Perception & Psychophysics
1988, 43, 137-146

Induction of rate-dependent processing by coarse-grained aspects of speech

PETER C. GORDON
Harvard University, Cambridge, Massachusetts

This study investigated the idea that human speech recognition can involve analyzing the speech signal at multiple levels of resolution, using the information obtained from relatively coarse levels of analysis as a context for interpreting detailed acoustic cues to segment identity. Three experiments examined the effectiveness of coarse-grained aspects of speech in inducing rate-dependent processing of closure duration as a cue to phonological voicing in a medial stop consonant (specifically, *rabid* vs. *rapid*). Experiment 1 showed that the rate of articulation of a severely filtered precursor phrase influenced voicing judgments about a segment in an unfiltered test word. Experiment 2 showed a similar effect when the amplitude envelope of the precursive speech was filled with a constant-frequency sine wave set at the fundamental of the test word. The contextual effects of these coarse-grained aspects of speech did not differ from those of acoustically detailed precursive speech. Experiment 3 showed that no context-dependent processing occurred when the amplitude envelope of the precursive speech was filled with white noise, indicating that the precursive sounds must have some acoustic continuity with the test word for integration to take place. The results support the idea that context-dependent processing can be based on coarse aspects of the speech signal.

Attempts to understand the recognition of spoken language are complicated by the fact that the acoustic cues to linguisitic units often depend on the context in which they appear. While there may be some context-*independent* cues to phonetic categories (Blumstein & Stevens, 1979; Cole & Scott, 1974; Stevens, 1980), it seems likely that a robust speech-recognition system must exploit context-*dependent* cues as well (Klatt, 1980). Human listeners certainly process many acoustic cues in a context-dependent fashion (Ladefoged & Broadbent, 1957; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Miller & Liberman, 1979). The present study explored one strategy for normalizing context-dependent temporal cues to phonetic segments. Specifically, it examined the question of whether relatively coarse aspects of the speech signal provide information about rate of articulation that can be used by human listeners as a context for interpreting detailed acoustic information about the identity of phonetic segments.

Many cues to the identity of phonetic segments consist of the durations of certain acoustic properties, or the rates at which acoustic properties change (Liberman, Delattre, Gerstman, & Cooper, 1956; Lisker, 1957; Lisker & Abramson, 1964; Peterson & Lehiste, 1960). These temporal properties are also influenced by factors unrelated to segment identity, such as overall speaking rate, position in a phrase, position in a word, stress level, and syllable structure (Klatt, 1976). A substantial body of literature indicates that listeners use *rate-dependent processing* (Miller, 1981, in press) to separate the segmental significance of temporal cues from these extrasegmental influences. The contexts that induce rate-dependent processing include pause rate and articulation rate (Miller & Grosjean, 1981), the duration of the syllable carrying the context-dependent cue (Miller & Liberman, 1979; Summerfield, 1981), and the rate of preceding speech (Port, 1979; Summerfield, 1981). The psychological processes underlying rate-dependent interpretation appear to operate automatically, independently of voluntary control (Miller, Green, & Shermer, 1984). While a good deal is known about rate-dependent processing (see Miller, 1981, in press, for reviews), some basic questions remain.

One question concerns the level to which listeners must process speech to obtain an effective context for interpreting an ambiguous acoustic cue. Because a variety of contextual factors influence the temporal properties of acoustic cues (e.g., overall speaking rate, phrasal position, word position, etc.; Klatt, 1976), a complete description of the linguistic factors influencing the temporal attributes of a segment would require identification of all of its surrounding segments and their durations. Such a strategy for context-dependent recognition is obviously unappealing. It poses the dilemma that, to recognize a segment, a listener must recognize its neighboring segments. Because the neighboring segments presumably exhibit context dependence as well, this presents a "classic chicken-egg problem" (Klatt, 1980). Recognizing neigh-

This research was supported by a small grant from the William F. Milton Foundation, and by funds provided by Harvard University. I would like to thank Tom Carrell for much helpful advice on all aspects of this work, Jean MacMillan, Joanne Miller, and Quentin Summerfield for helpful comments on an earlier version of this manuscript, and Jim Tobin and Julie White for testing subjects. Correspondence should be addressed to P. C. Gordon, Department of Psychology, Harvard University, William James Hall, 33 Kirkland St., Cambridge, MA 02138.

boring segments *can* provide useful contextual information by constraining a segment's possible identity through partial lexical access (e.g., Klatt, 1980; Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986). However, if we assume that the recognition of phonetic segments is sufficiently error-prone to require contextual support, then it does not seem prudent to look for that support solely in the recognition of other segments.

The present research examined whether relatively coarse levels of analysis, insufficient for segmental recognition, could provide listeners with a useful context for interpreting ambiguous temporal cues to segment identity. The advantages of an initial coarse-grained analysis as part of pattern recognition are straightforward. Analysis of the coarse aspects of a signal is relatively cheap computationally, and poses little threat of early misidentification of details that might preclude consideration of correct alternatives. In applying the idea of coarse-grained analysis to machine speech recognition, Zue and his colleagues (Lamel & Zue, 1984; Zue, 1985; Zue & Huttenlocher, 1983) have shown that a quite broad phonetic classification of the segments in an input sequence can dramatically narrow the search space of possible words that the sequence might contain.[1]

The current goal of examining *contextual* use of coarse-grained aspects of a signal has an additional advantage. If interpretation of an acoustic cue must be made in a context-dependent fasion, then accurate characterization of the context is essential. Otherwise, there is the potential for compounding errors; inaccurate identification of context leads to an incorrect basis for interpreting context-dependent cues. The present experiments tested the plausibility of coarse-grained analysis in human speech perception by determining whether coarse-grained representations of speech could induce rate-dependent processing of cues to segment identity.

EXPERIMENT 1

This experiment examined whether the low-frequency energy of a precursor phrase conveyed rate information that was integrated with the fine acoustic structure of a test word. Several studies had found that the overall articulation rate of precursive speech influenced the segment boundaries in a subsequent test word. This held true for phonological voicing in both syllable-initial stops (Diehl, Souther, & Convis, 1980; Miller et al., 1984; Summerfield, 1981) and medial stops (Miller & Grosjean, 1981; Port, 1979; Port & Dalby, 1982). In those earlier studies, the precursors consisted of acoustically detailed speech that presumably was easily recognized by the subjects. The present experiment examined the question of whether phonetic segments in the precursors would have to be recognizable for rate-dependent processing to be induced. This was done by using precursor phrases that were low-pass filtered to eliminate the acoustic information necessary to recognize phonetic segments. The phrases were filtered to reduce the precursor phrases by 50 dB at frequencies above 375 Hz, so that the recognizability of their constitutent segments was close to zero (French & Steinberg, 1947). The resulting precursors did, however, retain information that might convey speaking rate. In particular, the low-frequency energy roughly represented the amplitude envelope of voiced energy, which gives information about the timing of syllabic and/or vocalic nuclei (Lea, 1980). In addition, the low-frequency energy contained the F0 contour, which may reflect speaking rate (Cooper & Sorensen, 1981). The effectiveness of the filtered precursors in inducing rate-dependent processing was measured by their effect on the subjects' perception of phonological voicing in medial bilabial stop consonants (e.g., *rabid* vs. *rapid*). Unfiltered precursors were also used to provide a baseline for assessing the effects of filtering.

The medial voicing distinction has a number of acoustic correlates, including the presence of voicing during the closure, the duration of the preconsonantal vowel, and the duration of the consonantal closure (Lisker, 1957, 1978; Luce & Charles-Luce 1985). When other cues are ambiguous, an increase in the closure duration can change perception of the stop from voiced to voiceless (Lisker, 1978; Port, 1979).[2] In the poststress medial position, the closure duration is perceived relative to the duration of the preceding vowel. In this case, the vowel/consonant ratio may be the critical temporal cue for voicing (Port & Dalby, 1982). Since the ratio is unitless, it is potentially invariant with changes in speaking rate. However, the vowel/consonant ratio does not appear to be a consistent cue for voicing in syllable-final stops that vary in phrasal position (Luce & Charles-Luce, 1985) or for syllable-final voicing in sequences of two nonhomorganic stops (Repp & Williams, 1985). Even in situations where the consonant/vowel ratio is putatively an invariant cue, the medial voicing boundary has been found to shift as a function of the rate of articulation of a preceding phrase. This shift occurred even though the word in which the stop was embedded was held constant (Miller & Grosjean, 1981; Port, 1979; Port & Dalby, 1982).

Method

Subjects. Twelve Harvard University students served as subjects. They responded to a posted notice, and were paid $3 to participate in a single session that lasted approximately ½ h. They were tested in groups of 1 to 4.

Stimuli. The sentence "I'm trying to say rabid" (Port, 1979) served as a basis for the stimuli. It was spoken by a male native speaker of American English (the author) at three different rates (fast, medium, and slow). The utterances were spoken into a microphone (Shure SM 59), low-pass filtered at 4.7 kHz, and digitized at 10 kHz. Numerous recordings of the sentence were made at each speaking rate to find tokens that were roughly equal in pitch and loudness so that splicing would not sound unnatural.

The word "rabid" from the medium-rate utterance served as the basis for making the *rabid-rapid* stimulus series. The carrier from this sentence was not used in the experiment. The word *rabid* was excised from the carrier at a zero crossing close to its beginning, using a waveform editor. Not including the closure interval, the duration of the first syllable was 230 msec and the duration of the second syllable was 140 msec. The closure interval for the medial /b/ was replaced with silence of varying durations to create a *rabid-rapid* stimulus series. Informal listening indicated that this series failed to provide a very compelling voiceless medial consonant, even at farily long closure intervals. Therefore, two pitch periods (roughly 15 msec) were removed from the center of the preconsonantal vowel to make this cue to voicing more neutral. A new *rabid-rapid* series was made with closure intervals ranging from 30 to 120 msec in 5-msec steps. This series was pretested on 8 subjects, and it was determined that a range of intervals from 35 to 80 msec covered the transition from *rabid* to *rapid* for all of the subjects. None of the subjects characterized the short interval stimuli as containing flapped medial consonants (cf. Port, 1979), even when it was suggested that some of the stimuli might be heard as "ratted." Ten stimuli, with closure intervals ranging from 35 to 80 msec in 5-msec steps, were used in the experiment.

The carrier portion of the sentence ("I'm trying to say") was excised from the fast and slow utterances to provide fast and slow contexts for the voiced/voiceless series. The duration of carrier was 730 msec for the fast utterance and 1,920 msec for the slow utterance. Low-pass-filtered versions of these utterances were created digitally using the NERR algorithm (Kaiser & Reed, 1978). The filter had a roll-off of 50 dB and a pass-to-stop transition width of 250 Hz. The midpoint of the transition was 250 Hz. Each of the four carrier sentences (fast/slow × filtered/normal) was combined with each of the 10 stimuli in the *rabid-rapid* series to give 40 stimuli overall.

Procedure. The experiment was run in two blocks. The first block used filtered precursors, and the second used normal precursors. The precursor rate was varied within blocks. Each block consisted of 120 stimulus presentations (6 presentations of each stimulus × 20 stimuli) in a pseudorandom order.

The stimuli were output from the computer, low-pass-filtered at 4.7 kHz, and presented to subjects over Sennheiser HD-430 headphones at a comfortable listening level. There were 6 sec of silence between each stimulus presentation. Before the first block, the subjects were told that they were going to hear a series of stimuli and that they were to rate each stimulus on a scale of 1 to 6. The number 1 corresponded to an ideal *rapid* and 6 to an ideal *rabid*, and the numbers in between corresponded to the degree to which a stimulus sounded like one or the other. In addition, the subjects were told that they would hear something that sounded like muffled speech before each occurrence of the word to be judged. They were told to pay attention to this sound because the stimulus word would immediately follow it. Before the second block, the subjects were told that instead of the muffled speech, they would hear the sentence frame "I'm trying to say." The subjects wrote their ratings on a response sheet after each trial.

Results

The principal results are shown in Figure 1. The left panel shows the mean ratings at different closure durations with fast and slow precursors for the normal (unfiltered) precursors. The right panel shows the same results for the low-pass-filtered precursors. A three-way analysis of variance [precursor rate × precursor type (filtered/normal) × closure duration] was performed on these results. Consistent with the general downward trend of all of the curves, the analysis revealed a significant main effect of closure duration [$F(9,99) = 125.4, p < .001$]. It also showed a significant main effect of precursor rate: The slow precursor mean was 3.55, and the fast precursor mean was 3.41 [$F(1,11) = 6.3, p < .05$]. This effect is in the direction that would be expected if the closure duration was perceived relative to the rate of the precursor. A fast precursor would have caused the subjects to perceive the closure intervals as relatively long, and thus to perceive the stimuli as containing a medial /p/ as opposed to a medial /b/ (Port, 1979; Summerfield, 1981).

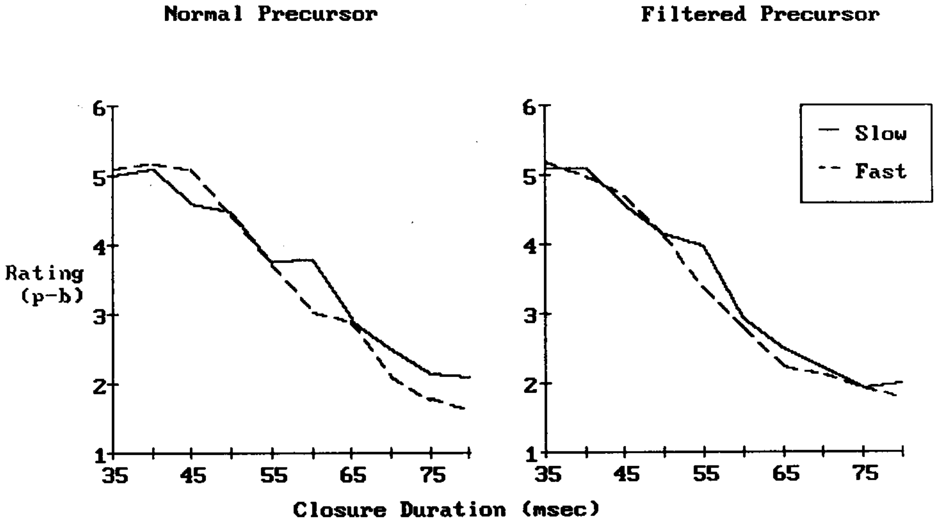

Figure 1. The results of Experiment 1. The subjects' judgments of phonological voicing are shown as a function of the intervocalic closure duration in the test word. The lower and upper endpoints of the 6-point rating scale correspond to *p* and *b*, respectively. The solid lines indicate ratings made for test words appended to slow precursive phrases; the dashed lines indicate ratings made for words appended to fast precursive phrases. The left panel shows the results for the normal (unmodified) precursors; the right panel shows the results for the severely filtered precursors.

In addition, there was no significant interaction between precursor rate and type of stimulus (normal vs. filtered). The mean slow/fast differences for normal and filtered precursors were 0.15 and 0.14, respectively [$F(1,11) < 1$].

There was a tendency for the filtered precursors (mean = 3.39) to yield lower ratings than the unfiltered precursors (mean = 3.57), but this trend fell short of conventional significance levels [$F(1,11) = 3.5$, $.10 > p > .05$]. There was a significant interaction of speaking rate and closure duration [$F(9,99) = 3.28$, $p < .01$]. A linear trends test showed that the effect of precursive speaking rate increased with closure duration [$t(11) = 2.64$, $p < .05$]. In addition, there was a significant interaction of type of stimulus and closure duration [$F(9,99) = 3.28$, $p < .01$]. The meaning of these interactions is not clear. There was no significant three-way interaction of type of stimulus, speaking rate, and closure duration [$F(9,99) = 1.73$, $p > .05$].

Discussion

These results support the idea that coarse aspects of the speech signal can provide listeners with an effective context for interpreting ambiguous acoustic cues. No difference was observed between the effectiveness of the severely filtered precursors and the acoustically detailed precursors in influencing subjects' perceptual judgments, thus indicating that rate-dependent processing need not rely on the identification of the phonetic segments in the rate-conveying context. Rather, it appears that simpler characteristics of the speech signal may suffice as a basis for interpreting the segmental significance of temporal cues.

Despite the constant timing information that was present in the test word, the precursive speech rate affected perception. This is consistent with the findings of previous studies (Miller et al., 1984; Miller & Grosjean, 1981; Port, 1979; Port & Dalby, 1982; Summerfield, 1981). Furthermore, this speech-rate effect occurred even though the precursive context was separated from the critical closure duration by the initial syllable of the test word (Miller & Grosjean, 1981; Port, 1979; Port & Dalby, 1982).

These findings indicate that intrinsic timing is not the sole source of rate effects in speech perception. Intrinsic timing models, as proposed by Fowler (1980) and Summerfield (1981), assert that the relevant timing information about a segment is completely specified by the articulatory gestures underlying the production of that segment. The effect of distant articulatory rate information on segment identification is inconsistent with intrinsic timing models (Miller, in press), even though this effect decreases with distance from the target segment (Port & Dalby, 1982; Summerfield, 1981).[3] The present finding indicates that the distant rate information derived from coarse-grained aspects of speech signals can affect the perceptual interpretation of local, detailed information about segment identity.

EXPERIMENT 2

Experiment 2 examined whether the amplitude envelope (AE) of speech could convey useful rate information that would help subjects to interpret durational cues to segment identity. The experiment employed precursor phrases, spoken at different rates, with AEs that had been filled with a constant-frequency sine wave. With regard to potential sources of information about speaking rate, these stimuli differed from the filtered precursors of the previous experiment, in that they eliminated information about F0 contour and conveyed the total amount of energy as a function of time, rather than simply conveying a portion of the voiced energy.

If these stimuli exerted a contextual influence on segmental identification, an important question can be posed about the kinds of information that can influence speech perception. The filtered precursors of Experiment 1 can rightly be considered as speech, albeit a very limited portion of what would usually be available. In contrast, the amplitude-varying sinusoids of the present experiment departed sharply from speech by concentrating all the acoustic energy of the original precusors at one frequency (the average fundamental frequency of the first syllable of the test word). Thus, rate-dependent processing would have occurred only if the subjects had integrated the nonspeech context with speech.

There has been at least one failure to find such an integration. Summerfield (1981) had subjects make phonological voicing judgments on the initial consonant of a syllable which followed the first four bars of the Christmas carol "Good King Wenceslaus." He found that the tempo of the tune had no effect on the voice-onset time boundary.

Summerfield (1981) was attempting to test the validity of "extrinsic" accounts of timing in speech perception (see Fowler, 1980). According to Summerfield, an extrinsic timing account would include a perceptual clock, adjusted for speaking rate, that was responsible for measuring segmentally significant durations. By Summerfield's reasoning, the tempo of the precursive tune should have affected the extrinsic timing mechanism, and hence the voicing judgment. His failure to find such an influence, as well as the Diehl et al. (1980) finding that rate effects are substantially reduced when the precursor and test syllable appear to have been produced by different speakers, was part of Summerfield's basis for rejecting extrinsic timing as unimportant in speech perception.

The present experiments tested the predictions of a theory that includes a kind of extrinsic timing.[4] Important timing information is thought to be extracted from the acoustic signal without the identification of phonetic segments. Some timing information is thus extrinsic to the segments. The current experiment focused on the pattern of total energy versus time in a speech signal (its AE). Although the auditory system performs a frequency analysis early in processing, the magnitudes of the different frequency components can be combined at higher levels to determine the overall loudness of a complex sound

(Moore, 1982). This combination of frequency components in speech recognition might produce loudness estimates that are statistically more reliable than the estimates for smaller component frequency bands. These estimates would remain relatively robust in the face of band-limited hearing loss or environmental distortion that might interfere with detailed phonetic analysis. The usefulness of performing the loudness estimates presupposes, of course, that important timing information can be extracted from the relatively simple patterns in the AE.

There are several reasons for believing that the AE can contribute to speech recognition. Work on automatic speech processing has shown that measures of total energy, as well as fairly wide-band energy, can be useful for puposes of syllabification and segmentation (Mermelstein, 1975; Reddy, 1966). In psychological experiments, it has been found that auditory information about the AE substantially improves concurrent speech-reading performance (Grant, Ardell, Kuhl, & Sparks, 1985). The usefulness of the AE in conveying temporal information has been demonstrated by two findings. First, Howell (1983) found that delaying auditory information about the AE disrupted articulation in a delayed auditory feedback paradigm. These disfluencies were attributed to a disruptive effect of misaligned information about the timing of speech output. Second, Howell (1984) found that nonspeech stimuli, with AEs similar to those of speech sounds, exhibited perceptual isochrony in a manner similar to speech (Marcus, 1981; Morton, Marcus, & Frankish, 1976).[5] Together, these findings demonstrate that the AE conveys timing information and can be useful in speech recognition. This suggests that the AE might be useful in interpreting durational cues to segment identity.

Method

Subjects, Stimuli, and Procedure. A new group of 12 individuals, from the same population as the previous experiment, served as subjects. The stimuli were constructed from the speech recordings used in the previous experiment. Versions of the fast and slow precursor phrases were constructed that preserved only information about the AE of the source precursor. These reconstructed precursor phrases consisted of an amplitude-varying sine wave set at the frequency of the average F0 in the first syllable of the *rabid* stimulus (134 Hz). The amplitude was varied so that each period of the sine wave matched the digital root mean square (DRMS) energy of the corresponding time slice in the original precursor.[6] Tokens of the *rabid-rapid* series were appended to the end of the fast and slow AE-only precursors. The experiment was run in two blocks, the first using AE-only precursors and the second using normal precursors. The composition of the blocks was the same as in the previous experiment. The instructions were also the same, with one exception. Before the first block, the subjects were told that they would hear "some tones" before each test word. (In Experiment 1, the subjects had been told that they would hear "muffled speech.")

Results

Figure 2 shows the subjects' ratings for the *rabid-rapid* series following various precursor phrases. A three-way analysis of variance [precursor rate × precursor type (AE-only vs. normal) × closure duration] was performed on these results. A significant main effect of closure duration was found [$F(9,99) = 112.4, p < .001$]; stimuli with longer closure durations were judged more *rapid*-like. Precursor rate had a significant effect on the subjects' judgments; the slow precursor mean was 3.66 and the fast precursor mean was 3.42 [$F(1,11) = 7.47, p < .025$]. The direction of the effect is the same as in the previous experiment, and is consistent with the idea that the closure interval is interpreted relative to the rate of the precursor. However, as in Experiment 1, there was no significant interaction between rate of articulation and precursor type (AE-only vs. normal). The mean slow/fast difference was .17 for AE-only precursors and .30 for normal precursors [$F(1,11) = 1.8, p > .10$].

There was a significant main effect of precursor type; the AE-only mean was 3.39 and the normal mean was

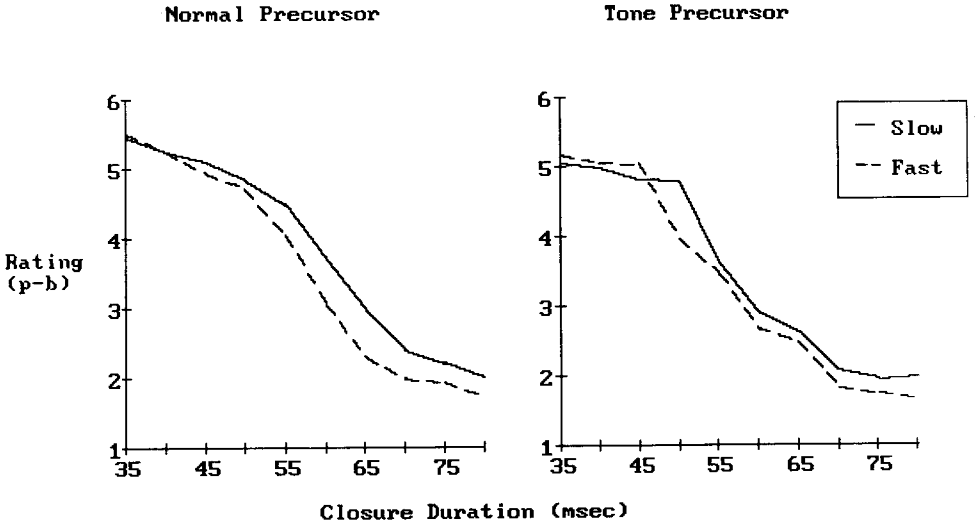

Figure 2. The results of Experiment 2. The subjects' ratings are displayed as in Figure 1. The left panel shows the results for the normal precursors; the right panel shows the results for the AE-only precursors.

3.69 [$F(1,11) = 9.06, p < .01$].[7] There was also a significant interaction of type of stimulus and closure duration [$F(9,99) = 2.77, p < .01$] and of speaking rate and closure duration [$F(9,99) = 2.02, p < .05$]. The three-way interaction of type of stimulus, closure duration, and speaking rate was also significant [$F(9,99) = 2.06, p < .05$]. The location of these interactions is not apparent from inspection of the data. Consequently, their meaning is not clear.

Discussion

The results of Experiment 2 indicate that the AE of speech provided a sufficient basis for subjects to interpret rate-dependent durational cues to segment identity, at least for the present stimuli. This reinforces the conclusion of Experiment 1 that coarse aspects of the speech signal, which lack the detail necessary for phonetic identification, may provide subjects with a useful context for interpreting ambiguous acoustic cues to segment identity. This finding also indicates that information about change in F0 was not necessary for subjects to extract rate information from a coarse representation of the speech signal.

The results also bear on Summerfield's (1981) theoretical formulation of rate effects. In his words, "This apparently radical stance [intrinsic timing] predicts that if Experiment 1 [demonstrating precursive effects] were to be repeated with a precursor devised so that temporal information were dissociated from information specifying articulatory maneuvers, then no effect on the perception of voicing contrasts should result" (p. 1088). This stance is apparently too radical. The AE-only precursors reflect some of the consequences of articulatory maneuvers. The rise and fall of amplitude roughly reflects the amount of glottal excitation, the amount of fricative excitation, and the degree of constriction in the vocal tract. However, the AE-only precursors certainly lack the information necessary to specify the articulatory gestures underlying the original precursive speech. It is apparent that some aspects of the timing of speech are extrinsic in the sense that they can be extracted and utilized by listeners in situations in which the specific underlying articulatory movements cannot be identified.

The present experiment demonstrated that, under some circumstances, perception of a phonetic segment was influenced by timing information that came from a seemingly different source. Furthermore, the size of the rate effects did not differ significantly between the AE-only precursors and the unmodified precursors. This finding contrasts with the results of Summerfield (1981) and Diehl et al. (1980), which indicated that precursive rate effects were reduced or eliminated when there was an apparent discrepancy between the source of the rate-bearing speech and the test syllable. It seems likely that many factors play a part in determining which sounds are integrated (or segregated) in perceiving speech. These include influences such as semantic continuity (Treisman, 1964), binaural localization of sources (Cherry, 1953), amplitude (Egan, Carterette, & Thwing, 1954), and continuity of acoustic energy (Bregman, 1978). It is likely that some combination of acoustic continuity and high-level attribution of source accounts for the differing results of Experiment 2 and those of Summerfield and Diehl et al.

Both the Summerfield (1981) and Diehl et al. (1980) studies used precursors with more identifiable sources than those of Experiment 2. Summerfield's precursors were probably correctly identified by his subjects as synthetically produced buzz-like versions of a tune. The Diehl et al. precursors were probably identified by their subjects as coming from a speaker other than the one who produced the test syllable. In contrast, it is less clear how subjects phenomenally characterized the precursors of Experiment 2. The amplitude-varying sinusoids did not sound like speech, at least to the author. In some respects, these stimuli are similar to ones used by Remez, Rubin, Pisoni, and Carrell (1981) in their studies of sinusoidal speech. They asked subjects to identify a variety of stimuli, some of which consisted of pure tones varying in frequency to match the variation in the center frequency of a formant. Their stimuli varied in frequency and the present stimuli varied in amplitude, but their findings may still be relevant. Very few of their subjects spontaneously identified their single-tone stimuli as speech, and, across subjects, these stimuli were not consistently labeled. On the basis of the Remez et al. findings, it seems likely that the stimuli in Experiment 2 did not suggest a source in a compelling fashion. This differs from the precursive stimuli of Summerfield and Diehl et al., which suggested a source distinct from the speech to be judged. If this was the case, then perhaps speech cues are more readily integrated with information from an unknown source than from a source with a conflicting identity.

There are acoustic reasons why the results of Summerfield (1981) might differ from the present ones. Summerfield's tune precursors were carried on a fundamental that was not continuous with the fundamental of the test syllable, and Diehl et al. (1980) showed that rate effects were reduced when precursors and test syllables differed in fundamental frequency. The tone precursors maintained acoustic compatibility with the F0 of the test word, without having the incompatible formant spacing of some of Diehl et al.'s stimuli. The finding of no difference in rate effects induced by the tone precursors and the plain precursors is consistent with the importance that has been attributed to F0 and its harmonics in isolating one speaker's voice from many (Darwin & Bethell-Fox, 1977), and in determining what acoustic energy contributes to phonetic perception (Broadbent & Ladefoged, 1957; Darwin & Gardner, 1986).

EXPERIMENT 3

As argued above, it seems likely that a variety of factors influence integration of information in speech recognition. This experiment attempted to establish some of the limits on the rate-dependent processing of the present test word by examining whether the form of information about the AE influenced its integration with a following test word. This was done by filling the AE of the precursor phrase with noise (Horii, House, & Hughes, 1971; Salasoo & Pisoni, 1985). The resulting sounds contained roughly the same information about energy variation as the amplitude-varying sinusoids of Experiment 2. However, they introduced a more marked acoustic discontinuity between the context stimulus and the test word. The periodic character of the sine waves was similar to the quasi-periodic character of the voiced portions of the test words. In addition, the two sounds had roughly the same pitch. These similarities did not exist between the noise-filled precursors and the test words. The resulting discontinuity may therefore have reduced the precursive rate effect, if acoustic continuity was an important factor in the perceptual integration of the precursor and the test word.

Method

Subjects, Stimuli, and Procedure. Twelve new subjects participated in the experiment. The stimuli were constructed from the speech recordings from Experiment 1. The fast and slow precursor phrases were filled with noise using a procedure developed by Horii et al. (1971), which involved changing the sign of a randomly selected 50% of the digitized speech samples. The resulting sounds preserved the temporal variation of DRMS energy in the original speech, but destroyed its spectral structure, resulting in something close to white noise. Tokens of the *rabid-rapid* series were appended to the end of the fast and slow noise precursors. The procedure was the same as in the previous two experiments except that subjects were told that they would hear some "radio static" before the test words in the first block.

Results

The results, shown in Figure 3, were analyzed as in the previous studies. As expected, a significant effect of closure duration on voicing judgments was obtained [$F(9,99) = 150.8, p < .001$]. A significant effect of precursor rate was not obtained [$F(1,11) = 2.54, p > .10$]. However, the interaction of rate and closure was significant [$F(9,99) = 4.79, p < .01$], as was the three-way interaction of precursor type (noise vs. normal), rate, and closure [$F(9,99) = 2.42, p < .025$]. Contrasts were evaluated at the middle four closure durations. A significant rate effect was found for the normal precursors, with a mean difference of .64 [$t(11) = 6.86, p < .001$], but not for the noise precursors, with a mean difference of .03 [$t(11) = 0.5, p > .10$]. A significant main effect of precursor type was obtained, with the noise precursors having a mean of 3.39 and the normal precursors a mean of 3.75 [$F(1,11) = 10.78$]. (See Note 7.) The interaction of precursor type and closure duration was also significant [$F(9,99) = 4.39, p < .01$].

Discussion

The results of this experiment illustrate one limit on the use of precursive rate information in influencing the processing of medial voicing as cued by closure duration. When the precursor envelopes were filled with noise, their rate had no effect on voicing judgments. The rate of the normal precursors continued to affect voicing judgments at the middle closure durations, where we would expect voicing identification to be most sensitive to contextual influences. Comparison of the results of this experiment with those of Experiment 2 indicates that the form of precursive information is crucial in obtaining rate effects with coarse-grained representations of speech. The relations of the sinusoidal and noise stimuli to the original speech were identical; both preserved the AE. However, there was much less continuity in acoustic energy between the noise precursors and the test words than between the sinusoidal precursors and the test words. The different pattern of results with these stimuli suggests that determination of the relevance of prior rate information in pho-

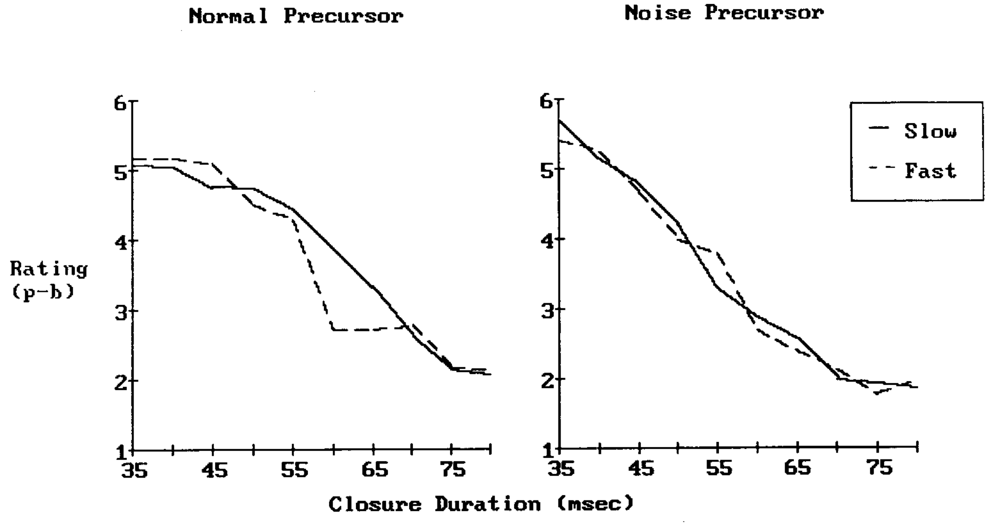

Figure 3. The results of Experiment 3. The subjects' ratings are displayed as in Figure 1. The left panel shows the results for the normal precursors; the right panel shows the results for the noise-filled precursors.

netic perception takes place, in part, at the level of auditory parsing of sounds (Bregman, 1978; Darwin, 1984).

GENERAL DISCUSSION

The research reported here demonstrates that the speech signal contains information about rate of articulation, independent of the identity of phonetic segments, and that listeners can exploit this information in rate-dependent processing. The findings are consistent with a general strategy of speech perception, in which the signal is processed at multiple levels of detail (Zue, 1985), with information at coarse levels providing a useful context for interpreting information at finer levels. As argued above, this strategy allows listeners to obtain information in a computationally cheap fashion, reduces the chance of early misidentification of detail, and bases contextual interpretation on aspects of the signal that may be easily, and reliably, identifiable. Certainly, it seems likely that the coarse representations of speech used here offer a more efficient basis for context-dependent interpretation than more derived measures of context, such as the identity of neighboring phonetic segments or the likelihood that a given word is being spoken.

There is a possibility that the rate effects observed in Experiments 1 and 2 were mediated by partial segmental recognition. The low-frequency stimuli of Experiment 1 preserved information about F0, which can serve as a cue to the identity of stop-consonant voicing (Haggard, Ambler, & Callow, 1970) and vowel height (Peterson & Barney, 1952). The amplitude-varying sine waves of Experiment 2 preserved some information, such as amplitude-onset characteristics of segments, which can also provide information about segment identity (Mack & Blumstein, 1983). Thus, the observed rate effects could have been mediated by a broad phonetic classification of the precursors (see Zue, 1985). Testing this possibility would require more refined methods of assessing the psychological encoding of coarse-grained speech information.

The results of Experiment 1 and 2 raise questions about the patterns of coarse-grained speech information that convey contextual information in general, and rate information in particular. Establishing the characteristics of these patterns requires acoustic analysis of a far greater variety of speech than was examined in the present study. However, the present study does offer some clues. The precursors of Experiments 1 and 2 both conveyed rate information to the subjects, even though they represented different aspects of the original speech precursors. This suggests that the coarse-grained patterns of contextual information are likely to be redundant. In this way, coarse-grained patterns would be similar to redundant, fine-grained acoustic cues to phonetic segments (Lisker, 1978). The ultimate value of the multiscale analysis advocated here depends on whether coarse speech patterns are more easily processed than fine-grained aspects of speech. Unless these patterns turn out to be reliably extractable, and relatively immune to the effects of casual articulation, environmental interference, sensorineural loss, and inattention, there would be little point to an initial coarse-grained analysis.

The value of coarse-grained analysis will also be determined by the kinds of phonetically significant information that can be conveyed by coarse-grained aspects of the signal. The present experiments showed that the overall speaking rate of a precursor sentence can be conveyed by the coarse-grained aspects of a signal. It is unclear whether local articulation rate (Klatt, 1976) can be so conveyed. In addition, it is possible that the segmental composition of an utterance influences the extent to which its coarse-grained attributes (e.g., the AE) convey timing information. Segments with quite similar coarse-grained attributes (e.g., vowels) can differ in intrinsic duration (Peterson & Lehiste, 1960), setting a potential limit on the usefulness of coarse-grained information. In such cases, a coarse-grained analysis might not be able to extract rate information, especially if the analysis is constrained to a small temporal window. However, an analysis that takes into account a relatively large stretch of speech may provide enough rate information to be useful as a context, independent of the intrinsic durations of its underlying segments.

An untested assumption of the procedure used here is that even under normal conditions, speech perception may involve extracting rate information by analyzing the signal at levels such as low-frequency energy (Experiment 1) or the AE (Experiment 2). The experiments did show that these levels of analysis conveyed rate information, and that speech perception processes could operate at these levels when more detailed information was absent. It has generally been found that speech recognition processes exploit most aspects of the speech signal that reflect the underlying articulatory gestures. Because the coarse-grained attributes of the signal reflect coarse attributes of articulation, it seems likely that listeners would exploit them and that coarse levels of analysis are operative with normal speech.

REFERENCES

Blumstein, S., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America*, **66**, 1001-1017.

Bregman, A. S. (1978). The formation of auditory streams. In J. Requin (Ed.), *Attention and performance VII* (pp. 63-75). Hillsdale, NJ: Erlbaum.

Broadbent, D. E., & Ladefoged, P. (1957). On the fusion of sounds reaching different sense organs. *Journal of the Acoustical Society of America*, **29**, 708-710.

Cherry, C. (1953). Some experiments on the reception of speech with one and two ears. *Journal of the Acoustical Society of America*, **25**, 975-979.

Cole, R. A., & Scott, B. (1974). Toward a theory of speech perception. *Psychological Review*, **81**, 348-374.

Cooper, A. M., Whalen, D. H., & Fowler, C. A. (1986). P-centers are unaffected by phonetic categorization. *Perception & Psychophysics*, **39**, 187-196.

Cooper, W. E., & Sorensen, J. M. (1981). *Fundamental frequency in sentence production*. Berlin: Springer-Verlag.

Darwin, C. J. (1984). Auditory processing and speech perception. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance*, *X: Control of language processes* (pp. 197-210). Hillsdale, NJ: Erlbaum.

Darwin, C. J., & Bethell-Fox, C. E. (1977). Pitch continuity and speech source attribution. *Journal of Experimental Psychology: Human Perception & Performance*, **3**, 665-672.

Darwin, C. J., & Gardner, R. B. (1986). Mistuning a harmonic of a vowel: Grouping and phrase effects on vowel quality. *Journal of the Acoustical Society of America*, **79**, 838-845.

Diehl, R. L., Souther, A. F., & Convis, C. L. (1980). Conditions on rate normalization in speech perception. *Perception & Psychophysics*, **27**, 435-443.

Egan, J., Carterette, E., & Thwing, E. (1954). Some factors affecting multichannel listening. *Journal of the Acoustical Society of America*, **26**, 774-782.

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

French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America*, **19**, 90-119.

Grant, K. W., Ardell, L. H., Kuhl, P. K., & Sparks, D. W. (1985). The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. *Journal of the Acoustical Society of America*, **77**, 671-677.

Green, D. M. (1976). *An introduction to hearing*. Hillsdale, NJ: Erlbaum.

Haggard, M. P., Ambler, S., & Callow, M. (1970). Pitch as a voicing cue. *Journal of the Acoustical Society of America*, **47**, 613-617.

Horii, Y., House, A. S., & Hughes, G. W. (1971). A masking noise with speech-envelope characteristics for studying intelligibility. *Journal of the Acoustical Society of America*, **49**, 1849-1856.

Howell, P. (1983). The effect of delaying auditory feedback of selected components of the speech signal. *Perception & Psychophysics*, **34**, 387-396.

Howell, P. (1984). An acoustic determinant of perceived and produced anisochrony. In M. P. R. Van den Broeck & A. Cohen (Eds.), *Proceedings of the 10th International Congress of Phonetic Sciences* (pp. 429-433). Dordrecht, Holland: Foris.

Kaiser, J. F., & Reed, W. A. (1978). Bandpass (bandstop) digital filter design routine. *Review of Scientific Instrumentation*, **48**, 1103-1106.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, **59**, 1208-1221.

Klatt, D. H. (1980). Speech perception: A model of acoustic-phonetic analysis and lexical access. In R. A. Cole (Ed.), *Perception and production of fluent speech* (pp. 243-288). Hillsdale, NJ: Erlbaum.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, **29**, 98-104.

Lamel, L. F., & Zue, V. W. (1984, March). Properties of consonant sequences within words and across word boundaries. *Proceedings of the ICASSP*, San Diego, CA.

Lea, W. A. (1980). Prosodic aids to speech recognition. In W. A. Lea (Ed.), *Trends in speech recognition* (pp. 166-205). Englewood Cliffs, NJ: Prentice-Hall.

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

Liberman, A. M., Delattre, P. C., Gerstman, L. J., & Cooper, F. S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology*, **52**, 127-137.

Lisker, L. (1957). Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, **33**, 42-49.

Lisker, L. (1978). *Rapid vs. rabid: A catalogue of acoustic features that may cue the distinction* (Status Report on Speech Research SR-54). New Haven, CT: Haskins Laboratories.

Lisker, L., & Abramson, A. S. (1964). A cross language study of voicing in initial stops: Acoustical measurements. *Word*, **20**, 384-422.

Luce, P. A., & Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *Journal of the Acoustical Society of America*, **78**, 1949-1957.

Mack, M., & Blumstein, S. E. (1983). Further evidence of acoustic invariance in speech production: The stop-glide contrast. *Journal of the Acoustical Society of America*, **73**, 1739-1750.

Marcus, S. (1981). Acoustic determinants of perceptual center (P-center) location. *Perception & Psychophysics*, **30**, 247-256.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1-86.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *Journal of the Acoustical Society of America*, **58**, 880-883.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39-74). Hillsdale, NJ: Erlbaum.

Miller, J. L. (in press). Rate-dependent processing in speech perception. In A. Ellis (Ed.), *Progress in the psychology of language* (Vol. 3). Hillsdale, NJ: Erlbaum.

Miller, J. L., Green, K., & Shermer, T. M. (1984). A distinction between the effects of sentential speaking rate and semantic congruity on word identification. *Perception & Psychophysics*, **36**, 329-337.

Miller, J. L., & Grosjean, F. (1981). How the components of speaking rate influence perception of phonetic segments. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 208-215.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, **25**, 457-465.

Moore, B. C. J. (1982). *An introduction to the psychology of hearing*. London: Academic Press.

Morton, J., Marcus, S. M., & Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Review*, **83**, 405-408.

Nittrouer, S., & Studdert-Kennedy, M. (1986). The stop-glide distinction: Acoustic analysis and perceptual effect of variation in syllable amplitude envelope for initial /b/ and /w/. *Journal of the Acoustical Society of America*, **80**, 1026-1029.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of vowels. *Journal of the Acoustical Society of America*, **20**, 528-535.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllabic nuclei in English. *Journal of the Acoustical Society of America*, **32**, 693-703.

Port, R. F. (1979). The influence of tempo on stop closure duration as a cue for voicing and place. *Journal of Phonetics*, **7**, 45-56.

Port, R. F., & Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Perception & Psychophysics*, **32**, 141-152.

Reddy, D. R. (1966). Segmentation of speech sounds. *Journal of the Acoustical Society of America*, **40**, 307-312.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, **212**, 947-950.

Repp, B. H., & Williams, D. R. (1985). Influence of following context on perception of the voiced-voiceless distinction in syllable-final stop consonants. *Journal of the Acoustical Society of America*, **78**, 445-457.

Salasoo, A., & Pisoni, D. B.(1985). Interaction of knowledge sources in spoken word identification. *Journal of Memory & Language*, **24**, 210-231.

Shinn, P., & Blumstein, S. E. (1984). On the role of the amplitude envelope for the perception of [b] and [w]. *Journal of the Acoustical Society of America*, **75**, 1243-1252.

Shinn, P., Blumstein, S. E., & Jongman, A. (1985). Limitations of context conditioned effects in the perception of [b] and [w]. *Perception & Psychophysics*, **38**, 397-407.

Stevens, K. N. (1980). Acoustic correlates of some phonetic categories. *Journal of the Acoustical Society of America*, **68**, 836-842.

Summerfield, Q. (1981). On articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 1074-1095.

Treisman, A. (1964). Verbal cues, language, and meaning in attention. *American Journal of Psychology*, **77**, 206-214.

Wardrip-Fruin, C. (1985). The effect of signal degradation on the status of cues to voicing in utterance-final stop consonants. *Journal of the Acoustical Society of America*, 77, 1907-1912.

Zue, V. W. (1985). The use of speech knowledge in automatic speech recognition. *Proceedings of the IEEE*, 73, 1602-1615.

Zue, V. W., & Huttenlocher, D. P. (1983, December). *Computer recognition of isolated words from large vocabularies: Lexical access using partial phonetic information.* Paper presented at the International Conference on Advanced Automation, Taiwan.

NOTES

1. The usefulness of a strategy for machine recognition does not, of course, indicate that it is employed by humans.

2. Shinn and Blumstein (1984; Shinn, Blumstein, & Jongman, 1985) argued that the importance of context-dependent processing in human speech perception may have been overestimated because of heavy reliance on stylized synthetic speech sounds that lack putatively invariant acoustic correlates present in naturally produced phonetic segments. Shinn et al. found that previously obtained rate dependencies on the stop-glide distinction (Miller & Liberman, 1979) were greatly reduced when more natural amplitude-onset characteristics were present in the stimuli. However, Nittrouer and Studdert-Kennedy (1986) showed that amplitude onset was less perceptually salient than other, presumably context-dependent, cues to the stop-glide distinction. In addition, Wardrip-Fruin (1985) showed that a context-dependent temporal cue to syllable final voicing—vowel duration—was given greater perceptual weight when stimuli were presented in noise versus silence. This suggests that some context-dependent cues may be more robust under difficult listening conditions than some of the putatively invariant cues, thus bolstering the importance of context-dependent processing of temporal cues.

3. Although Summerfield (1981) acknowledged that the effect of distant rate information was inconsistent with intrinsic timing as a sole account in speech recognition, he argued that the relatively small size of the effects indicated that *extrinsic* timing was unimportant. However, assessment of the relative contributions of distant and near timing information involves stimuli in which there is an artificial conflict between distant and near information. That distant rate information should have any effect at all under those circumstances attests to listeners' tendency to use that information. It seems likely that listeners would use distant rate information even more when it did not conflict with local rate information.

4. I am reluctant to claim that the approach I am proposing amounts to an extrinsic timing theory, since properties have been attributed to extrinsic timing models that I do not advocate. For example, Fowler (1980, p. 113), in a critical review, defines theories of extrinsic timing in speech production as ones that "exclude timing from representation in the talker's articulatory plan for his utterance." Given that properties such as "intrinsic durational characteristics" of syllabic nuclei (Peterson & Lehiste, 1960) or "voice-onset time" (Lisker & Abramson, 1964) have long been discussed as distinguishing characteristics of phonetic segments, a theory of extrinsic timing as defined by Fowler is clearly untenable. Rather, it seems that speech production and perception must be characterized in terms of intrinsic and extrinsic timing.

5. A. M. Cooper, Whalen, and Fowler (1986) reported several replications of a finding by Marcus (1981) that the "P-centers" of syllables are not influenced by phonetic categorization A. M. Cooper et al. argued that their findings, as well as Marcus's findings, disprove Howell's (1984) claim that variation in the AE can account for variation in the location of P-centers (in the perceptual sense), independent of the perceiver's knowledge of articulatory dynamics. However, all of the stimulus manipulations that A. M. Cooper et al. and Marcus used to shift P-center location involved changes to the AE, which is essentially Howell's claim.

6. Equating the energy of the sine periods with the time-slices of speech does not necessarily equate their loudness. The sine waves are at a frequency (134 Hz) with lower loudness sensitivity than much of the energy in speech (i.e., in the range of 1000-5000 Hz). However, because energy is present at only one frequency in the sine wave, there is no within-stimulus masking (Green, 1976). In any case, the pattern of loudness over time is roughly the same for the sine waves and the speech. Furthermore, it is consistent with the general idea of coarse-grained analysis in speech recognition that a small divergence between the properties of a coarse-grained representation of speech and those of real speech should not have great perceptual implications.

7. The significant difference between stimuli following the noise-filled precursors and those following the normal precursors was in the same direction as the significant difference between the sine-filled precursors and the normal precursors (Experiment 2), and as the not-quite-significant difference ($p < .10$) between the filtered precursors and the normal precursors (Experiment 1). In all three cases, the modified precursors yielded lower ratings (more P-like) than the normal precursors. One explanation for this difference is that the modified precursors were perceived as having been spoken more slowly than the normal precursors, and thus that the two kinds of stimuli did not convey exactly the same information about speaking rate. This explanation, however, is far from certain. Another possible explanation stems from the fact that stimulus type was confounded with order of presentation. The stimuli with modified precursors were always presented first so as not to bias the subjects toward regarding them as modified versions of speech that they had heard prevously. It is therefore possible that this main effect was due to some change in judgments over time. More generally, it seems hard to interpret *absolute* differences between the modified and unmodified precursors, because the stimuli consisting of both modified and real speech may have elicited overall response strategies that were different from those elicited by the stimuli consisting entirely of real speech. It seems safer to focus on the ability, or lack thereof, of the modified precursor to convey the *relative* speaking rate of the precursors from which they were derived.

(Manuscript received March 5, 1987;
revision accepted for publication July 21, 1987.)