UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

DOCUMENTATION PAGE

AD-A243 979

DEC 24 1991

1b RESTRICTIVE MARKINGS

3 DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release; distribution Unlimited.

4 PERFORMING ORGANIZATION REPORT NUMBER(S)

5 MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-TR- 91 0293

6a NAME OF PERFORMING ORGANIZATION	6b OFFICE SYMBOL (If applicable)	7a NAME OF MONITORING ORGANIZATION
Harvard University		Air Force Office of Scientific Research/NL

6c. ADDRESS (City, State, and ZIP Code)	7b ADDRESS (City, State, and ZIP Code)
Department of Psychology 33 Kirkland Street Cambridge, Mass. 02138	Building 410 Bolling AFB, DC 20332-6448

8a. NAME OF FUNDING / SPONSORING ORGANIZATION	8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
AFSOR	NL	AFOSR #89-0461

8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS			
	PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
Building 410 Bolling AFB, DC 20332-6448	61102f	2313	A4	

11 TITLE (Include Security Classification)
Perception and the Temporal Properties of Speech

12. PERSONAL AUTHOR(S)
Peter C. Gordon

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
~~Final~~ Technical Report	FROM 7/89 TO 7/91	November 6, 1991	

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	speech perception, prosody, context effects, phonetic segments, fricatives, stress, Attention

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

Four experiments addressing the role of attention in phonetic perception are reported. The first experiment shows that the relative importance of two cues to the voicing distinction changes when subjects must perform an arithmetic distractor task at the same time as identifying a speech stimulus. The voice onset time cue loses phonetic significance when subjects are distracted, while the F0 onset frequency cue does not. The second experiment shows a similar pattern for two cues to the distinction between the vowels /i/ (as in "beat") and /I/ (as in "bit"). Together these experiments indicate that careful attention to speech perception is necessary for strong acoustic cues to achieve their full phonetic impact, while weaker acoustic cues achieve their full phonetic impact without close attention. Experiment 3 shows that this pattern is obtained when the distractor task places little demand on verbal short-term memory. Experiment 4 provides a large data set for testing formal models of the role of attention in speech perception. Attention is shown to influence the signal-to-noise ratio in phonetic encoding. This principle is instantiated in a network model in which the role of attention is to reduce noise in the phonetic encoding of acoustic cues. Implications of this work for understanding speech perception and general theories of the role of attention in perception are discussed.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b TELEPHONE (Include Area Code)	22c OFFICE SYMBOL
Dr. Alfred Fregly	(202) 767-5021	NL

DD FORM 1473, 84 MAR

All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

91 1223 134

20 NOV 1991

Statement of Work

Work on this project will extend previous work on the context-dependent nature of temporal cues to the identity of phonetic segments, and on the role of coarse-grained aspects of the speech signal in facilitating segment recognition. These extensions will address the following questions: Do adjacent segments exhibit mutual dependencies resulting in perceptual ambiguity that can be overcome by contextual information present in coarse-signal characteristics? Can coarse-grained aspects of the speech signal, lacking sufficient information for segment identification, convey speaking rate independently of variation in the inherent durations of the underlying segments? Do coarse-grained aspects of precursive speech contribute contextual information that is used early in the timecourse of segment recognition? Can coarse-grained aspects of the speech signal direct attention to the location of upcoming stressed syllables?

Word on the project will directly study the nature of coarse-grained aspects of the signal and their relation to processing the suprasegmental temporal aspects of speech. New techniques will be developed for creating coarse-grained representations of speech that eliminate information about segment identity but preserve prosodically-relevant aspects of the speech signal. These techniques will permit control over degree of resolution in the short-time spectrum of speech. Perceptual studies, involving direct judgments on stimuli with varying amounts of spectral resolution, will be performed to determine what the amount of spectral detail that is necessary for perceiving important temporal components of prosody.

As part of the project a computer simulation will be developed that will test the computational adequacy of the processes that are hypothesized to underlie human perception of the temporal properties of speech. This model will address three related issues: the segmentation of speech into syllables, the use of temporal relations between syllables to generate expectancies about the temporal properties of upcoming syllables, and the contextual modulation of feature analyzers for processing temporal cues to segment identity.

Status of Research

Publications

Gordon, P.C., Schaeffer, C.P., & Kennison, S.M. (in press). Disambiguation of segmental dependencies by extended phonetic context. **Language and Speech.**

Yaniv, I., Meyer, D.E., Gordon, P.C., Huff, C.A., & Sevald, C.A. (1990). Vowel similarity and syllable structure in motor programming of speech. **Journal of Memory and Language.**

Eberhardt, J.L., & Gordon, P.C. (1990). Effects of attention on the perception of phonologically natural forms. **Journal of the Acoustical Society of America**, 88, Suppl. 1.

Gow, D.W., & Gordon, P.C. (1990). Perceptual and acoustic measures of stress shift. **Journal of the Acoustical Society of America**, 88, Suppl. 1.

Eberhardt, J.L., & Gordon, P.C. (1989). The effects of attention on the phonetic integration of acoustic information. **Journal of the Acoustical Society of America**, 86, Suppl. 1.

Gow, D.W., & Gordon, P.C. (1989). Two paradigms for examining the role of phonological stress in sentence processing. **Journal of the Acoustical Society of America**, 86, Suppl. 1.

Manuscripts under review

Gordon, P.C., Eberhardt, J.L., & Rueckl, J.G. Attentional modulation of the phonetic significance of acoustic cues.

Gow, D.W., & Gordon, P.C. Syllable stress in the processing and representation of spoken sentences.

Attentional Modulation of the Phonetic Significance of Acoustic Cues

Peter C. Gordon Jennifer L. Eberhardt Jay G. Rueckl

Harvard University

Running Head: Paying Attention to Phonetic Perception

Send Correspondence to:

Peter C. Gordon
Department of Psychology
Harvard University
33 Kirkland St.
Cambridge, MA 02138

pcg@wjh12.harvard.edu

Abstract

Four experiments addressing the role of attention in phonetic perception are reported. The first experiment shows that the relative importance of two cues to the voicing distinction changes when subjects must perform an arithmetic distractor task at the same time as identifying a speech stimulus. The voice onset time cue loses phonetic significance when subjects are distracted, while the F0 onset frequency cue does not. The second experiment shows a similar pattern for two cues to the distinction between the vowels /i/ (as in "beat") and /I/ (as in "bit"). When distracted, listeners attach less phonetic significance to formant patterns while there is a net increase in the phonetic significance attached to vowel duration. Together these experiments indicate that careful attention to speech perception is necessary for strong acoustic cues (voice-onset time and formant patterns) to achieve their full phonetic impact, while weaker acoustic cues (F0 onset frequency and vowel duration) achieve their full phonetic impact without close attention. Experiment 3 shows that this pattern is obtained when the distractor task places little demand on verbal short-term memory. Experiment 4 provides a large data set for testing formal models of the role of attention in speech perception. Attention is shown to influence the signal-to-noise ratio in phonetic encoding. This principle is instantiated in a network model in which the role of attention is to reduce noise in the phonetic encoding of acoustic cues. Implications of this work for understanding speech perception and general theories of the role of attention in perception are discussed.

A basic goal of research in speech perception is to understand the relation between characteristics of the acoustic signal and our phonetic percepts. The range of possible acoustic-phonetic relations is fundamentally shaped by the nature of the human vocal tract and auditory system as well as the distribution of sounds within a language. Even given these constraints, the relation is not constant, but may vary as a function of factors such as environmental noise (e.g., Wardrip-Fruin, 1985) and hearing ability (e.g., Lindholm, Dorman, Taylor, & Hannley, 1988). In this paper, we demonstrate that the relation also varies as a function of the amount of attention that is given to speech perception. Examination of two different phonetic contrasts shows that the importance of weak acoustic cues increases relative to that of strong acoustic cues when subjects are prevented from devoting full attention to speech stimuli. The results are quantitatively well accounted for by a model in which information from different acoustic cues is combined independently (Oden & Massaro, 1978). When this model is interpreted in terms of statistical decision theory, the shift in cue importance can be seen as resulting from increased noise in encoding the phonetic significance of acoustic cues when listeners can not pay close attention to them. This interpretation is instantiated as a stochastic interactive activation model (McClelland, 1991) in which the role of attention is to reduce noise in a pattern recognition network.

Roles of Attention in Speech Comprehension

It is commonly said in introductory lectures that speech perception is a subjectively easy yet computationally difficult task. The computational complexity of speech recognition is hard to dispute, but the notion that it is subjectively easy conflicts with the readily available intuition that at least in some circumstances (e.g., noisy environments and unfamiliar accents) recognition of speech is subjectively demanding. This experience is consistent with our professional experience listening analytically to speech. When careful attention is not given to this task, important aspects of the acoustic-phonetic pattern may escape notice. This suggests that perceiving the phonetic significance of some acoustic cues may require attention and therefore that the ultimate phonetic perception of a complex of acoustic cues may depend on how much attention is given to the stimulus. The goal of the present investigation is to test the validity of this suggestion by examining whether the relative importance of acoustic cues to the identity of phonetic segments varies as a function of attention and to develop a computational model of the role of attention in perceptual processing.

The operation of attention in the comprehension of spoken language has been studied from many perspectives. Our discussion of its roles will be organized around four related topics: (1) aspects of language that facilitate the attentional selection of a specific speech signal, (2) the timecourse of attentional selection, (3) capacity and bottleneck explanations of attention, and (4) attentional effects on basic auditory processes that may precede speech recognition. The extensive nature of attentional effects that have been demonstrated suggest that attention may also be operative in the recognition of phonetic segments – the domain of present interest. However, there appears to be little previous evidence that bears directly on this issue.

The study of the attentional selection of a single speech signal from a background of competing signals and noise played a fundamental part in the development of modern theories of attention. This work, inspired by Cherry's (1953) shadowing technique, has provided a great deal of evidence about the kinds of distinctiveness that can form a basis for attentional selection. Distinctiveness at the following levels of language have been found to facilitate selection: location of source as cued by binaural disparity (Cherry, 1953), amplitude (Egan, Carterette, & Thwing, 1954), fundamental frequency (Darwin & Bethell-Fox, 1977), and semantic continuity (Treisman, 1964). These results, summarized clearly in Bregman (1990), suggest that attention can seize on many low level aspects of an acoustic signal as well as high-level semantic aspects of language.

The timecourse of selection provides additional evidence about the operation of attention in language comprehension. Using shadowing methodology, this issue has been studied by abruptly stopping the language input, and asking the listener to report as much as possible from the unattended channel (e.g., Bryden, 1971; Glucksberg & Cowen, 1970). Listeners are able to report accurately a few seconds of material from the unattended channel, indicating that the input signal - in some form - is stored temporarily before being lost due to lack of attentive processing. The speech signal is thought to be stored in a temporary auditory memory or Precategorical Acoustic Store (PAS, Crowder & Morton, 1969). Phonetic recognition is thought of as a labeling of the information that is held in this memory. Research on this topic using delayed discrimination tasks (e.g., Crowder, 1982; Pisoni, 1973; 1975) and the suffix effect (e.g., Crowder & Morton, 1969) has generally been concerned with the accessibility and persistence of information in auditory form rather than with the process of phonetic labeling and its possible dependence on attentive processing. However, the results from dichotic listening tasks which show that information can be recalled from the unattended channel suggest that attention-dependent processing plays a role at some processing step between auditory memory and the report of a linguistic label.

The above characterization can be seen as implying that language comprehension reaches some level of processing without attention, but that there is a critical point beyond which attention is necessary. This, of course, raises a classic question in attention research: Are attentional limits well characterized by a bottleneck in processing or by some general capacity limits? This question has been central to large literatures on the subject of attention (see e.g., Kahneman, 1973; Pashler, 1989). Some studies in both the bottleneck and capacity traditions have direct relevance to the question at hand.

Bottleneck explanations of attention immediately provoke the question of whether the limit is early or late in processing (Broadbent, 1958; Deutsch & Deutsch, 1963; Duncan, 1980; Treisman & Geffen, 1967; and on). The occurrence of late selection could be taken as evidence against the idea that attentive processing is necessary at the stage of acoustic-phonetic mapping. If it were, then an unattended signal would never reach a lexical level of encoding that is dependent on some segmental pattern recognition. In fact, research in the late-selection tradition by Shiffrin, Pisoni and Castaneda-Mendez (1974) seems to suggests that attention has no effect on the perception of phonetic segments. They used the simultaneous vs. successive technique of Shiffrin and Gardner (1972) to see whether recognition of consonants improved when listeners knew the ear to which a stimulus would be presented. This knowledge, available in the successive condition but not in the simultaneous condition, had no effect on performance. Shiffrin et al. (1974) interpreted this finding as indicating that attentional limitations are post-perceptual and involve control processes associated with short-term memory; a conclusion that is consistent with other findings using the simultaneous-successive procedure (Duncan, 1980; Shiffrin & Gardner, 1972; Shiffrin & Grantham, 1974). However, this finding can not be taken as definitive for at least a couple of reasons. First, attention may play a role in phonetic recognition at a level other than the selection of the physical channel (ear) to which a stimulus is presented. Second, subsequent research using the simultaneous-successive paradigm (Kleiss & Lane, 1986) has shown that successive advantages are found when the required perceptual discriminations are very fine, calling into question the general conclusion from the original work that attentional limits are post-perceptual.

Conceptualizations of attention as limited capacity also provide relevant information concerning the role of attention in speech comprehension. Luce, Feustel and Pisoni (1983) found that recalling synthetic speech placed greater demands on the central capacity associated with short-term memory than did recalling natural speech. While the complexity of the task used by Luce et al. (1983) makes it difficult to pin down the exact nature of the increased demands of the synthetic speech, it only differed from the natural speech in its acoustic quality, which suggests that the observed effect was in part perceptual. The idea of capacity limitations

has also motivated studies examining whether prosody guides attention to significant locations in the speech signal so that they receive more extensive processing (Martin, 1972). A variety of experiments (Buxton, 1983; Cutler, 1976, Metzler, Martin, Mills, Imhoff & Zohar, 1976; Shields, McHugh & Martin, 1974, cf. Mens & Povel, 1986) have shown that segments in prosodically predictable locations (mostly stressed syllables) are recognized more rapidly in phoneme-monitoring tasks. As with the results of Luce et al. (1983), these findings are consistent with the idea that attention plays a role phonetic perception, but their complexity makes it difficult to pin down the exact nature of the observed effects.

A final area of attentional research to consider is the role of attention in basic auditory detection. It has been known for some time that listeners can focus attention at a pre-specified frequency, and that they are very poor at detecting near-threshold tones that are more than a critical band away from the frequency at which they are attempting to detect a tone (Greenberg & Larkin, 1968; Scharf, Quigley, Aoki, Peachey & Reeves, 1987; Swets, 1963; 1984). The potential relevance of this phenomenon to the role of attention in phonetic recognition depends on one's view of the relation between speech perception and audition. A prominent view of this relation holds that the processes underlying speech perception are distinct from those of basic auditory perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman, 1982; Liberman & Mattingly, 1985). However, a substantial number of researchers have begun to argue that the nature of auditory processing does affect phonetic perception (e.g., Diehl, 1987; Klatt, 1982; Lindblom, 1986). If this view is accepted, then effects of attention on auditory detection show that attention operates at a level of processing earlier than phonetic recognition. This complements the bulk of findings reviewed earlier that can be interpreted conservatively as indicating that attentional resources are used in the post-perceptual processing of speech. Thus, existing research points to roles for attention in phases of spoken language processing that occur both earlier and later than the recognition of phonetic segments, without providing compelling evidence about whether it plays a role at that level.

Relative Importance of Acoustic Cues to Segment Identity

Acoustic-phonetic research has been concerned in large measure with determining what acoustic cues have phonetic significance perceptually. It has long been clear that phonetic distinctions are not cued by a single acoustic characteristic, but rather that many aspects of the acoustic signal contribute to people's perception of speech sounds. In a famous example, Lisker (1978) listed 16 acoustic characteristics that may contribute to the perception of the voicing distinction in inter-vocalic stop consonants. Accompanying this sort of enumeration, there has been considerable research and debate about what acoustic features are most important in the recognition of certain phonetic distinctions. One difficulty in assessing the relative importance of acoustic cues stems from methodological difficulties in constructing stimuli that allow assessment of the phonetic importance of acoustic cues. A major point of Lisker's (1978) paper was that some perceptual impact could be found for nearly any acoustic correlate of a phonetic distinction if all the other correlates of the distinction were neutralized. However, effects of some of these acoustic dimensions could not be found if other acoustic dimensions were given more realistic values. For example, Shinn, Blumstein & Jongman (1985) have argued that context-dependent cues contribute little to perception if context-invariant cues are present in the experimental stimuli. However, Nittrouer and Studdert-Kennedy (1986) cogently challenged the naturalism of the Shinn et al. (1985) stimuli, providing further reason to believe that results obtained in this sort of experiment may be quite situation-specific. Other debates about the perceptual importance of various acoustic-phonetic relations can be seen as resulting in good part from the difficulty of preserving the natural interdependencies among acoustic dimensions while systematically manipulating those dimensions in experimental stimuli

The present investigation of the effect of attention on the phonetic encoding of acoustic cues raises additional questions about the generality of experimental results in speech perception. Most research of this sort involves listening conditions that are near optimal in terms of the amount of attention that is given to phonetic encoding of acoustic cues. Subjects are typically asked to do nothing but listen to the speech sounds and must usually attend only to a specific segment. This situation contrasts considerably with the conditions under which people often perceive speech, where they may be simultaneously performing another task (such as driving a car) and where they are almost certainly focusing on the meaning of a communication rather than the identity of a single pre-specified phonetic segment. This prompts the concern that some of the effects that have been observed with close listening conditions are limited to laboratory conditions. However, it also prompts the hope that studying the change in relative phonetic importance from focused to unfocused attention will provide information about what cues are naturally more salient in conditions of unstudied listening.

A more fundamental reason for studying the role of attention in phonetic encoding is that it is a general, higher-level perceptual process that may play a role in shaping the acoustic-phonetic patterns of languages. As noted earlier, a number of factors have been found to influence the relative phonetic importance of acoustic cues. These include noise (e.g., Wardrip-Fruin, 1985), hearing disability (e.g., Lindholm et al., 1988), early development (e.g., Bernstein, 1983) and late development (e.g., Price & Simon, 1984). However, the operation of attention differs from these other factors in that it is an always present property of a person. In contrast, noise is not a human property, nor is it always present. Hearing disabilities and development are human dimensions but they are not always operative. With regard to these properties, attention is more on a par with the structure of the vocal tract or the basic auditory system which have long been considered to have a fundamental role in shaping the acoustic-phonetic pattern of languages. The present research investigates whether attention level plays a role in shaping acoustic-phonetic relations by examining whether it differentially affects the phonetic importance of strong and weak acoustic cues to the identity of phonetic segments. It seems likely that differences in the inherent strength of cues reflect the cumulative influences on acoustic-phonetic patterns, including possible consequences of naturally-occuring variation in attention level.

Experiment 1

This experiment examines whether the amount of attention that is allocated to speech perception influences the relative importance of two acoustic cues to the voicing distinction between the consonants /b/ and /p/. The amount of attention available for speech perception is manipulated by sometimes having subjects perform a visually-presented arithmetic distractor task while the speech stimulus is presented. The two cues to consonant voicing are voice-onset time (VOT) and the onset frequency of the fundamental (F0). VOT is the time between the release of a consonant and the onset of phonation. Voiced stop consonants like /b/ have short VOTs (0 to 10 msec) while voiceless sounds like /p/ have long VOTs (50 to 70 msec) (Lisker & Abramson, 1964). In addition, voiced consonants tend to have a lower onset frequency of F0 than do voiceless consonants.

In comparing the importance of these two cues, Abramson and Lisker (1985) have argued that VOT is the primary cue to voicing because the onset frequency of F0 has a strong effect on perceptual judgments only when VOT is ambiguous. Further evidence of the greater importance of VOT comes from Bernstein's (1983) finding that perceptual judgments by young children are consistently influenced by VOT but not by F0. Thus, the use of these two cues allows us to examine the effect of attention on the relative importance of acoustic cues when one of the cues is strong and the other is weak. Examining the perception of these cues under

low levels of attention allows us to see whether weak cues achieve their modest phonetic contribution because of the close attention demanded of subjects in the typical speech experiment or whether strong cues achieve their large impact because of close attention.

Method

Subjects. Twelve students at Harvard University were tested individually in a single hour-long session. They received base pay of $4.00 plus a bonus of up to $3.00 depending on their speed and accuracy in the arithmetic distractor task. All subjects were native speakers of English who reported having normal hearing.

Stimuli. The stimuli were created with the Klatt (1980) synthesizer and varied along two dimensions: VOT and onset frequency of F0. A silent VOT interval, following an initial burst, ranged from 0 to 70 msec in 10 msec steps. The formant transition characteristics were appropriate for a labial place of articulation, and the steady state formant frequencies were appropriate for /a/. Two onset frequencies of F0, 100 Hz and 150 Hz, were used. This frequency was changed in a linear fashion to 125 Hz over the first 50 msec of voicing. All characteristics of these stimuli, other than F0, were taken from McClaskey, Pisoni and Carrell (1983).

Design. The experiment included one practice block and ten experimental blocks of 32 trials each. In the practice block, subjects performed only the arithmetic distractor task. Half of the experimental blocks were conducted in the *distractor condition* in which subjects had to both perform the arithmetic task and recognize a speech sound. The other blocks were conducted in the *no-distractor condition* where subjects only had to recognize the speech sound. The experimental blocks alternated between the distractor and no-distractor conditions, with all subjects starting in the distractor condition. Each experimental block included four presentations of each of the eight VOT values in a random sequence. F0 onset frequency was manipulated across pairs of distractor and no-distractor blocks. Half of the subjects began with a low-frequency F0 onset and the other half with a high-frequency F0 onset.

Procedure. Subjects initiated a trial by clicking a mouse button. At the start of the trial two fixation lines appeared on the computer screen followed by the visual test stimulus. For the practice block and for blocks in the distractor condition, the visual stimulus consisted of three two-digit numbers which were all multiples of ten. Subjects were asked to decide whether the difference between the first and second numbers was the same as the difference between the second and third numbers. They were told to respond as quickly and accurately as possible by clicking an appropriate mouse button. The number of trials requiring affirmative and negative responses was equal for each block. Immediately after each distractor block, feedback was given on speed and accuracy of response in the arithmetic task, and points were awarded based on speed and accuracy. The amount of bonus money that subjects received was based on the number of points they earned.

In distractor blocks other than the practice block, a speech sound was presented 500 msec after the appearance of the numbers.[1] The speech sound was presented over headphones at a comfortable listening level. After the subjects had made a response in the number task, they were prompted to identify the speech sound by a "b" and a "p" appearing on the computer screen. The subjects were told to identify the speech sound as accurately as possible and that speed was not important. Subjects were told that their bonus would depend only on their performance on the arithmetic task and that they should treat it as primary. At the end of each distractor block, the experimenter showed the subject his or her number of errors and mean RT for the block. The experimenter then encouraged the subject to try hard. Feedback was given for the arithmetic task only.

The next experimental test block was presented in the no-distractor condition. In this condition, three pairs of zeros appeared on the computer screen as the speech sound was presented. The duration that the visual stimulus was displayed was derived from the subject's average response time to the number task in the previous block of the distractor condition. After the visual stimulus, the subjects were prompted to respond to the auditory stimulus by a "b" and a "p" appearing on the computer screen. Subjects were again told to try to respond accurately, although speedy responses were not necessary.

Results

Distractor Task Performance. Figure 1 shows the mean response times (of correct responses) and accuracies in the distractor task as a function of the characteristics of the speech sound presented on a trial. Response times varied significantly as a function of the interaction of VOT and F0 onset frequency; $F(7,77) = 15.6$, $p < .001$. The fastest response times occurred when VOT and F0 provided congruent cues to phonetic segment identity, $t(11) = 5.7$, $p < .001$ for the linear interaction. This occurs for stimuli with short VOTs and low F0, and for stimuli with long VOTs and high F0. Response times were longest when VOT and F0 provided incongruent cues to segment identity, i.e., short VOTs paired with high F0 and long VOTs paired with low F0.

Identification of Speech Stimuli. Figure 2 shows listeners' identifications of the speech stimuli as a function of VOT, F0 onset frequency, and distractor condition. For the distractor blocks in this and subsequent experiments, speech identification responses were excluded if the response in the distractor task was incorrect. As would be expected given previous results, there were significant main effects of VOT ($F(7,77) = 87.0$, $p < .001$) and F0 onset frequency ($F(1,11) = 170.2$, $p < .001$) on judgments of the stimuli. As would also be expected, there was a significant interaction of VOT and F0 onset frequency; $F(7,77) = 29.0$, $p < .001$. F0 onset frequency had a greater impact at intermediate values of VOT (20 to 50 msec) than near the endpoints (0,10, 60 and 70 msec), $t(11) = 12.3$, $p < .001$.

A significant interaction was found between distractor condition and the effect of VOT on judgments of the speech stimuli; $F(7,77) = 12.9$, $p < .001$. As can be seen by comparing the left and right panels of Figure 2, the effect of VOT on identification was stronger in the no-distractor condition than in the distractor condition. The distractor task did not have a similar impact on the F0-onset-frequency cue to voicing. In the no-distractor condition, stimuli with an F0 onset of 100 Hz produced 32 percent more /b/ responses than did those with a 150 Hz F0 onset, while in the distractor condition, the analogous difference was 36 percent. However, this increase in the importance of F0 onset frequency in the distractor condition was not significant $F(1,11) = 1.6$, $p > .20$.

A significant interaction was found between the three factors of distractor condition, VOT and F0 onset frequency; $F(7,77) = 2.8$, $p < .02$. Figure 3 makes the form of this interaction apparent. It shows the difference in the proportion /b/ identifications between stimuli with F0 onset frequencies of 100 Hz and 150 Hz, broken down by VOT and distractor task. This difference is an indication of the phonetic importance of fundamental frequency. The bowed shape of the two lines makes it clear that F0 has its greatest significance at intermediate levels of VOT. This pattern, however, is more pronounced for the no-distractor condition than for the distractor condition. In particular, F0 onset-frequency had a greater effect at the VOT endpoints (0 and 70 msec) in the distractor condition than in the no-distractor condition, $t(11) = 3.2$, $p < .01$.

Discussion

The results of the experiment showed that performance on each task was significantly influenced by the other task. This was somewhat unexpected in the case of the speed of the arithmetic distractor task depending on the congruence of the two acoustic cues to the speech sound. This finding was unexpected because the instructions to the subject emphasized that the arithmetic task was primary and that they should devote their full effort to it. However unexpected, the finding is consistent with the idea that the speech identification and distractor tasks involve a shared limited capacity. The finding that the magnitude of the interference depended on the congruence of the two acoustic cues to voicing indicates that the process of identifying phonetic segments consumes more capacity when there is conflicting stimulus information, and that this influences a concurrent task. Unfortunately, not much can be made of this specific finding because it is not repeated in the next three experiments.

The manner in which the distractor task influenced the phonetic importance of the VOT and F0-onset-frequency cues to voicing is more central to the present concern. The effectiveness of VOT as a cue to the voicing distinction of /ba/ vs. /pa/ was reduced in the distractor condition as compared to the no-distractor condition. One interpretation of this finding is that the importance of VOT for phonetic perception is reduced when close attention can not be given to the stimulus. A less interesting possibility is that the ability to identify the speech sound was disrupted in general by the distractor task. This possibility is ruled out by the results obtained for F0 onset frequency. The phonetic importance of F0 was not diminished by simultaneous performance of the distractor task; in fact, there was a non-significant increase in its importance. This indicates that the distractor task did not simply produce an overall decrement in listeners' ability to identify the speech stimulus, but rather that the decrement in the importance of VOT was specific to that aspect of the stimulus. The phonetic significance of F0 onset frequency does not appear to depend on the ability to pay close attention to the stimulus. In addition, F0 onset frequency had a significant effect on judgments of voicing even when VOT was unambiguous (cf. Abramson & Lisker, 1985). This effect of F0 onset frequency at the VOT endpoints increased when listeners were prevented from devoting full attention to the speech stimulus.

The change in the relative importance of VOT and F0 onset frequency provides a first answer to our question concerning the importance of attention in the encoding of strong and weak phonetic cues. On the basis of these results, it appears that strong cues achieve their commanding phonetic importance through careful attention to the stimulus. Weak cues can achieve their modest contribution even without careful attention. Of course, this interpretation presumes that there are not characteristics particular to the processing of VOT and F0 onset frequency that make VOT more dependent than F0 onset frequency on attentive processing.

Experiment 2

This experiment has two goals: to test the generality of the previous result concerning the greater dependence of strong cues than weak cues on attentive processing and to test a specific hypothesis about how the lack of attention affects phonetic encoding. This is done by examining two acoustic cues to the distinction between the vowels /i/ (as in "beat") and /I/ (as in "bit"). The two cues are formant pattern (for /i/ the first formant is lower and the second and third formants higher than for /I/) and duration (/i/ tends to be longer than /I/). Several sources of evidence indicate that formant pattern can be considered the stronger or primary cue, while duration can be considered a weaker, secondary cue. Formant pattern depends on the shape of the vocal tract, which is the articulatory characteristic most uniquely related to vowel identity. Vowel duration depends on the amount of time that a vocal tract shape is maintained, or more likely on the rate at which the vocal tract approaches and moves away from a target configuration. While vowels do differ on average in their inherent durations (Peterson & Lehiste, 1960), these durations also depend on a large number of other factors such as overall

speaking rate, prosodic patterns, and identity of neighboring segments (Gordon, 1989; Klatt, 1976; Peterson & Lehiste, 1960). Formant patterns are also subject to a variety of contextual effects (e.g., coarticulation, reduction and vocal-tract differences between speakers), but these influences are less drastic than those that operate on inherent vowel duration. Perceptual data support this analysis of formant pattern as the stronger cue. Vowel duration tends to influence subject's perceptual judgments only when formant pattern is ambiguous (Pisoni, 1975). If the pattern of results in the previous experiment has been correctly interpreted as indicating a dependence of strong, but not weak, cues on attentive processing, then we would expect that the phonetic importance of formant pattern ought to decrease relative to that of duration as attention to the speech stimulus is decreased.

Our second goal was to test the hypothesis that a low attention level influences speech perception by delaying phonetic access to an auditory representation of the speech stimulus. If this were the case, we would expect a greater reduction in the phonetic importance of formant pattern for short duration vowels than for long duration vowels, because there would be less time available to access the auditory representation. In addition to the actual physical differences in duration between the stimuli, it has been argued that the persistence of short-term auditory memory increases with the duration of a stimulus, with such a process accounting for why short-duration vowels are perceived in a more categorical fashion than long-duration vowels (Fujisaki & Kawashima, 1969; 1970; Pisoni, 1971; but see Pisoni, 1973; 1975; Crowder, 1981). If restricted attention influences speech perception in this way, then we would expect that shifts in the relative phonetic importance of acoustic cues would be determined by their duration and their relative persistence in auditory memory, not by differences in inherent phonetic strength.

Method

Subjects. Twelve new subjects, drawn from the same pool as the previous experiment, participated in a single hour-long session. Pay was the same as for the previous experiment.

Stimuli. The speech stimuli were created on the Klatt (1980) synthesizer and were closely modeled after those of Pisoni (1975). A seven-member series of formant patterns combined with two vowel durations (300 msec and 50 msec) yielded a total of 14 vowel stimuli. The formant series was constructed by varying the center frequencies of the first three formants in equal logarithmic steps from /i/ to /I/ (see Table 1). The fourth and fifth formants were held constant at 3500 Hz and 4500 Hz respectively. The bandwidths of the first three formant frequencies were fixed at 60, 90, and 150, respectively. The 300 msec and 50 msec vowels differed in their rise and decay times. The rise and decay times were 50 msec for the 300 msec vowels and 10 msec for the 50 msec vowels. For the 300 msec vowels, fundamental frequency fell from 125 Hz at onset to 80 Hz at offset while for the 50 msec vowels fundamental frequency fell from 125 Hz to 100 Hz.

Design and Procedure. The design was the same as the previous experiment with the following exceptions. There was one practice block and eight experimental blocks. Each block had 35 trials, five of each formant pattern. Duration of the speech sounds was manipulated between blocks. The procedure for the previous experiment was modified so that subjects were told to identify the speech sounds as /i/ as in "beet" or /I/ as in "bit", and were prompted to respond by an "e" or "i" on the computer screen.

Results

Distractor Task Performance. Figure 4 shows mean response times and accuracies for the arithmetic distractor task as a function of the formant pattern and duration of the concurrently presented vowel. For response times, the main effect of vowel duration was not significant [$F(1,11) = 2.3$, $p > .15$], the main effect of formant pattern was significant [$F(6,66) = 5.8$,

$p < .001$, and the interaction of duration and formant pattern was not significant [$F(6,66) < 1$]. Based on the results of the previous experiment a planned test was performed on the linear interaction of formant pattern and duration. This test was not significant; $t(11) = .4$, $p > .2$.

Identification of Speech Stimuli. Figure 5 shows the mean proportion of /i/ responses as a function of formant pattern, vowel duration, and distractor condition. There were significant main effects of formant pattern [$F(6,66) = 235.9$, $p < .001$] and vowel duration [$F(1,11) = 47.4$, $p < .001$], as well as a significant interaction of these two factors [$F(6,66) = 18.2$, $p < .001$]. Duration had a greater impact at the intermediate formant patterns than at the endpoints of the continuum, $t(11) = 7.3$, $p < .001$.

A significant interaction was found between distractor condition and the effect of formant pattern on identifications of the speech sounds; $F(6,66) = 17.3$, $p < .001$. Formant pattern had a greater influence in the no-distractor condition than in the distractor condition. There was also a significant interaction between the effects of distractor manipulation and vowel duration on speech identifications; $F(1,11) = 13.4$, $p < .005$. However, in contrast to what was observed for formant pattern, the effect of duration was greater in the distractor than in the no-distractor condition; 24.1% more /i/ responses occurred for 300 msec vowels than for 50 msec vowels in the distractor condition, while the difference was only 15.9% in the no-distractor condition.

The three-way interaction of vowel duration, formant pattern, and distractor task was significant; $F(6,66) = 2.57$, $p < .05$. While vowel duration always had its greatest impact when formant pattern was intermediate, it had an effect on the extreme formant patterns in the distractor condition but not in the no-distractor condition. A planned contrast showed that the effect of vowel duration on the endpoint formant patterns was greater in the distractor condition than in the no-distractor condition, $t(11) = 3.6$, $p < .005$.

Discussion

Performance on the distractor task was not systematically related to the acoustic characteristics of the concurrently presented speech sound. While a significant effect of formant pattern was observed, the differences underlying this effect are not clearly related to the progression of formant patterns. More importantly, there was not a significant interaction between formant pattern and vowel duration. This contrasts with the interaction observed in the previous experiment where the two cues to voicing, VOT and F0 onset frequency, had an interactive impact on response times in the distractor task; faster times where observed when the phonetic significance of the cues was congruent than when they were incongruent. This discrepancy could be due to some difference between the way in which the cues to voicing and the cues to vowel identity are processed or integrated. More likely, it is due to some unintended difference in instructional emphasis in the two experiments. As the distractor task performance is not central to the goals of this project, these possibilities are not pursued further.

The results of the speech perception task support and extend the findings of Experiment 1. The relative importance of two acoustic cues was found to change when listeners could not devote their full attention to the speech perception task. The strong acoustic cue of formant pattern decreased in phonetic importance when listeners were simultaneously performing the distractor task. This is analogous to the effect observed for VOT in the previous experiment. For the weak cue of vowel duration, there was a significant increase in importance when subjects performed the distractor task. This is a similar but stronger effect than the non-significant increase observed for the weak cue of F0 onset frequency in the last experiment. Taken together, the results of these experiments support the following view of the relation between cue strength and attention. A stimulus must be carefully attended to for a strong acoustic cue to realize its full impact on phonetic categorization. When listeners are prevented

from doing so, the importance of such cues will diminish. In contrast, weak cues depend less on attention in order to achieve their phonetic impact, and their net contribution to listeners' identifications is not diminished and may actually increase when attention is diverted by a competing task.

The second goal of this experiment was to test the hypothesis that performing the distractor task influenced speech perception by delaying phonetic access to a decaying auditory representation of the stimulus. That hypothesis lead to the expectation that the distractor task would impair encoding of formant information for 50 msec vowels more than for 300 msec vowels. Examination of Figure 5 shows that this did not occur. In the distractor condition, formant pattern had at least as big an effect on listeners' identifications for the 50 msec vowels as it did for the 300 msec vowels. Therefore, this experiment provides no support for the hypothesis that access to a decaying auditory representation is delayed by the distractor task.

Experiment 3

The goal of this experiment is to show that the phonetic encoding of acoustic cues can be affected by distractor tasks other than the arithmetic one used in the previous experiments. In addition to increasing methodological generality, this experiment will determine whether an effect on speech perception can be observed when the role of verbal short-term memory in the distractor task is reduced. This provides an initial step toward determining the locus of processing at which the distractor task competes with speech perception. The experiment combines the vowel identification task of the previous experiment with a new line-length discrimination task.

In the previous arithmetic distractor task, the stimuli consisted of a sequence of three numbers, and subjects had to make a speeded judgment about whether the difference between the first two was equal to the difference between the second two. At least initially, this probably involved verbal encoding of the numbers, calculation of the two differences, and comparison in short-term memory. As subjects became practiced in the task, it is possible that some of this became automatic and that the role of verbal short-term memory was reduced. The line-length discrimination used in the present experiment was designed so as to reduce as much as possible the role of verbal processing of the distractor stimuli. This was done by presenting subjects with two vertical lines and asking them to make a speeded judgment as to whether the one on the left or the right was longer. As the relevant stimulus characteristics were difficult to encode verbally and the stimuli were present until a response was made, it seems likely that this task placed fewer demands on verbal encoding or short-term memory than the arithmetic task did.

Method

Subjects. Twelve new subjects from the same population as the previous experiment served as paid subjects in a single hour-long session.

Stimuli, Design and Procedure. The speech stimuli were the vowel sounds used in Experiment 2. The general procedure was the same as in the previous experiment except for the new distractor task. For this task, a central fixation mark appeared followed by two vertical lines, one to either side of fixation, and the subject had to make a speeded keypress indicating the longer of the two lines. The mapping between stimuli and responses was compatible, the left key indicated the left line and the right key indicated the right line. The absolute length of the lines was varied across trials, and the difference in length between the short and long lines was roughly proportional to absolute line length. The centers of the two lines were at the same heights, but the horizontal position of each line within the hemifield was randomly determined on each trial. The parameters of the visual stimuli were explored during pilot work to find

values that made the task as difficult as possible while still allowing an attentive subject to accurately determine which line was longer.

Results

Distractor Task Performance. Figure 6 shows the mean response times and accuracies for the line-length discriminations as a function of vowel duration and formant pattern. Response times were significantly longer when the vowel sound was 300 msec than when it was 50 msec, $F(1,11) = 8.1$, $p < .02$, however, there were significantly fewer errors for the longer vowels than the shorter vowels [$F(1,11) = 16.0$, $p < .005$] suggesting a speed-accuracy tradeoff. Formant pattern also had a significant effect on response times [$F(6,66) = 3.0$, $p < .02$], as did the interaction between duration and formant pattern [$F(6,66) = 2.6$, $p < .05$]. The linear interaction of formant pattern and vowel duration was not significant; $t(11) = .66$, $p > .2$. In this experiment, as in the last one, the effects on distractor task performance of the speech sounds are weakly related to the phonetic significance of the acoustic cues.

Identification of Speech Stimuli. Figure 7 shows the mean proportion of /i/ responses as a function of formant pattern, vowel duration, and distractor condition. There were significant main effects of formant pattern [$F(6,66) = 128.0$, $p < .001$] and vowel duration [$F(1,11) = 41.9$, $p < .001$], as well as a significant interaction of these two factors [$F(6,66) = 10.5$, $p < .001$]. Duration had a greater impact at the intermediate formant patterns than at the endpoints of the continuum, $t(11) = 6.7$, $p < .001$.

Identifying the speech sounds while simultaneously performing the distractor task diminished the phonetic significance of the formant pattern; $F(6,66) = 6.56$, $p < .001$. The net effect of duration increased from 13.5% in the no-distractor condition to 18.0% in the distractor condition, but this difference was not significant; $F(1,11) = 0.74$. The three-way interaction of vowel duration, formant pattern, and distractor task also failed to reach significance; $F(6,66) = 1.0$. However, a planned contrast showed that the effect of vowel duration on the endpoint formant patterns was greater in the distractor condition than in the no-distractor condition, $t(11) = 2.32$, $p < .05$.

Discussion

The effect of the line-length discrimination task on the speech identifications was similar to, but weaker than, the effect of the arithmetic task observed in Experiment 2. Both tasks caused significant decreases in the effectiveness of formant pattern as a cue to vowel identity. Both tasks resulted in a net increase in the importance of duration as a vowel cue, though this effect was not significant with the line-length task. Both tasks caused significant increases in the effect of duration for the formant patterns at the endpoints of the continuum. The effects of the line-length task indicate that tasks other than speeded mental arithmetic can influence speech perception, and thus satisfy the goal of increasing methodological generality across distractor tasks. These effects also indicate that a heavy verbal short-term memory component is not a necessary condition for observing that a concurrent task affects speech perception.

The smaller impact of the line-length task as compared to the arithmetic task could result from several factors. Perhaps a portion of the impact of the arithmetic task was due to its reliance on short-term memory or some other processes that were called on less in the line-length task. Alternatively, the line-length task may have had a smaller impact because it was easier than the arithmetic task. The mean response time and accuracy for the line-length task were 877 msec and 90% while they were 1434 msec and 94% for the arithmetic task in the previous experiment.[2] The greater ease of the line-length task may have caused it to draw less on general processing resources and therefore to have had less impact on concurrent speech perception.

Experiment 4

The goal of this experiment is to provide data that can be used to test formal models of the effect of attentiveness on the phonetic encoding of acoustic cues. The initial modeling effort will be based on a class of models in which information from different sources is combined independently in making perceptual judgments. This idea is embodied in both signal detection theory and Luce's choice theory, which have very similar structures and which often make quantitatively similar predictions (see McClelland, 1991 for a cogent and relevant review). These models have enjoyed successful application in far-flung domains such as visual letter recognition (Oden, 1979), judgments of social traits (Anderson, 1974), and semantic judgments (Oden, 1977). Most relevant to the present endeavor, the *Fuzzy Logical Model of Perception*, which incorporates Luce's choice rule, has provided excellent accounts of many results in speech perception including the phonetic integration of distinct acoustic information and the role of context in speech perception (e.g., Oden & Massaro, 1978, Massaro, 1989). The present modeling of speech perception, though based on these successes, will employ the alternative signal detection formalism because it is quite naturally implemented as a stochastic interactive activation model (McClelland, 1991) which we will show provides a good account of the role of attention in recognizing phonetic segments.

A key feature of these independent cue models is that each perceptually significant feature of a stimulus is encoded independently of the other features of the stimulus that are simultaneously present. Therefore, when fitting the model to data the number of parameters in the model equals the sum of the number of feature values for the different features, while the number of observations in a factorial design equals the product of the number of feature values for the different features. Our previous experiments have therefore not allowed a meaningful test of the model because of the low ratio of observations to parameters. The present experiment tackles this issue by using the vowel stimuli of Experiments 2 and 3 but increasing the number of duration levels from two to five, resulting in 35 speech sounds. The experiment uses the arithmetic distractor task of Experiments 1 and 2 because of its large impact on phonetic encoding.

Method

Subjects. Thirty-five individuals served as paid subjects in a single experimental session.

Stimuli, Design and Procedure. The vowel stimuli were the same as in Experiments 2 and 3 except that three additional vowel durations of 90, 120, and 190 msec were added to the previously used durations of 50 and 300 msec. Combined factorially with the seven formant frequencies, this yielded a total of 35 vowel stimuli. A session consisted of one block of the distractor task alone for practice followed by 8 experimental blocks. These alternated between the no-distractor and distractor conditions. Each block had 35 trials involving one presentation of each vowel stimulus in a random order. This means that the vowel stimuli at the different durations were mixed rather than blocked as in the previous experiments. The procedure was otherwise the same as in Experiments 2 and 3.

Results

The mean proportions of /i/ identifications for the speech sounds in the no-distractor and distractor conditions are shown in Figure 8. As expected, there were significant main effects of formant pattern [$F(6,204) = 263.4$, $p < .001$] and of vowel duration [$F(4,136) = 76.7$, $p < .001$]. There was a significant interaction between formant pattern and whether or not the distractor task was performed, [$F(6,204) = 20.6$, $p < .001$]. This was due in good part to a decrease in the importance of formant pattern when the distractor task was being performed as shown by the linear interaction of formant pattern with distractor task, [$F(1,34) = 55.2$,

$p < .001$]. There was also a significant interaction between the distractor task and vowel duration; [$F(4,136) = 13.0$, $p < .001$]. A significant interaction between the linear effect of vowel duration and distractor condition [$F(1,34) = 13.1$, $p < .001$] showed that this is due in good part to increasing importance of vowel duration when the distractor task was being performed. A planned test showed that the linear effect of vowel duration was greater at the formant pattern endpoints in the distractor condition than in the no-distractor condition, [$F(1,34) = 24.2$; $p < .001$].

The above analyses show that the pattern of effects observed in this experiment is very similar to that observed in the previous three experiments.

Independent Cue Models

In an independent cue model within the signal detection framework, the perceptual significance of each level of a feature is expressed as a z-score. This score represents the distance between a decision criterion and the mean encoding value for a feature in units of standard deviations of the encoding distribution. When a stimulus contains more than one feature, its overall perceptual value is given by the sum of the z-scores for the features it contains. Therefore, fitting an independent cue model to the present data involves finding the set of z-scores, one for each level of formant pattern and vowel duration, that minimizes the squared deviations from the observed response probabilities when the predicted probabilities are given by the sum of the cue values in a stimulus, as shown in the following:

$$p(/i/|S_{DjFk}) = ZCDF(Z_{Dj} + Z_{Fk}).$$

Here, the probability of responding /i/ given a stimulus S with duration j (D_j) and formant pattern k (F_k) equals the value of the cumulative normal distribution function (ZCDF) for the sum of the cue values (expressed in z-scores) for duration j (Z_{Dj}) and formant pattern k (Z_{Fk}).

As an initial step, separate fits were found for the no-distractor and distractor conditions. The resulting parameter values are shown in Table 2. Looking first at the parameters for the no-distractor condition, it is apparent that the range of values for levels of formant pattern is greater than that for vowel duration. This reflects the larger perceptual role of formant pattern as compared to vowel duration. The fit of these parameters to the data has a root-mean square (RMS) error of .027, indicating a high correspondence between the predicted and observed response probabilities. By comparison, the parameters for the distractor condition show a similar pattern but are for the most part reduced in absolute value. This reduction in cue values in the distractor condition reflects a reduction in the signal-to-noise ratio in encoding phonetic cues when attention level is diminished. Thus, attention can be characterized as affecting the signal-to-noise ratio in encoding phonetic information. When attention is focused on phonetic encoding, as in the no-distractor condition, higher signal-to-noise ratios are obtained. When attention is not focused on phonetic encoding, as in the distractor condition, lower signal-to-noise ratios are obtained. As in the no-distractor condition, the distractor condition shows a greater range of values for formant patterns than for vowel durations. However, the difference is less than for the no-distractor condition, indicating the relative importance of vowel duration has increased. Another noteworthy aspect of these results is that the fit for the distractor condition has an RMS error of .051 which is larger than that of the no-distractor condition. This increased error probably reflects increased variability stemming from the concurrent tasks and should not be taken as evidence against the model.

The fits shown in Table 2 demonstrate that a model based on independently combining information from the formant pattern and vowel duration cues provides a viable quantitative framework for the present results. However, because separate models were fit for the no-distractor and distractor conditions, the results do not provide a unified account of the effect of diminished attention on the phonetic encoding of acoustic cues. Our strategy for formulating

such an account involves exploring how the cue values obtained in the no-distractor condition must be modified in order to fit the data from the distractor condition. Different ways of modifying the cue values will be assessed in terms of how well they fit the results from the distractor condition. Because no modification of the parameters obtained in the no-distractor can provide a better fit than was obtained when the distractor results were fit separately, a ceiling on the fit is given by the RMS error of .051 that was observed for that model.

The most straightforward way to conceive of the effect of attention within this framework is that its withdrawal has separate effects on the phonetic encoding of formant pattern and vowel duration. This can be implemented simply in a model described by the equation:

(1) $p(/i/|S_{DjFk}) = ZCDF(A_D Z_{Dj} + A_F Z_{Fk} + K)$.

Here, the phonetic value of the vowel durations are scaled by one attention parameter A_D and those of the formant patterns are scaled by a second attention parameter A_F. These attention parameters can increase or diminish the perceptual significance of the cues that they modify. The resulting modifications in perceptual significance consist of linear changes in the signal-to-noise ratios for phonetic encoding of the two kinds of acoustic information. The constant K in the equation must be included in the model because the phonetic values that are modified (i.e., those from the first model fit to the no-distractor condition, top of Table 2) include an arbitrary constant. When the phonetic values for formant patterns and vowel durations are multiplied by different attention factors the value of the constant is no longer arbitrary and must be included in the model.[3] When fit to the data from the no-distractor condition, this model has an RMS error of .069. The attention parameter is .709 for formant pattern, while it is .936 for vowel duration. The values of these parameters convey important information about what happens to the encoding of the two kinds of acoustic cues when listeners are not able to pay close attention to the speech stimulus. With regard to formant pattern, the value of .709 indicates that the distinctive information conveyed by different levels of formant pattern is considerably reduced under low attention levels. With regard to vowel duration, the value of .936 indicates that the distinctive information conveyed by different levels of this acoustic cue is also reduced somewhat. The analysis of the effect of attention on vowel duration at this level is thus quite different from one that looked simply at the observed response probabilities. There, it would seem that the importance of vowel duration actually increased when listeners could not pay close attention to the speech stimulus. The modeling result shows that this increase in importance is not due to increased distinctiveness of vowel duration but rather due to reduced competition from formant pattern because of the apparently greater dependence of that cue on attentive processing.

While the analysis using separate parameters yields some important information, it is not entirely satisfying because it does not offer any clues as to why attentive processing might be more important for formant pattern than for vowel duration. A model that treats the two kinds of cues identically could potentially provide a more complete account of the role of perception in recognizing phonetic segments. The two kinds of cues would have equal prior standing in a model with only one attention parameter, such as:

(2) $p(/i/|S_{DjFk}) = ZCDF(AZ_{Dj} + AZ_{Fk} + K)$

where A has the same value for vowel duration and formant pattern. Information is lost from the cue values obtained in the focused attention condition when A is less than 1, and the response probabilities move toward the constant response bias given by the parameter K. This model fits the data with an RMS error of .080 and a value for the attention parameter of .689. This fit is not as good as for the previous model, but it is achieved with one fewer free parameter. The implication of this finding is that the effect of diminished attention can be characterized, at least in part, as a general diminution of phonetic distinctiveness independent of its acoustic source. The reason that this model can fit as well as it does stems from a

somewhat unintuitive feature of the way that sums of z-scores map onto probabilities. The impact on response probabilities of the constant phonetic value of a vowel duration diminishes as the absolute magnitude of the phonetic value of the associated formant pattern increases. As we have noted several times, formant pattern has substantial phonetic distinctiveness leading to relatively large phonetic cue values. When these phonetic cue values are diminished by the attention factor, they may be brought into a range in which the phonetic cue value of a vowel duration has a greater impact on response probabilities even though it is diminished by the same factor.

To illustrate this point, consider what happens to the predicted response probabilities for the two stimuli given by combining the first formant pattern with the first two vowel durations. As shown by Table 2, the cue value for the first formant pattern is 2.04, while it is -.838 and -.374 for the first two vowel durations respectively. Summing these cue values gives z-scores of 1.20 and 1.66 which yield response probabilities of .886 and .952. Therefore, the net effect of the vowel duration cue on response probability is .066. When these response z-scores are multiplied by the attention factor of .689, the cue value for the formant pattern becomes 1.41 and those for the two vowel durations become .58 and .26. These yield sums for the stimuli of .83 and 1.15, resulting in response probabilities of .796 and .875. Thus, one effect of reduced attention in this model is to increase the net effect of vowel duration on the identification of these two stimuli from .066 to .079.

The above model suggests the possibility that attention may influence vowel duration and formant pattern in similar ways, but that the magnitude of their phonetic distinctiveness may be the basis of the difference apparent in the response probabilities. The next model takes this possibility one step further. In it, the effect of attention on cue strength is not linear, but rather is proportional to the absolute magnitude of the cue as shown below:

$$(3)\quad p(/i/|S_{DjFK}) = ZCDF(Z_{Dj} - A|Z_{Dj}|Z_{Dj} + Z_{Fk} - A|Z_{Fk}|Z_{Fk} + K).$$

Here, the cue strength for each feature value is decreased by the product of an attention factor, the absolute value of the cue strength, and the cue strength. This model fits the data with an RMS error of .074 and a value of the attention parameter (A) of .170. This improvement over the previous model indicates that low attention causes an accelerating loss of information as the significance of a cue increases. This model with a single attention parameter achieves a fit (.074) that is not far off that (.069) which was achieved in the model that used two attention parameters to separately fit vowel duration and formant pattern. This latter model (Equation 3) seems preferable because of its greater parsimony and because it offers the possibility of a unified account of the effect of attention on both vowel duration and formant pattern.

The effect of attention has been incorporated in the above models by changing the phonetic values of the acoustic cues. The phonetic values are given in a scale (z-scores) that expresses the distance between a decision criterion and the mean of the encoding distribution in terms of the width of the encoding distribution (which is assumed to be normal). The diminished phonetic values due to low attention can therefore be interpreted mechanistically as being due to a lessening of the distance between the mean of the encoding distribution and the decision criterion, or an increase in the variability of the encoding of phonetic values, or both. This leads to the intuitively appealing idea that the role of attention in recognizing speech patterns is to optimize the signal-to-noise ratio in the phonetic encoding of acoustic cues. We pursue a specific mechanistic implementation of this idea below.

Stochastic Interactive Activation

In a recent paper, Massaro (1989) showed that interactive-activation models (McClelland & Rumelhart, 1981, McClelland & Elman, 1986, Rumelhart & McClelland, 1982), as they were

originally put forth, were not capable of accounting for the large number of additive cue effects that have been found in perceptual (and other kinds of) research. He argued that the interactive processing in such models precluded additive integration of information. However, McClelland (1991) showed that the structural and dynamical assumptions of the models were compatible with such effects, but not in conjunction with the decision rule that had originally been used in the models. The original rule was based on the relative activation levels of the output units. Interactive-activation models can exhibit additive cue effects if the decision rule is changed to one of simply selecting the most active output unit. Further, it must be assumed that there is variability in either the stimulus input or in the transmission of activation between network units. This noise in the processing network means that there can be variability across trials in which output node is most highly activated and consequently in identifications of a stimulus. At least under some circumstances, such a model will exhibit additive cue effects. These *stochastic interactive activation* models (McClelland, 1991) constitute an important advance in connectionist models because they show that interactive activation models can exhibit additive effects and because the assumption of noisy processing in the network is consistent with assertions of the biological plausibility of network simulations (Rumelhart & McClelland, 1986). The models implement additive statistical decision models in a fairly straightforward way. The information value of an activation level is relative to the (assumedly normal and constant) variability in the network. Thus, activation levels can easily be related to z-scores. For present purposes, stochastic interactive activation offers a well specified mechanism that exhibits the kind of additive cue effects that have been observed here. In addition, the structural assumptions of interactive-activation models may be useful in accounting for the particular way in which attention appears to operate in phonetic encoding.

Bounded Activations. According to the model shown in Equation 3, attention has a nonlinear effect on the signal-to-noise ratio of phonetic encoding, the loss of attention causes accelerating distortion (loss of information) as the cue values increase. One way in which a processing mechanism might exhibit this increased loss of fidelity at high cue values would be if there were some limits on the representational capabilities of the processing units. Such limits exist within interactive-activation models in the form of the bounds on the activation levels that can be attained by nodes in a network. The assumption of bounded activations is a very common one and is computationally important for many of the attractive properties of multi-stage and recurrent networks (Rumelhart, Hinton, & McClelland, 1986). Thus, the presence of bounds on activation is independently motivated and they provide a mechanism that might produce the nonlinear dependence of phonetic significance on attention level that we have observed.

The accelerating loss of information in the distractor condition could occur because the increased variability of phonetic encoding due to reduced attention would produce a greater number of extreme activation values that would be clipped by the bounds on activation levels. Figure 9 illustrates the workings of this process. The top part of the graph shows density functions for the phonetic encoding of two acoustic cue values, one moderate and the other large. The variance of these distributions is selected so that the occurrence of a value that exceeds the upper or lower bounds is very unlikely. Thus the means of the distributions roughly equal their modes. The bottom part of the graph shows the effect of encoding the same two acoustic cues with greater variability. Both functions now bump into the upper bound to some extent, which produces a clipping of the distribution. That would cause the means of the distributions to shift to the left of their modes (toward the neutral point). This occurs to a greater degree for the stronger cue than the weaker cue, resulting in accelerating information loss as the modal phonetic value of an acoustic cue increases.

An Interactive Model. A stochastic-interactive activation model, simulating this process, was fit to the speech identifications from the distractor condition. The simulation was essentially analogous to the model given in Equation 3 except that the accelerating loss of information was caused by bounds on activation levels rather than by algebraic fiat. Figure 10

shows the organization of the network that was used in the simulation. It has a phonetic property level and a phonetic segment level. The property level contains nodes that receive excitatory external input and encode the phonetic significance of the two acoustic dimensions, formant pattern and vowel duration, for both /i/ and /I/. These property nodes have reciprocal excitatory connections with their parent segment nodes. The two segment nodes are mutually inhibitory. The design of this small network was meant to capture as simply as possible the stimulus dimensions and response options present in the task, while building in interactive processing similar to that of earlier interactive models (McClelland & Rumelhart, 1981; McClelland, 1991; Rumelhart & McClelland, 1982). The use of separate property nodes for each phonetic segment is analogous to the design of the McClelland and Elman (1986) application of interactive activation models to speech.

The operation of the network, with one key exception, is the same as McClelland's (1991) stochastic interactive activation model. All nodes start a trial at their resting level and external activations are applied to the property nodes. The activations of all nodes are then updated through a series of cycles based on continuing external stimulation and the activations of nodes to which they have weighted connections. After a specified number of cycles, the response for the trial is given by the phonetic segment node that has the highest running average activation level. Details of the operation of the network are given in the Appendix. The innovation of McClelland's (1991) model was that in addition to the regular sources of activation, the updating of each node on a cycle included a noise term generated from a normal distribution with a mean of zero. The presence of this noise causes variability across trials in the output node that has the highest activation. The magnitude of the external activations relative to the noise determines the network's performance as a statistical decision process. The use of noise in the present model differed from McClelland's in that it was added to a node's output rather than to its input. The output of a node, including the added noise, was constrained to fall between zero and the node's maximum activation rate of one (as it was in McClelland's model). If the output exceeded either of those bounds, it was set to the bound. This provided the mechanism for clipping extreme activations.

The simulation was fit to the data from the no-distractor condition in the following way. The external input to the phonetic property nodes was derived from the cue values for the no-distractor condition shown in Table 2. These values were linearly re-scaled into positive activation values that were distributed around .5 rather than 0. The input for /I/ property nodes was set to 1 minus the input to the /i/ property nodes. The slope of the scaling function was a free parameter in the model and played a critical role in allowing activation bounds to influence network output. If the slope were very small, then all of the external activations would be clustered tightly around the midpoint (.5) and far from the lower and upper bounds on node outputs (0 and 1 respectively). Thus, a small slope would produce little or no clipping However, a larger slope brings the activation levels closer to the bounds and would allow clipping to play a role. The other free parameters in the model were the standard deviation of the noise and a bias parameter. Changes in the standard deviation of the noise affect the signal-to-noise ratio of perceptual encoding. The bias parameter is analogous to the constant in Equation 3 and was implemented by adjusting the relative resting levels of the two phonetic segment nodes. The response of the network on a trial was based on which phonetic segment node had the highest running average after 20 cycles. The network was run through a 1000 trials with a given external input in order to compute response probabilities. The best possible fit to the observed response proportions in the distractor condition was determined by embedding the network in a minimization algorithm (O'Neil, 1971) that searched the parameter space for the optimum configuration.

The simulation fit the observed data with an RMS error of .075. This is nearly identical to the fit of .074 obtained through Equation 3 which allowed for accelerating information loss, and is better than the fit obtained by Equation 2 which did not allow for accelerating information loss. The fit yielded a slope for the activation scaling parameter of .17 and a standard deviation

for the noise distribution of .24. As a consequence, the activation values for the phonetic cue values of the extreme formant patterns were subject to a substantial amount of clipping. For example, the most extreme cue value (see Table 2) was a z-score of 2. When this is transformed into an activation level by adding .5 (the middle of the activation range) and multiplying by the scaling factor of .17, the result is a modal activation of .84. Given that the standard deviation of the noise distribution is .24, this is .66 standard deviations away from the upper bound on the node's output activation. Thus, the mean output activation of nodes driven by this input will be less than their modal value, given the clipping process produced by the upper bound.

In order to show that clipping by the bounded activations contributed to the observed fit, the network was re-run with the slope of the scaling function fixed at .02 rather than at the optimum value of .17. The use of this smaller slope means that clipping of the output activations by the activation bounds were very rare. The resulting simulation fit the data with an RMS error of .08. This fit is identical to that obtained in the algebraic model (Equation 2) in which information loss was proportional to cue value. The difference between the fit of this model and the previous one demonstrates that clipping by the activation bounds is the source of the accelerating information loss that occurs in the distractor condition.

In addition to showing that clipping enhances the fit to the distractor data, it is important to show that clipping does not impair simulation of the data from the no-distractor condition. To do so, the slope of the scaling function was fixed at the value of .17 obtained from the best fit to the distractor data, and the network was fit to the no-distractor data by changing the standard deviation of the noise distribution. The resulting simulation fit the no-distractor data with an RMS error of .028, not far off the fit of .027 that was obtained when the no-distractor data were fit independently by an algebraic model that included no bounds. The standard deviation of the noise distribution for the no-distractor condition was .18, which is a third smaller than for the distractor condition. This smaller amount of noise meant that there was less clipping than in the distractor simulation, and enabled the network to accurately model the no-distractor data.

The above simulations show that stochastic interactive activation models can do an excellent job of instantiating statistical decision models that additively combine different sources of information (McClelland, 1991; cf. Massaro, 1989). Further, they show that bounds on activations provide a viable mechanism for producing the pattern of accelerating information loss that had been shown by our algebraic models to provide a parsimonious account of the effect of reduced attention. By providing an independently motivated basis with which to account for the accelerating loss of information, the structural and dynamic characteristics of interactive-activation models clearly take us beyond the statistical decision models in our understanding of the role of attention in phonetic perception.

Mechanisms of Enhancing Signal-to-Noise Ratios. In the algebraic models, attention level influences the signal-to-noise ratio in encoding phonetic information, reduced attention lowers the signal-to-noise ratio while increased attention enhances the signal-to-noise ratio. In the simulations described above, phonetic information was represented as the activation level of network units relative the noise in the network. Changes in signal-to-noise ratio due to attention level were achieved by holding the modal level of activation associated with an acoustic cue constant and varying the amount of noise in the network. The simulations have not motivated this choice of how signal-to-noise ratio is changed, nor have they specified the mechanism by which attention level influences the amount of variability in encoding. These issues raise interesting questions even though the present data do not lead to definitive answers.

One way attention might influence noise would be if phonetic encoding on a trial involved multiple samples of a stimulus with the sampling process having a constant variance,

independent of attention level, but with attention level influencing the extent of the sampling. As sampling increased, the variance of the mean of the samples would decrease. Because it is the variance of the mean of the phonetic value for an acoustic cue that determines its information value, this mechanism could produce changes in signal-to-noise ratio without attention directly affecting the level of noise in the network.

An interesting way in which this might occur would be if attention controlled the duration of the link-up between acoustic cues and phonetic encoding units. There would be less variability in the mean output of the phonetic unit stimulated over a long period of time than over a short period of time. Experiment 2 was designed to test something like this possibility by looking for a differential effect of reduced attention in the phonetic encoding of spectral cues to formant pattern in short (50 msec) and long (300 msec) vowels. The short vowels would presumably offer less opportunity for sampling than the long vowels, which could compound problems in attention-based temporal hook-up between phonetic units and acoustic stimuli or their auditory representations. The results of Experiment 2 (as well as Experiments 3 and 4) showed that encoding formant information was not differentially difficult in the short-duration vowels, thereby providing no support for the hypothesis. However, it is possible that the temporal linkage is more dependent on auditory memory than on the physical duration of the stimulus. This possibility was discussed earlier and it was pointed out that several researchers have argued that the duration of an auditory memory is positively related to the duration of the acoustic stimulus (e.g., Fujasaki & Kawashima, 1969; 1970). However, this analysis was based on inferences drawn from a specific model of the relation between discrimination and categorical perception. Subsequent research has called this model into question (Repp, Healy, & Crowder, 1979). Further, Pisoni (1973) showed that auditory memory for short and long vowels showed similar persistence as indicated by the rate at which performance in an AX discrimination task decreased with increasing delay within the pair of stimuli. Thus, Experiment 2 may not have effectively manipulated the persistence of information in auditory memory, leaving open the possibility that the duration of acoustic stimulation of phonetic units may play an important role in determining the variability of phonetic encoding.

A second (and not exclusive) way that attention might influence signal-to-noise ratios in perceptual encoding is by modulating the amplification of signal characteristics at a stage of processing before additional noise is encountered in transmitting stimulus information. Servan-Schreiber, Printz and Cohen (1990) have investigated the way in which catecholamines might be related to signal detection behavior. These neuroactive substances have been found to increase the responsiveness of individual neurons and, in studies involving drugs or pathological conditions, to influence observable signal detection performance. Servan-Schreiber et al. (1990) have pointed out that an increase in a cell's input-output gain can not of itself lead to improved signal-to-noise performance. The increased gain will apply to noise in the input as well as the signal. However, increased gain by a unit can improve a network's signal-to-noise performance if that gain occurs before additional noise is added to processing. While Servan-Schreiber et al. (1990) show that this truth holds for any strictly increasing gain function, they investigate it in detail for logistic gain functions, which they treat as a model of neural activity. These functions are S-shaped, and as the gain parameter increases the function becomes sharper, approaching a square wave. This family of gain functions can be used to model the operation of attention in phonetic encoding. However, its performance will be nearly identical to the algebraic models developed here which were based on an assumption of normally-distributed noise (see McClelland, 1991). Adjusting the gain parameter of a logistic function is directly analogous to adjusting the attention scaling parameter (A) in Equation 2, which modeled the effect of attention as a proportional loss (or gain) in the information value of the acoustic cues. Such models do not produce the accelerating loss of information that occurs in Equation 3 or in the network model developed above. Of course, the Servan-Schreiber et al (1990) analysis applies to any strictly increasing function. In its domain of current application, Equation 3 produces a strict increase in information for the no-distractor as compared to the distractor condition and thus could be taken as a gain function underlying attentional amplification of signals. However,

at least as currently developed, models of attention based on moderating the gain of a network unit's response do not offer much insight into the exact form of information modulation that was observed in the present studies and is captured in Equation 3.

Summary of the Modeling

The analyses above demonstrate that the role of attention in phonetic perception is well accounted for by models in which attention level affects the signal-to-noise ratio in the phonetic encoding of acoustic cues. This feature is shared by the three statistical decision models (Equations 1 - 3) and the network simulation. These models differ in how changes in the precision of phonetic encoding due to attention interact with specific acoustic-phonetic relations. In the model expressed by Equation 1, attention has differential effects on the formant pattern and duration cues to vowel identity. In the subsequent models, attention has equivalent effects on the underlying processing of these two cues, but the inherent strength of the acoustic-phonetic relations produces the attentional differences apparent in the response probabilities. The effect of reduced attention in the first of these models (Equation 2) is to cause a proportional loss of information in the underlying cue representations. In the next model (Equation 3), which achieves a better fit, reduced attention causes accelerating loss of information in the underlying cue representations. The network simulation produces this accelerating loss of information because increased noise in network activity due to reduced attention results in the bounds on the representational capability of network units having a significant contribution to the network's performance.

The extent to which attention level produces nonlinear changes in perceptual significance is important in assessing the relative merits of the different models. The models incorporating this feature (Equation 3 and the network simulation) give better fits to the data than does the model involving proportional change in perceptual significance (Equation 2). Further, one of the main virtues of implementing the statistical decision model as a stochastic interactive activation model was to see whether principles intrinsic to such models might account for the nonlinear effect of attention level on phonetic significance that was described but unmotivated by the algebraic model. The existence of bounds on the maximum and minimum activation rates of network units provided an independently motivated mechanism that can produce just such an effect. This demonstration has the further merit of indicating an important circumstance in which stochastic interactive activation models will not behave as additive statistical decision models. An understanding of the ability of parallel network models to exhibit both additive and non-additive effects is essential to the evaluation of connectionist models of perception (Massaro, 1989, McClelland, 1991). For these reasons, models incorporating nonlinear effects of attention are very interesting and have been explored in detail. However, it is important to note that the improvement in fit to the data of these models compared to the proportional information loss model is not large. Application of these principles to other data sets will be necessary in evaluating the ultimate value of this account of the effect of attention on perceptual encoding.

General Discussion

The experiments reported here have shown that attention plays a role in the perception of phonetic segments and that the relative importance of acoustic cues depends on the amount of attention that is devoted to the speech stimulus. Experiment 1 showed that the strong voicing cue of voice-onset time decreased in phonetic importance under low attention levels, while the weak voicing cue of F0 onset frequency maintained its phonetic contribution. Experiment 2 showed that the strong formant pattern cue to the distinction between /i/ and /I/ also decreased in phonetic importance under low attention levels, while the weak cue of vowel duration actually increased its net contribution to phonetic perception. Both these experiments

used an arithmetic distractor task that very likely placed demands on short-term verbal memory. Experiment 3 showed that the general pattern of speech perception results could be obtained when the distractor task consisted of a line-length discrimination that placed less demand on short-term memory. Experiment 4 provided a large data set with which to test quantitative models of the role of attention in the phonetic encoding of acoustic cues. Below, we consider the implications of these findings for understanding speech perception. Then, various facets of the current modeling are reviewed and their implications for general issues in theories of attention are discussed.

Implications of the Findings for Theories of Speech Perception

The clearest implication of the present results is that patterns of phonetic cue importance obtained under conditions of focused listening should not be taken as definitive. It seems very likely that listeners ordinarily focus less attention on phonetic perception than they do in the laboratory and that under low attention conditions, there is an increased contribution of weak cues relative to strong cues. This finding forces the conclusion that speech perception is more dependent on multiple cues than was previously believed, and that it is unlikely that a single strong cue is generally dominant in recognition.

This conclusion may shed some light on the difficulties listeners have in understanding synthetic speech (Pisoni & Hunnicut, 1980). Pisoni (1981; discussed in Luce et al., 1983) has suggested that part of this difficulty stems from the limitations of synthesis-by-rule systems in generating the large variety of acoustic cues that are found in natural speech. Under the model outlined here, all relevant acoustic cues contribute to phonetic perception. A single strong cue can lead to high levels of recognition, but only if careful attention is given to the stimulus. Thus, synthetic speech which successfully encoded only the strong acoustic cues to segment identity would place heavy demands on attentional resources in order to be recognized successfully – an analysis that is consistent with the findings of Luce et al., (1983).

In a less artificial vein, these findings may also have implications for theories of sound change. Ohala and others (Javkin, 1979; Ohala, 1988) have articulated a model of sound change as arising from propagation of error in speech communication considered as a transmission line. According to this view, differences may occur between a speaker's phonetic intention and a listener's phonetic perception based on various kinds of distortions in the transmission process. When the listener then takes a turn as a speaker, these distorted phonetic impressions may then be introduced into the process as novel elements of phonetic intentions. Javkin (1979) identifies three sources of transmission distortions. articulatory errors, acoustic interference, and biases in auditory perception. He points out that the great majority of phonological analysis has focused on the role of articulatory error, though he produces interesting evidence to support a role for auditory processes in phonological change. The present analysis of the role of attention in phonetic perception offers additional clues as to how auditory perception might participate in phonological change.

The view of phonetic perception adopted here considers phonetic encoding of acoustic cues as a noisy process, and shows that the amount of encoding noise increases under low attention. This process has the interesting consequence of increasing the net contribution to a phonetic percept of weak cues relative to strong cues. Low attention would therefore be a factor that promotes equalization of the phonetic values of acoustic cues. Of course, this process can not explain how a dominant acoustic cue is created from a weak one, but it does point to how an originally insignificant acoustic correlate might achieve enough phonetic status as a cue that it might be acted upon by other forces of sound change. As analyzed here, the effect of attention on phonetic perception would be neutral with regard to the direction of sound change; the addition of noise is symmetric and does not push the perception of a phonetic segment in any particular direction. However, noisy phonetic encoding due to low

attention could combine with directional forces to accelerate sound change. These directional forces could be internal to the perceiver, as in the cases of perceptual biases studied by Javkin (1979). Or, they could derive from effects of phonetic context on the acoustic structure of phonetic segments.

Our analysis of the present results has characterized attention as operating at a phonetic rather than auditory level of processing. This characterization results from the finding that the dependence of an acoustic cue on attentive processing relates to its potential phonetic strength. This provides the most straightforward unification of our results concerning the VOT and F0 onset frequency cues to stop consonant voicing and the formant pattern and duration cues to vowel identity. In both cases, the differential effects of attention were related to the magnitude of the potential phonetic strength of the acoustic cues. The strong cues, VOT and formant pattern, required close attention to achieve their commanding phonetic significance. The weak cues, F0 onset frequency and vowel duration, did not require careful attention in order to maintain their potential phonetic significance. Further, several models of the results of Experiment 4 (Equations 2 and 3 and the network simulation) show how attentional processing at a phonetic level could interact with phonetic strength in order to produce the observed results. Given the present findings, this phonetic level of explanation is much more complete than the alternative of accounting for the role of attention in the auditory processing of each of the individual acoustic cues. For this reason, we prefer a phonetic level of explanation at present.

Still, it is worth considering the possibility that properties specific to the individual acoustic cues account for the differential effect of attention. The best fit to the distractor data of Experiment 4 was obtained by the model (Equation 1) that used different attention parameters to scale the information loss for the formant pattern and duration cues to vowel identity. This is not surprising since the model included an additional free parameter compared to the latter models. Further, a difference in the processing of the two cues can not be explained by such a model *per se* because the model builds in differential treatment of the cues. Motivation for this differential treatment must come from outside the model. This leads to the question of whether there are any psychoacoustic factors that might make formant pattern and VOT depend heavily on close attention, whereas vowel duration and F0 onset frequency do not. On the face of it, there is little to suggest that this is the case. Of the attention-dependent cues, formant pattern is primarily a spectral cue while VOT is primarily a temporal cue. Similarly, for the cues that are not attention dependent, F0 onset frequency is a spectral cue while vowel duration is a temporal cue. In making this comparison, we do not mean to imply that formant pattern and F0 onset frequency are processed by one auditory mechanism for spectral processing while vowel duration and VOT are processed by a separate auditory mechanism for temporal processing. We only mean to point out that dependence on attentive processing does not relate to the most obvious dimension of psychoacoustic similarity among the acoustic cues that we studied. Unless developed in unforeseen directions, this kind of explanation does not compete with explanations based on potential phonetic strength.

There is, however, one other dimension of analysis that correlates with degree of attention dependence. In addition to being cues to segment identity, both vowel duration and F0 onset frequency are prosodic cues, while formant pattern and VOT are not. A great deal of research indicates that prosodic patterns play an important role in the process of selecting a speech signal because they provide some continuity of the signal over time (Bregman, 1990). Accordingly, a first goal of the speech perception system may be to latch onto prosodic cues because they provide a basis for continuing signal selection (in general, if not in the laboratory perception of monosyllables). If this were the case, prosodic cues might be processed for a longer time than non-prosodic cues, giving them an advantage for accurate encoding (recall that one way in which the signal-to-noise ratio of perceptual encoding can be reduced is by increasing the size of the sample of the signal). When attention can be focused on speech

perception, this early advantage may be overcome by sustained sampling of the signal which would allow the full significance of the non-prosodic cues to be encoded. When attention is not devoted exclusively to speech perception, the non-prosodic cues may not be encoded with the precision necessary to allow them to overcome their initial disadvantage relative to the prosodic cues. As noted above, this account is speculative. Its main appeal is that it potentially unifies two facets of attention: selection and processing facilitation. However, until it receives further investigation we prefer the phonetic level of explanation advanced earlier.

One characteristic of all of the models described above is that attention operates early in perceptual processing. This characteristic is shared with other work indicating a role of attention in early selection of perceptual inputs (Broadbent, 1958; Kleiss & Lane, 1986; Pashler, 1984). It is not shared by theories that stress that attentional selection occurs late in processing (Deutsch & Deutsch, 1963; Duncan, 1980; Shiffrin & Gardner, 1972). The findings that motivate the current modeling are not consistent with the notion that the extraction of perceptual features is preattentive (Neisser, 1967). That idea is embodied in current theories that propose that featural information is extracted automatically and in parallel, and that the role of attention in perception is to integrate features into objects (Treisman & Gelade, 1980). The present work indicates that the accuracy with which features are encoded depends on the available attentional resources and describes a mechanism that manifests this dependence. We believe that this demonstrates some good reasons for paying attention to phonetic perception.

REFERENCES

Abramson, A.S., & Lisker, L. (1985). Relative power of cues: F0 shift versus voice timing. In V.A. Fromkin, **Phonetic linguistics: Essays in honor of Peter Ladefoged** (pp. 25-33). New York: Academic Press.

Anderson, N.H. (1974). Information integration theory: A brief survey. In D.H. Krantz, R.C. Atkinson, R.D. Luce, & P. Suppes (Eds.), **Contemporary developments in mathematical psychology (Vol 2)**. San Francisco, Freeman.

Bernstein, L.E. (1983). Perceptual development for labeling words varying in voice onset time and fundamental frequency. **Journal of Phonetics, 11**, 383-393.

Broadbent, D.E. (1958). **Perception and Communication.** London: Pergamon.

Bregman, A.S. (1990). **Auditory scene analysis: The perceptual organization of sound.** Cambridge, MA: MIT Press.

Bryden, M.P. (1971). Attentional strategies and short-term memory in dichotic listening. **Cognitive Psychology, 2**, 99-116.

Buxton, H. (1983). Temporal predictability in the perception of English speech. In A. Cutler & D.R. Ladd (Eds.), **Prosody: Models and Measurements**, Berlin: Springer-Verlag, 111-122.

Cherry, E.C. (1953). Some experiments on the recognition of speech with one and two ears. **Journal of the Acoustical Society of America, 26**, 975-979.

Crowder, R.G. (1982). Decay of Auditory Memory in Vowel Discrimintation. **Journal of Experimental Psychology: Learning, Memory, and Cognition, 8**, 153-162.

Crowder, R.G. (1981). The role of auditory memory in speech perception and discrimination. In T.Myers, J.Laver, & J.Anderson (Eds.), **The cognitive representation of speech.** Amsterdam: North-Holland, 167-179.

Crowder, R.G., & Morton, J. Precategorical acoustic storage (PAS), **Perception & Psychophysics, 5**, 365-373.

Cutler, A. (1976). Phoneme monitoring reaction time as a function of preceding intonation contour. **Perception & Psychophysics, 20**, 55-60.

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

Deutsch, J.A. & Deutsch, D. (1963). Attention. Some theoretical considerations. **Psychological Review, 70**, 80-90.

Diehl, R.L. (1987). Auditory constraints on speech perception. In M.E.H. Schouten (Ed), **The psychophysics of speech perception.** Boston, MA. Martinus Nijhoff, 210-219.

Duncan, J. (1980). The locus of interference in the perception of simultaneous stimuli. **Psychological Review, 87**, 272-300.

Egan, J., Carterette, E., & Thwing, E. (1954). Some factors affecting multichannel listening. **Journal of the Acoustical Society of America, 79,** 838-845.

Fujisaki, H., & Kawashima, T. (1969). On the modes and mechanisms of speech perception. **Annual Report of the Engineering Research Institute, Vol. 28,** Faculty of Engineering, University of Tokyo, Tokyo, 67-73.

Fujisaki, H., & Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. **Annual Report of the Engineering Research Institute, Vol. 29,** Faculty of Engineering, University of Tokyo, Tokyo, 207-214.

Glucksberg, S., & Cowen, G.N., Jr. (1970). Memory for nonattended auditory material. **Cognitive Psychology, 1,** 149-156.

Gordon, P.C. (1989). Context effects in recognizing syllable-final /z/ and /s/ in different phrasal positions. **Journal of the Acoustical Society of America, 86,** 1698-1707.

Greenberg, G.Z., & Larkin, W.D. (1968). Frequency-response characteristics of auditory observers detecting signals of a single frequency in noise: The probe-signal method. **Journal of the Acoustical Society of America, 30,** 904-911.

Javkin, H.R. (1979). **Phonetic universals and phonological change.** Unpublished doctoral dissertation. University of California, Berkeley.

Kleiss, J.A., & Lane, D.M. (1986). Locus of persistence of capacity limitations in visual informatin processing. **Journal of Experimental Psychology: Human Perception and Performance, 12,** 200-210.

Kahneman, D. (1973). **Attention and Effort.** Englewood Cliffs, NJ: Prentice-Hall.

Klatt, D.H. (1976). Linguistic use 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, NUJ: Erlbaum.

Liberman, A.M. (1982). On finding that speech is special. **American Psychologist, 37,** 148-167.

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

Liberman, A.M., & Mattingly, I.G. (1985). The motor theory of speech perception revised **Cognition, 21,** 1-36.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J.J. Ohala, & J.J. Jaeger (Eds.), **Experimental phonology.** New York, NY: Academic Press, 13-44.

Lindholm, J.M., Dorman, M., Taylor, B.E., & Hannley, M.T. (1988). Stimulus factors influencing the identification of voiced stop consonants by normal-hearing and hearing-impaired adults. **Journal of the Acoustical Society of America, 83,** 1608-1614.

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

Luce, P.A., Feustel, T. C., & Pisoni, D.B. (1983). Capacity demands in short-trm memory for synthetic and natural speech. **Human Factors, 25**, 17-32.

Martin, J.G. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior. **Psychological Review, 79**, 487-509.

Massaro, D.W. (1989). Testing between the TRACE Model and the Fuzzy Logical Model of speech perception. **Cognitive Psychology, 21**, 398-421.

McClaskey, C.L., Pisoni, D.B., & Carrell, T.D. (1983). Transfer of training of a new linguistic contrast in voicing. **Perception & Psychophysics, 34**, 323-330.

McClelland, J.L. (1991). Stochastic interactive processes and the effect of context on perception. **Cognitive Psychology, 23**, 1-44.

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

McClelland, J.L., & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. **Psychological Review, 88**, 375-407.

Meltzer, R.H., Martin, J.G., Mills, C.B., Imhoff, D.L.,& Zohar, D. (1976). Anticipatory coarticulation and reaction time to phoneme targets in spontaneous speech. **Phonetica, 37**, 159-168.

Meltzer, R.H., Martin, J.G., Mills, C.B., Imhoff, D.L.,& Zohar, D. (1976). Anticipatory coarticulation and reaction time to phoneme targets in spontaneous speech. **Phonetica, 37**, 159-168.

Mens, L.H., & Povel, D.J. (1986). Evidence against a predictive role for rhythm in speech perception. **The Quarterly Journal of Experimental Psychology, 38A**, 177-192.

Neisser, U. (1967). **Cognitive psychology**. New York: Appleton-Century-Crofts.

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.

Oden, G.C. (1977). Fuzziness in semantic memory: Choosing exemplars of subjective categories. **Memory & Cognition, 5**, 190-204.

Oden, G.C. (1979). A fuzzy logical model of letter identification. **Journal of Experimental Psychology: Human Perception and Performance, 5**, 336-352.

Oden, G.C., & Massaro, D.W. (1978). Integration of featural information in speech perception **Psychological Review, 85**, 172-191.

O'Neill, R. (1971). Function minimization using a simplex procedure. Algorithm AS 47. **Applied Statistics**, 338.

Pashler, H. (1984). Evidence against late selection: Stimulus quality effects in previewed displays. **Journal of Experimental Psychology: Human Perception and Performance**, 10, 429-448.

Pashler, H. (1989). Dissociations and dependencies between speed and accuracy. Evidence for a two-component theory of divided attention in simple tasks. **Cognitive Psychology**, 21, 469-514.

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

Pisoni, D.B. (1971). On the nature of categorical perception of speech sounds. Unpublished doctoral thesis. University of Michigan.

Pisoni, D.B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. **Perception & Psychophysics**, 13, 253-260.

Pisoni, D.B. (1975). Auditory short-term memory and vowel perception. **Memory & Cognition**, 3, 7-18.

Pisoni, D.B., & Hunnicut, S. (1980). Perceptual evaluation of MITask: The MIT unrestricted text-to-speech system. In **1980 IEEE International Conference Record on Acoustics, Speech and Signal Processing**. New York: IEEE.

Price, P.J., & Simon, H.J. (1984). Perception of temporal differences in speech by "normal-hearing" adults: Effects of age and intensity. **Journal of the Acoustical Society of America**, 76, 405-410.

Rumelhart, D.E., & McClelland, J.L. (1982). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. **Psychological Review**, 89, 60-94.

Rumelhart, D.E., Hinton, G.E., & McClelland, J.L. (1986). A general framework for parallel distributed processing. In Rumelhart, D.E., & McClelland, J.L. (Eds.), **Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations**. Cambridge, MA: MIT Press.

Repp, B.H., Healy, A.F., & Crowder, R.G. (1979). Categories and context in the perception of isolated steady-state vowels. **Experimental Psychology: Human Perception and Performance**, 5, 129-145.

Scharf, B., Quigley, S., Aoki, C., Peachey, N., & Reeves, A. (1987). Focused auditory attention and frequency selectivity. **Perception and Psychophysics**, 42, 215-223.

Servan-Schreiber, D., Printz, H., & Cohen, J.D. (1990). A network model of catecholamine effects: Gain, signal-to-noise ratio, and behavior. **Science**, 249, 892-895.

Shields, J.L., McHugh, A., & Martin, J.G. (1974). Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. **Journal of Experimental Psychology**, 102, 250-255.

Shiffrin, R.M., & Gardner, G.T. (1972). Visual processing capacity and attentional control. **Journal of Experimental Psychology, 93,** 72-82.

Shiffrin, R.M., & Grantham, D. (1974). Can attention be allocated to sensory modalities? **Perception & Psychophysics, 15,** 460-474.

Shiffrin, R.M., Pisoni, D.B., & Castaneda-Mendez, K. (1974). Is attention shared between the ears? **Cognitive Psychology,** 6, 190-215.

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.

Swets, J.A. (1963). Central factors in auditory frequency selectivity. **Psychological Bulletin, 60,** 429-440.

Swets, J.A. (1984). Mathematical models of attention. In R. Parasuraman & D.R. Davies (Eds.), **Varieties of Attention** (pp. 183-242). New York: Academic Press.

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

Treisman, A.M., & Gelade, G. (1980). A feature-integration theory of attention. **Cognitive Psychology, 12,** 97-136.

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.

Woodworth, R. (1938). **Experimental psychology.** New York: Henry Holt and Company.

Appendix

The operation of the network follows the general principles discussed in Rumelhart et al. (1986) and followed in McClelland (1991). At the beginning of a trial, all nodes are set to their resting levels and appropriate external activations are applied to relevant nodes. At each time step in processing, the input to a node i is computed as follows:

$$net_i = \sum_j w_{ji} o_j + ext_i.$$

Here, w_{ji} is the weight of the connection from node j to node i, o_j is the output of node j, and ext_i is the external input. Given the net input, the activation of a node is updated using the Rumelhart-McClelland rule:

If ($net_i > 0$) then:

$$\Delta(a_i) = I(M - a_i)net_i - D(a_i - r)$$

or else:

$$\Delta(a_i) = I(a_i - m)net_i - D(a_i - r)$$

The constant M is the maximum activation rate, m is the minimum activation rate, r is the resting activation level, and I and D respectively scale the effects of input and decay. The values for the various constants were taken from McClelland (1991) who characterized them as generic. Excitatory connections were set at weights of 1, while inhibitory connections had weights of -1, $M = 1$; $m = -.2$; $r = -.1$; $I = .1$ and $D = .1$. The final response was selected as the output unit with the highest "running average", using the following formula (McClelland, 1991):

$$a_i(t) = (\lambda)o_i(t) + (1 - \lambda)a_i(t-1)$$

where a equals the running average, t equals the time step, o is a unit's output, and the parameter λ is set to 0.05.

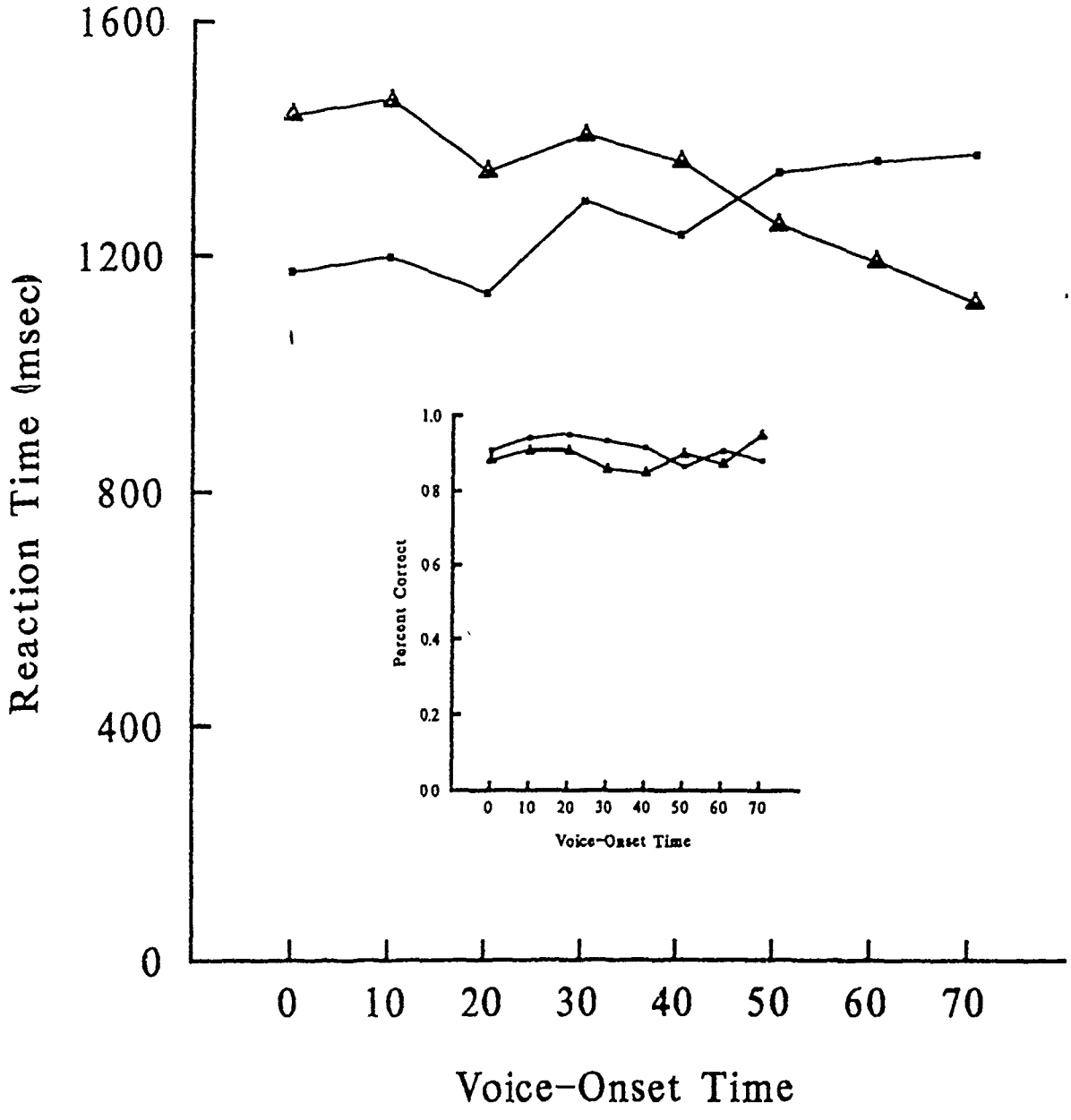
Author Notes

The research reported in this paper was supported by grant 80-0416 from the Life Sciences Directorate of the Air Force Office of Sponsored Research to Harvard University (P.C Gordon, Principal Investigator). Direct correspondence to P.C. Gordon, Department of Psychology, Harvard University, 33 Kirkland St., Cambridge, MA 02138.

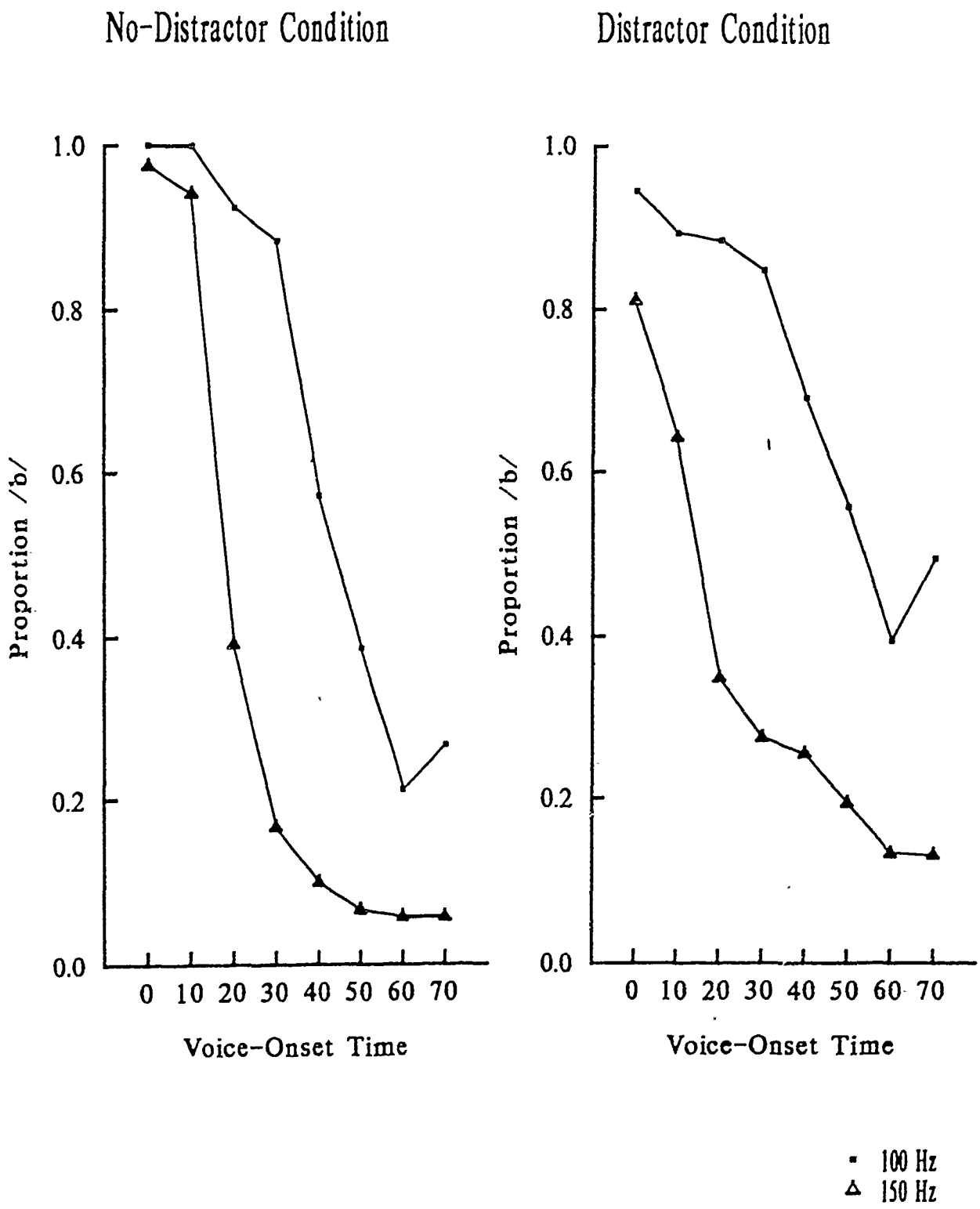
Table 1: Vowel Formant Frequencies. These are the center frequencies of the first three formants of the seven-member continuum from /i/ to /I/. The fourth and fifth formants were set at 3500 Hz and 4500 Hz respectively. The stimuli were closely modeled after ones used by Pisoni (1975).

		Formant Frequencies (Hz)		
	Stimulus Number	F1	F2	F3
/i/	1	270	2300	3019
	2	285	2262	2960
	3	298	2226	2902
	4	315	2180	2836
	5	336	2144	2776
	6	353	2103	2719
/I/	7	374	2070	2666

Fig 1



▪ 100 Hz
△ 150 Hz


No-Distractor Condition
Distractor Condition
Proportion /b/
1.0
0.8
0.6
0.4
0.2
0.0
0 10 20 30 40 50 60 70
Voice-Onset Time
100 Hz
150 Hz

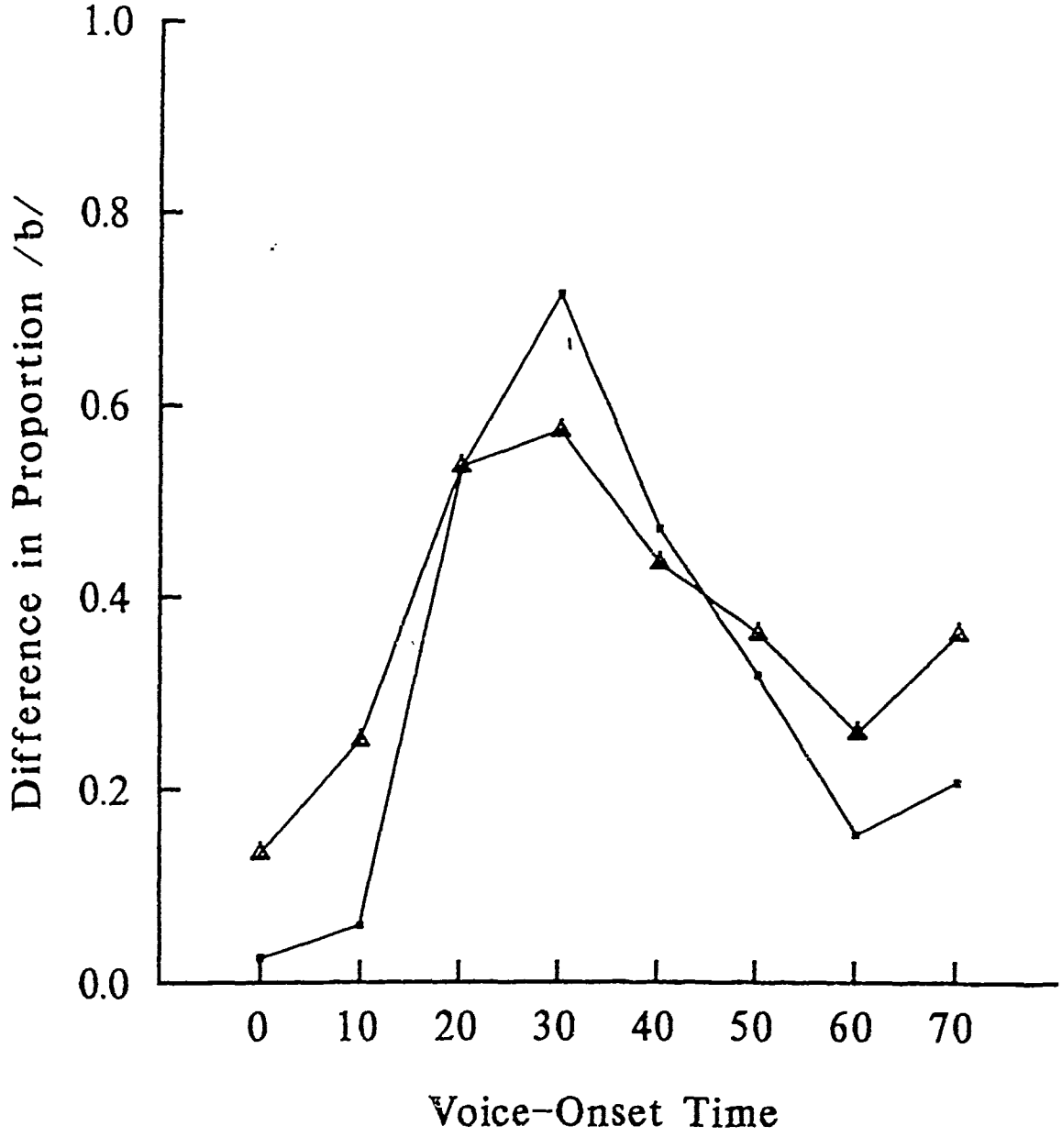
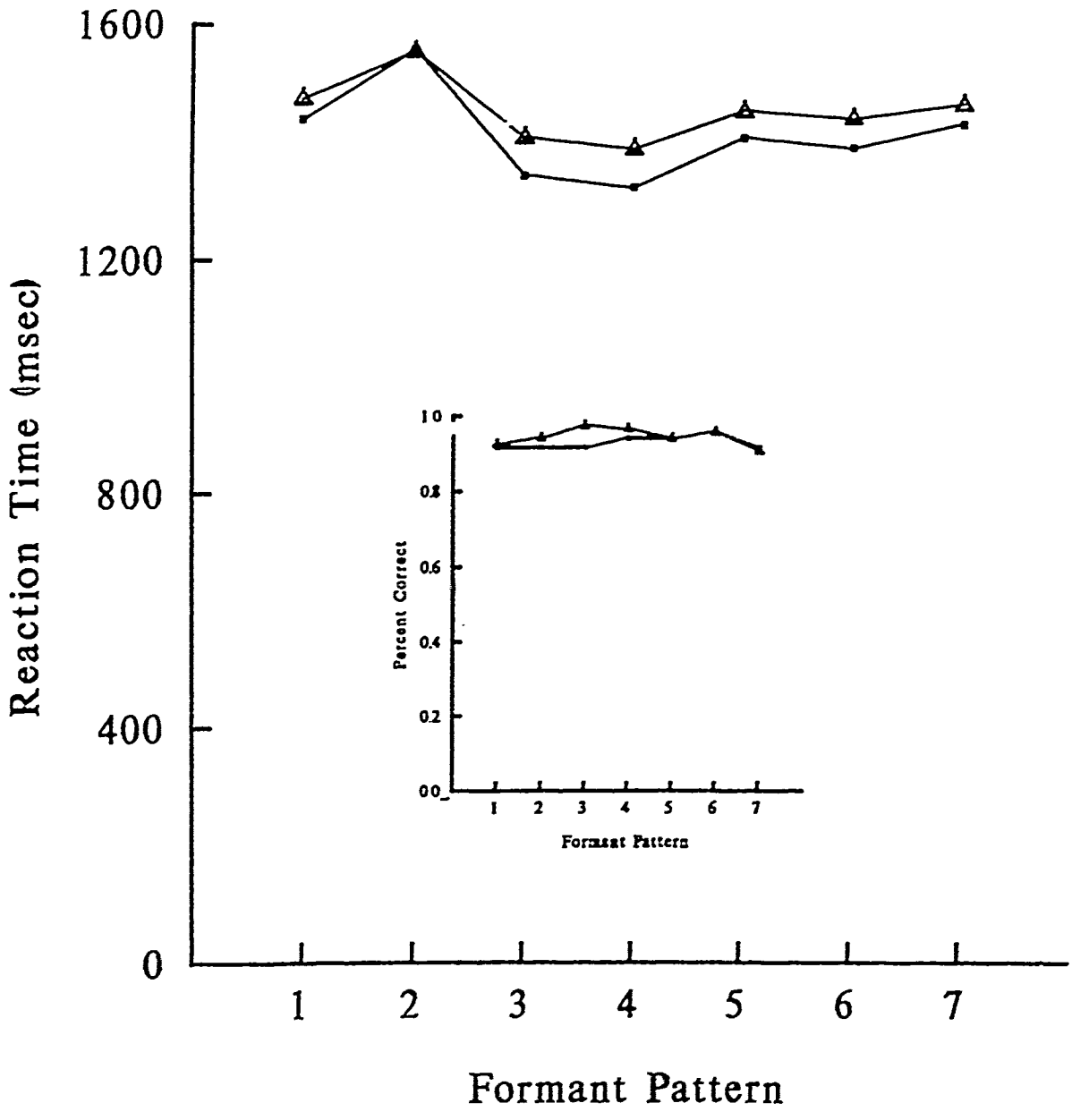

1.0
0.8
0.6
0.4
0.2
0.0
Difference in Proportion /b/
0
10
20
30
40
50
60
70
Voice-Onset Time
No-Distractor
Distractor

Fig 4



▪ 60 msec
△ 300 msec

Fig 5

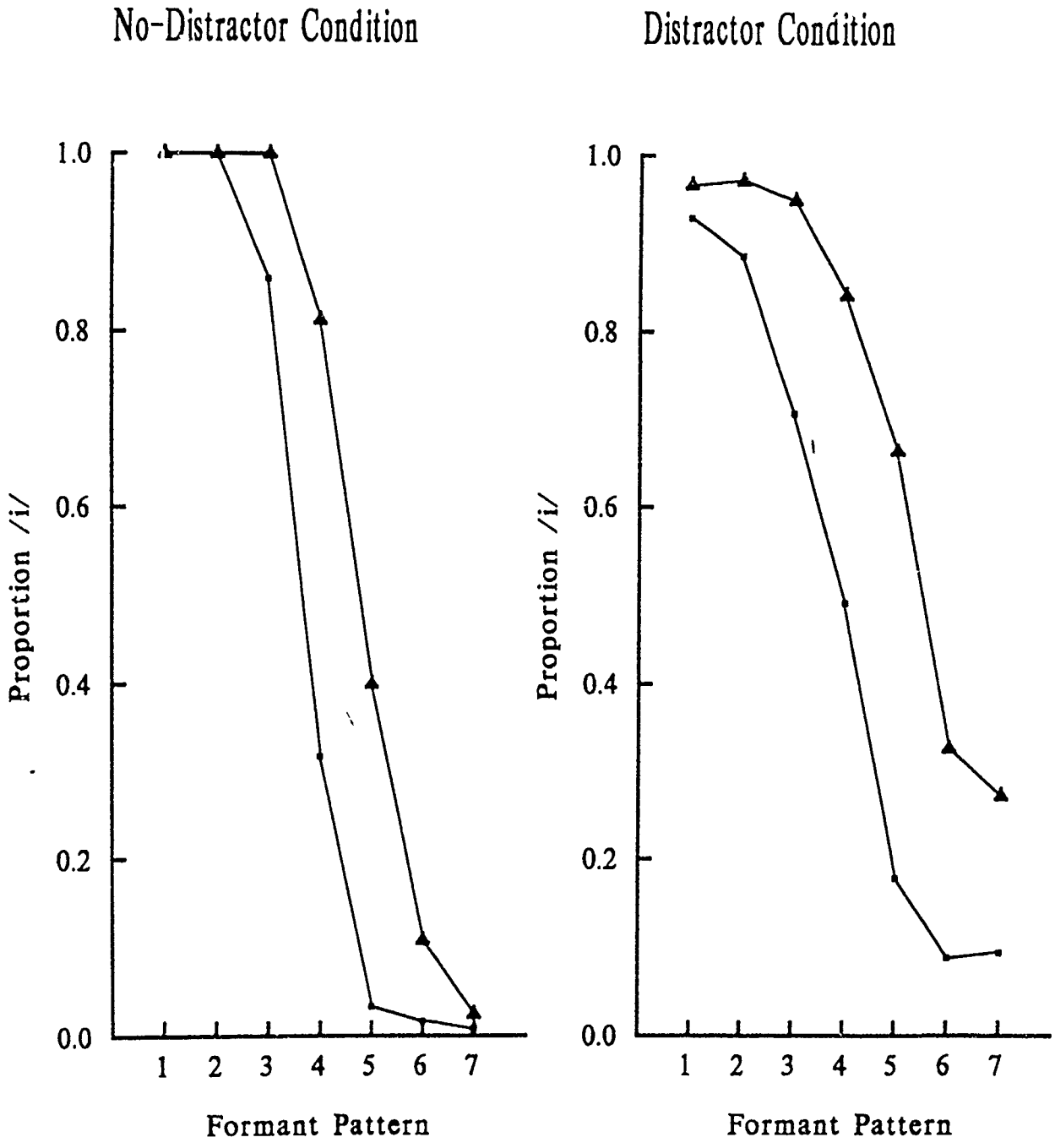
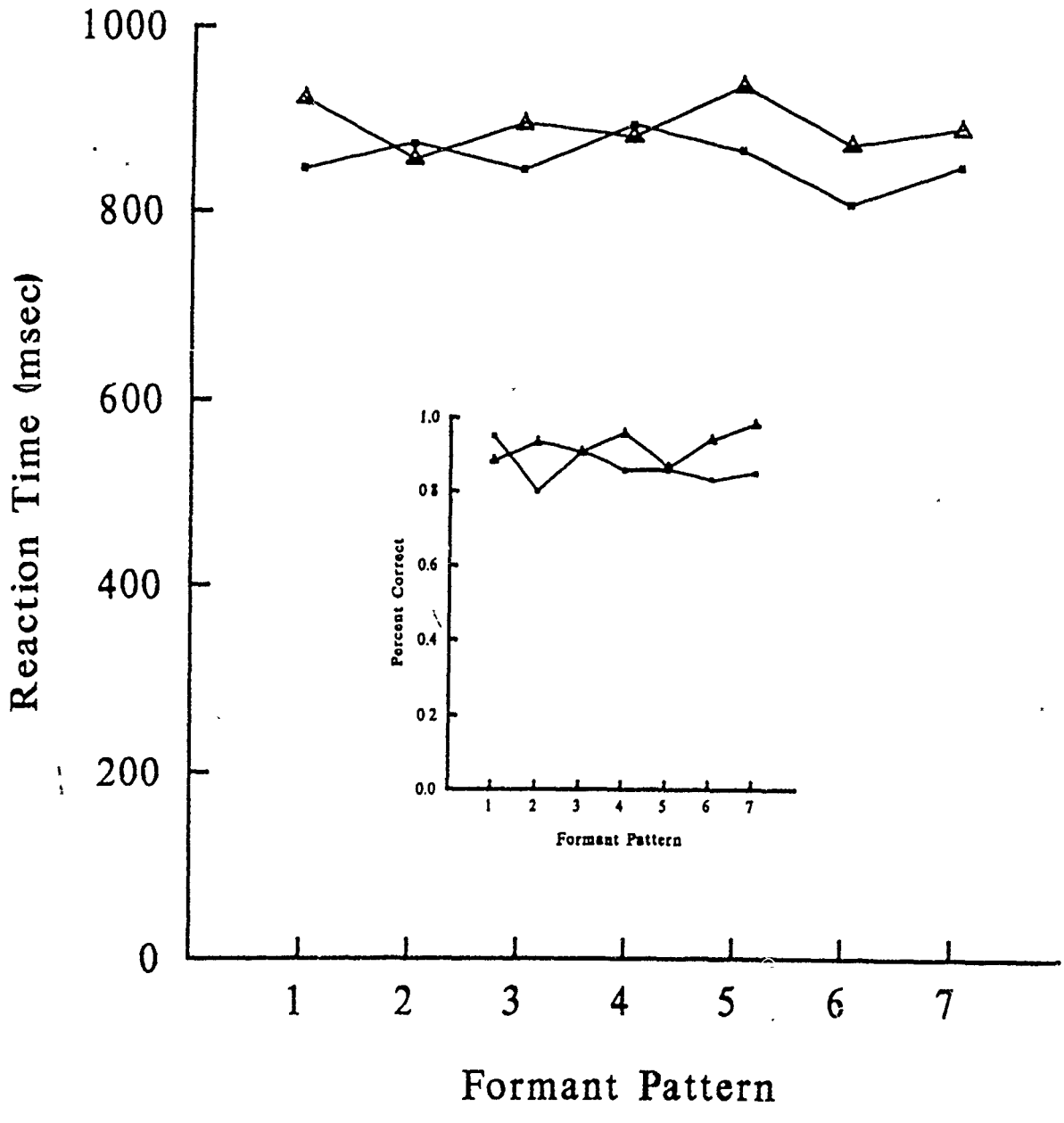

No-Distractor Condition
Distractor Condition
1.0
0.8
0.6
0.4
0.2
0.0
Proportion /i/
1 2 3 4 5 6 7
Formant Pattern
50 msec
300 msec

Fig 6



▪ 60 msec
△ 300 msec

Fig 7

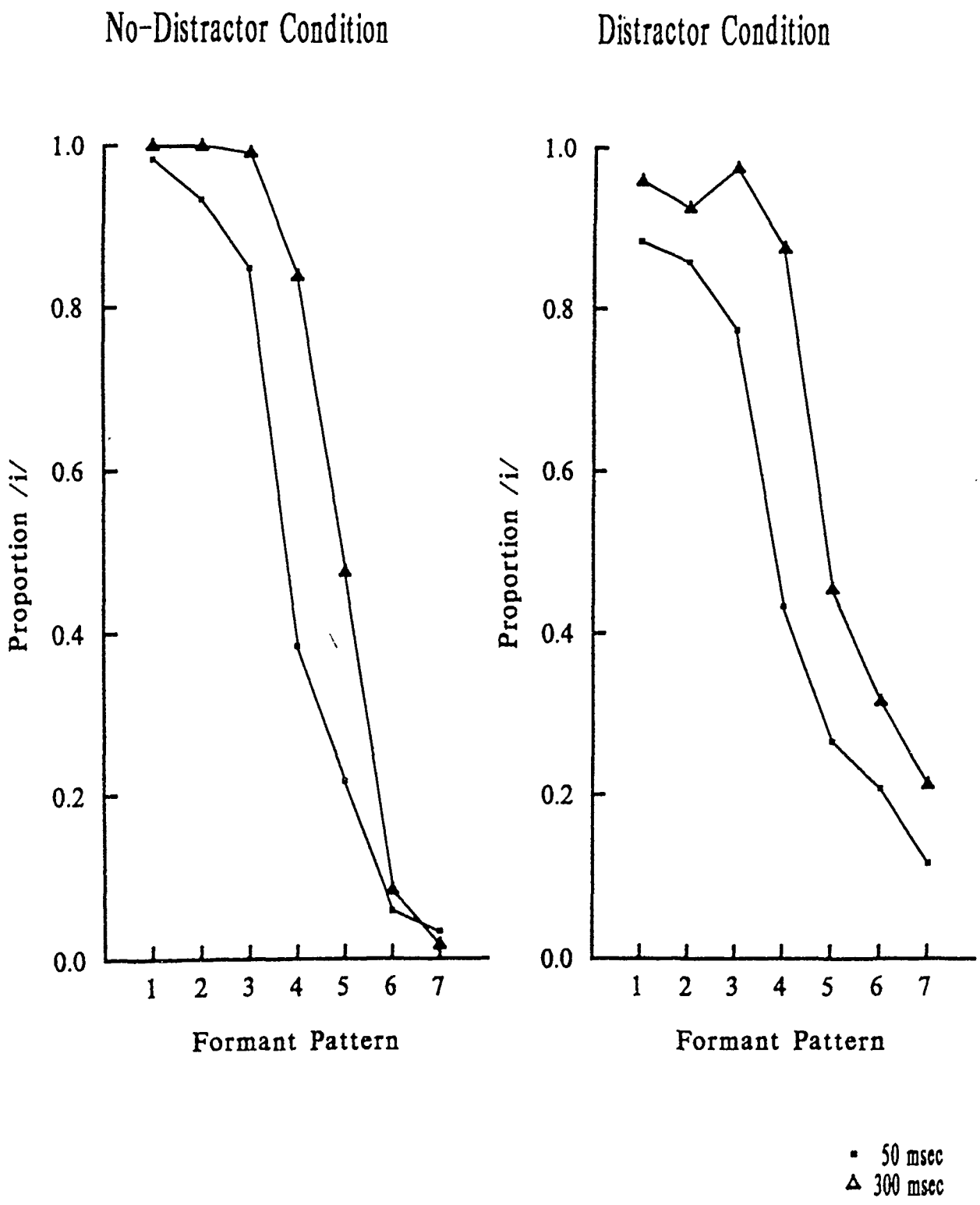

No-Distractor Condition
Distractor Condition
Proportion /i/
1.0
0.8
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7
Formant Pattern
50 msec
300 msec

Fig 8

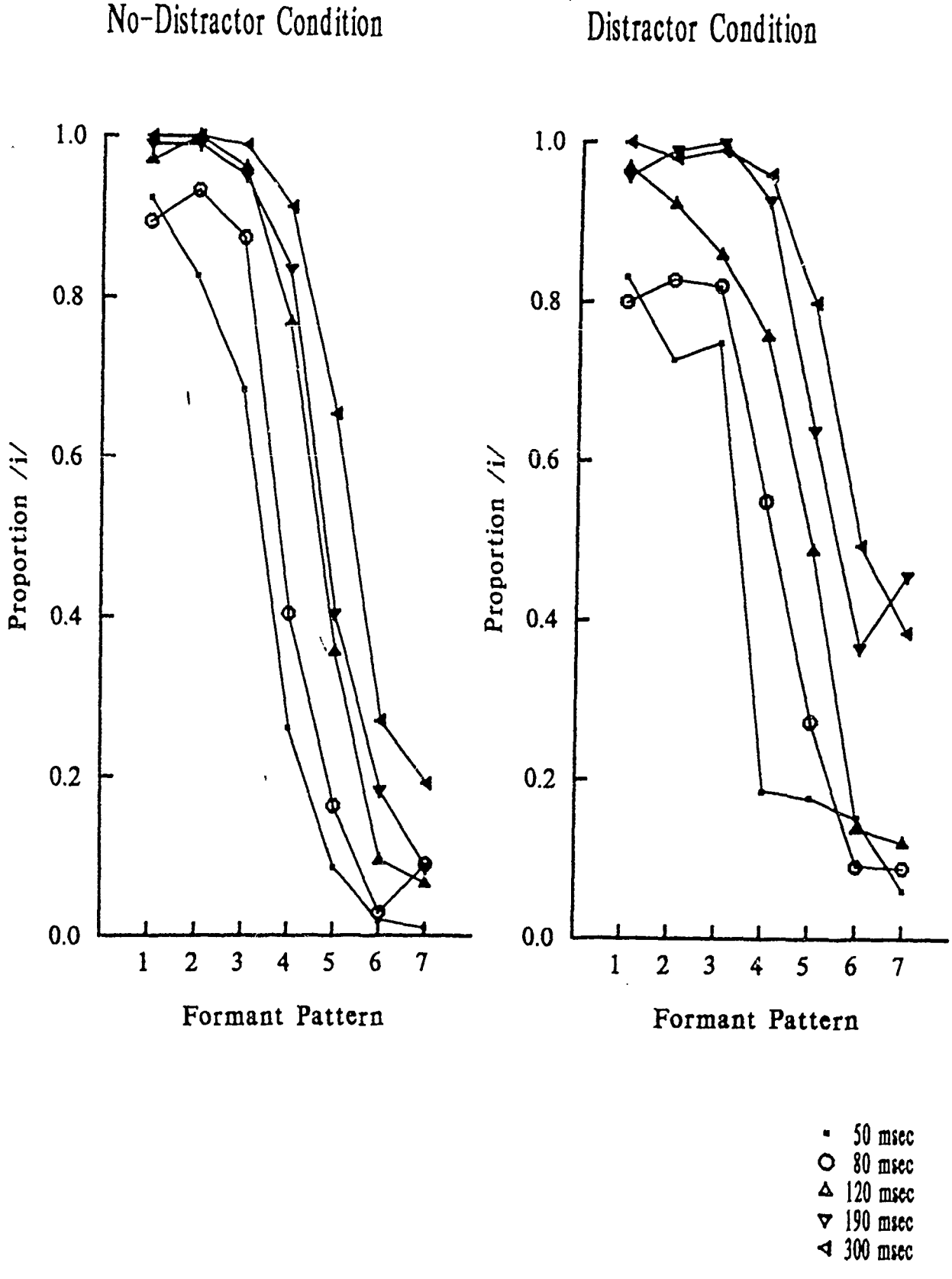
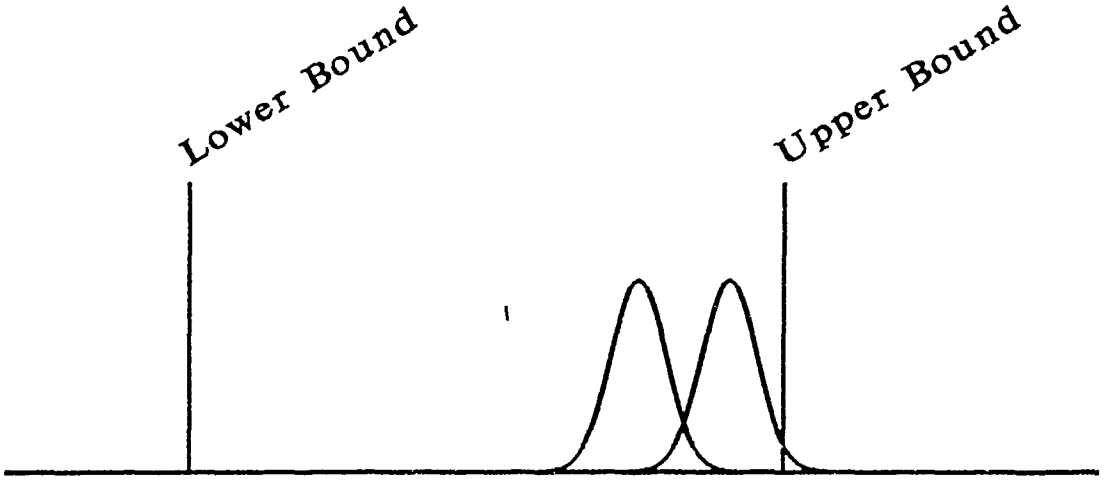
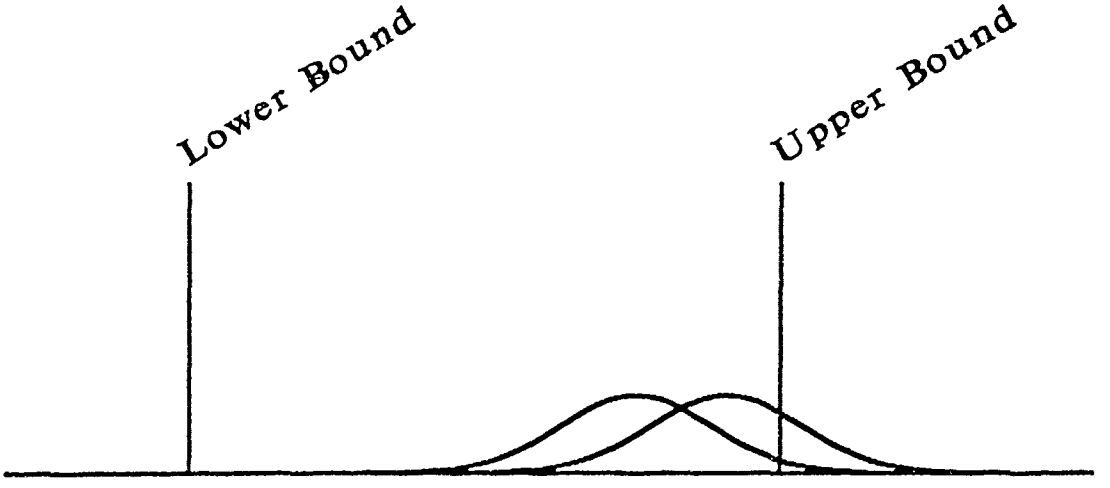

No-Distractor Condition
Distractor Condition
1.0
0.8
0.6
0.4
0.2
0.0
Proportion /i/
1 2 3 4 5 6 7
Formant Pattern
50 msec
80 msec
120 msec
190 msec
300 msec

Fig 9

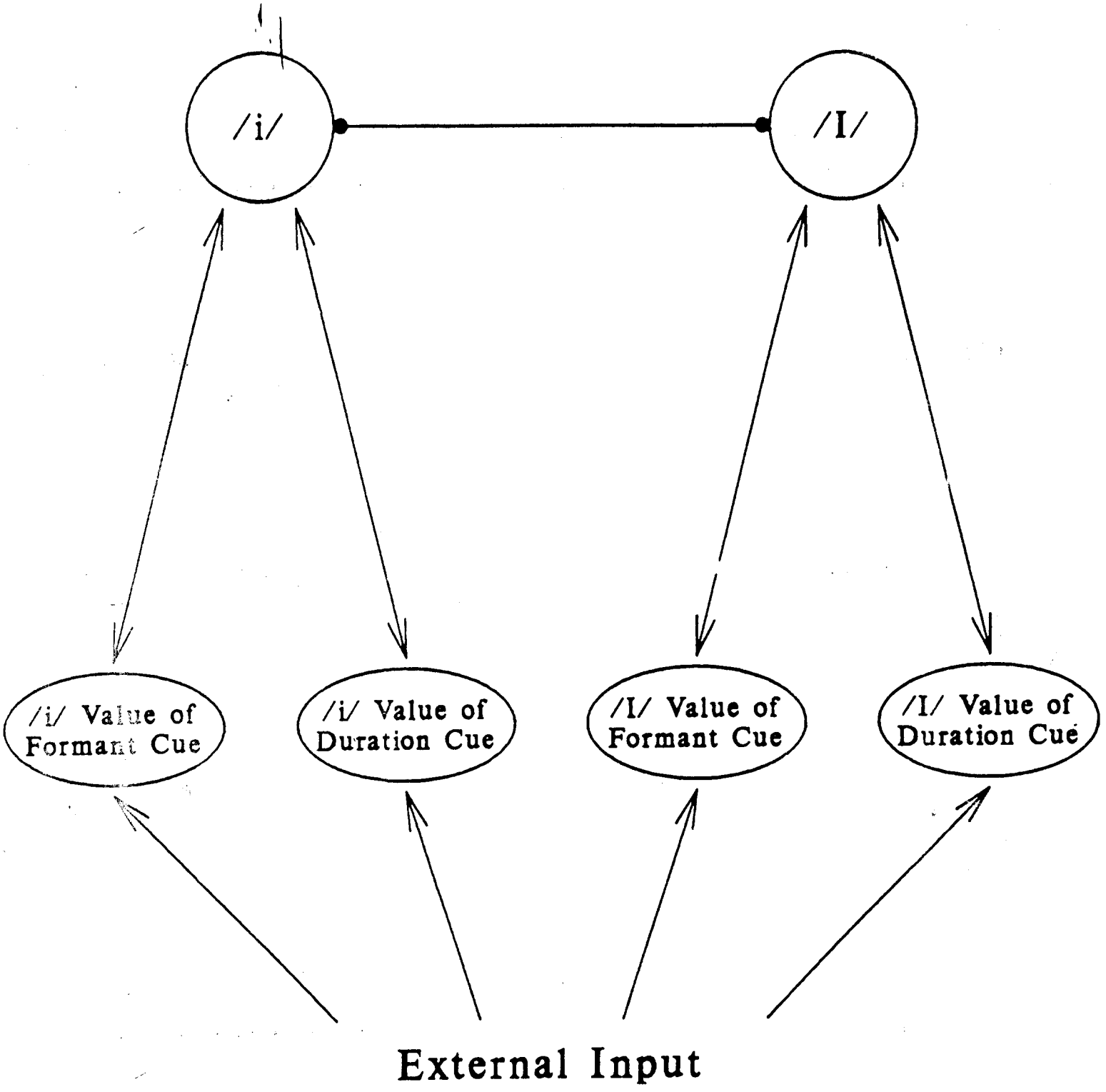


Phonetic Value



Phonetic Value

Fig 1C


/i/
/I/
/i/ Value of Formant Cue
/i/ Value of Duration Cue
/I/ Value of Formant Cue
/I/ Value of Duration Cue
External Input