Memory & Cognition
1975, Vol. 3 (1), 7-18

Auditory short-term memory and vowel perception*†

DAVID B. PISONI
Indiana University, Bloomington, Indiana 47401

The distinction between categorical and continuous modes of speech perception has played an important role in recent theoretical accounts of the speech perception process. Certain classes of speech sounds such as stop consonants are usually perceived in a categorical or phonetic mode. Listeners can discriminate between two sounds only to the extent that they have identified those stimuli as different phonetic segments. Recently, several findings have suggested that vowels, which are usually perceived in a continuous mode, may also be perceived in a categorical-like mode, although this outcome may be dependent upon various experimental manipulations. This paper reports three experiments that examined the role of auditory short-term memory in the discrimination of brief 50-msec vowels and longer 300-msec vowels. Although vowels may be perceived in a categorical-like mode, differences still exist in perception between stop consonants and steady state vowels. The findings are discussed in relation to auditory and phonetic coding in short-term memory.

A basic assumption underlying recent theoretical work in speech perception has been that the perception of speech sounds involves processes and mechanisms that are somehow basically different from the processes involved in the perception of other auditory stimuli (Liberman, Mattingly, & Turvey, 1972; Studdert-Kennedy, 1973). One line of evidence cited in support of this view concerns the identification and discrimination of various classes of synthetic speech sounds (see Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Some classes of speech sounds, such as stop consonants, have been found to be perceived in a more nearly categorical mode; listeners can discriminate between two acoustically different stop consonants only to the extent that they can identify the stimuli as different on an absolute basis (Liberman, Harris, Hoffman, & Griffith, 1957; Liberman, Harris, Kinney, & Lane, 1961; Mattingly, Liberman, Syrdal, & Halwes, 1971; Pisoni, 1971). In contrast, other classes of speech sounds such as steady state vowels have been found to be perceived in a more nearly continuous mode; listeners can discriminate among many more vowels than would be expected on the basis of absolute identification alone (Fry, Abramson, Eimas, & Liberman, 1962). It should be emphasized that these differences between consonants and vowels are not absolute. For example, some degree of categorical perception has been noted in vowels (Stevens, Liberman, Studdert-Kennedy, & Ohman, 1969; Fujisaki & Kawashima, 1969; Pisoni, 1971) and, as we shall see, this presents an interesting theoretical problem.

*This research was supported in part by a PHS Biomedical Sciences Grant (S05 RR 7031) to Indiana University, NIMH Research Grant MH 24027, and in part by a grant from NICHD to Haskins Laboratories. I am grateful to Michael Studdert-Kennedy, A. M. Liberman, and R. M. Shiffrin for their advice on this project. I would also like to thank D. L. Glanzman, J. Tash, J. Sawusch, and C. Lewis for their help in running experiments and analyzing data.

†A shorter report of some of these findings was presented at the 85th meeting of the Acoustical Society of America, Boston, April 11, 1973.

Differences between consonants and vowels have also been implicated in several recent studies dealing with immediate recall. For example, Crowder (1971, 1973a, b) has reported that, for lists of stop-consonant vowel syllables presented auditorily, a recency effect is observed in immediate serial recall if the syllables in the list contrast only on vowels (e.g., /bi/, /ba/, /bu/); however, the recency effect is absent if the syllables contrast only on the stop consonants (e.g., /ba/, /da/, /ga/). The recency effect describes an advantage in recall of the last serial position over the second to last serial position in a list of items.

Crowder (1971, 1973a, b) also reports two other differences in immediate memory for stop-consonant vowel stimuli: a modality effect and a suffix effect. The modality effect refers to the advantage of auditory over visual presentation for recall of items from later serial positions of a list. This modality effect has been observed for vowels but not for consonants. On the other hand, the suffix effect refers to a decrease in performance for items at the end of a list when a redundant word is presented after the last item in that list. The suffix effect has also been found with vowels but not with stop consonants. All three findings—the recency effect, the modality effect, and the suffix effect—are characteristic of a form of auditory memory called precategorical acoustic storage (PAS) by Crowder and Morton (1969). They have argued that this form of memory holds some relatively unanalyzed representation of an acoustic stimulus for approximately 2 sec.

In the original study by Crowder (1971), information about the vowel and consonant was confounded by their position within the syllables. The stop consonants were in initial position in the syllable and the vowels were in final position. Similar results, however, have been reported by Cole (1973), who found that consonants show less of a recency effect than vowels, regardless of the initial or final position in the syllables of the critical to-be-remembered information (e.g., /ba/ vs /ab/).

Crowder (1973a) recently replicated these results in a study which controlled for position of the information within the syllable. Both Crowder (1971, 1973a) and Cole (1973) have attempted to explain the differences in recency effects for consonants and vowels in terms of differences in auditory memory for these two classes of speech sounds. They assume that the recency effect for the vowels is due to retrieval of some auditory representation for vowels from a sensory memory store such as Crowder and Morton's PAS system. Crowder (1971, 1973a, b) has been somewhat more specific and further suggests that auditory information for vowels, but not for stop consonants, is represented in PAS.

Crowder (1971, 1973a, b), Liberman et al (1972), and Cole (1973) have all noted the parallel between differences in perception of stop consonants and vowels (the categorical vs continuous distinction) and differences in serial recall for these two types of stimulus vocabularies. All of these investigators have suggested that the differences in immediate recall may, in fact, be due to differences in perceptual processing for these two classes of speech sounds. For example, in discussing the recency effect Liberman et al (1972) state: "... the difference in recency effect between the stops and vowels is exactly what we would expect.... the special process that decodes the stops strips away all auditory information and presents to immediate perception a categorical linguistic event the listener can be aware of only as (b,d,g,p,t,k). Thus, there is for these segments no auditory, precategorical form that is available to consciousness for a time long enough to produce a recency effect. The relatively unencoded vowels, on the other hand, are capable of being perceived in a different way. Perception is more nearly continuous than categorical.... the auditory characteristics of the signal can be preserved for a while [p. 329]."

The explanation provided by Liberman et al (1972) appears to be consistent with the assumption that the differences may be perceptual in nature, presumably occurring at a relatively early stage of perceptual analysis. However, many of the earlier perceptual studies dealing with the identification and discrimination of consonants and vowels have not been very specific about either where the differences between consonants and vowels arise during perceptual processing or the nature of these differences. Moreover, although one might want to argue that the recall findings are due to differences in perceptual processing for consonants and vowels, several recent findings seem to indicate that vowels may also be perceived categorically, much like stop consonants. If vowels are perceived categorically in much the same way as stop consonants, we are clearly faced with somewhat of a dilemma in trying to account for the serial recall data by reference back to the perceptual findings. One way to deal with this problem would be to demonstrate that the categorical perception findings for the vowels are basically different from those obtained for the stop consonants.

In several previous reports, we have suggested that the major differences in discrimination between stop consonants and steady state vowels are to be found in an examination of within-phonetic category comparisons (Pisoni, 1971, 1973). Discrimination performance for the putative categorially perceived vowels is well above chance within phonetic categories, suggesting an auditory as well as a phonetic basis for the decision in discrimination. The situation for the stop consonants is quite different. Under identical experimental conditions, Ss apparently cannot retrieve the auditory information needed for a correct within phonetic category decision with the consonants (Pisoni, 1973). The present paper describes a revised model of the perceptual processes involved in the ABX discrimination test based on Fujisaki and Kawashima (1970) and then reports a series of experiments dealing with the discrimination of steady state vowels. One purpose of these studies is to make explicit some of the differences between the type of categorical-like perception recently observed with vowels and the type of categorical perception typically observed with stop consonants. A second purpose is to emphasize the role that short-term memory plays in speech perception.

AUDITORY AND PHONETIC MEMORY CODES

Since Fujisaki and Kawashima's (1970) findings on categorical-like perception of vowels are central to a number of theoretical efforts in speech perception (Pisoni, 1973; Studdert-Kennedy, 1973), we consider some of their results and a modified version of their original model of the perceptual processes involved in the ABX discrimination test.

Fujisaki and Kawashima (1968, 1969, 1970) proposed a two-stage model of categorical perception, a model based on a distinction between auditory and phonetic information in short-term memory (STM).[1] This model assumes that differences in discrimination between classes of speech sounds are due to the degree to which auditory and phonetic information is employed in the decision process in discrimination. Although not explicitly described by Fujisaki and Kawashima, we assume, following Studdert-Kennedy (1973), that auditory information is coded in short-term memory (STM) subsequent to an analysis of the acoustic waveform into a set of time-varying psychological dimensions such as pitch, loudness, and timbre. Similarly, we assume that phonetic information is coded as abstract phonetic features in STM after the "auditory" dimensions have made contact with some type of representation generated from synthesis rules residing in long-term memory (LTM).

The basic model proposed by Fujisaki and Kawashima (1969, 1970) is shown with several additions and modifications in Fig. 1. As shown, the model applies to discrimination exclusively within the ABX discrimination format, but the same assumptions could

be adapted to other discrimination procedures. It is assumed that encoding of speech sounds involves information about both the phonetic features of the stimulus and the auditory properties of the acoustic input. Furthermore, auditory information at relatively early stages of processing may be lost more rapidly from STS than the higher order phonetic information. According to this view, both an auditory and a phonetic representation *are* present in STS; the comparison process in ABX discrimination will entail the retrieval of either the auditory trace or the phonetic code for a correct decision.

According to this model, when a listener is required to discriminate between two *different* phonetic types, the decision in the discrimination task is based on phonetic information coded in STS. These derived phonetic properties or features of the auditory stimuli are maintained in phonetic short-term store. In this case, the listener determines whether the first two stimuli (i.e., A and B) are different phonetic segments. Since A and B were identified as different phonetic segments, the listener's decision about X is based exclusively on a comparison of the phonetic information coded in STS. Thus, he compares X with B and X with A and then determines which is the correct match.

However, the situation is somewhat different when the listener is required to discriminate between two *identical* phonetic types, that is, two stimuli that are acoustically different but that have been drawn from the same phonetic category. Now the listener must rely exclusively on the auditory information for each stimulus coded in STS. In order to arrive at a correct decision in the discrimination task, the listener must retrieve and compare with stimulus X the auditory representations of the two stimuli in auditory short-term store, since the two stimuli, A and B, were not originally identified as different phonetic segments. The listener must make a comparative judgment, based on auditory information of the acoustic properties of these stimuli, rather than an absolute judgment based only on the phonetic features of the two stimuli.

The basic model, which was first developed by Fujisaki and Kawashima (1969, 1970) and expanded here, predicts that categorical perception is related to the degree to which auditory and phonetic information in STS can be employed in the decision process during ABX discrimination. It has been reported that the major differences in discrimination between stop consonants and steady state vowels appear to be related to retrieval of *auditory* rather than *phonetic* information from STS (Fujisaki & Kawashima, 1970; Pisoni, 1971, 1973). But the extent to which auditory and phonetic information is encoded in STS and later retrieved for use in a discrimination task will depend on a number of factors: for example, the duration of the critical information in the signal, the acoustic environment or context of the cues, whether the acoustic cues are steady state or transient, and the particular information processing task. All these factors should presumably influence the way auditory and phonetic information is used by the decision process in discrimination.

The experiments reported in this paper are concerned with three related questions about vowel discrimination and the role of auditory STM in speech perception. First, what effect does duration play in vowel discrimination? Fujisaki and Kawashima (1970) found that isolated steady state vowels of very brief duration (e.g., 50 msec) tend to be perceived in a categorical-like mode; there was a peak across the phonetic boundary and a trough within phonetic categories. However, although Fujisaki and Kawashima showed that perception of vowels was more nearly categorical at

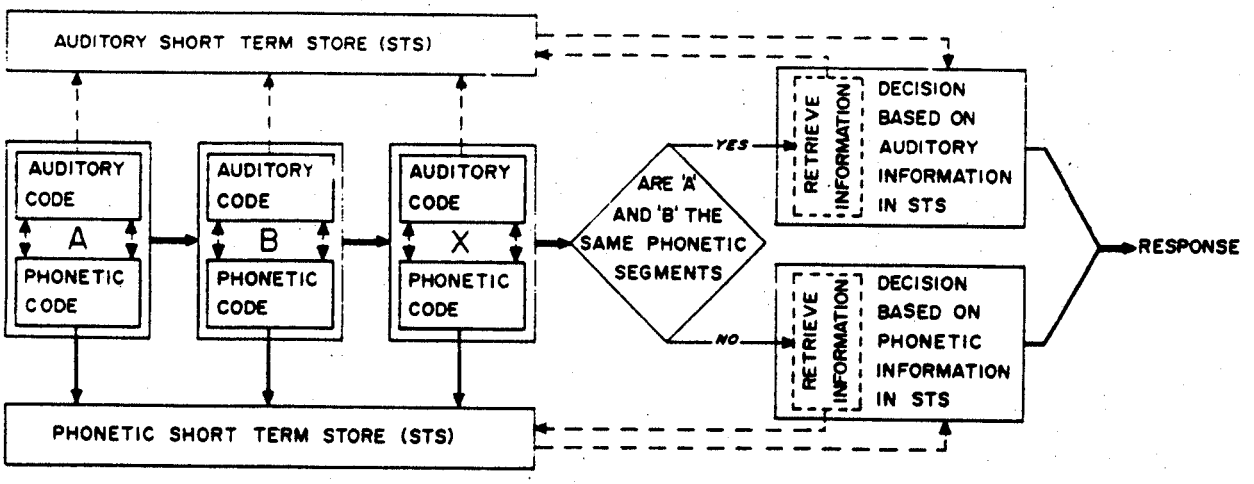

Fig. 1. A schematic representation of the perceptual and decision processes in an ABX discrimination task. Test stimuli are presented and two representations are stored in STS, an auditory code and a phonetic code. If the first two stimuli, A and B, have been recognized and encoded as different phonetic segments, then the comparison of stimulus X with A or B is based on phonetic information in STS. Otherwise, the comparison of X with A or B is based on auditory information in STS.

short durations, they did not employ stimuli with durations that were comparable to the earlier vowel perception studies conducted by Haskins Laboratories (Fry et al, 1962; Stevens et al, 1969). It is possible that the longer vowels of 300 msec also may be perceived in a somewhat categorical-like mode. Without the necessary control condition, it is not possible to argue that categorical perception is due primarily to the duration of the critical information in the stimulus.

The second question deals with the ABX discrimination test that Fujisaki and Kawashima (1969, 1970), among others, have employed in their experiments on vowels. Is the categorical-like discrimination found with vowels in the ABX test also found more generally with other discrimination procedures (Pisoni, 1971)? We consider discrimination performance with vowels in another test procedure, the 4IAX test of paired similarity. This discrimination test has previously been shown to be more sensitive to acoustic differences between speech stimuli (Pisoni, 1971; Pisoni & Lazarus, 1974). If there are large differences between the ABX and the 4IAX test for both short 50-msec vowels and longer 300-msec vowels, we will have additional evidence for suspecting that the type of categorical perception observed for vowels is somehow different from that observed for the stop consonants. In previous studies we have found that different discrimination procedures have relatively little effect on consonant discrimination (Pisoni, 1971, 1973).

The third question deals with the effect of context. What role does the immediately surrounding acoustic environment have on vowel discrimination? Stop consonants always occur in syllabic context. Moreover, there is a relatively complex relation between perceived phonetic segment and its representation in the acoustic signal; the essential acoustic cue for the stop consonants is a rapidly changing spectrum (F1 and F2 transitions) which is both short in duration (50 msec) and transient in nature (Liberman, Delattre, Cooper, & Gerstman, 1954). In contrast, the major acoustic cue for vowels, the frequencies of the first three formants, has a relatively long duration and remains uniform over the entire length of the stimulus (Delattre, Liberman, Cooper, & Gerstman, 1952). Stevens (1968), Sachs (1969), and Fujisaki and Kawashima (1969, 1970) have found that vowels tend to be perceived more categorically when they appear in a fixed context than when the same stimuli are presented in isolation. Fujisaki and Kawashima suggested that the context served as a "perceptual reference or anchor," thus permitting more stable phonetic identification of the vowel. However, it could be argued that the context selectively interfered with retention of the auditory information in target vowels. If this interference hypothesis is correct, vowel discrimination should be poorer when a reference context follows a target vowel (retroactive interference) than when it precedes it (proactive interference). The retroactive context may act to interrupt the processing of the target vowel. In addition, the amount of interference should be related to the acoustic similarity of the target sound and context. For example, vowels should suffer more interference from other similar vowels than from tones or white noise (Darwin, 1971).

EXPERIMENT I

In this experiment we compare discrimination of short (50-msec) vowels with longer (300-msec) vowels. The major aim of the study was to replicate and extend the findings of Fujisaki and Kawashima (1970) and Pisoni (1971), who reported that vowels are perceived as more nearly categorical at short durations.

Method

Subjects. Eighteen undergraduate students at Indiana University served as Ss. They were obtained from the Psychology Department's S pool and received either 1 h course credit or $1.50 for each session. All Ss were right-handed native speakers of English with no history of a hearing or speech disorder. None of the Ss had heard any synthetic speech before the present experiment.

Materials. *Stimuli.* Two sets of seven steady state vowels were synthesized on the vocal tract analogue synthesizer at the Research Laboratory of Electronics, Massachusetts Institute of Technology. Table 1 provides the frequencies of the first three formants for both sets of vowels. The fourth and fifth formants were fixed at 3500 Hz and 4500 Hz, respectively. The seven stimuli were arranged so that the first three formants varied in equal logarithmic steps from the English vowels /i/ through /I/. The formant frequencies chosen were identical to those used by Stevens et al (1969) in their cross-language study of vowel perception.

One set of stimuli had a steady state duration of 300 msec (the equivalent of approximately 30 glottal pulses), with a rise and decay time of 50 msec. The second set of stimuli had a steady state duration of 50 msec (the equivalent of five glottal pulses), with a rise and decay time of 10 msec. Both sets of stimuli had identical formant frequency values and had a falling fundamental frequency. F_0 fell linearly from 125 Hz to 80 Hz for the long vowels and from 125 to 100 Hz for the short vowels. The bandwidths of the first three formants were fixed at 50, 80, and 110 Hz, respectively. The stimuli were originally recorded on magnetic tape at MIT and then digitized on the PCM system at Haskins Laboratories, where the waveforms were stored on disk for test preparation.

Experimental Tapes. All the experimental tapes were produced under computer control from the digital values of these stimuli. A 1000-Hz tone was placed at the beginning of each tape to insure that the playback levels would be uniform throughout the testing sessions.

Four different 70-item identification test sequences were prepared for each set of vowel stimuli. Each identification test contained 10 different randomizations of an entire set of seven stimuli. The stimuli were recorded singly with a 4-sec interval between presentations and an 8-sec interval after every 10 trials.

Four different 88-trial ABX discrimination tapes were also constructed for each set of stimuli. All possible one- and two-step comparisons of the seven stimuli appeared twice within each tape. The tapes were balanced, with the restriction that each ABX triad occur equally often within each half of the test. Stimuli within each triad were separated by 1 sec, while successive triads were separated by 4 sec. There was an 8-sec pause after every 10 trials.

Procedure. The experimental tapes were reproduced on a high-quality tape recorder (Ampex AG-500) and were presented binaurally through Telephonics (TDH-30) matched and

Table 1
Formant Frequencies for Vowel Stimuli

Stimulus Number	Formant Frequency (Hz)		
	F_1	F_2	F_3
1	270	2300	3019
2	285	2262	2960
3	298	2226	2902
4	315	2180	2836
5	336	2144	2776
6	353	2103	2719
7	374	2070	2666

calibrated headphones. The gain of the tape recorder playback was adjusted to give a voltage across the earphones equivalent to 70 dB SPL re 0.0002 dynes/cm² for a 1000-Hz calibration tone. To compensate for the differences in loudness between the 300-msec and 50-msec vowels due to stimulus duration, the gain for the calibration tone on the 50-msec vowel tapes was adjusted by means of decade attenuators to be +8 dB above the 300-msec vowels. Measurements were made on a VTVM (Hewlett Packard Model 1051) before presentation of each experimental tape. Ss were run in two counterbalanced groups of nine Ss each. They were tested in a small experimental classroom. All Ss in a given session heard the same stimuli in the same order.

E read aloud to Ss a set of instructions which explained the nature of the experiment. Ss also had a set of printed instructions before them. Ss were told that this was an experiment dealing with speech perception. For the identification tests, Ss were required to identify each stimulus as either the vowel /i/, as in "beet," or /I/, as in "bit." In the ABX discrimination tests, Ss were told that the stimuli would be arranged in groups of three and that their task was to decide whether the third sound was more like the first or the second sound. Ss were told to guess if they were not sure, but to respond on every trial. Judgments were recorded in prepared response booklets.

Ss were run for an hour a day on 4 consecutive days. On the first 2 days one group received the 300-msec vowels and the other group received the 50-msec vowels. The conditions were reversed for each group on the last 2 days. An identification test for a given stimulus condition was always followed immediately by the corresponding ABX discrimination tests. When the data are combined over the four sessions, each S provided 40 identification responses to each of the seven stimuli in both the 300- and 50-msec vowel conditions. Each S also provided 32 judgments for each of the AB discrimination comparisons in each stimulus condition.

Results

The probabilities of identification averaged over the 18 Ss for each stimulus condition are given in Table 2.

The identification probabilities for both stimulus conditions are quite sharp and consistent. Examination of Table 2 shows that the probabilities of identification for the two vowel conditions are very nearly exact complements of each other. There is a slight shift in crossover point or phonetic boundary between /i/ and /I/ as stimulus duration is reduced from 300 msec to 50 msec; the boundary shifts predictably in favor of the short lax vowel /I/.

The average one- and two-step obtained ABX discrimination functions are shown in Fig. 2 for both stimulus conditions. The predicted discrimination functions, which were derived from the identification probabilities according to the Haskins model of categorical perception (Liberman et al, 1957; Pollack & Pisoni, 1971) are also plotted in Fig. 2. The predicted functions represent what would be expected under the strong categorical perception assumption: that discrimination is no better than absolute identification.

Table 2
Probabilities of Identification Averaged Over 18 Subjects for 300-Msec and 50-Msec Stimuli

	Stimulus Number						
	1	2	3	4	5	6	7
	300-Msec Vowels						
/i/	.998	.997	.968	.635	.141	.033	.018
/I/	.002	.003	.032	.365	.859	.967	.982
	50-Msec Vowels						
/i/	.980	.979	.871	.419	.071	.025	.021
/I/	.020	.021	.129	.581	.929	.975	.979

The obtained discrimination functions for both vowel conditions show peaks at the phonetic boundary and troughs within phonetic categories. Analysis of variance indicates that discrimination performance is significantly better on the 300-msec vowels than on the 50-msec vowels [$F(1,16) = 9.59$, $p < .01$], but only for the one-step comparisons. This finding is consistent with Fujisaki and Kawashima (1970) and Pisoni (1971). The two-step obtained discrimination functions did not differ significantly from each other. There was a significant difference between obtained and predicted discrimination scores for both the one- and two-step comparisons [$F(1,16) = 77.27$, $p < .001$ and $F(1,16) = 343.12$, $p < .001$, respectively].

We may obtain a better quantitative idea of these results by comparing the obtained discrimination functions to those predicted from the model of categorical perception. We assume that in the ideal case of categorical perception there will be an exact mapping

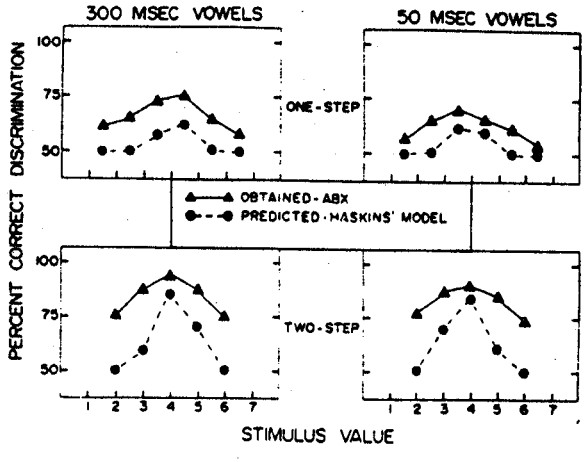

Fig. 2. ABX vowel discrimination functions for long (300-msec) and short (50-msec) stimuli averaged over 18 Ss under blocked presentation. The dashed lines show the predicted discrimination functions derived from the Haskins model of categorical perception.

of the discrimination functions predicted from identification and the functions obtained in ABX discrimination. Although an exact mapping is rarely found, since the obtained functions are usually higher than the predicted, we can use this assumption to our advantage for comparative purposes.

We assume that the difference between the obtained and predicted discrimination functions for a given condition represents a measure of the degree to which that particular condition deviates from the predictions of the idealized categorical perception model. Hence, it follows that the smaller the discrepancy between the obtained and predicted functions, the closer the obtained discrimination function will be to the categorical model. If short vowels are more categorical than longer vowels, we expect a smaller difference between the obtained and predicted functions for the 50-msec condition than for the 300-msec condition. The analyses reported below were carried out by first calculating difference scores on the obtained and predicted data for each S and then performing separate analyses of variance on the one- and two-step scores. Of greatest interest is the comparison between the long and short vowel conditions.

Analysis of variance on the one-step difference scores revealed a significant effect for stimulus duration [$F(1,16) = 12.80$, $p < .005$]. The difference between the obtained and predicted scores was greater for the longer vowels than for the shorter vowels. There was also a significant main effect for stimulus comparisons along the continuum [$F(5,80) = 3.68$, $p < .01$]. None of the interactions reached statistical significance.

A similar analysis on the two-step data failed to find a significant difference for the main effect of vowel duration, although the stimulus comparison did reach significance again [$F(4,64) = 9.44$, $p < .001$]. In addition, the Vowel Duration by Stimulus Comparison interaction was significant [$F(4,64) = 3.79$, $p < .01$].

Discussion

The results of this experiment indicate that vowels of both long and short duration may be perceived in a categorical-like mode. Differences in discrimination as they are related to stimulus duration are revealed only in the one-step comparisons. This finding is consistent with the results previously reported by Fujisaki and Kawashima (1970). Although there was no overall effect of vowel duration for the two-step data, differences restricted to particular types of stimulus comparisons along the continuum did occur. These results appear to be due to the apparent differences in the location of the phonetic boundary between /i/ and /I/ under the two duration conditions. Since Fujisaki and Kawashima (1970) employed only one-step stimulus comparisons, the present two-step data have little bearing on their results or conclusions.

The major outcome of this experiment may appear to be somewhat at variance with previous studies of vowel discrimination, particularly the original vowel perception studies conducted by investigators at Haskins Laboratories (Fry et al, 1962; Stevens et al, 1969). In these studies, vowel discrimination was described as more nearly continuous than categorical. However, Stevens et al (1969) did find some evidence for peaks in the discrimination functions which were correlated with changes in identification, but the troughs in the discrimination functions were well above chance when compared with the discrimination data typically found with the stop consonants. Although the discrimination functions, particularly the two-step data, appear by inspection to be categorical, we note that performance within phonetic categories is in fact well above chance. An auditory, nonphonetic basis for discrimination is available to the listener.

One of the major weaknesses of the original Haskins' model of categorical perception is its failure to account for within-category discrimination performance that may be at a level well above chance. In the original model (Liberman et al, 1957; Liberman et al, 1961; Pollack & Pisoni, 1971), it was assumed explicitly that if a listener identifies two stimuli as the same he can discriminate them only by chance.

However, in the model developed by Fujisaki and Kawashima (1970), performance that is above chance on within-category comparisons is assumed to reflect the underlying contribution of auditory short-term memory to ABX discrimination. Two assumptions are implicit in the model described in Fig. 1. First, discrimination will be based on phonetic information if the first two members of an ABX triad (A and B) are judged by the listener to be *different* phonetic segments. Second, discrimination will be based on auditory short-term memory if the first two members of an ABX triad have been judged to be the *same* phonetic segments. It is the contribution of auditory short-term memory that is of most interest.

Following Fujisaki and Kawashima, the predicted correct ABX discrimination score may be expressed by the following two components:

$$C_{ABX} = C_{A \neq B} + C_{A=B}$$

$$= C_{A \neq B} + M_{A=B} \cdot P_{A=B}$$

where $C_{A \neq B}$ is the probability that a correct discrimination occurs on the basis of phonetic identification, $C_{A=B}$ is the probability that a correct discrimination occurs on the basis of auditory short-term memory, $P_{A=B}$ is the probability that stimuli A and B are identified as the same phonetic segments, and $M_{A=B}$ is the conditional probability that a correct discrimination takes place when A and B are identified

as the same phonetic segments. This latter quantity indicates the degree to which judgments are based on auditory short-term memory and is equal to the asymptotic value of C_{ABX} at the extremes of the stimulus range (within-category comparisons).

These components are related according to the following equations:

$$C_{A \neq B} = \tfrac{1}{2}[(P_1 - P_2) + P_1(1 - P_2) + P_2(1 - P_1)]$$

$$C_{A=B} = [P_1 P_2 + (1 - P_1)(1 - P_2)] \cdot M_{A=B}$$

P_1 and P_2 represent the probabilities that stimuli A and B in the triad are identified as the same phonetic segments in an absolute identification test.

A new set of predicted ABX discrimination scores was obtained from the model outlined above. Figure 3 shows the obtained one- and two-step discrimination functions along with the new predicted functions derived from Fujisaki and Kawashima's model. Examination of this figure indicates that the new predicted discrimination functions match the obtained functions much more closely than the traditional Haskins' predictions. However, it should be pointed out that the better fit of the obtained data is to be expected, since one parameter from the obtained data has been used in the predictions. The simplicity and advantage of the Haskins' model of categorical perception lies in the fact that no additional assumptions or data are required to predict discrimination performance under the strong categorical assumption.

To summarize, this experiment has shown categorical-like discrimination functions for both short (50-msec) vowels and longer (300-msec) vowels. Although there were peaks in the discrimination functions at the category boundary, the level of within-category discrimination was well above chance expectation. When the contribution of auditory short-term memory is included in the predicted discrimination functions, according to Fujisaki and Kawashima's model, relatively better fits are obtained for the observed discrimination scores for both vowel conditions. These results suggest two conclusions. First, the role of stimulus duration taken alone appears to have relatively little effect on the shape of the discrimination functions. Second, the type of categorical perception observed with these vowels appears to be different from that observed in previous studies with the stop consonants. The obtained discrimination functions for the consonants usually match the predicted functions fairly well. Moreover, discrimination within categories is very nearly close to chance.

It is possible that the peaks and troughs in the discrimination functions in the present study might be due to the nature of the ABX test procedure. The arrangement of stimuli in this test format may prevent listeners from retrieving the auditory information needed for discrimination and subsequently may force them to rely more heavily on phonetic coding in short-term memory. The categorical-like discrimination observed with vowels might be due more to phonetic coding in STM than to earlier perceptual processes. The next experiment examines this possibility.

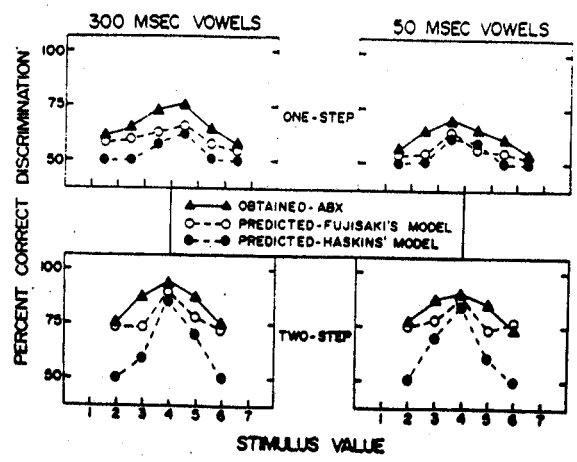

Fig. 3. ABX vowel discrimination functions along with the predicted functions derived from Fujisaki's model. The Haskins functions are also plotted for comparison.

EXPERIMENT II

In this experiment short and long duration vowels are compared under two discrimination procedures: the traditional ABX test and the 4IAX test of paired similarity. If the categorical-like discrimination observed with these vowels in Experiment I is due mainly to the nature of the ABX test, we should expect to find differences between these two types of discrimination procedures. Moreover, since the differences in vowel discrimination appear to be due primarily to the availability of auditory information, we anticipate advantages in discrimination to reveal themselves on within rather than between phonetic category comparisons.

Figure 4 shows the arrangement of stimuli in the traditional ABX test and the 4IAX test of paired similarity. In the ABX test, pairs of stimuli are arranged in triads; the first two stimuli are always different and the third stimulus is identical to the first or second. The Ss' task is to match the third stimulus with either the first (A) or the second (B) stimulus. This discrimination procedure requires that S encode and store each of the three stimuli over a relatively long time (e.g., several seconds) before arriving at a decision.

In the 4IAX test, two pairs of stimuli are presented on every trial; one pair is always the same and one pair is always different. The Ss' task is to determine which pair contains the same stimuli, the first pair or the second pair. We assume that the 4IAX is more sensitive to purely auditory information, since a correct decision can be made on a pairwise comparison. The first two stimuli are compared and a difference, d_1, is obtained and stored in short-term memory. The second pair of stimuli

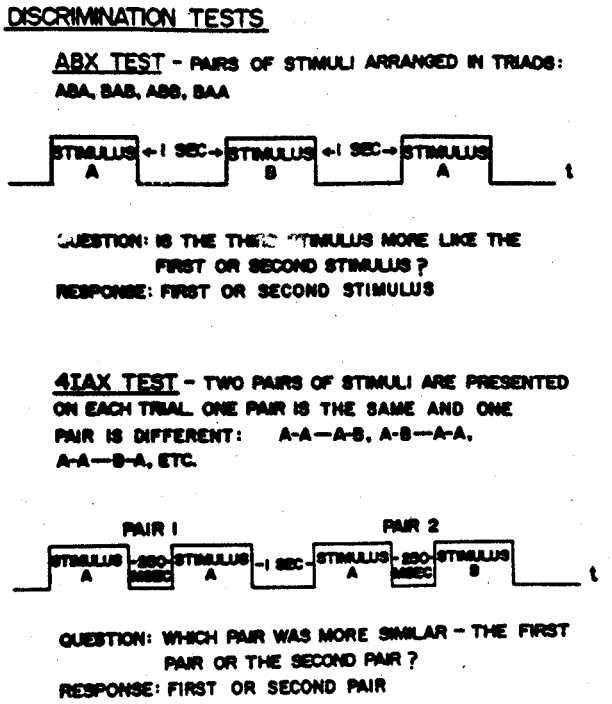

Fig. 4. Details of the two discrimination procedures, the standard ABX test and the 4IAX test of paired similarity.

is compared and the difference, d_2, is also obtained and stored. A final decision may be obtained when the two differences are later retrieved and compared.

Method

Subjects. Fourteen undergraduate students served as Ss. They were either paid for their services or received the equivalent in credit hours for their participation as part of a course requirement. They met the same requirements as Ss used in the previous experiments.

Materials. *Stimuli.* The 50-msec short vowel continuum and the 300-msec long continuum from Experiment I were used.

Experimental Tapes. The same identification tapes and ABX discrimination tapes from Experiment I were also used here. In addition, a new set of discrimination tapes was prepared in 4IAX format for both vowel conditions. All possible one- and two-step comparisons of the seven stimuli in each continuum were employed and arranged in the following 4IAX sequences: AA-AB, AA-BA, AB-AA, and BA-AA. Four different 88-item discrimination tapes were produced under computer control. The stimuli within each pair were separated by 150 msec, and stimulus pairs were separated by 1 sec. Successive trials were separated by 5 sec. After every 10 trials there was an extra 10-sec pause.

Procedure. The fourteen Ss were run in two groups of seven Ss each. One group was assigned to the long (300-msec) vowel condition, the other group was assigned to the short (50-msec) vowel condition. Thus, vowel duration was a between-Ss variable and the discrimination test type was a within-Ss variable.

On each day Ss first received the standard identification test for a given vowel condition. This was followed by both types of discrimination tests. Four Ss in each group received the discrimination tests in one order, while the other three Ss were presented with the reverse arrangement.

The instructions for identification and ABX discrimination were identical to those used in Experiment I. For the 4IAX discrimination test, Ss were told that they would hear two pairs of sounds on each trial and that their task was to determine which pair sounded more similar, the first pair or the second pair.

Results and Discussion

Table 3 shows the average probabilities of identification for each stimulus condition. The data are averaged over the seven Ss in each vowel condition. These data are almost identical to the probabilities obtained in the first experiment.

Figure 5 shows the obtained discrimination functions for ABX and 4IAX discrimination for the two vowel conditions. Inspection of this figure reveals relatively large and consistent differences in discrimination between the two types of test procedures. Performance is much better at every stimulus comparison for the 4IAX test than for the ABX test. This is true for both vowel conditions, although the effects are most noticeable for the long 300-msec vowels. The difference between the two discrimination tests was highly significant for both the one- and two-step comparisons [$F(1,12) = 36.10$, $p < .001$ and $F(1,12) = 21.85$, $p < .001$, respectively]. The main effect of vowel duration did not reach significance in either the one- or two-step analysis.

The most interesting result, however, is the interaction between type of discrimination test and stimulus comparison along the continuum. Both the one-step and two-step interactions were significant [$F(5,60) = 3.79$, $p < .01$ and $F(4,48) = 3.76$, $p < .01$]. This result, taken together with the main effect of test type, suggests not only that discrimination performance is better in the 4IAX test format but also that the shapes of the two discrimination functions are quite different. This result may be seen most clearly in the two-step discrimination functions for the 300-msec vowels. A distinct advantage of the 4IAX test over the ABX test for within phonetic category comparisons may be seen in this data. Discrimination in the 4IAX test may be thought of as more nearly continuous than categorical with these stimuli.

We conclude that the advantage in discrimination with the 4IAX test is due to the retrieval of auditory information from STM. As noted earlier, the ABX test forces Ss to rely more extensively on phonetic rather

Table 3
Probabilities of Identification Averaged Over Seven Subjects in Each Vowel Condition

	Stimulus Number						
	1	2	3	4	5	6	7
	300-Msec Vowels						
/i/	1.000	1.000	.967	.681	.157	.005	.010
/I/	.000	.000	.003	.319	.843	.995	.990
	50-Msec Vowels						
/i/	.967	.971	.824	.338	.119	.043	.029
/I/	.033	.029	.176	.662	.881	.957	.971

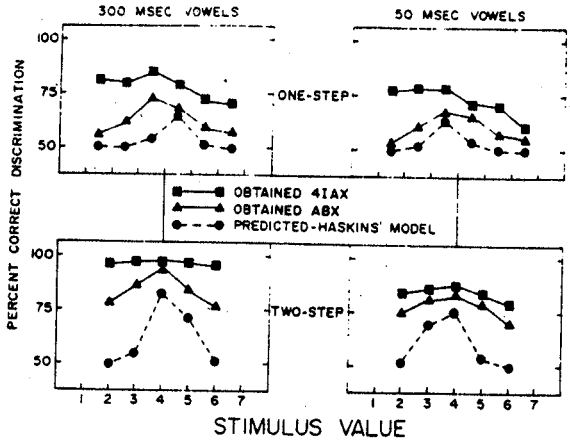

Fig. 5. Average discrimination functions for long and short vowels under ABX and 4IAX test conditions. The functions are based on seven Ss in each vowel duration condition.

than auditory coding in STM. Moreover, these results suggest that the categorical-like discrimination observed in Experiment I for both long and short vowels was probably due to the particular constraints of the ABX test rather than to some inherent property of the stimuli employed or to a limitation on the sensory capacities of the Ss.

More generally, it would appear that the form of categorical discrimination observed with the vowels is in fact different from that observed with the stop consonants. Comparable manipulations of the experimental procedures (i.e., use of the 4IAX test) with a stop-consonant continuum have thus far failed to show equivalent changes in either the overall level or the shape of the discrimination functions (Pisoni, 1971). Figure 6, which is taken from a recent paper by Pisoni and Lazarus (1974) shows the results obtained under ABX and 4IAX discrimination with a synthetic consonant continuum. The stimuli varied in voice onset time from /ba/ through /pa/ and were collected under the same experimental conditions as the vowel data in the present study. Ss first took an absolute identification test and then received either an ABX test or a 4IAX discrimination test. The obtained discrimination functions show a slight advantage in favor of the 4IAX test, but overall the obtained functions still match the functions predicted from the categorical perception model fairly well.

These findings do not preclude the possibility that auditory information can be employed in consonant discrimination; rather, it is assumed that auditory information from the earliest stages of processing tends to be lost from STM more rapidly than phonetic information. As a result, decisions that require a comparison of phonetic information will be more accurate and reliable than decisions that require a comparison of auditory information in STM. Since auditory information must be maintained in STM for a brief period of time, it might be possible to interfere with the retention of this information by the presentation of additional stimuli. The following experiment was conducted to study possible interference effects in vowel discrimination.

EXPERIMENT III

Stevens (1968), Sachs (1969), and Fujisaki and Kawashima (1970) have reported that vowels presented in fixed contexts (i.e., embedded in an acoustic environment) are perceived more categorically than the same vowels presented in isolation. Fujisaki and Kawashima (1970) suggested that the added context served as a perceptual anchor or reference. However, the context could interfere selectively with the retention of both auditory and phonetic information. If the perceptual anchor hypothesis is correct, it should be of little consequence where the reference context is placed (i.e., before or after the target vowel). However, if the context does selectively interfere with the encoding or retention of auditory information, then temporal position of the reference or interference sound should show differential effects on discrimination. In addition, the similarity of the context and target vowels should be directly related to the amount of interference produced.

Method

Subjects. Twenty undergraduate students served as Ss. They were paid at the rate of $2/h for their services and met the same requirements as those Ss used in the previous experiments.

Materials. *Stimuli.* The 50-msec short vowel continuum from the first experiment was used as the basic stimulus set. Four types of interfering stimuli were then constructed and used as contexts for each of the original seven stimuli. The interfering stimuli were 50 msec in duration and equal in overall intensity to the original vowels. The stimuli consisted of the following: (1) a 1000-Hz pure tone, (2) a burst of white noise, (3) the vowel /a/, and (4) the vowel /ɛ/. Each type of interfering context either preceded the target vowel (proactive interference condition) or followed the target vowel (retroactive interference condition).

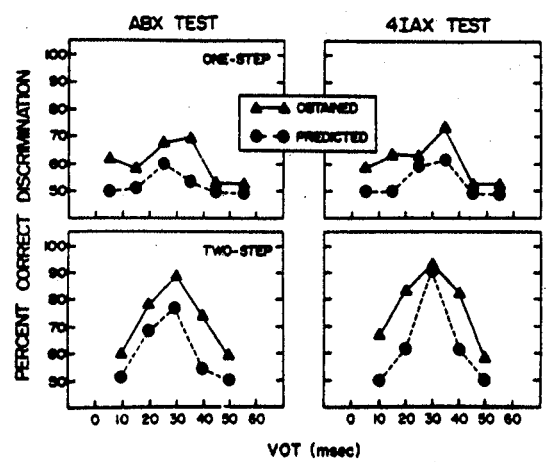

Fig. 6. Average discrimination functions obtained with the ABX and 4IAX tests for a synthetic stop-consonant continuum varying in voice onset time from /ba/ to /pa/. Data are taken from Pisoni and Lazarus (1974).

All stimuli were initiated and terminated at the temporal baseline so that transients were avoided. Except for this constraint, the ISI between target sound and interfering context was nominally zero. The original set of seven vowels was also presented alone as a control and will be referred to as the silent condition.

Experimental Tapes. Two types of identification and discrimination tests were prepared for each of the four types of interfering contexts. Two different 70-item identification tests were prepared for each of the proactive and retroactive interference conditions. In addition, four different 88-item ABX discrimination tapes were also constructed for each of these conditions. The identification and discrimination tapes for the silent condition were the same as those used in the previous two experiments. Test orders and timing sequences were similar to those described in Experiment I.

Procedure. The procedure was similar to that used in the previous experiment except for the following differences. Ss were run in four separate groups of five Ss each. Each group received one of the four interference conditions. Under a given condition, Ss received identification and discrimination tests for the silent (control) condition and for the proactive and retroactive interference conditions.

The instructions were the same as those used previously, except that when Ss were run under proactive or retroactive interference conditions they were told to ignore the interfering sound and to try to concentrate on only the target vowels, /i/ and /I/.

Ss were run for 1½ h a day on 2 consecutive days. On each day Ss received the silent vowel condition first, followed then by the proactive and retroactive conditions in differing order. Before each ABX discrimination test, Ss received the corresponding identification test, silent, retroactive, or proactive condition.

Results and Discussion

Table 4 shows the average one- and two-step percent correct discrimination scores for the silent, proactive, and retroactive context conditions for each of the four groups. These scores have been summed over the stimulus comparisons. Discrimination performance is generally lowest in the retroactive condition and highest in the silent condition for each type of interference. Analyses of variance were performed separately on the one- and two-step scores for retroactive and proactive context conditions. None of the main effects or interactions were significant for the one-step analysis. However, the main effect of context position was significant for the two-step analysis [$F(1,12) = 12.94$, $p < .005$]. Discrimination was lower in the retroactive condition than in the proactive condition. The main effect for type of interference did not reach significance in this analysis nor did the interaction of Type of Interference by Context Position. However, Newman-Keuls tests on the individual means showed that, in the retroactive condition, the /ɛ/ vowel context differed significantly from all other contexts. In the proactive condition, the /ɛ/ vowel context differed significantly from only the tone context condition.

The major outcome of this study is predicted by the interference assumption: There is more retroactive interference than proactive interference in vowel discrimination. Moreover, as shown in the two-step data in Table 4, there is more interference for a more similar vowel (e.g., /ɛ/) than a less similar vowel (e.g., /a/) or nonspeech stimuli. These results are not entirely surprising, since we would not expect noise or tonal stimuli to have any substantial effect on the initial encoding process for the target vowels used here.

The results of this study provide evidence for interfering effects in the discrimination of vowels in the ABX test paradigm. These effects are greater when the interfering context follows a target vowel than when it precedes the vowel. Moreover, the effect of similarity between context and target vowel appears to be related to interference with the initial encoding process when both auditory and phonetic information are registered in short-term store.

The results of this experiment argue against Fujisaki and Kawashima's general "perceptual anchor" hypothesis because they indicate relatively specific interference effects in vowel discrimination due to the temporal relations and similarity of other stimuli. These findings are also related to the results reported recently by Pisoni (1973), who showed that vowel discrimination is subject to decay over time. Using an A-X delayed comparison recognition memory paradigm, Pisoni showed that both auditory and phonetic information in vowels decay as the interval between the standard and comparison stimulus is increased. Both the interference and decay effects found in vowel discrimination agree with the results found more generally in other

Table 4
Average Percent Correct for Context Position for Each of Four Types of Interference

		One-Step Discrimination Context Position			Two-Step Discrimination Context Position		
		Silent Control	Proactive	Retroactive	Silent Control	Proactive	Retroactive
I	White Noise	59	60	58	78	75	70
II	1000-Hz Tone	62	59	54	79	82	76
III	Vowel /a/	60	60	57	85	75	73
IV	Vowel /ɛ/	64	53	55	78	71	59
Mean		61	58	57	80	76	70

short-term memory related tasks. We may conclude that short-term memory plays an intimate role in speech perception as well as other perceptual processes.

SUMMARY AND CONCLUSIONS

The experiments reported in this paper have been concerned with the role of auditory short-term memory in vowel perception and, more generally, with the relationship between auditory and phonetic coding in speech perception. The main findings of these studies indicate that vowels of both short (50-msec) and longer (300-msec) duration appear to be discriminated in a categorical-like mode; there is a peak in the ABX discrimination functions for stimulus comparisons selected from different phonetic categories and a trough in these discrimination functions for comparisons selected from within the same phonetic category. Stimulus duration per se was shown to play a relatively minor role in contributing to the shape and level of the discrimination functions. The categorical-like discrimination for the vowels was assumed to reflect the greater dependence on phonetic rather than auditory coding in the ABX format. Support for this conclusion was obtained in two additional experiments. One study showed that vowel discrimination could be improved substantially when auditory information in STM was made more readily available for use in discrimination by changing the discrimination paradigm. The other experiment demonstrated specific temporal and similarity interference effects in ABX discrimination, thus providing evidence for the existence of auditory memory codes for vowels in short-term memory. All three findings suggest that the type of categorical perception found with vowels is distinctly different from that found with stop consonants.

A major issue in speech perception has been the distinction between categorical and continuous modes of processing as reflected in the differences in discrimination between consonants and vowels. Despite several recent findings, we conclude that meaningful and theoretically important differences still exist between consonants and vowels. Moreover, we suggest that differences between categorical and continuous modes of discrimination are primarily due to a failure of retrieval of auditory information in STM. The earliest stages of auditory processing of speech sounds tend to be lost from subsequent processing. Loss of this information may be due to interference from succeeding acoustic events, the decay of auditory information over time, and the particular information processing task confronting the Ss.

REFERENCES

Cole, R. A. Different memory functions for consonants and vowels. Cognitive Psychology, 1973, 4, 39-54.

Crowder, R. G. The sound of vowels and consonants in immediate memory. Journal of Verbal Learning & Verbal Behavior, 1971, 10, 587-596.

Crowder, R. G. Representation of speech sounds in precategorical acoustic storage. Journal of Experimental Psychology, 1973a, 98, 14-24.

Crowder, R. G. Precategorical acoustic storage for vowels of short and long duration. Perception & Psychophysics, 1973b, 13, 502-506.

Crowder, R. G., & Morton, J. Precategorical acoustic storage (PAS). Perception & Psychophysics, 1969, 5, 365-373.

Darwin, C. J. Dichotic backward masking of complex sounds. Quarterly Journal of Experimental Psychology, 1971, 23, 386-392.

Delattre, P. C., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. Observations on one- and two-formant vowels synthesized from spectrographic patterns. Word, 1952, 8, 195-210.

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. The identification and discrimination of synthetic vowels. Language & Speech, 1962, 5, 171-189.

Fujisaki, H., & Kawashima, T. The influence of various factors on the identification and discrimination of synthetic speech sounds. Reports of the 6th International Congress on Acoustics, Tokyo, August 1968.

Fujisaki, H., & Kawashima, T. On the modes and mechanisms of speech perception. Annual Report of the Engineering Research Institute, 1969, 28, 67-73.

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

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

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychological Monographs, 1954, 68, 1-13.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology, 1957, 54, 358-368.

Liberman, A. M., Harris, K. S., Kinney, J., & Lane, H. The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. Journal of Experimental Psychology, 1961, 61, 379-388.

Liberman, A. M., Mattingly, I. G., & Turvey, M. T. Language codes and memory codes. In A. W. Melton and E. Martin (Eds.), *Coding processes in human memory.* New York: Winston, 1972.

Mattingly, I. G., Liberman, A. M., Syrdal, A. K., & Halwes, T. Discrimination in speech and nonspeech modes. Cognitive Psychology, 1971, 2, 131-157.

Pisoni, D. B. On the nature of categorical perception of speech sounds. *Status report on speech research (SR-27).* New Haven: Haskins Laboratories, 1971. P. 101.

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

Pisoni, D. B., & Lazarus, J. H. Categorical and non-categorical modes of speech perception along the voicing continuum. Journal of the Acoustical Society of America, 1974, 55, 328-333.

Pollack, I., & Pisoni, D. B. On the comparison between identification and discrimination tests in speech perception. Psychonomic Science, 1971, 24, 299-300.

Sachs, R. M. Vowel identification and discrimination in isolation vs. word context. *Quarterly progress report No. 93.* Cambridge, Mass: Research Laboratory of Electronics, Massachusetts Institute of Technology, 1969. Pp. 220-229.

Stevens, K. N. On the relations between speech movements and speech perception. Zeitschrift fur Phonetik, Sprachwissenschaft und Kommunikationsforschung, 1968, 21, 102-106.

Stevens, K. N., Liberman, A. M., Studdert-Kennedy, M., & Ohman, S. E. G. Cross-language study of vowel perception. Language & Speech, 1969, 12, 1-23.

Studdert-Kennedy, M. The perception of speech. In T. A. Sebeok (Ed.), *Current trends in linguistics.* Vol. XII. The Hague: Mouton, 1973.

NOTE

1. Within phonology there are two levels of representation, phonetic and phonemic. In current generative phonology, a phonetic (i.e., systematic phonetic) representation of an

utterance consists of a sequence of discrete segments that differ from each other in a limited number of ways. The composition of these segments is usually described by a phonetic matrix where the columns indicate phonetic symbols and the rows specify the phonetic features in the segment. The representation of the input at this stage is already abstract in the sense that segments and categories are already present. A phonemic (i.e., systematic phonemic) representation is more abstract than a phonetic representation. The listener applies the phonological rules of his language to the phonetic matrix, thus eliminating predictable and redundant phonetic details. For example, segments that may be different at the systematic phonetic level (e.g., the first segment of [p^hit] and the second segment of [spit] differ in terms of aspiration) may be treated as the same at the more abstract systematic phonemic level (see Studdert-Kennedy, 1973). In this work we assume that auditory and phonetic information is processed at two distinct stages of perceptual analysis.

(Received for publication February 1, 1974;
revision accepted May 6, 1974.)