Perception & Psychophysics
1997, 59 (7), 1141-1152

1057

Perceiving the causes of coarticulatory acoustic variation: Consonant voicing and vowel pitch

JENNIFER S. PARDO
Yale University, New Haven, Connecticut
and Haskins Laboratories, New Haven, Connecticut

and

CAROL A. FOWLER
Yale University, New Haven, Connecticut
Haskins Laboratories, New Haven, Connecticut
and University of Connecticut, Storrs, Connecticut

Coarticulatory acoustic variation is presumed to be caused by temporally overlapping linguistically significant gestures of the vocal tract. The complex acoustic consequences of such gestures can be hypothesized to specify them without recourse to context-sensitive representations of phonetic segments. When the consequences of separate gestures converge on a common acoustic dimension (e.g., fundamental frequency), perceptual parsing of the acoustic consequences of overlapping spoken gestures, rather than associations of acoustic features, is required to resolve the distinct gestural events. Direct tests of this theory were conducted. These tests revealed mutual influences of (1) fundamental frequency during a vowel on prior consonant perception, and (2) consonant identity on following vowel stress and pitch perception. The results of these converging tests lead to the conclusion that speech perception involves a process in which acoustic information for coarticulated gestures is parsed from the stream of speech.

In attempting to understand the nature of speech perception, it is necessary to address problems that are general to perception. One central problem for perception is that of informational variability. Perceivers are successful at identifying and interacting with the distal objects and events in the environment. This is accomplished in the face of seemingly overwhelming variability in proximal stimulation. How do perceivers deal with such variability in stimulation to arrive at stability in perception? One way to approach this problem is to consider all observed departures from idealized invariant proximal stimulation as irrelevant to an event's identity, serving only as noise for the perceptual system to overcome. Alternatively, stimulus variability that is caused by, and therefore is information for, distal events can be hypothesized to serve as specifying information for the perceptual system to use. If such systematic variability in stimulation does specify the event causing it, perception is a process of detection of the relevant proximal information for distal object identity.

This research was supported in part by NICHD Grant HD-01994 to Haskins Laboratories. The authors would like to thank Bob Crowder and John Kihlstrom for comments on earlier versions of the manuscript, Bob Crowder for generously sharing lab space, Randy Diehl and an anonymous reviewer for thoughtful critique, Subhobrata Mitra for his work on stimulus materials for Experiment 1, and Robert Remez for help with Figure 1. Correspondence should be addressed to J. S. Pardo, Yale University, Department of Psychology, Box 208205, New Haven, CT 06520 (e-mail: pardo@minerva.cis.yale.edu).

Acoustic Complexities of Speech

The search for invariance in proximal stimulation from which to perceive objects and events in the environment is no less difficult for consonants and vowels than for anything else (see, e.g., Stevens & Blumstein, 1981). In speech production, the articulatory gestures of the vocal tract occur rapidly and overlap one another in time, resulting in tremendous coarticulatory acoustic diversity that appears to correspond poorly with the underlying syntactic and phonemic structure of an utterance. For example, in the waveform corresponding to the spoken phrase, "A stitch in time saves nine," it is not obvious where acoustic information for each phonetic segment begins and ends. Likewise, the vowels in "time" and "nine," although perceptually equivalent, will be signaled by different acoustic patterns due in part to the differences in their surrounding phonetic contexts. This nondiscrete property of the acoustic speech signal—the consequence of coarticulation that leads to considerable variability—has even been considered necessary for the effective transmission of speech (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967).

How does the perceiver deal with the variability in acoustic structure that is due to coarticulation? If phonetic segments are specified not only by their particular spectral characteristics, but also by their temporal properties, then overlapping phonetic gestures can be perceived as just that—physical events that occur over time and that overlap, rather than merely influence, one another (Fowler, 1980, 1983; Fowler & Saltzman, 1993).

Thus, it is possible to arrive at some understanding of the kind of information that can be useful for speech perception—information specifying the linguistically significant gestures of the vocal tract.

Defining a Gesture

A *gesture* is a linguistically significant action of the vocal tract that is implemented by a transiently achieved coordinative structure or synergy (Fowler & Saltzman, 1993). A prototypical and well-studied example is bilabial closure for /b/, /p/, or /m/, which is achieved by a transiently established coordinative relation among the articulators of jaw, upper lip, and lower lip (Kelso, Tuller, Vatikiotis-Bateson, & Fowler, 1984). As a consequence of the coordinative relation among the articulators, closure is achieved flexibly and equifinally. "Flexibility" means that the specific contribution that each articulator makes to the closure gesture will vary with context—for example, with coarticulatory demands on it. Accordingly, the jaw will contribute less, and the lips correspondingly more, to closure during /ba/ than during /bi/ due to the overlapping influences of the vowels' articulators. "Equifinality" means that despite competing coarticulatory demands on each articulator that affect the nature of the contribution each makes to achievement of a gesture, the coordinative relation among the articulators ensures invariant achievement of the macroscopic gestural goal, here, bilabial closure. Thus, the gesture that constitutes a phoneme or part of a phoneme involves multiple articulators, some of which may also be involved in other phonemes' gestures.[1]

The gestures that are relevant to the present research are the devoicing gesture of the larynx for a voiceless obstruent and the gesture or gestures that achieve contrastive stress accent. The devoicing gesture is achieved by opening and stiffening the vocal folds during achievement of consonantal closure. The contrastive accent is achieved by a variety of laryngeal and respiratory means that, among other consequences, cause the vocal folds to open and close more rapidly than on unaccented syllables (see Fowler, 1995). Although the larynx is involved in both gestures, the gestures differ in three ways that will give rise to distinguishable acoustic consequences. First, they are qualitatively distinct actions—the devoicing gesture opens the vocal folds; the accentual gesture modulates the rate of opening and closing of the folds. Second, their time courses are distinct—the devoicing gesture is brief and is tied temporally to production of the unvoiced consonant; contrastive accent has a syllable as its domain. Third, contrastive accents appear to involve the respiratory system as well as the larynx (Fowler, 1995); accordingly, there are correlated effects on fundamental frequency, amplitude, and syllable duration that may jointly serve as an acoustic signature of this type of accent. When devoicing for a consonant and contrastive stress gestures are coarticulated, they have converging effects on a common acoustic dimension, fundamental frequency. How can the perceiver derive the underlying phonological structure of an utterance when the acoustic manifestations of phonemes vary with coarticulatory context? We intend to show that this is accomplished by listeners' use of acoustic signatures of gestures to specify the underlying gestural events, in this case, consonant devoicing and contrastive stress accent.

A Theory of Segmental Parsing

Fowler and other researchers (Fowler & Saltzman, 1993; Fowler & Smith, 1986) have outlined a specific proposal for understanding how acoustic variation due to coarticulation can specify gestural events. They have

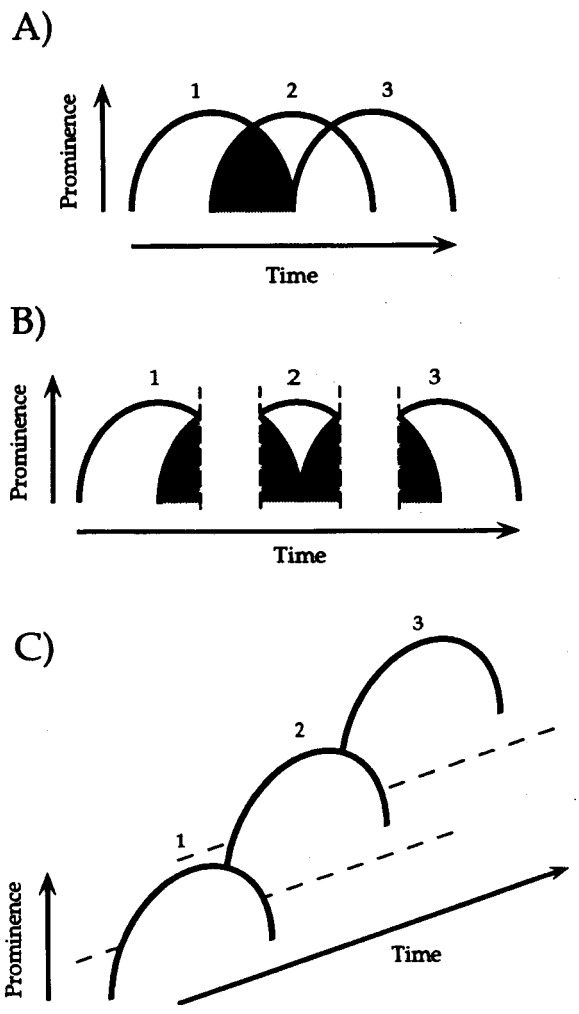

Figure 1. Illustration of the gestural parsing approach to speech perception outlined in Fowler and Smith (1986), which was modified by and adapted from Remez (1994). The theory of gestural parsing arises from the consideration of the prominence relationship shown in Figure 1A: The influence of a gesture grows and subsides over time, with portions of consecutive gestures overlapping, as outlined by the gray area between segments 1 and 2. In Figure 1B, perceivers represent separate contextualized segments, designated by relative prominence shifts at the dotted lines, thereby requiring an additional process to remove the resulting attached gray area. In Figure 1C, a semiparallel analysis of segmental vectors along gestural lines, perceptual parsing, automatically removes coarticulatory acoustic variation.

observed that gestures occur in a smoothly graded fashion in which the relative prominence of a segment waxes and wanes continuously over time, as illustrated in Figure 1A. From this analysis, there will be times during the acoustic realization of a gesture when it dominates the signal, and times when it has an influence on, but is less prominent than, another gesture's acoustic manifestation. In order to understand what was said, the perceiver must find some way to extract acoustic information for the separate gestures from the coarticulated acoustic waveform.

One way to divide the acoustic signal into intervals that should be maximally informative about individual phonetic segments is to draw lines perpendicular to the axis of time, as in the context-sensitive acoustic fragments of Figure 1B. Listeners must be supposed to do something similar to this in any perceptual theory in which invariance at the gestural level is denied and perception occurs in two broad stages: (1) an initial auditory analysis that is not special to spoken events, and (2) classification, based on the context-sensitive results of that analysis, into phonological consonant and vowel categories. The initial auditory analysis must yield a context-sensitive signal because, in the absence of gestural invariants, there is no way for general auditory processes to evade the context sensitivity. This is the case whether these auditory processes are presumed to segment the signal (e.g., as in Figure 1B) or not (e.g., as in Figure 1A). Our context-invariant (with respect to coarticulation) alternative approach is that perceivers parse the acoustic signal into acoustic signatures of gestures, as in Figure 1C. This account is to be distinguished from other accounts of speech perception in invoking context-invariant gestures as perceptual events, as opposed to context-sensitive auditory representations mapped onto discrete phonological categories.

Some recent empirical studies of speech perception have focused on the acoustic detail and context sensitivity of spoken signals. Sawusch and Gagnon (1995) proposed and concluded that both phonetic and nonphonetic categorization of sounds is based on the same intermediary auditory representation. Likewise, Samuel (1981) and Samuel and Newport (1979) have argued for a primary role for purely acoustic properties in phonetic categorization. Furthermore, other researchers (e.g., Diehl & Kluender, 1989a, 1989b; Kluender, Diehl, & Killeen, 1987; Kuhl, 1987) have referred to prototype or exemplar-based segmental categories, which are context-sensitive representations of consonant and vowel segments. Thus, the acoustic/auditory approach predicts that at some level, context-sensitive acoustic correlates of phonetic segments are represented by the perceiver and should play a part in further phonetic or nonphonetic perceptual tasks.

Although Elman and McClelland (1986) did not propose that articulatory gestures are extracted in their TRACE model of speech perception, they did incorporate a notion similar to gestural parsing by having acoustic feature nodes in different time slices map onto multiple phonemes at the phonemic level. This allows adjacent overlapping phonemes to influence the relative activation strengths of the acoustic features of their neighboring phonemes. The influence serves to cancel out coarticulatory acoustic effects and leaves open the possibility for context-invariant representation with respect to coarticulation. Their illustration of overlapping connections between phonemes and acoustic feature nodes is quite similar to our account of gestural parsing of acoustic information, provided in Figure 1C.

The theory of perceptual parsing makes predictions about perception that are distinct from those of a theory of purely acoustically governed perception, exemplified in Figure 1B. If perceivers parse the acoustic signal along gestural lines, rather than into context-sensitive acoustic segments, effectively removing acoustic influences among overlapping segments from the start, gesturally parsed segments should sound different from acoustically partitioned segments in some of their acoustic/auditory properties.

Testing the Theory

At this point, there is some promising empirical research examining this prediction of the theory of gestural parsing (Fowler, 1981, 1984; Fowler & Smith, 1986). In the first study, the subjects were asked to choose which of two pairs of trisyllabic VCəCV nonsense words contained more similar medial vowels. The medial vowels were either acoustically identical or different. For example, comparison pairs of Type A, /abə$_a$ba/–/ibə$_i$bi/, contained acoustically different medial schwa vowels (due to their different original coarticulatory contexts) that were spliced into appropriate coarticulatory contexts. (The subscripts on the medial vowels indicate original flanking vowel context.) Comparison pairs of Type B, /abə$_i$ba/–/ibə$_i$bi/, contained acoustically identical medial schwa vowels spliced into both an inappropriate and an appropriate coarticulatory context. If perceivers parse along gestural lines, removing the acoustic effects of flanking vowels from the medial schwas, then the acoustically different schwas in their appropriate contexts (Type A pairs) should sound alike. In contrast, if perceivers use acoustically chopped segments (see Figure 1B), then the schwas in Type A pairs should sound different, and the acoustically identical schwas in Type B pairs should sound alike. Fowler (1981) found that subjects rated schwas in Type A pairs as more similar than those in Type B pairs, thereby suggesting that perceivers parsed the acoustic signal according to its gestural causes.[2]

These findings were extended in Fowler's (1984; Fowler & Smith, 1986) later two studies of differences in choice response times for vowel identification in appropriate versus inappropriate coarticulatory contexts. Overall, listeners were faster at identifying final vowels in əCV disyllables when the schwas provided appropriate coarticulatory information for the vowel. This implies that listeners were using the information in the schwa vowel to anticipate the final vowel. When this information was inconsistent, as in the inappropriate con-

text condition (e.g., ∂_iCa), listeners were misled, and vowel identification was slowed. Whereas the earlier discrimination results showed that coarticulatory effects on the schwa did not cause the schwa to sound context sensitive when it appeared in its proper context, the response time task showed that this was not because the coarticulatory effects are inaudible. Rather than serving as information for schwa quality, they serve as information for their distinct source, the coarticulating vowel.

The response time and discrimination results suggest that information for a particular phonetic gesture is parsed from that of its neighbors. This leads to slower identification when anticipatory gestural information is misleading and to perceived similarity of acoustically different schwas when gestural parsing should eliminate the distinctness. The study reported here further addresses the gestural parsing proposal in light of the following, more general, findings on segmental production and perception.

In a speech production study, Silverman (1987) compared the acoustic effects of voiceless versus voiced consonants on the *f*0 of overlapping adjacent vowels in CV syllables. The voicing feature distinguishes such consonants as /g/, /d/, and /b/ (voiced) from /k/, /t/, and /p/ (voiceless). His overall findings were that the fundamental frequency (*f*0) of vowels following voiceless consonants always fell from a higher frequency than did those following voiced consonants.[3] An explanation for this acoustic difference is provided in a study by Löfqvist, Baer, McGarr, and Seider Story (1989), who found that when talkers open the vocal folds to devoice a consonant, they tense the cricothyroid muscle (CT). This in turn stiffens the vocal folds to keep them apart as air from the lungs rushes through the glottis. When talkers adduct the folds for the overlapping vowel portion of the utterance, the residual stiffening of the folds raises *f*0.[4]

Silverman (1986, 1987; see also Diehl & Molis, 1995; Haggard, Summerfield, & Roberts, 1981; Whalen, Abramson, Lisker, & Mody, 1990) next sought evidence that such *f*0 information produced during a vowel is used in consonant voicing perception. He hypothesized and found that imposing a falling *f*0 contour on a vowel following a consonant (consistent with the effect on *f*0 of voiceless consonants) created a shift in the identification curves for a voiced to voiceless consonant continuum toward more voiceless responses. These findings are in line with predictions made by the parsing theory in that some of the *f*0 information typically associated with a postconsonantal vowel is due to, and should therefore influence, perception of the consonant.

Silverman (1987) also examined the interaction of intrinsic fundamental frequency (I*f*0) of vowels with perceived intonation. Vowel I*f*0 reflects the regular pattern of *f*0 variation across different vowels: In general, close vowels such as /i/ have higher *f*0s than do open vowels such as /a/. To the degree that greater perceived prominence, or stress, relies on the perception of higher *f*0 information heard as higher pitch (see, e.g., Lehiste, 1970), one might expect that close vowels would be perceived as relatively more stressed than open vowels in a sentence. On the contrary, Silverman hypothesized and found that listeners adjusted for I*f*0 in making relative stress judgments. Therefore, this study provides complementary, but not converging, evidence for the gestural parsing theory.

As a result of these findings, the following tests were devised for the theory of perceptual parsing. In the first experiment, we attempted to replicate the findings that acoustic information during a vowel influences the perception of a prior overlapping consonant. This is a necessary but not uniquely sufficient condition for our parsing theory. It could be that no parsing occurs; rather, hearing a vowel as higher in pitch cues a preceding voiceless consonant. To resolve this, we performed a second experiment to test for a reciprocal influence of consonant identity on the perception of the pitch of a following overlapping vowel. Taken together, these experimental findings may provide additional insight into how the perceptual system deals with coarticulatory acoustic variation.

EXPERIMENT 1
Do Perceivers Use Vowel *f*0 Information to Disambiguate Prior Overlapping Consonant Identity?

One kind of prediction that a theory of gestural parsing makes concerns the information that signals consonantal identity. That is, because segments are coarticulated, some acoustic consequences of consonant production will occur during the time that acoustic consequences of a following vowel are most prominent in the speech signal. These consequences of consonant production should serve as perceptual information for the prior consonant. In this experiment, we attempted to replicate Silverman's (1986, 1987) finding that information during a vowel affects the perception of a prior overlapping consonant. Because we know that *f*0 typically falls steeply after voiceless as opposed to voiced consonants (see, e.g., Hombert, 1978; Silverman, 1987), and following Löfqvist et al. (1989), we ascribe this to residual vocal fold tension—a fall in *f*0 after consonants that are ambiguous with respect to voicing ought to foster perception of the consonant as unvoiced if listeners are parsing the consonant gestural information from the overlapping vowel portion.

Method

Subjects

Twenty-six Yale University undergraduates were tested and received introductory psychology credit for their participation. All were native speakers of English and reported normal hearing.

Materials

The test materials consisted of 24 resynthesized /amaCa/ tokens designed to vary in their degree of perceptual ambiguity between /amaga/ and /amaka/, with stress on the final syllable. This was accomplished by taking /amaka/ and /amaga/ utterances (spoken by C.A.F.), measuring the natural closure durations and voice onset

times (VOT) and creating a continuum varying both closure duration and VOT within these natural values. The continuum was created by digital editing (using the HADES software package developed at Haskins Laboratories for a DEC VAXStation; see Rubin, 1995, for more detail) of the original /amaka/ to shorten closure duration successively in 5-msec steps from 45 msec, and VOT in 5-msec steps from 25 msec for five steps, leaving a total of six different items. Appendix A provides closure duration and VOT values for these tokens.

To test for perceptual parsing of consonant information from overlapping vowel information, we also varied the $f0$ during the final /a/ to give it four possible falling contours ($f0$ ramp) within each step of the continuum: We resynthesized the tokens and introduced a flat $f0$, a 10-Hz fall, a 20-Hz fall, or a 30-Hz fall in $f0$ on the postconsonantal vowel. The initial /ama/ portions were assigned flat $f0$ contours. In the resynthesis procedure, the original spectral values of the utterances are used to generate new tokens with designated $f0$ values and contours. We used the ILS software package for a DEC VAXStation to perform the resynthesis.

Procedure

In order to obtain more complete information about the subjects' perception of these tokens, the task was constructed so that we could obtain not only information about consonant identity, but also ratings of perceived consonant goodness. Therefore, on each trial the subjects heard one of the /amaCa/ tokens and circled a number from 1 to 5, indicating both consonant identity and goodness (1 = *clear "ga,"* 2 = *less clear "ga,"* 3 = *completely ambiguous between "ga"* and *"ka,"* 4 = *less clear "ka,"* 5 = *clear "ka"*). Listeners were given sufficient time in which to make their choices (4 sec per trial). The ratings for each token were then averaged across the five randomly distributed presentations of each item across the 120-trial test. These data were subjected to a two-way repeated measures analysis of variance (ANOVA) to test for the effects of $f0$ ramp (different falling contours) and token step (variations in closure duration and VOT). The subjects were tested individually or in pairs from a cassette tape over headphones.

Results and Discussion

As illustrated in Figure 2, the effect of $f0$ ramp on mean ratings of consonant goodness and identity was to create an overall shift in the response curves for the token steps from more good "ga" ratings to more good "ka" ratings. First, increases in closure duration and VOT led to increases in "ka" responses. Second, higher $f0$s also increased "ka" responses. Third, the pattern of influence of token step and $f0$ ramp (interaction) was concentrated at the "ga" end of the continuum. Thus, the traditional influence of closure duration and VOT variation on consonant identification seen at a flat $f0$ ramp was influenced by adding the more voiceless-consistent information to the overlapping vowel in the 10-, 20-, and 30-Hz $f0$ ramp conditions. An ANOVA confirmed the hypothesized influence of $f0$ contour during a vowel on prior consonant perception, revealing main effects of $f0$ ramp [flat, 10-, 20-, and 30-Hz; $F(3,75) = 84.43, p < .0001$] and token step (1–6; $F(5,125) = 303.15, p < .0001$], as well as an interaction [$F(15,375) = 10.78, p < .0001$]. This finding is consistent with the idea that listeners attribute $f0$ information for voicelessness occurring during the vowel-prominent portion of the syllable to the previous overlapping consonant in perceiving consonant identity and goodness.

Although we replicated Silverman's (1986, 1987) finding that $f0$ information in the domain of a vowel is sufficient to influence prior consonant identification, it does not necessarily follow from this experiment that listeners are removing this information from the vowel and attributing it to the consonant, as we predict with gestural parsing. It could be that listeners first hear the vowel as higher in pitch in an auditory representation of the signal. In memory, they have associated high pitch in a vowel with voicelessness of a preceding consonant, and they use this association as the basis for their response. This would not be in line with the theory of segmental parsing that we are proposing as a mechanism for perception; however, it is consistent with the data. Thus, converging evidence is necessary to test whether listeners actually (1) separate the acoustic consequences of a consonant from those of an overlapping vowel because they are caused by distinct gestural events (Figure 1C) or (2) associate particular kinds of vowels with particular kinds of consonants as the basis for perceiving segmental identity (Figure 1B).

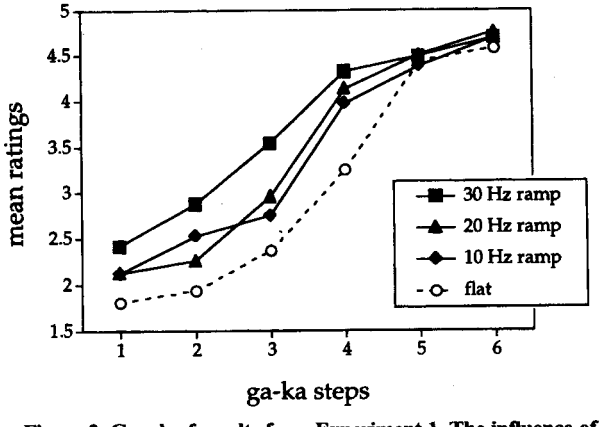

Figure 2. Graph of results from Experiment 1. The influence of vowel $f0$ information on the prior consonant is seen in the separation of the ramped from the flat ratings curves.

EXPERIMENT 2
Do Perceivers Parse Consonant $f0$ Information From Overlapping Vowel $f0$ Information?

Another prediction of the theoryof segmental parsing is that acoustic information for a consonant is perceptually removed from overlapping vowel information. Having shown in Experiment 1 that the $f0$ information for consonant voicing that occurs during a vowel can influence identification of a prior consonant for out next step we tested whether that information is removed from the acoustic consequences of the vowel. In this experiment, we tested the reverse influence of consonant voicing on perception of the pitch of an overlapping vowel. As discussed earlier, consonant voicing and intonational accent gestures are due to the action of multiple, partially shared articulators. We have focused on the acoustic consequences of CT activation as part of the devoicing gesture

in voiceless consonants. Although there are other influences on vowel $f0$, the effect of the devoicing gesture is to raise $f0$ during the vowel-prominent portion of a syllable. We predict that when a portion of the $f0$ information during a vowel can be attributed to a prior consonant's devoicing gesture, it will not contribute to the perception of the pitch of the vowel.

Here, we tested for differences in the perception of the pitch of the vowel, /a/, in different consonantal voicing contexts in two different conditions. If subjects parse consonant $f0$ information from the vowel, as they do with vowel I$f0$ from intonation (Silverman, 1987), then vowels following voiceless consonants, such as /k/, ought to sound lower in pitch than vowels following voiced consonants, such as /g/, when the two are actually identical in $f0$. Likewise, vowels following voiceless consonants ought to sound equal in pitch to vowels following voiced consonants when they are actually higher in $f0$. Furthermore, such observations would not be expected for context-sensitive representations of acoustic segments (Figure 1B), where we would expect listeners to use unparsed $f0$ information in a vowel following a voiceless consonant to hear a higher pitched vowel. Finally, we were interested in an additional consideration following from Silverman's findings—that vowel pitch might best be assessed within a sentence intonational contour. Thus, we tested for a difference in perceptual parsing of vowel pitch between vowels presented both in and out of a sentence context.

Method

Subjects

Twenty-one Yale University and University of Connecticut undergraduates were tested and received introductory psychology credit for their participation. All were native speakers of English and reported normal hearing.

Materials

The test materials for both conditions closely paralleled those of Silverman's (1987) I$f0$ study and started with two natural sentences (produced by C.A.F.) that differed only in the ordering of the two nonsense words, /amaga/ and /amaka/. They were (1) "I said amaga not amaka today," and (2) "I said amaka not amaga today," with contrastive stress on the /ga/ and /ka/ syllables. The /amaga/ and /amaka/ nonsense words were then spliced out of the sentence and resynthesized to assign $f0$ contours. The initial /am/ portions of the four tokens were given their original $f0$ values. In order to eliminate any anticipatory effects of /g/ and /k/ on the prior medial /a/, the $f0$ values assigned to that portion were equivalent, contoured values (180 Hz for the first 60 msec and a 25-Hz fall over the final 40 msec).

The final /ga/ and /ka/ syllables of the nonsense words received special attention because they are the focus of comparison. Pilot studies indicated that successful resynthesis requires the most natural $f0$ values and contours possible. Therefore, these syllables all had the same natural overall $f0$ contour, with the main differences being the dominant central, or base, values of the contours during the critical final vowel. The $f0$ values fell for 10 msec from 5 Hz above the base value of the syllable, then remained at the steady base value for 50% of the remainder of the syllable, and tapered to 4 Hz below the base value for the final 50% of the syllable. The syllables averaged 150 msec in duration. The initial fall of 5 Hz was chosen to be within the range of normal $f0$ values for both voiced and voiceless consonant contexts for this speaker. The base $f0$ values of the /ga/ and /ka/ syllables that appeared first in their sentences (Sentences 1 and 2, respectively) were resynthesized to match the average for the speaker at 200 Hz. These syllables served as a fixed basis for comparison with their counterparts later in their sentences. The base $f0$ values of the /ga/ and /ka/ syllables appearing second in their sentences were given the same overall contour, with values ranging in 5-Hz steps from 20 Hz below to 20 Hz above, inclusively, the 200-Hz base $f0$ value of the initial comparison syllables. A schematic diagram of the comparison pairings appears in Figure 3. The resynthesized $f0$ values for all the tokens are listed in Appendix B. The natural VOT values were not changed and averaged 20 msec for /ga/ and 60 msec for /ka/ syllables. Digital editing and resynthesis techniques were carried out using the HADES and ILS software packages on a DEC VAXStation.

Figure 3. Comparison trial structure for Experiment 2. The initial /amaCa/ token, whether /amaga/ or /amaka/, contains a single base $f0$ value, to be compared with one of nine possible base $f0$ values for the second /amaCa/ token. The comparisons range in 5-Hz steps from 20 Hz above to 20 Hz below the $f0$ value of the initial token.

In the sentences condition, the tokens were inserted back into their original sentence contexts to closely parallel the design of Silverman's (1987) study of I$f0$. In the pairs condition, the nonsense trisyllables were presented in pairs, with the same 400-msec interval between items, without their sentence context, to determine whether the sentence context is necessary for the evaluation of vowel pitch. Overall, 18 different counterbalanced pairings (two orderings of the /amaga/ and /amaka/ nonsense words by nine difference in hertz comparisons) appeared in the sentences and pairs conditions.

Procedure

Sentences. Because we were testing whether or not perceivers remove overlapping consonant $f0$ information from the following vowel portions of syllables, we asked subjects to make comparisons between the /ga/ and /ka/ syllables in the carrier sentences on the basis of relative stress. They were instructed to listen carefully to each sentence, focus on the /ga/ and /ka/ syllables, and tell us which one sounded more stressed in the sentence by sounding higher in pitch. Each sentence was repeated five times in random order across the 90-trial test, and subjects had 4 sec between trials to circle their answers. The "ka" responses were then collapsed across sentence order to yield average percent "ka" preference scores for each difference in hertz comparison. An analogous procedure yielded percent "ga" preference scores for each difference in hertz comparison. These preference scores were subjected to a two-way repeated measures ANOVA to test for the effects of difference in hertz and syllable difference. We then obtained a more direct measure of the amount of perceptual parsing of $f0$ information from the vowel. The /ka/ preference data were subjected to probit curve-fitting analyses to determine the difference in hertz

between /ga/ and /ka/ at the 50% crossover point for /ga/ to /ka/ preference. A one-tailed t test was performed on these data to test for a difference from a 0-Hz crossover point. Evidence for parsing would constitute finding a positive difference in hertz between /ka/ and /ga/ at the 50% crossover point for preference. That is, /ka/ must be higher in $f0$ than /ga/ to sound the same in pitch. The subjects were tested individually or in groups (up to 4) over headphones. Stimuli were presented from a cassette tape.

Pairs. In this condition, there were only two main procedural departures from the sentences condition: (1) The subjects were instructed to make comparisons directly on the basis of pitch (as opposed to stress mediated by pitch), and (2) a new random ordering of comparison pairs was used. All other testing details were the same as in the sentences condition.

Each subject performed in both conditions, counterbalanced for ordering of the conditions. Finally, a two-tailed t test was performed to test for a difference in the amount of observed parsing between these two conditions. If the sentence context is necessary for the evaluation of vowel pitch, we should find a difference between the amount of parsing between these two conditions.

Results and Discussion

Sentences. Figure 4 plots percent judgments that /ga/ or /ka/ sounded more stressed in the sentence as a function of difference in hertz between the syllables. For the /ga/ curve in the figure, the x-axis represents the difference in Hz (/ga/ minus /ka/) between the /ga/ and /ka/ syllables in a sentence, and likewise (/ka/ minus /ga/) for the /ka/ curve. The two curves are therefore completely predictable one from the other. That is, the percent /ka/ preference at −20 Hz is 100% minus the percent /ga/ preference at +20 Hz.

The curves for /ga/ and /ka/ form separate ogives. At all differences in hertz, /ga/ was perceived as more stressed or higher pitched than /ka/. This tendency was particularly strong in the middle of the curves, where the difference in hertz between /ka/ and /ga/ syllables was smallest. These data indicate that, as predicted, when /ga/ and /ka/ are physically equivalent in $f0$, /ka/ is perceived as being lower in pitch than /ga/. An ANOVA confirmed the hypothesized influence of consonant identity on following vowel pitch perception, revealing a main effect of difference in hertz [20, 15, 10, 5, 0, −5, −10, −15, −20; $F(8,160) = 97.86$, $p < .0001$] and of syllable difference [/ka/ vs. /ga/; $F(1,20) = 6.56$, $p < .019$], as well as a marginal interaction [$F(8,160) = 1.97$, $p < .054$]. (A prior three-way ANOVA that included syllable order showed that it had no effect, so only the two-way ANOVA data are reported.) The main effect of syllable appears to reflect perceptual parsing from /ka/ syllables of the higher $f0$ normally caused by devoicing consonants. More compelling evidence for this is seen in the next set of data analyses.

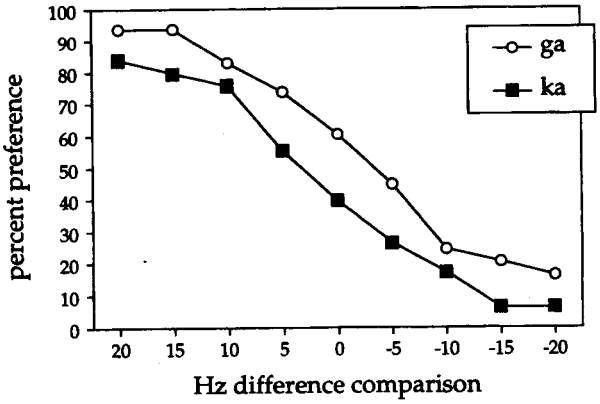

Figure 4. Graph of results from the sentence condition of Experiment 2. The influence of consonant identity on vowel pitch perception is seen in the separation of the /ga/ from the /ka/ preference curves at equivalent difference in hertz comparisons.

To obtain a quantitative estimate of perceptual parsing, each subject's /ka/ preference data were analyzed to determine the difference in hertz at the 50% crossover point of the ogival curve: Positive difference in hertz scores indicate parsing that lowers the pitch of the vowel, whereas negative difference in hertz scores indicate the opposite, and a 0-Hz difference at the 50% crossover point indicates a failure to find parsing. In performing the test, 1 subject's data were rejected by the probit analyses because the percent values did not form anything close to an ogive.[5] The one-tailed t test for a difference from zero on the mean parsing score of the remaining subjects indicated parsing that lowers the pitch of vowels in /ka/ syllables [$M = 3.66$, $SD = 7.76$; $t(19) = 2.16$, $p < .025$]. That is, subjects appeared to be removing about 4 Hz from /a/ vowels following the voiceless consonant /k/.

Pairs. Figure 5 plots percent judgments that /ga/ or /ka/ is higher in pitch across the pairs as a function of the difference in hertz between them. The two ogival curves are again distinct, replicating the main overall tendency for /ga/ syllables to sound higher pitched than /ka/ syllables. However, the curves are closer together than in the sentences condition. An ANOVA confirmed the hypothesized influence of consonant identity on following vowel pitch perception, revealing an effect of difference in hertz [20, 15, 10, 5, 0, −5, −10, −15, −20; $F(8,160) = 192.40$, $p < .0001$] and of syllable difference (/ka/ vs. /ga/; $F(1,20) = 4.86$, $p < .039$], as well as an interaction [$F(8,160) = 220.90$, $p < .018$]. (The interaction was significant here because the curves converge at the endpoints.)

The probit analyses for these data revealed numerical parsing, but the mean difference in hertz between /ka/ and /ga/ at the point of subjective pitch equality was only marginally different from zero in a one-tailed t test [$M = 1.94$, $SD = 5.28$; $t(20) = 1.68$, $p < .06$]. Although the effect did go in the predicted direction—that is, /ka/ syllables must be higher to sound equal to /ga/ syllables—it is not significantly different from no parsing in these analyses.

Finally, the two-tailed t test for the difference in the parsing measures between the sentences and pairs conditions was nonsignificant [$M = 2.75$, $SD = 7.42$; $t(19) = 1.66$, $p < .11$]. (Note that these data also exclude the data for the subject who had to be eliminated from the sentences condition in the parsing analyses.)

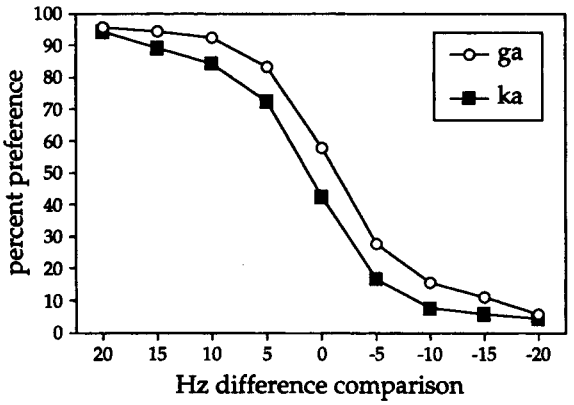

Figure 5. Graph of results from the pairs condition of Experiment 2. A replication of the influence of consonant identity on vowel pitch perception is seen in the separation of the /ga/ from the /ka/ preference curves at equivalent difference in hertz comparisons for a slightly different task.

The results of both ANOVAs showed, as predicted, overall effects of voiced versus voiceless consonants on syllable preference in both sentences and pairs; however, the parsing measure derived from probit analyses fared less well. Although there was significant parsing, which lowered the vowel following the voiceless consonant in the sentence context, the measure of parsing across pairs was only marginally different from zero. Moreover, the difference in the parsing measure between these two context conditions was also nonsignificant. Thus, we cannot conclude that subjects were doing anything differently across the two conditions, only that the assessment of parsing provided by the probit analyses failed to indicate significant parsing without a sentence context.

The magnitude of our significant measure of parsing itself is unexpectedly low, considering that Silverman (1987) observed that vowels following voiceless consonants were on average 7 Hz higher than those following voiced consonants at consonant release. Lehiste and Peterson (1961) provided evidence for an average difference in $f0$ of /a/ vowels after /k/ and /g/ of 12 Hz. Finally, Hombert's (1978) observations indicated a 15-Hz difference in $f0$ values at the release of voiceless as opposed to voiced consonants. One might therefore expect subjects to parse at least 7 Hz and possibly up to 15 Hz from a vowel in the context of a voiceless consonant. Perhaps the failure of the probit measure of parsing in the pairs context was due to a more general problem with this measure of parsing, even in the sentence context condition. We will return to these issues as we move on to a more general discussion of these studies.

GENERAL DISCUSSION

In these experiments, we tested a theory that listeners perceive the causes of coarticulatory acoustic variation. First, we developed the theory that listeners perceive linguistically significant gestures of the vocal tract whose acoustic consequences can be extracted via gestural parsing. The specific predictions from the theory were that $f0$ information during a vowel would influence the perception of voicing of a prior overlapping consonant, and that consonant voicing would likewise influence the perception of the pitch of the following overlapping vowel. In general, these predictions were borne out by our tests, although not without some unexpected observations.

The results of Experiment 1 are clear. Following Silverman (1986, 1987), we tested the hypothesis that $f0$ information during a vowel would influence the perceived identity and goodness along the voicing continuum of a prior overlapping consonant. This is exactly what we observed in the data: The goodness/identity ratings were markedly influenced by our manipulation of $f0$ during the vowel: Overlapping vowels with $f0$ information consistent with /ka/ increased /ka/ ratings, even at the /ga/ end of the continuum. Although this can be interpreted as evidence for the parsing theory, it does not exclude other possible explanations. A context-sensitive acoustic theory can account for the finding by positing associative links between higher pitched vowels heard as such and preceding voiceless consonants. However, if listeners do not hear these vowels as being higher in pitch, this explanation can be ruled out. Furthermore, an observation that listeners do not hear the vowels as higher in pitch would serve as additional direct evidence for the parsing theory. This is the kind of evidence that we sought in conducting the next experiment.

Experiment 2 was conducted to test the collateral hypothesis that using $f0$ information during a vowel to perceive overlapping consonant voicing identity would influence the perception of the pitch of a following vowel. We were also interested in the effect of the presence of a sentence context on perception of a vowel's pitch. As Silverman (1987) pointed out, intonational context is very important for the perception of segmental pitch, and this could have an impact on our results. If listeners are parsing segments along gestural, rather than purely acoustic, lines (the difference between Figures 1C and 1B), then the effects on $f0$ of a consonant should not be heard as part of an overlapping vowel's pitch. The results of Experiment 2 provided evidence that listeners do not always perceive vowels following voiceless consonants as higher pitched than vowels following voiced consonants when their acoustic manifestations follow that pattern. The significant separation of the /ga/ and /ka/ preference curves at equivalent levels of difference in hertz was one kind of supporting evidence. As a quantitative measure of parsing, the probit analyses provided additional evidence for the conclusion, but reliably only in the sentence context condition. Possibly, the sentence context is necessary for parsing to be measured, although the analysis did not show that subjects were doing anything different when the sentence was not present. However, in both contexts, the magnitude of measured parsing was smaller than would be expected given what is

known in speech production about the typical effect on *f*0 of voiceless as opposed to voiced consonants. Next, we discuss explanations for this particular outcome.

First, it is necessary to rule out the possibility that the parsing results are spurious. If there were only the probit analyses to consider, this suggestion might have some merit, at least for words spoken in isolation. However, both ANOVAs revealed significant curve separation, thereby replicating each other. It is unlikely that such reliable findings as these are spurious; therefore, we could question the probit analyses. Yet these analyses do incorporate the very same data as the ANOVAs, with low standard errors. Thus, we must conclude that there is nothing out of the ordinary with the data-analytic techniques; the problem must be somewhat more interesting.

Is it that our listeners were really only parsing a small portion of typical consonant effects from overlapping vowels during perception? This could be true in two ways: First, our testing materials may not have lent themselves very well to the parsing mechanism, and second, our task may not have been sufficiently sensitive to reflect the greater perceptual parsing that our subjects might show outside the laboratory. Both these possibilities may be playing a role here. Although the stimuli were intelligible and natural sounding, the resynthesis technique used to control the *f*0 values may have created an unusual situation for the listeners. After all, the *f*0 differences experienced in this study were not really caused by a talker, perhaps making the subjects parse less than they might have given completely natural input. The materials could have been rejected, or the resulting parsed product could have been degraded due to the insufficiency of the materials for parsing. Moreover, the task of hearing vowel pitch could have been performed on portions of vowels in which little or no consonant influences occur, at some point in the vowel after the consonant gesture had ended. Therefore, both the limitations in the materials we used and the sensitivity of our task may have reduced the accuracy of our parsing measure from greater expected levels in the sentences to nonsignificant levels in the isolated pairs. Another study by Fowler and Brown (1997) has reported similar underestimates of expected parsing with vowel I*f*0 using paired words. This idea is especially appealing if we consider the isolated pairs condition a larger departure from typical speech than even the sentences condition, and note that its measure of parsing was numerically lower. Thus, the relative adequacy of these materials and measures reflect a necessary compromise between the ordinary world of the perceiver and the confines of the laboratory.

A possible qualification of the coarticulatory relationship among consonants and vowels is illustrated in Figure 6, where typical consonantal influences are depicted to occur in more initial portions of vowels. In light of the results of Experiment 1, in which the effect of *f*0 during the vowel on consonant identification occurred to a larger extent than did the reciprocal influences observed in Experiment 2, perhaps this last explanation more accurately characterizes the problems encountered in Experiment 2. That is, if consonants overlap with and influence the acoustic signal in relatively shorter portions of vowels than the reverse, then parsing of consonant information from vowel information may be relatively more difficult to measure accurately with our pitch comparison task.

To confirm this, we need to know how persistent the consonant voicing gesture is in its influence on *f*0 during a following vowel. Looking back at Silverman's (1987) production data, we see that the raising of *f*0 following voiceless consonants may persist throughout the vowel (although it is ultimately reduced to around 2.5 Hz), but as he pointed out, his data were confounded by the consonantal context following these vowels as well. Hombert (1978) showed that voiceless consonant effects on *f*0 may persist as long as 100 msec after release into a vowel (but they are reduced to around 4 Hz at this point). However, these data are likewise difficult to interpret in relation to the present study because he used only 5 talkers, and the vowel in which measurements were made was /i/ as opposed to our use of /a/. Lehiste (1970) and Lehiste and Peterson (1961) provided average *f*0 values for only one portion of the signal—at the peak intonation contour of syllable nuclei—and were unclear in describing exactly where in the signal these values were obtained. Thus, their work provides no useful information for the persistence of *f*0 departures. Our measurements of our talkers' original utterances show relatively brief consonantal ef-

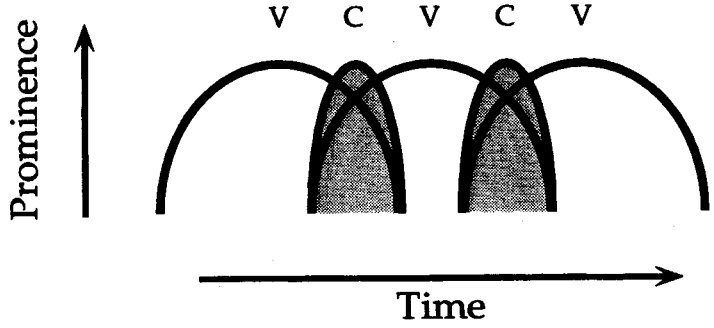

Figure 6. A possible qualification of the nature of segmental parsing. Consonants are represented as having relatively smaller influences on vowels than the reverse, possibly resulting in difficulties with parsing measurement.

fects on *f*0, and these served as the basis for imposing brief *f*0 falls on the /a/ vowels we used in our tests. However, this is also based on very limited information gleaned from materials not designed specifically to examine this question. In summary, although the relative influences of consonants and vowels both gesturally and acoustically are unclear, we tentatively propose that the depiction in Figure 6 may accurately represent the relationship. If this is the case, our pitch judgment task performance probably reflects both parsing of overlapping consonant information from vowel information and perception of the unparsed portion of the vowel, leading to lower estimates of *f*0 parsing than would be expected on the basis of production measures alone.

There is a final issue that a reviewer's comment provokes us to address. We have proposed that listeners parse unitary acoustic dimensions such as *f*0 when distinct linguistic gestures have converging effects on them. Further, we have proposed that parsing of the sort we observed in our experiments occurs because listeners detect the acoustic signatures of gestures as a means of identifying the gestures themselves, which constitute the speaker's phonological message. However, there may be an alternative interpretation of these findings. As most speech theorists propose, ourselves excluded, phonological categories may not be gestural or, for that matter, acoustic; rather, they may be abstract mental categories. For reasons that largely are not addressed in these theories (but see, e.g., Diehl & Kluender, 1989a, 1989b; Kingston & Diehl, 1994, and Kluender, 1994, who do address them), the categories have become associated with a constellation of often diverse acoustic "cues" that listeners use to identify the categories in speech. To signal a phonological category to a listener, therefore, a speaker has to produce its associated constellation of acoustic cues. Sometimes, due to coarticulation, cues from different constellations may converge on a common acoustic dimension such as *f*0, and to recover each constellation, listeners must parse. From this perspective, speakers articulate in order to produce the acoustic cue constellations of abstract phonological categories, and this explains the tight correspondence between articulation and constellations; it is not that the constellations serve to specify gestures to listeners.

There are theoretical grounds on which we have argued for our alternative proposal (see, e.g., Fowler, 1996). Here, however, we focus on some empirical grounds that we believe can distinguish these views. Diehl and colleagues have proposed that the diverse acoustic cues associated with distinct abstract phonological categories tend to be selected to serve in constellations because they are mutually auditorily enhancing. These investigators have emphasized the degree of independent control over the articulators that speakers can, in principle, exert to produce constellations of mutually enhancing cues (see, e.g., Diehl & Kluender, 1989a, 1989b). In contrast, our understanding of the literature on speech production (for a review, see, e.g., Fowler & Saltzman, 1993) is that, regardless of what anyone may argue in principle, in reality, speech is like other intentional actions (e.g., Turvey, 1990) in that it involves a high degree of coordination among articulators and therefore considerable loss of independence. The jaw and lips may be independent in principle, but they are not independent when they jointly contribute to the coarse-grained gestural goal of bilabial closure, for example. Likewise, vocal fold abductors and tensors may, in principle, be independent, but not when they jointly contribute to the coarse-grained gestural goal of devoicing.

Accordingly, our theory would require that acoustic cues serve in constellations only when they are products of the same gesture or coupled gestures in the ways we have observed in our experiments—both in providing information for the gesture (or phonological category) and in being parsed from other such cues that converge on a common acoustic dimension. For us, cues cannot serve in common constellations when they are products of independent gestures. Gestures, as defined in the introduction, comprise coordinations among articulatory contributors, and so the components of constellations (i.e., the acoustic cues) are not independently produced. Therefore, they cannot be independently controlled to provide maximal acoustic distinctiveness. In light of this consideration, our interpretation of the present findings depends on vocal fold abductions and stiffening being coupled components of a devoicing gesture, as Löfqvist and colleagues (Löfqvist et al., 1989; Löfqvist, McGarr, & Honda, 1984) appear to have shown. It depends on their not being independent actions—one to devoice a consonant and the other to enhance the distinguishability of unvoiced and voiced consonants, as Kingston and Diehl (1994) proposed (in part on the basis of an erroneous critique of Löfqvist et al., 1989, as argued in note 3). An important direction for future research to take, then, is to test a case of each sort—cases in which acoustic cues are believed to be independently produced and cases in which the cues are believed to be joint consequences of a gesture (the present case, we argue).

Further explorations of the perception of coarticulatory acoustic variation could focus on these final issues: (1) the methodological questions surrounding the construction of testing materials and measures, (2) the question of the duration of acoustic devoicing effects on *f*0, and (3) the details of articulatory coupling and independence that will constrain our gestural (versus the acoustic dispersion) account. Further production studies are necessary (1) to determine both how large and how persistent consonant acoustic influences are on following overlapping vowels and (2) to determine the extent to which components of gestures can be independently controlled. Such findings can both inform material and task construction in attempting to measure gestural parsing ef-

fects and constrain the theory of parsing. However, we are more interested in stressing this study's relevance as additional evidence for the theory of perceptual parsing. This perspective on speech perception suggests that we should not ask how listeners *overcome* coarticulatory acoustic variation, but rather how they *use* it as information for its gestural causes. This new conceptualization of variability leads to new possibilities for understanding both speech perception and perception in general.

REFERENCES

DIEHL, R. L., & KLUENDER, K. R. (1989a). On the objects of speech perception. *Ecological Psychology*, **1**, 121-144.

DIEHL, R. L., & KLUENDER, K. R. (1989b). Reply to commentaries. *Ecological Psychology*, **1**, 195-225.

DIEHL, R. L., & MOLIS, M. R. (1995). Effect of fundamental frequency on medial [+voice]/[−voice] judgments. *Phonetica*, **52**, 188-195.

ELMAN, J. L., & MCCLELLAND, J. L. (1986). Exploiting lawful variability in the speech wave. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 360-385). Hillsdale, NJ: Erlbaum.

FOWLER, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, **8**, 113-133.

FOWLER, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech & Hearing Research*, **46**, 127-139.

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

FOWLER, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception & Psychophysics*, **36**, 359-368.

FOWLER, C. A. (1995). Acoustic and kinematic correlates of contrastive stress accent in spoken English. In F. Bell-Berti & L. Raphael (Eds.), *Producing speech: Contemporary issues* (pp. 355-373). New York: American Institute of Physics.

FOWLER, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, **99**, 1730-1741.

FOWLER, C. A., & BROWN, J. M. (1997). Intrinsic *f*0 differences in spoken and sung vowels and their perception by listeners. *Perception & Psychophysics*, **59**, 729-738.

FOWLER, C. A., & SALTZMAN, E. (1993). Coordination and coarticulation in speech production. *Language & Speech*, **36**, 171-195.

FOWLER, C. A., & SMITH, M. R. (1986). Speech perception as vector analysis: An approach to the problem of invariance and segmentation. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 123-139). Hillsdale, NJ: Erlbaum.

HAGGARD, M. [P.], SUMMERFIELD, [A.] Q., & ROBERTS, M. (1981). Psychoacoustical and cultural determinants of phoneme boundaries: Evidence from trading F_0 cues in the voiced–voiceless distinction. *Journal of Phonetics*, **9**, 49-62.

HOMBERT, J. M. (1978). Consonant types, vowel quality, and tone. In V. A. Fromkin (Ed.), *Tone: A linguistic survey* (pp. 77-112). New York: Academic Press.

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

KINGSTON, J., & DIEHL, R. L. (1994). Phonetic knowledge. *Language*, **70**, 419-454.

KLUENDER, K. R. (1994). Speech perception as a tractable problem in cognitive science. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 173-217). San Diego: Academic Press.

KLUENDER, K. R., DIEHL, R. L., & KILLEEN, P. R. (1987). Japanese quail can learn phonetic categories. *Science*, **237**, 1195-1197.

KUHL, P. K. (1987). The special-mechanisms debate in speech research: Categorization tests on animals and infants. In S. Harnad (Ed.), *Categorical perception* (pp. 355-386). New York: Cambridge University Press.

LEHISTE, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

LEHISTE, I., & PETERSON, G. E. (1961). Some basic considerations in the analysis of intonation. *Journal of the Acoustical Society of America*, **33**, 419-475.

LIBERMAN, A. M., COOPER, F. S., SHANKWEILER, D. P., & STUDDERT-KENNEDY, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431-461.

LÖFQVIST, A., BAER, T., MCGARR, N. S., & SEIDER STORY, R. (1989). The cricothyroid muscle in voicing control. *Journal of the Acoustical Society of America*, **85**, 1314-1321.

LÖFQVIST, A., MCGARR, N. S., & HONDA, K. (1984). Laryngeal muscle and articulatory control. *Journal of the Acoustical Society of America*, **76**, 951-954.

REMEZ, R. E. (1994). A guide to research on the perception of speech. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 145-172). San Diego: Academic Press.

RUBIN, P. E. (1995). HADES: A case study of the development of a signal analysis system. In A. Syrdal, R. Bennett, & S. Greenspan (Eds.), *Applied signal technology* (pp. 501-520). Boca Raton, FL: CRC Press.

SAMUEL, A. G. (1981). The role of bottom-up confirmation in the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 1124-1131.

SAMUEL, A. G., & NEWPORT, E. L. (1979). Adaptation of speech by nonspeech: Evidence for complex cue detectors. *Journal of Experimental Psychology: Human Perception & Performance*, **5**, 563-578.

SAWUSCH, J. R., & GAGNON, D. A. (1995). Auditory coding, cues, and coherence in phonetic perception. *Journal of Experimental Psychology: Human Perception & Performance*, **21**, 635-652.

SILVERMAN, K. E. A. (1986). *F*0 segmental cues depend on intonation: The case of the rise after voiced stops. *Phonetica*, **43**, 76-91.

SILVERMAN, K. E. A. (1987). *The structure and processing of fundamental frequency contours*. Unpublished doctoral dissertation, Cambridge University.

STEVENS, K. N., & BLUMSTEIN, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 1-38). Hillsdale, NJ: Erlbaum.

TURVEY, M. T. (1990). Coordination. *American Psychologist*, **45**, 938-953.

WHALEN, D. H., ABRAMSON, A. S., LISKER, L., & MODY, M. (1990). Gradient effects of fundamental frequency on stop consonant voicing judgments. *Phonetica*, **47**, 36-49.

NOTES

1. In some cases, gestures constitute phonemic segments themselves. In other cases (e.g., the unvoiced consonants under examination), two or more gestures may constitute a phonemic segment: For the unvoiced consonants, an oral constriction gesture and a devoicing gesture.

2. For Type B pairs, the consequences of the parsing of an /a/ vowel's gesture from the medial /∂_1/ made it sound more like a higher vowel, /ɪ/, than a schwa.

3. Perceptually, *f*0 is heard to a first approximation as the overall pitch of an utterance. Because *f*0 physically corresponds to the rate at which the vocal cords open and close, thereby fueling the spectral harmonic structure of spoken sounds, its acoustic realization is an important attribute of natural speech.

4. Kingston and Diehl (1994) have argued that the cricothyroid activity found by Löfqvist et al. (1989) during production of voiceless obstruents cannot explain the increase in *f*0 during a vowel following the voiceless consonant. They characterized the findings as showing that the elevation of cricothyroid activity for the voiceless consonant "occurs" at the end of a vowel preceding the consonant. This can be a variable interval from the onset of the vowel showing the increase in *f*0 (because consonants vary in intrinsic duration and may participate in clusters). Thus, longer intervals from the ostensible cause of the *f*0 in-

crease should be associated with smaller increases in $f0$. Indeed, for vowels farthest from the occurrence of cricothyroid elevation in Löfqvist et al.'s (1989) or Löfqvist, McGarr, and Honda's (1984) stimuli, Kingston and Diehl judged that they were too far away to show any effects of consonantal cricothyroid activity on $f0$. In contrast to these expectations, in the data the magnitudes of $f0$ elevation during vowels were invariant over these variable intervals.

Kingston and Diehl's (1994) description of the findings of Löfqvist et al. (1989) is misleading. An important mistake they made was to confuse the *onset* of cricothyroid activity with its "occurrence." Muscular activity does not come on and go off instantaneously. Consistent with the function that Löfqvist et al. (1989) ascribed to the cricothyroid activity during voiceless consonants—that of stiffening the vocal folds to keep them open during the constriction interval—the activity of the cricothyroid in their data did have its onset at the offset of a vowel preceding a voiceless consonant, but it also continued until consonant release. The relevant measure for estimating expected elevation of $f0$ during a following vowel, then, is the interval between the *offset* of cricothyroid activity and the onset of the following vowel, and this interval was quite short. Accordingly, it is quite plausible, as Löfqvist et al. concluded, that residual tension in the vocal folds during vowels that follow voiceless consonants explains the raised $f0$ during following vowels.

5. The program could not fit a curve to this subject's data in the sentences condition and crashed at every attempt. After examining the data, we decided that they were too deviant to attempt to include in these analyses. However, the data are included in the previous ANOVA, and this subject's data for the pairs condition were accepted by the program and included in those analyses.

APPENDIX A
Closure Durations and Voice Onset Times (VOT) for Tokens in Experiment 1

Step	Closure Duration	VOT
1 (/ga/)	20	0
2	25	5
3	30	10
4	35	15
5	40	20
6 (/ka/)	45	25

Note—Durations and VOTs are in milliseconds.

APPENDIX B
Fundamental Frequency Values for Tokens in Experiments 2A and 2B

	Token									
	a	m	a	g	a	a	m	a	k	a
First in sentence	217	210	180–155		205–200–196	196	180	180–155		205–200–196
Second, +20 Hz	190	175	180–155		225–220–216	190	180	180–155		225–220–216
Second, +15 Hz	190	175	180–155		220–215–211	190	180	180–155		220–215–211
Second, +10 Hz	190	175	180–155		215–210–206	190	180	180–155		215–210–206
Second, +5 Hz	190	175	180–155		210–205–201	190	180	180–155		210–205–201
Second, 0 Hz	190	175	180–155		205–200–196	190	180	180–155		205–200–196
Second, −5 Hz	190	175	180–155		200–195–191	190	180	180–155		200–195–191
Second, −10 Hz	190	175	180–155		195–190–186	190	180	180–155		195–190–186
Second, −15 Hz	190	175	180–155		190–185–181	190	180	180–155		190–185–181
Second, −20 Hz	190	175	180–155		185–180–176	190	180	180–155		185–180–176

(Manuscript received June 10, 1996,
revision accepted for publication November 13, 1996.)