthongs. An analysis of variance comparing the two groups on the original, spliced, and cross-spliced stimuli yielded a highly significant effect of splicing condition [$F(2,30) = 39.09$, $p < .001$], but no effect of subject group and no interaction (both $Fs < 1$). It seems that whether or not subjects experience the anticipatory vowel information in the burst as a glide, it serves them as information for a vowel and, in cross-spliced stimuli, subjects are misled by it.

4IAX

Table 3 provides the outcome of the 4IAX test collapsed over subjects in Groups 2 and 3. The data were collapsed over the groups because analyses performed on the individual groups did not differ.

As predicted, on A trials, listeners reliably chose syllables with acoustically identical bursts in their proper contexts as more similar than syllables with acoustically different bursts in identical contexts [/gu/, $t(23) = 18.34$, $p < .001$; /gi/, $t(23) = 11.84$, $p < .001$]. This verifies that the anticipatory coarticulatory information is audible in the context of a syllable. Of greater interest is performance on B trials. On these trials, listeners compared syllables with acoustically different bursts, each in their proper coarticulatory contexts (e.g., [$g_{i}i$]-[$g_{u}u$]), with syl-

Table 3
Outcome of the 4IAX Test Collapsed Over Subjects in Groups 2 and 3

	Cross-Spliced Syllable*		Response Selection**	
Trials	$g_u i$	$g_i u$	Acoustically Identical	Acoustically Different
A	.94	.82	2.48	1.91
B	.34	.27	1.86	2.25

**Proportion of A and B trials in which listeners selected syllables having acoustically identical bursts as more similar than syllables having acoustically different bursts. **Confidence judgments.*

lables with acoustically identical bursts, one in its original context and one not (e.g., [g_ii]-[g_iu]). As predicted, on these trials, in contrast to A trials, listeners reliably selected the syllables with different bursts as more similar than those with identical bursts [/gu/, $t(23) = -3.26$, $p = .004$; /gi/, $t(23) = -3.96$, $p < .001$].

As shown in Table 3, confidence judgments mirror the response selections. The confidence judgments in Table 3 are collapsed over syllable type. This was necessary because subjects occasionally had no responses either in the "acoustically different bursts" category on A trials or in the "acoustically identical bursts" category on B trials. No subject had missing data when the data were collapsed over [g_i] and [g_u] trials. On A trials, subjects are more confident of their (correct) judgments that syllables with identical bursts are more similar than those with different bursts. On B trials, their confidence reverses. A two-way analysis of variance [trial type (A,B) × judgment (syllables with acoustically identical bursts, those with different bursts)] was performed on the confidence judgments. In that analysis, the effect of trial type [$F(1,23) = 5.02$, $p < .03$] and the interaction [$F(1,23) = 39.33$, $p < .001$] were significant. The significant interaction reflected the effect of interest. On A trials, listeners were more confident of their correct selections of syllables having acoustically identical bursts than of their errors [$F(3,23) = 9.31$, $p < .001$]; on B trials, they were less confident of their selection of those having acoustically identical bursts than of their selection of different bursts in their proper contexts [$F(3,23) = 4.28$, $p = .02$].

Response selection by individuals hearing diphthongs was examined separately from individuals hearing monophthongs. The average performance on B trials of the 8 subjects reliably hearing diphthongs did not differ from that of the 10 subjects hearing monophthongs [$t(16) = 1.07$, $p = .30$].

It is also of interest to look separately at subjects for whom cross splicing changed the identity of the consonant to /b/ or /d/ and those for whom it did not. For subjects of the first type, the 4IAX task confronts them with an easy between-category discrimination. For subjects in the second category, the task is one of within-category discrimination.

For these analyses, data from Groups 2 and 3 were pooled. In all, there were 15 subjects who reported the syllables with the cross-spliced burst reliably as /b/- or /d/-initial in at least one syllable. All but two of these were subjects in the condition with cross-spliced [g_i]. Across Groups 2 and 3, there were 19 subjects reliably reporting /g/ in at least one syllable. Performance differences were significant between these two groups, as expected from the general findings that between-category discrimination is easier than within-category discrimination [$t(32) = 2.92$, , $p < .01$]. However, subjects hearing /b/ or /d/ were not wholly responsible for the outcome on B trials. Of those 15 subjects, 13 had performance levels below .5 [$t(14) = -5.76$, $p < .001$]. Of the 19 subjects hearing /g/, 12 showed the predicted direction of difference [$t(18) = -2.02$, $p = .056$]. We conclude, then, that although the within-category discrimination is much more difficult than the between-category discrimination, it is not qualitatively different from between-category discrimination. Overall, subjects hear syllables with acoustically different bursts in their proper coarticulatory contexts as more similar than those with acoustically identical bursts, one in its proper context and one not; making the discrimination at all is facilitated if the segmentation process leads the cross-spliced burst to fall into a phonemic category different from its original one.

DISCUSSION

In this study, as in the earlier research reported by Fowler and Smith (in press), subjects' choice reaction time and discrimination performances reflect the segmentation strategy of Figure 1c more closely than that of Figure 1b. Listeners use coarticulatory information as information for the influencing segment, and they do not integrate it into their perceptual experience of the segment with which it co-occurs in time. The present study extends the findings of Fowler and Smith (in press) to anticipatory coarticulatory influences and to coarticulatory relationships of consonants and vowels.

The segmentation of speech that our research supports closely resembles that achieved by a recent computer model of speech perception described by Elman and McClelland (1983). In their model (cf. McClelland & Rumelhart, 1982), features, phonemes, and words are represented by "nodes" interconnected by excitatory and inhibitory links. In general, excitatory connections link nodes that are mutually consistent; inhibitory connections link nodes that are inconsistent. (For example, phoneme nodes excite words of which they are constituents; word nodes inhibit each other.) Acoustic information input to the model activates features compatible with it; in turn, the features activate phonemes consistent with them, and phonemes activate words. Of particular interest here is the segmentation of the acoustic signal that the model achieves over time as it identifies phonetic segments from an acoustic speech signal. Over time, the acoustic signal first provides stronger and then weaker evidence for the presence of a particular phonetic segment. I have called that waxing and waning of information the "prominence" pattern for a segment. In the model of Elman and McClelland (1983), the activation pattern for a phonetic segment

tracks the waxing and waning of information for the segment in the acoustic signal. Due to coarticulation, in most time frames, the model receives featural information consistent with two phonetic segments concurrently—for example, a syllable-initial consonant and a following vowel. When that happens, two phonemes are highly activated concurrently. Eventually information for the first segment dies out, leaving the highly activated second segment. The activation patterns for a sequence of phonemes, therefore, resemble the prominence curves represented in Figure 1a. Thus, in the model, although there is no explicit segmentation process separate from the process of identifying phonetic segments, nonetheless, a segmentation of the signal is achieved, and it is precisely the segmentation that I have found characteristic of human listeners. In the present study, listeners begin using acoustic information for a segment as such whenever it occurs in the speech signal. This leads to a reaction time advantage for original and spliced over cross-spliced stimuli in the choice reaction time study. If the information is coarticulatory, the listeners do not integrate it with information for a segment with which it co-occurs in time. This leads to the findings in the 4IAX study.[4]

The model of Elman and McClelland would not achieve the segmentation it does if the acoustic signal did not support it. It has not been obvious that the signal does support this segmentation, however, because visible displays of the signal do not invite it; indeed, acoustic analysis guided by visible displays have not achieved it. [This is true not only of segmentations used in the phonetics literature as described in the introduction; it also appears to characterize segmentations described by naive subjects learning to read spectrograms based on a whole-word training procedure (Greene, Pisoni, & Carrell, 1984).] It will be important for future research to make explicit the relationship between the listeners' and the model's segmentation of the acoustic speech signal, on the one hand, and the support for it that the signal provides, on the other.

One step further back in the chain of communication, the acoustic speech signal could not reliably give rise to the segmentation it does without support from the talkers' articulations. That is, the gestures corresponding to a given phonetic segment must, in some sense, cohere in articulation, and those corresponding to different segments must, in the same sense, be separable.[5] This line of reasoning, in turn, suggests that Hockett's (1955) often-cited Easter-egg analogy is misleading. Hockett compared the effects of coarticulation on phonetic segments to a process of sending a row of Easter eggs through a wringer. His analogy reflected the view, still current (cf. MacNeilage & Ladefoged, 1976), that coarticulation destroys both the coherence of individual phonetic segments and their separation one from the other. Necessarily, then, the acoustic signal cannot be supposed to provide sufficient information, in itself, to support perception of the segments; rather, phonetic identifications must be interpretations imposed on the signal by a listener (cf. Studdert-Kennedy, in press). The present findings and the behavior of Elman and McClelland's computer model render this perspective on articulation doubtful, however. In view of that, it is not surprising that research on articulation suggests a picture tidier than Hockett's analogy implies. For example, research by Barry and Kuenzel (1975), Butcher and Weiher (1976), Carney and Moll (1971), and Ohman (1966) agree in showing that vowel-to-vowel movements of the tongue-body occur before, throughout, and after the production of an intervocalic consonant in a VCV production. Ohman's interpretation is that, in VCVs, consonantal gestures are superimposed on on-going diphthongal vowel-vowel gestures. In this type of utterance, then, coarticulation does not destroy the coherence of features of individual phonetic segments or the separation among distinct segments as the Easter-egg analogy implies. Indeed, rather than being an irrecoverable smearing of consonantal and vocalic gestures, in these utterances coarticulation is the overlapping occurrence of two distinct types of gestures—one for the vowel-to-vowel movements and one for the consonantal gestures.

This research on C-V coarticulation converges with other production research, in which segment durations are measured. In that literature, vowels are measured to shorten as consonants are added to a syllable, and, in similar fashion, stressed vowels are measured to shorten as unstressed vowels are added to a word or stress foot (e.g., Fowler, 1977, 1981b; Lindblom & Rapp, 1973). Data in Fowler (1981b) suggest, however, that at least some of the measured shortening is not articulatory shortening in fact, but rather reflects the sort of articulatory overlap reported by Ohman and others (and illustrated in Figure 1). It is identified as shortening only because measurement conventions do not include, as part of a vowel's duration, the parts of its coarticulatory extent where another segment predominates in the signal. Together with the articulatory measures, the shortening measures further support the hypothesis that consonants and vowels (and stressed and unstressed vowels) are nondestructively overlapped in production in a way consistent with the perceptual segmentation of the acoustic signal that the present research and that of Fowler and Smith (in press) suggest.

The production research just described and our interpretation of the present findings both predict that the perceived duration of a phonetic segment and should exceed its measured duration, instead corresponding approximately to its coarticulatory extent. A similar expectation can be derived from Liberman and Studdert-Kennedy's (1978) discussion of reasons why coarticulation may be necessary for perceivers. In their view, talkers have to produce speech that meets two competing requirements. Because meanings of grammatical utterances have to be extracted from grammatically coherent groups of words, and cannot be determined word by word, speech may have to be transmitted at a rapid rate. The listener has to be able to remember the beginning of a syntactic phrase at the time the end of it is produced. Second, however, the rate cannot exceed that at which listeners are no longer able to determine the order of sequences of sounds

(Warren, 1976). Liberman and Studdert-Kennedy point out that coarticulation allows relatively long-duration segments to occupy relatively short intervals of time. However, this would be a perceptual advantage only to a listener who heard coarticulatory overlap as overlap rather than as context-sensitivity of discrete phonetic segments. In recent work, I have found some evidence that the perceived duration of a vowel does, indeed, exceed its measured duration (Fowler, 1983a).

Together, the research and theoretical considerations outlined here suggest a coherent perspective on the production and perception of speech (cf. Fowler, 1983a, 1983c). Talkers produce phonetic segments in overlapping time frames. The articulatory overlap, however, does not smear the segments; rather, it preserves the coherence of the temporally extended parts of an individual phonetic segment and the separation of distinct segments. Compatibly, the acoustic signal provides information for the separation of overlapping segments and the coherence of temporally extended parts of a segment. Finally, listeners segment the signal realistically, recovering the segments that talkers produce.

REFERENCES

ABBS, J., & GRACCO, V. (in press). Control of complex motor gestures and orofacial muscle responses to load perturbations of the lips during speech. *Journal of Neurophysiology.*

ALFONSO, P., & BAER, T. (1982). Dynamics of vowel articulation. *Language and Speech*, **25**, 151-173.

BARRY, W., & KUENZEL, H. (1975). Co-articulatory airflow characteristics of intervocalic voiceless plosives. *Journal of Phonetics*, **3**, 263-282.

BELL-BERTI, F., & HARRIS, K . (1976). Some aspects of coarticulation. *Haskins Laboratories Status Reports on Speech Research*, **SR-45/46**, 197-204.

BUTCHER, A., & WEIHER, E. (1976). An electropalatographic investigation of coarticulation in VCV sequences. *Journal of Phonetics*, **4**, 59-74.

CARNEY, P., & MOLL, K. (1971). A cinefluorographic investigation of fricative consonant-vowel coarticulation. *Phonetica*, **23**, 193-202.

COLE, R., & SCOTT, B. (1974). The phantom in the phoneme: Invariant cues for stop consonants. *Perception & Psychophysics*, **15**, 101-107.

DORMAN, M., STUDDERT-KENNEDY, M., & RAPHAEL, L. (1977). Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics*, **22**, 109-122.

ELMAN, J., & MCCLELLAND, J. (1983). *Speech perception as a cognitive process: The interactive activation model* (ICS Report No. 8302). San Diego: University of California, Institute of Cognitive Science.

FOWLER, C. A. (1977). *Timing control in speech production.* Bloomington: Indiana University, Linguistics Club.

FOWLER, C. A. (1981a). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech and Hearing Research*, **24**, 127-139.

FOWLER, C. A. (1981b). A relationship between coarticulation and compensatory shortening. *Phonetica*, **38**, 35-50.

FOWLER, C. A. (1983a). Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in sequences of monosyllabic stress feet. *Journal of Experimental Psychology: General*, **112**, 386-412.

FOWLER, C. A. (1983b, May). *Perceiving the natural structure of coarticulated speech.* Paper presented at the 105th meeting of the Acoustical Society of America, Cincinnati.

FOWLER, C. A. (1983c). Realism and unrealism: A reply. *Journal of Phonetics*, **11**, 303-322.

FOWLER, C. A., & SMITH, M. (in press). Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In J. Perkell (Ed.), *Invariance and variability of speech processes.* Hillsdale, NJ: Erlbaum.

FROMKIN, V. (Ed.) (1973). *Speech errors as linguistic evidence.* The Hague: Mouton.

GREENE, B., PISONI, D., & CARRELL, T. (1984). Recognition of speech spectrograms. *Journal of the Acoustical Society of America*, **76**, 32-43.

HOCKETT, C. (1955). *Manual of phonology* (Publications in Anthropology and Linguistics, No. 11). Bloomington: Indiana University.

HOCKETT, C. (1960). The origin of language. *Scientific American*, **203**, 89-96.

KELSO, J. A. S., TULLER, B., BATESON, E., & FOWLER, C. (1984). Articulatory adaptation to jaw perturbations during speech: Evidence for functional synergies. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 812-832.

KLATT, D. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, **3**, 129-140.

LEHISTE, I., & SHOCKEY, L. (1972). On the perception of coarticulatory effects in VCV syllables. *Journal of Speech and Hearing Research*, **15**, 500-506.

LIBERMAN, A., DELATTRE, P., & COOPER, F. (1952). The role of stimulus variables in the perception of stop consonants. *American Journal of Psychology*, **65**, 497-516.

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

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

LISKER, L. (1972). On time and timing in speech. In T. Sebeok (Ed.), *Current trends in linguistics* (Vol. 12). The Hague: Mouton.

MACNEILAGE, P., & LADEFOGED, P. (1976). The production of speech and language. In E. Carterette & M. Friedman (Eds.), *Handbook of perception: Language and speech* (Vol. 7). New York: Academic Press.

MANN, V. (1980). Influence of a preceding liquid on stop consonant perception. *Perception & Psychophysics*, **28**, 407-412.

MANN, V., & REPP, B. (1980). Influence of vocalic context on perception of the [š]-[s] distinction. *Perception & Psychophysics*, **28**, 213-228.

MARSLEN-WILSON, W., & WELSH, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.

MARTIN, J., & BUNNELL, T. (1982). Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel sequences. *Journal of Experiment Psychology: Human Perception and Performance*, **8**, 473-488.

MCCLELLAND, J., & RUMELHART, D. (1981). An interactive activation model of context effects in letter perception, Part 1. *Psychological Review*, **88**, 375-407.

MCNEILL, D., & LINDIG, K. (1973). The perceptual reality of phonemes, syllables, words and sentences. *Journal of Verbal Learning and Verbal Behavior*, **12**, 419-430.

MELTZER, R., MARTIN, J., MILLS, C., IMHOFF, D., & ZOHAR, D. (1976). Reaction time to temporally-displaced phoneme targets in continuous speech. *Journal of Experimental Psychology: Human Perception and Performance*, **2**, 277-290.

OCHIAI, K., & FUJIMURA, O. (1971). *Vowel identification and phonetic contexts. (Reports 22-2, pp. 103-111). Tokyo: University of Electrocommunications.*

OHMAN, S. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, **39**, 151-168.

PETERSON, L., & LEHISTE, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, **32**, 693-703.

PISONI, D. (1971). On the nature of categorical perception of speech