Quarterly Journal of Experimental Psychology (1976) **28**, 361–378

PERCEPTUAL CATEGORIES FOR MUSICLIKE SOUNDS: IMPLICATIONS FOR THEORIES OF SPEECH PERCEPTION

JAMES E. CUTTING,*

Wesleyan University and Haskins Laboratories

BURTON S. ROSNER

University of Pennsylvania

AND CHRISTOPHER F. FOARD

Yale University and Haskins Laboratories, New Haven, Conn., U.S.A.

Sawtooth acoustic stimuli of different rise times are identified as coming from a plucked string instrument (*pluck*) or a bowed one (*bow*). Like stop consonants, these sounds are perceived categorically—discrimination is poor for stimuli identified as belonging to a single class but good for those identified as members of different classes. Varying the interval between two successive musiclike stimuli hardly alters discrimination. Sawtooth stimuli lasting 750 ms are clearly perceived categorically; those lasting 250 ms are not. Prolonged exposure to a *pluck* or *bow* stimulus can shift the rise-time boundary between categories. Shifts due to such selective adaptation decrease as adapting and test stimuli share fewer characteristics. Adaptation of postulated "feature detectors" therefore may occur in input systems prior to the detectors themselves. Our findings contradict previous claims that categorical perception and selective adaptation are manifestations of psychological processes unique to speech perception.

Introduction

Recent investigations have led to proposals that speech perception involves unique psychological processes (see Liberman, 1970). Wood (1975), in particular, names six phenomena which supposedly distinguish auditory from phonetic processing. (Auditory processing is said to operate on linguistic and nonlinguistic sounds alike, while only speech sounds undergo phonetic processing.) Two of the items in Wood's list are categorical perception and selective adaptation. Categorical perception entails accurate discrimination along certain acoustic continua between stimuli identified as different phonemes, but poor discrimination between sounds of equal acoustic difference identified as the same phoneme (Liberman, Harris,

* Requests for reprints should be sent to J. E. Cutting, Haskins Laboratories, 270 Crown Street, New Haven, Conn. 06511, U.S.A.

Hoffman and Griffith, 1957; Mattingly, Liberman, Syrdal and Halwes, 1971; Pisoni, 1973). Selective adaptation refers to a shift along such an acoustic continuum of the boundary between two phonemic categories, following repeated exposure to an unambiguous exemplar of either phoneme (Eimas and Corbit, 1973; Eimas, Cooper and Corbit, 1973).

Cutting and Rosner (1974) have studied the perception of musiclike sounds differing in rise time. These stimuli simulate a guitar or plucked violin when rise times are less than 30 ms and a violin played with a bow when rise times exceed 50 ms. The two kinds of stimuli accordingly are identifiable as *pluck* or *bow*. In a three-item (ABX) discrimination task, they satisfy the strictest criteria for categorical perception previously applied to speech (Studdert-Kennedy, Liberman, Harris and Cooper, 1970). The present study first ascertains whether discrimination of musiclike sounds meets a further new criterion for categorical perception, established by studies on speech. It then proceeds to test whether musiclike sounds are similar to phonemes in undergoing selective adaptation.

Experiment I: perception of stop consonants, vowels and musiclike stimuli

Several reports besides that of Cutting and Rosner (1974) have described categorical perception of nonlinguistic sounds. Lane (1965) claimed that pretrained observers categorically perceived stimuli whose identifiability as syllables had been destroyed by inversion of their spectrograms along the frequency axis; after reconsidering his data, however, Studdert-Kennedy *et al.* (1970) concluded that Lane's demonstration had failed. Locke and Kellar (1973) uncovered evidence of categorical perceptions in trained musicians judging triadic chords whose middle note varied. By changing the time between onsets of a hiss and a buzz, Miller, Wier, Pastore, Kelly and Dooling (1976) simulated aperiodic and periodic aspects of different stop consonants. The resulting sounds were perceived categorically.

Pisoni (1973) recently demonstrated a new property of categorically perceived speech sounds. He used a two-item (AX) discrimination paradigm and varied the delay between members of a stimulus pair. Offset–onset delays ranging from 0·0 to 2·0 s did not affect accurate discriminations between synthetic stop consonants identified as different phonemes or poor discriminations between acoustically different items falling into the same phoneme category. The stops were synthesized in consonant–vowel syllables containing identical vowels. Pisoni obtained quite different results for synthetic isolated vowels, which ABX discrimination tests had shown were not categorically perceived. As intrapair interval increased to 2·0 s, AX discrimination worsened between acoustically different items, regardless of their phonemic identities.

These results provide another criterion for categorical perception: interstimulus delays of up to a few seconds should not alter the probability of correct discrimination between categorically perceived stimuli. The results also reveal a weakness in previous demonstrations of categorical perception of nonlinguistic sounds. Locke and Kellar (1973) and Cutting and Rosner (1974) used stimuli of about 1 s

between stimuli. Pisoni's findings for vowels indicate that the resulting 2 s onset–onset delays in these experiments could have reduced discrimination between stimuli identified alike to a chance level, thereby giving a false impression of categorical perception.

These considerations led to Experiments I and II, which directly compare the perception of stops, vowels and musiclike sounds in identification, ABX, and variable-interval AX discrimination tasks. The issue is whether the musiclike sounds are perceived like the categorical stops or like the much less categorical vowels. In addition, Cutting and Rosner had compared perception of musiclike sounds with that of fricatives and affricates only. Liberman, Cooper, Shankweiler and Studdert-Kennedy (1967) have noted that the latter consonants are "less encoded" and less categorical than stops; Darwin (1971) has demonstrated in a dichotic listening task that fricatives can yield a small ear advantage more like that for vowels than for stops. Thus, comparison of musiclike sounds with synthetic stops in Experiment I affords a stronger standard for categorical perception than Cutting and Rosner utilised.

Method

Stimuli

Two seven-item arrays of speech stimuli and one nine-item array of musiclike stimuli were generated. One speech array consisted of synthetic consonant–vowel syllables varying systematically in direction and extent of second- and third-formant transitions. These items were identifiable as /bæ/ as in *bad* or /dæ/ as in *dad* and are designated hereafter as *consonants*. They were synthesized on the parallel-resonance system at Haskings Laboratories. The other array of speech stimuli consisted of three-formant steady-state *vowels* differing in formant frequencies and identifiable as either /i/ as in *eat* or /I/ as in *it*. These items were produced on the vocal-tract analog synthesizer at the Research Laboratory of Electronics, Massachusetts Institute of Technology. Pisoni (1973) had used both sets of stimuli in 300 ms long versions. The nine musiclike items were generated on the Moog synthesizer at the Presser Electronic Music Studio, University of Pennsylvania. They consisted of sawtooth waves at 294 Hz differing in rise time by 10 ms increments from 0 to 80 ms and decaying back to zero amplitude over the 1 s immediately following completion of rise time. The stimuli accordingly lasted between 1020 and 1100 ms. The item with 0 ms rise time reached full amplitude in one quarter cycle.

The speech and musiclike stimuli had been digitized and stored on disc file using the pulse code modulation system at Haskins Laboratories (Cooper and Mattingly, 1969). To achieve identical and short onset–onset times for the delayed AX task, all stimuli were retrieved in their original versions, truncated at 250-ms duration, and stored in new files. Thus, every stimulus had an abrupt offset.

Tapes and procedures

Audio tapes were prepared with the 250 ms duration stimuli, using the pulse code modulation system. For each stimulus array, we made one identification, one delayed AX discrimination, and one ABX discrimination tape.

Each *identification tape* began with the full-duration stimuli from the two endpoints of the appropriate array, repeated five times each in alternation; Pisoni, and Cutting and Rosner had originally used these stimuli. Subjects readily agreed that the identification responses appropriate to each stimulus array were easily applied. A similar alternating sequence of the endpoint items for the 250 ms duration stimuli was presented next. Most subjects reported that identifiability was unimpaired. The identification task then followed. Tapes for the two arrays of speech stimuli each contained a random sequence of 70 items:

7 stimuli per array × 10 observations per stimulus. After hearing an item, subjects wrote *B* or *D* to identify consonants or *EE* or *IH* to identify vowels. The identification tape for the musiclike stimuli had 90 items: 9 stimuli × 10 observations per stimulus. After hearing an item, subjects wrote *P* for *pluck* or *B* for *bow*. A 3 s pause occurred between successive items on all identification tapes.

Variable-interval *AX discrimination tapes* were patterned after those of Pisoni (1973). Stimuli in each speech array were numbered 1 through 7 in order and those in the sawtooth array were numbered 0 through 8. Stimuli numbered 1, 3, 5 and 7 were then selected from each array to form AX pairs. Those numbered 1 and 3 within an array fell into one identification category, while those numbered 5 and 7 fell into the other. Within an array, selected items were paired with themselves to produce four AA pairs (1–1, 3–3, 5–5, 7–7). Each item also was paired with its immediately adjacent selected neighbor(s) to produce six AB/BA pairs (1–3, 3–1, 3–5, 5–3, 5–7, 7–5). To equalize occurrences of within-category and between-category comparisons, the 3–5 and 5–3 pairs were represented twice and all other AB/BA or AA pairs only once, giving a block total of 12 pairs. The interval between the offset of A and the onset of X within a trial was 250, 750, or 1800 ms. Each AX tape therefore contained a random sequence of 72 trials: 12 pairs per block times 3 time-delays times 2 observations per pair. Each trial began with a 100 ms 1000 Hz tone, followed by 750 ms of silence, stimulus A, the variable silent interval, and then stimulus X. Four seconds intervened between trials. After hearing a pair, listeners wrote *S* for *same* if they thought the two items were identical and *D* for *different* if they were not.

Three *ABX discrimination tapes* were prepared with stimuli numbered 1 through 7 from each array. AB comparisons were constructed by pairing each stimulus with the next item either one or two steps higher in the array. This yielded for each array six one-step and five two-step comparisons, for a total of 11. Every AB comparison then appeared in four ABX arrangements: ABA, ABB, BAA, and BAB. An ABX tape contained a random sequence of 88 triads: 11 comparisons × 4 ABX arrangements × 2 observations per comparison, with 1·0 s between successive members of a triad and 4·0 s between triads. After hearing a triad, subjects wrote *A* or *B* to indicate whether the third stimulus seemed identical to the first or the second item.

Subjects and apparatus

Subjects were 16 young adults who were secretaries, undergraduates or graduate students at the University of Pennsylvania. Each was a right-handed native American speaker of English with no history of hearing difficulty. The subjects listened in groups of four to audio tapes reproduced on an Ampex AG500 tape recorder. Signals were sent through a listening station and presented diotically at 80 dB SPL over matched Telephonics earphones (Model TDH39). Subjects performed the identification, AX, and ABX tasks in that order within an array of stimuli (consonants, vowels or musiclike sounds). The order of presentation of arrays followed a balanced designed across subjects.

Results

Consonants and vowels

The consonant–vowel syllables yielded categorical perception as defined by Liberman *et al.* (1957) and Studdert-Kennedy *et al.* (1970). The upper left panel of Figure 1 displays ABX discrimination results superimposed upon identification data. Discrimination was clearly best around the boundary between /b/ and /d/. Both one-step and two-step ABX functions showed significantly better performance there than elsewhere ($F(5,75) = 4{\cdot}61$; $P < 0{\cdot}01$; and $F(4,60) = 19{\cdot}3$, $P < 0{\cdot}001$, respectively). Results from the variable interval AX task appear in the lower left panel of the figure. As in the ABX task, the comparisons between stimuli 3 and 5 were more successful than between stimuli

1 and 3 or between 5 and 7 ($F(1,15) = 75{\cdot}2$, $P < 0{\cdot}001$). The duration of the silent interval between the stimuli did not influence the judgements.

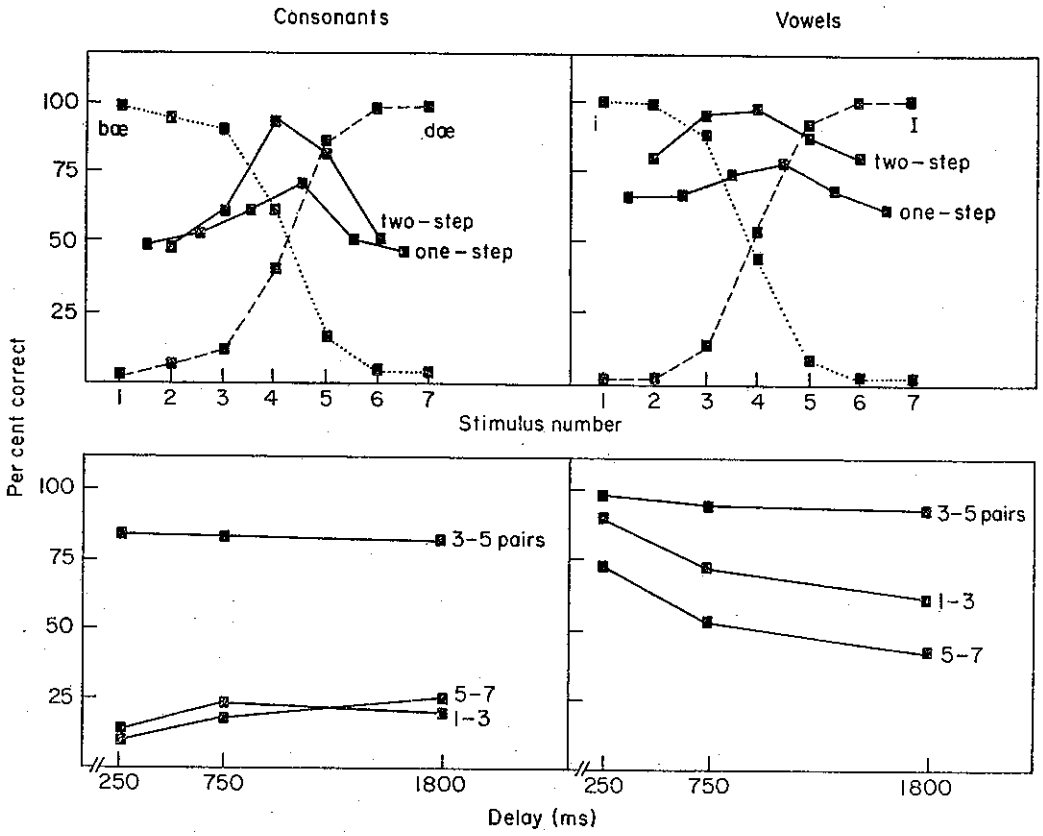

FIGURE 1. Mean identification, ABX discrimination, and variable-interval AX discrimination functions for consonants and vowels.

The right half of Figure 1 displays corresponding identification and discrimination data for vowels. The one-step and two-step ABX functions are superimposed upon the identification data and have significant peaks at the vowel boundary ($F(5,75) = 3{\cdot}71$, $P < 0{\cdot}01$; $F(4,60) = 6{\cdot}57$, $P < 0{\cdot}01$, respectively). Discrimination was better for vowels than for consonants ($F(1,15) = 38{\cdot}2$, $P < 0{\cdot}001$; $F(1,15) = 39{\cdot}1$, $P < 0{\cdot}001$, for one-step and two-step comparisons respectively). One-step functions for vowels and consonants did not differ significantly in shape; that is, there was no stimulus class by ABX comparison interaction. The two-step function for vowels however, lacked the troughs of that for consonants ($F(4,60) = 6{\cdot}84$, $P < 0{\cdot}001$). This result once again emphasizes an essential feature of categorical perception: discrimination functions should have troughs where within-category performance falls almost to chance. Consonants but not vowels yield this outcome consistently.

The delayed AX results indicate that perception of vowels and perception of consonants follow different time courses. For both types of speech segments, between-category comparisons involving stimuli 3 and 5 were made more accurately than within-category comparisons for stimuli 1 and 3 or stimuli 5 and 7, taken together ($F(1,15) = 52{\cdot}3$, $P < 0{\cdot}001$). Unlike the consonants, however,

vowels significantly decreased in discriminability as the silent interval lengthened between the A and X stimuli ($F(2,30) = 8{\cdot}25$, $P < 0{\cdot}01$). The stimulus class by AX comparison by delay, interval interaction also was significant ($F(2,30) = 12{\cdot}9$, $P < 0{\cdot}001$). All these results for consonants and vowels essentially reproduce Pisoni's. We also analysed our results as he had by calculating d' for each comparison. No patterns emerged besides those already apparent in the lower half of Figure 1. False alarms were evenly distributed across AA pairs and occurred at an overall rate of less than 8%.

250-ms sawtooth stimuli

Results for these stimuli deviated markedly from those reported by Cutting and Rosner (1974). Identification functions in the present experiment were more shallow and crossed at about 30-ms rise time, compared to the previously reported boundary of 40 ms. The one-step ABX function was quite flat. The two-step one seemed to peak at around 30 ms, but this maximum was not statistically reliable. Variable-interval AX functions showed no change in discrimination with length of delay. Since the ABX functions were not clearly indicative of categorical perception, these AX results are difficult to interpret. The conservative conclusion is that the 250-ms sawtooth stimuli failed to yield categorical perception.

This conclusion is based on data obtained from all 16 subjects. Inspection of individual results revealed that six listeners, despite their avowals to the contrary before testing began, could not identify the sounds consistently as *pluck* or *bow*. These subjects also performed both discrimination tasks at chance levels. The other 10 listeners gave consistent evidence of categorical perception, performing in the identification and discrimination tasks as had all subjects for consonants. Rather than following the precedent of eliminating data from subjects who gave poor identification performance (Liberman *et al.*, 1957; Lane, 1965), we conducted a new experiment which deliberately included such listeners. This experiment examined the possibility that duration of sawtooth stimuli influences categorical perception, since the 250-ms duration sawtooth stimuli which gave inconsistent results across subjects differed from those used by Cutting and Rosner in being much briefer.

Experiment II: categorical perception of 750-ms sawtooth stimuli

Method

The sawtooth items originally used by Cutting and Rosner were now truncated at 750 ms. Informal listening suggested that items of this duration were much easier to identify than the 250-ms stimuli. Identification, variable-interval AX, and ABX tapes for the 750-ms simuli were recorded in the same configurations as for Experiment I. Eight subjects from that experiment were recalled and paid to perform the same tasks with the new stimuli as they had with the old ones. The eight subjects included four who had not consistently identified the shorter stimuli in Experiment I and four who had. They listened individually to test tapes reproduced over a Revox tape recorder. Signals were presented diotically at 80 dB SPL over matched Telephonics earphones (Model TDH39). The subjects responded as in Experiment I.

Results

The left half of Figure 2 displays the results for the 750-ms sawtooth sounds. Data from Experiment I for the same eight subjects judging 250-ms musiclike stimuli appear in the right half of the figure. Identification functions in Figure 2 for the 750-ms stimuli were steeper than those for the shorter terms. In addition, the category boundary defined by the intersection of the identification functions approximated the 40 ms reported by Cutting and Rosner. The difference in boundaries for 250-ms and 750-ms items (roughly 30 and 35 ms respectively) is significant ($F(8,56) = 3{\cdot}68$, $P < 0{\cdot}01$).

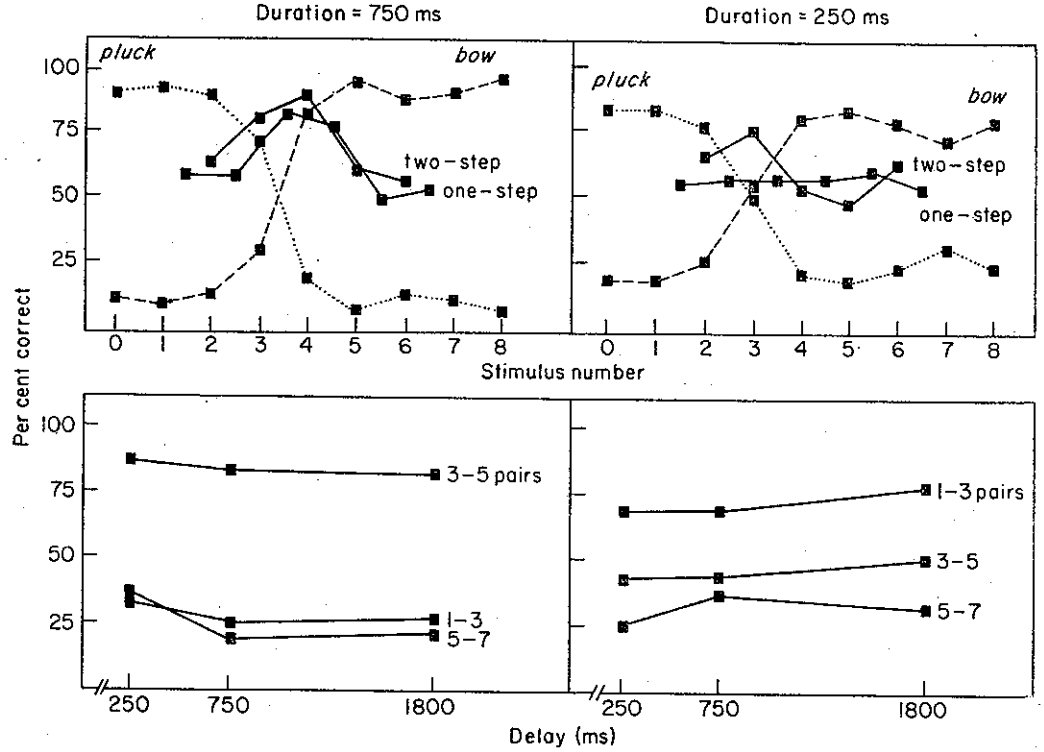

FIGURE 2. Mean identification, ABX and AX discrimination functions for sawtooth items truncated at 750 and at 250 ms.

The one-step and two-step ABX functions indicate categorical perception of the 750-ms stimuli. Both showed significant peaks ($F(5,35) = 4{\cdot}94$, $P < 0{\cdot}01$; $F(4,28) = 10{\cdot}26$, $P < 0{\cdot}001$, respectively). Both differed significantly from those for shorter stimuli ($F(5,35) = 2{\cdot}66$, $P < 0{\cdot}05$; $F(4,28) = 4{\cdot}83$, $P < 0{\cdot}01$, respectively).

The AX discriminations followed suit. The 1–3 pairs were easiest to discriminate for the 250-ms stimuli, but the 3–5 pairs were easier than the 1–3 or 5–7 pairs in Experiment II. This change in arrangement of pair discriminability was significant ($F(2,14) = 11{\cdot}78$, $P < 0{\cdot}001$). Thus, the longer sawtooth items yielded a pattern of results for delayed AX discrimination like that in Figure 1 for stop consonants. Unlike the consonants, however, the 750-ms sawtooth stimuli showed a small decay of discriminability over time ($F(2,14) = 3{\cdot}44$, $P < 0{\cdot}05$), a trend slightly like that seen for vowels. But matched comparisons across the

subjects of Experiment II showed that the sawtooth items did not differ significantly from consonants in the time course of AX discrimination.

Discussion

The results of Experiment II indicate that information used for within-category or between-category discrimination decreases very little with time intervals of up to 1800 ms for non-linguistic sounds differing in rise time. However, one might still challenge any inference that these stimuli are perceived categorically. The sawtooth items in Experiment II lasted 750 ms, so that a 250-ms offset–onset interval produced a 1000-ms onset–onset delay. Discrimination within a category conceivably could decline to an asymptote within 1 s. Musiclike sounds, then, would seen superficially to be perceived like stops, although they might actually be perceived like vowels. Two facts contradict this argument. First, discrimination performance on within-category sawtooth pairs at a 1000-ms onset–onset delay is well below that which we and Pisoni have found for vowels at identical or longer delays. Second, both sets of data on vowels imply that considerable drops should occur in within-category discrimination performance for sawtooth items across the range of delays used here. The decrease in performance for intervals of 250 and 750 ms shown in Figure 2 is smaller than the data on vowels would predict. Thus, the results of Experiment II reinforce the conclusion of Cutting and Rosner: sawtooth stimuli of different rise times are perceived categorically. In fact, comparison of these results with those from Experiment I shows that the perception of such sounds is functionally identical with that of the most categorical of speech sounds, namely, stop consonants.

Role of stimulus duration

The next question is why 750-ms sawtooth stimuli are perceived categorically while 250-ms ones are not. To begin with, this difference runs counter to Pisoni's (1973) finding that longer lasting, 300-ms synthetic vowels were perceived less categorically than briefer, 50-ms vowels. Pisoni (1973, 1975) explained his results by invoking a model of categorical perception offered by Fujisaki and Kawashima (1968, 1970). According to this model, troughs and peaks in ABX discrimination functions for speech result from the relative strengths of information in two different memory stores. Troughs would reach chance if within-category information resided in phonetic memory exclusively. They may remain above chance, however, partly because a separate auditory memory retains some relatively crude acoustic information which differentiates between stimuli within a category. Peaks reflect the retention of differential phonetic and auditory representations of stimuli which straddle a category boundary. Peaks would remain prominent, however, even if all between-category acoustic information were lost. Pisoni argued that briefer vowels produce weaker representations in auditory short term memory than do longer ones. This blurs within-category differentiation for briefer vowels and contributes to lowered troughs characteristic of categorical perception. Troughs for longer vowels remain well above chance, due to stronger auditory traces. The briefer vowels are still identifiable, however, so that between-category differentiation produces a peak in the ABX discrimination function.

No similar argument can explain our diametrically opposite findings with sawtooth stimuli. Our results might have arisen because the abrupt offset of each 250-ms stimulus was disruptive and masked critical information about onset envelope. The time course of auditory backward masking seems at least one order of magnitude too short to produce such effects (see Duifhuis, 1973, for a review). The simplest explanation of our findings is that trimming the sawtooth stimuli to 250 ms weakened the timbre associated with a stringed instrument. Listeners remarked that the shorter stimuli often sounded like quick toots on a harmonica. This change in quality diminished differential identifiability of stimuli with different rise times, including those at the extremes of the array. Consequently, the complementary identification functions for the 250-ms stimuli moved towards each other and became shallower than those for the 750-ms items. The latter functions in turn were somewhat shallower than those reported by Cutting and Rosner for full-duration items lasting about 1 s. A decrease in the rise time boundary between categories and a loss of the peak in the one-step ABX function accompanied this flattening of identification functions for 250-ms stimuli. The shift in rise time boundary is *not* a mathematically necessary consequence of the flattening. Enough categorical information apparently remained, however, to generate a small peak in the two-step function.

Differences in rise time can produce differences in timbre between otherwise identical stimuli, but the results shown in Figure 2 suggest that continued acoustic stimulation immediately after completion of the onset envelope is necessary for that cue to be effective. Synthetic sawtooth waves must last at least another 250 ms after the onset envelope in order to be categorized as *pluck* or *bow*.* Stringed instruments that are played faster than four notes per second, however, seem to retain their timbre: audiences apparently can discriminate rapidly bowed passages from those played quickly by *pizzicato*. These instruments may provide a greater variety of cues for timbre than can be realized in synthetic sawtooth waves, or contextual cues may help maintain constancy of timbre in the concert hall.

Models for categorical perception

The model of Fujisaki and Kawashima (1968, 1970) for categorical perception attributes reduced within-category discrimination to rapid fading of stimulus representations from auditory short-term memory. This same theory can explain poor discrimination between sawtooth stimuli identified alike. The model, however, makes phonetic memory responsible for peaks at category boundaries. The distinction between *pluck* and *bow* sounds is not phonetic, however, so that the Fujisaki–Kawashima model as it stands cannot explain peaks in the discrimination functions for these stimuli. There are two possible modifications of the model for conquering this difficulty. In both of them, linguistic and non-linguistic stimuli would share access to auditory short term memory. In one modification, both sorts of stimuli would have access to a common, second type of memory which

* The interaction between onset time and immediately subsequent acoustic information in determining discrimination of *pluck* from *bow* has a parallel in speech perception. For example, formant transitions for stop consonants are heard in isolation as chirps (Mattingly *et al.*, 1971). These cues require subsequent steady-state vocalic formants (as well as first-formant transitions) in order to generate percepts of speech sounds.

is reserved for information coded at a more abstract level. The other modification would postulate a phonetic memory for encoding speech sounds and a separate non-phonetic one for encoding nonlinguistic sounds. Studies of interference between the two classes of stimuli might decide between the alternative modifications.

Another view of categorical perception has emphasized the relationship between perception and production (Liberman *et al.*, 1967). On this view, discrete perceptions of sounds distributed continuously along an acoustic array such as /bæ/-to-/dæ/ partly reflect an inability to produce any speech sound intermediate between /b/ and /d/. Whatever its relevance to speech, this sort of theory is inappropriate for musiclike sounds. It would imply that difficulties in producing sounds intermediate between plucked and bowed notes on stringed instruments encouraged perception to become correspondingly dichotomous during evolution over many generations. These instruments, however, only appeared a few thousand years ago.

Extent of categorization

The categorical perception of speech sounds is a manifestation of a much broader process: the segmentation of a continuous acoustic stream into discrete phonetic elements (Studdert-Kennedy, 1975). Perceiving the initial phoneme in the syllable /bæ/ requires choices between /b/ and its voiceless counterpart /p/, between /b/ and its alveolar neighbour /d/, and between /b/ and its nasal relative /m/. The first two distinctions are categorical (Pisoni, 1973) and the third may be. Thus, at least three binary decisions must be made about a syllable which can be uttered in 100 ms or less. Liberman, Mattingly and Turvey (1972) have estimated that the rate of coding of continuous speech reaches 40 bits/s. Comparable estimates are not available for music (but see Chamberlain, 1974). The rate of discrete categorization of sounds may be much higher for speech than for music. Thus, while categorical perception may not be unique to speech, it may operate there at the highest speeds.

Experiment III: selective adaptation

The acoustical properties of continuous speech offer no evident explanation of categorical perception. Two well-known theories of speech perception, however, attempt to account for the phenomenon. One is the "motor theory of speech perception" (Liberman *et al.*, 1967; Cooper, 1972) and the other is a more explicit "analysis-by-synthesis" model posed by Stevens (1960, 1972; Stevens and House, 1972). Both models grew out of difficulties in finding for each English phoneme a characteristic set of invariant acoustic cues which would permit rapid segmentation of the speech stream. The theories directly state (Liberman *et al.*, 1967) or imply (Stevens and House, 1972) that efferent processes mediate speech perception. Production of speech sounds, especially prevocalic stops, seems to contain far more invariant features than does the resulting speech wave. If the production of speech mediated its perception, the problem of finding invariant cues would be largely solved. Either theory then could explain categorical perception as the consequence of different motoric events involved in the production of items

identified as different phonemes and of very similar ones for items placed in the same phonemic category.

We noted previously that this type of theory cannot be extended plausibly to explain categorical perception of musiclike sounds. An alternative approach has recently resulted from studies of speech perception in infants and adults. Although infants do not possess the vocal tract configuration for producing many adult speech sounds (Lieberman, Crelin and Klatt, 1972), they nevertheless appear to perceive and segment these sounds in the same way as adults (Eimas, Siqueland, Jusczyk and Vigorito, 1971; Eimas, 1974; Cutting and Eimas, 1975). Perception thus precedes production in the infant. This fact is at variance with the view that articulatory processes mediate speech perception (see Palermo, 1975, for a review). Eimas and his colleagues (Eimas and Corbit, 1973; Eimas, Cooper and Corbit, 1973) accepted this argument and went on to suppose that speech perception involves analysis-by-feature-detection. They sought evidence for such feature detectors in adults, by modifying an adaptation paradigm borrowed from vision research. Starting from the phenomenon of categorical perception, they exposed listeners to several hundred repetitions of a synthetic consonant–vowel syllable at one end of an acoustic array such as /pa/-to-/ba/. This procedure could temporarily change the acoustic boundary between the phonemic categories in the array. Identification functions shifted so that previously ambiguous items along the array now were placed regularly into the category adjacent to that for the adapting stimulus. Furthermore, the peaks of discrimination functions moved towards the loci of the new boundaries. Eimas and his co-workers concluded that they had adapted "linguistic feature detectors", a result which would support an analysis-by-feature-detection theory of speech perception. This theory provides a straightforward explanation of categorical perception as reflecting activation of one or another particular type of detector.

A feature-detector theory of categorical perception could explain our results for musiclike sounds. We therefore asked whether these sounds would display the selective adaptation which their theory predicts. A positive outcome also would show that selective adaptation is not unique to phonetic processing, contrary to the previously mentioned assumption of Wood (1975). We designed our experiments to assess the effects on selective adaptation, if it occurred, of using adapting stimuli differing in frequency or waveform from the test stimuli for which we obtained identification functions. Such cross-adaptation experiments can help to define the stimulus properties to which particular "feature detectors" are tuned. Similar studies have been done on speech sounds (Eimas and Corbit, 1973; Eimas, Cooper and Corbit, 1973; Ades, 1974*a*,*b*; Cooper, 1974*a*,*b*, 1975; Diehl, 1975).

Method

Stimuli

Stimuli previously employed by Cutting and Rosner were used. They were musiclike sounds varying in duration from 1020 to 1100 ms as rise time increased in 10-ms increments from 0 to 80 ms. Items selected for identification before and after adaptation were 440-Hz sawtooth waves at all times. Eight adapting stimuli were formed from orthogonal combina-

tions of 0 or 80 ms rise time, 294 or 440 Hz frequency, and sawtooth or sinusoidal waveform. Thus, the 440-Hz sawtooth adapting stimuli permitted direct mesurement of adaptation; the 440-Hz sinusoidal items gave adaptation across waveform; the 294-Hz sawtooth items adaptation across frequency; and the 294-Hz sinusoidal items adaptation across both waveform and frequency.

Tapes

Eight adaptation tapes which followed an identical pattern were recorded at Haskins Laboratories, using the pulse code modulation system. The first adaptation sequence contained 100 repetitions of the adapting stimulus and then seven stimuli for identification from the nine-stimulus 440-Hz sawtooth array. Five subsequent sequences each contained 50 repetitions of the adapting stimulus and then seven items for identification. In all six sequences, 600 ms intervened between repetitions of the adapting stimulus, 2 s between its last presentation and the first identification stimulus, 2 s between successive identification stimuli, and 5 s betwen the last identifiction item and the next adapting series. One run through a tape yielded 42 postadaptation identifications. The five important midrange stimuli for identification with rise times of 20 through 60 ms each occurred six times. The other four occurred three times each. Every subject heard each tape twice in succession, so there were 12 observations per subject for every midrange item and six for every other item, in each adaptation situation. A preadaptation identification tape also was recorded, consisting of 90 stimuli in a random sequence: nine items in the identification array times 10 observations per item.

Subjects and procedure

Subjects were the eight remaining listeners from Experiment I who did not participate in Experiment II. They were paid for their services. Each was tested individually with the equipment of Experiment II and heard the signals diotically at 80 dB SPL. Each of the eight sessions per subject began with the preadaptation identification tape, followed by two exposures to one particular adapation tape. Sessions were 24 h apart on the average. The order in which subjects heard the eight tapes followed a balanced design. Subjects responded to each identification item by writing *P* for *pluck* or *B* for *bow*.

Results

The four panels of Figure 3 summarize the data. The figure shows identification functions only for *pluck* responses, omitting the redundant (complementary) functions for *bow*. The identification functions are as steep as those described by Cutting and Rosner. Each panel in Figure 3 contains two postadaptation functfunctions, one for identifications following adaptation with a *pluck* stimulus (0-ms rise time) and the other for those following adaptation with a *bow* stimulus (80-ms rise time). These two curves lie astride the mean of the two preadaptation functions, which hardly differed and were therefore combined. Nevertheless, this averaging could obscure small differences which might affect quantitative assessment of postadaptation boundary shifts. We therefore measured the postadaptation shift (PoAS) for each subject in a given adaptation situation through the following formula:

$$\text{PoAs} = 10 \cdot \left[\sum_i P_{i,\text{post}} \text{ (pluck)} - \sum_i P_{i,\text{pre}} \text{ (pluck)} \right],$$

where $P_{i,\text{post}}$(pluck) and $P_{i,\text{pre}}$(pluck) are the probability of a *pluck* response to the *i*th stimulus after and before adapatation, respectively. The formula subtracts the area under the curve for *pluck* identifications before adaptation from that under

the curve following adaptation in a given session. Since adjacent items in the test array differed in rise time by 10-ms steps, multiplying the difference in areas by 10 gives the shift in ms. Table I shows mean adaptation shifts (PoAS) across subjects for each adapting condition, along with the results of a Wilcoxon matched-pairs signed-ranks test for significance of the shift.

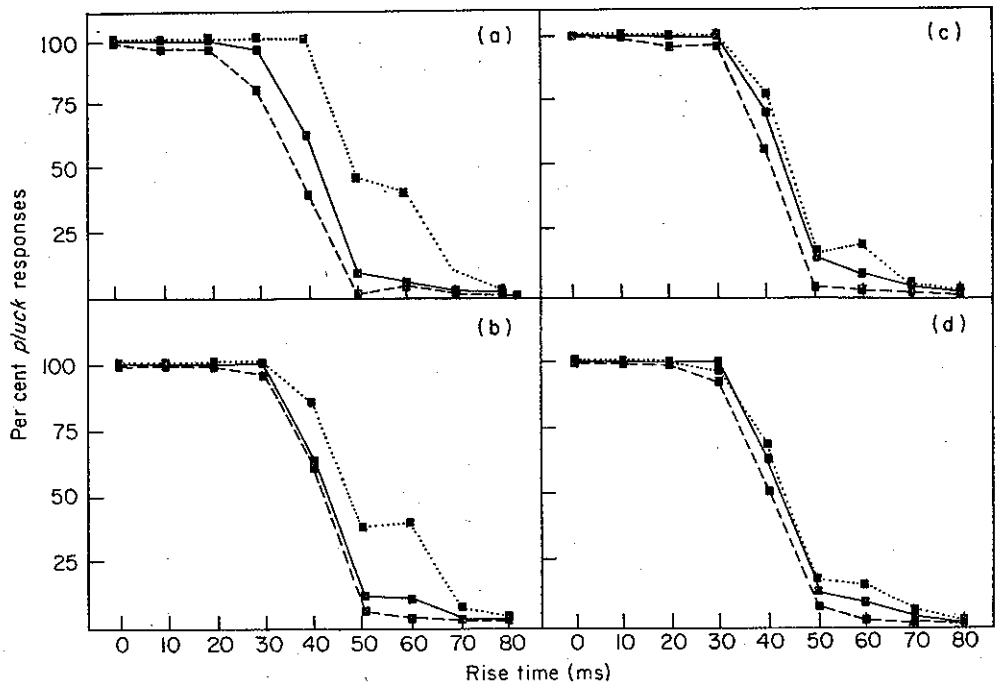

FIGURE 3. Mean selective adaptation functions in eight situations: adaptation (a), adaptation across frequency (b), adaptation across waveform (c), adaptation across waveform and frequency (d) for adaptation with *pluck* (▪ – – – ▪), adaptation with *bow* (▪ · · · ▪) and pre-adaptation (▪ —— ▪).

TABLE I

Postadaptation boundary shifts (PoAS) for nonlinguistic auditory stimuli differing in rise time, in eight adaptation situations

Condition	PoAS in ms	
	Adapt to 0 ms (*pluck*)	Adapt to 80 ms (*bow*)
Adaptation	−2·9 (T(8) = 1, $P < 0{\cdot}02$)	10·0 (T(8) = 0, $P < 0{\cdot}01$)
Adapt across frequency	0·3 (T(8) = 19, $P > 0{\cdot}10$)	6·6 (T(8) = 0, $P < 0{\cdot}01$)
Adapt across waveform	−5·5 (T(8) = 1, $P < 0{\cdot}02$)	3·5 (T(7) = 1, $P < 0{\cdot}05$)
Adapt across waveform and frequency	−2·2 (T(8) = 3, $P < 0{\cdot}05$)	2·2 (T(8) = 10, $P > 0{\cdot}10$)

Negative numbers and positive numbers indicate shifts towards 0 and 80 ms rise time respectively.

Seven of the eight adapting situations yielded boundary shifts in the predicted direction, towards the rise time of the adapting stimulus. Six of these were significant. The boundary shifts in Table I are sufficiently large so that differences between shifts following adaptation to *pluck* as against *bow* were always significant

($T(8) = 0$, $P < 0{\cdot}01$; $T(7) = 2$, $P < 0{\cdot}05$; $T(8) = 1$, $P < 0{\cdot}02$; $T(8) = 0$, $P < 0{\cdot}01$; across each row of Table I respectively). For 440-Hz sawtooth adapting stimuli, adaptation with a *bow* stimulus produced a larger absolute magnitude of adaptation shift than did adaptation wih a *pluck* stimulus ($T(8) = 4$, $P < 0{\cdot}05$). This result seems at least partly related to an inherent characeristic of stimuli such as the musiclike sounds: onset envelopes cannot have less than a 0-ms rise time but can take far more than 80-ms to rise. Adaptation shifts with speech stimuli have shown similar asymmetries (Cooper, 1975).

Table I shows important differences between postadaptation shifts for different steady-state parameters of the adapting stimuli. In general, the fewer the parameters shared between adapting and test stimuli, the smaller the adaptation shift. After exposure to *pluck* stimuli, shifts were greater when the adapting and test stimuli had the same frequency than when they did not ($T(8) = 1$, $P < 0{\cdot}02$). This difference across frequency approached significance following adaptation to *bow* stimuli. After adaptation to *bow* stimuli, shifts were larger when adapting and test stimuli shared waveform than when they did not ($T(8) = 1{\cdot}5$, $P < 0{\cdot}02$). Waveform, however, did not influence this shift after adaptation with *pluck* stimuli.

Discussion

The occurrence of selective adaptation has been used as evidence for "feature detectors" in speech. By the same token, our results along with those of others such as Kay and Matthews (1972) suggest that feature detectors exist in humans for non-linguistic sounds. Thus, selective adaptation in audition and the notion of feature detectors are not confined to speech. Moreover, their existence can no longer be cited as support for a distinction between auditory processing which works on all sounds and phonetic processing which applies exclusively to linguistic stimuli. Feature-detection theories of speech perception must now address the issue of whether various detectors are non-linguistic or specifically linguistic. The entire gamut of possibilities is open, ranging from all speech detectors being linguistic to their all being non-linguistic, auditory precursors to phonetic labelling. The former extreme encounters a problem of parsimony, if detectors for speech and non-linguistic sounds perform the same general functions. Cutting and Rosner (1974) already have demonstrated one such case: rise time signals the distinction both between certain consonants and between musiclike sounds of different timbre.

Our results on cross-adaptation between musiclike sounds suggest a possible structure for auditory feature detectors. The largest postadaptation boundary shifts occur when adapting and test stimuli share frequency and waveform; smaller shifts occur when they share only one or neither of these properties. One explanation for these findings would assume that the auditory system analyses such stimulus properties as frequency and waveform and sends the results along to the feature detectors. Information merely about stimulus rise time would not trigger the detectors. They would need the results of the prior analysis of other stimulus properties as well. The feature detectors could not produce a categorical perceptual decision about *pluck* vs. *bow* unless prior processing of frequency

and waveform provided them with certain indications. This would prevent brief transients such as clicks from activating the detectors and would explain why timbre detectors also need acoustic information about events immediately subsequent to onset envelope in order to work. Now suppose that selective adaptation reduced activity in all input systems which inform the detectors about stimulus properties, including channels which transfer unanalysed information about rise time. The extent of cross-adaptation would then depend on the number of properties shared by adapting and test stimuli. The larger this number, the greater the extent of cross-adaptation, since more input channels to a detector would lose their responsiveness. This loss could occur entirely prior to the detectors themselves or at the entry junctions to the detectors or both.

This arrangement is consistent with results from selective attention tasks which show that the speed of categorical perceptual decisions about *pluck* vs. *bow* depends at least partly on variation in stimulus properties other than rise time (Blechner, Day and Cutting, 1976). Identical selective attention tasks have yielded similar results on speech stimuli (Wood, 1974; Wood and Day, 1975). If there are phonetic feature detectors separate from purely auditory ones, these findings suggest that the arrangement of inputs follows a similar pattern in both cases. Selective adaptation of "linguistic feature detectors" then could occur entirely before the detectors themselves or at their input junctions. This hypothesis would explain why degree of adaptation for speech sounds depends on vowel environment (Cooper, 1974*b*). Posner (1976) has suggested a pattern recognition system for vision which resembles the one just described; Cutting (1976) discusses the auditory case further.

Experiments I and II demonstrate that the strictest criteria for categorical perception apply to musiclike sounds as well as to stop consonants. Recent work by Jusczyk, Rosner, Cutting, Foard and Smith (1975) strongly supports this result; they used the method devised by Eimas *et al.* (1971) to study perception of speech sounds by infants. The results of Jusczyk *et al.* (1975) indicate that infants categorize sawtooth stimuli of different rise times in the same way as adults. The results of Experiment III establish that musiclike sounds undergo selective adaptation like stop consonants and may activate nonlinguistic auditory feature detectors. Still another result once thought unique to speech perception is selective interference with categorical discrimination by irrelevant variation of stimuli along an auditory, nonlinguistic dimension. Blechner *et al.* (1976) found virtually identical interference in categorical perception of rise time when musiclike stimuli were varied along an irrelevant dimension. It is evident that the arsenal of empirical findings which once distinguished speech perception as a unique type of auditory perception is being steadily depleted.

This research was supported partly by National Institute of Child Health and Human Development Grant HD-01994 to Haskins Laboratories. We thank Michael Studdert-Kennedy for comments on earlier drafts of this paper, Nancy C. Waugh for help leading to its present form, David B. Pisoni for use of his stimuli, and Ruth S. Day for technical assistance. B. S. Rosner was on leave at the University of Oxford during preparation of this paper. C. F. Foard is now at the University of Pennsylvania.

References

ADES, A. E. (1974*a*). Bilateral component in speech perception? *Journal of the Acoustical Society of America,* **56**, 610–16.

ADES, A. E. (1974*b*). How phonetic is selective adaptation? Experiments on syllable position and vowel environment. *Perception and Psychophysics,* **16**, 61–6..

BLECHNER, M. J., DAY, R. S. and CUTTING, J. E. (1976). The processing of two dimensions in nonspeech stimuli: the auditory–phonetic distinction reconsidered. *Journal of Experimental Psychology: Human Perception and Performance,* **2**. In press.

CHAMBERLAIN, P. J. (1974). Pitch and duration in recognition of musiclike structures. *Perceptual and Motor Skills,* **38**, 419–28.

COOPER, F. S. (1972). How is language conveyed by speech? In KAVANAGH, J. F. and MATTINGLY, I. G. (Eds). *Language by Ear and by Eye.* Pp. 25–45. Cambridge, Mass.: M.I.T. Press.

COOPER, F. S. and MATTINGLY, I. G. (1969). Computer-controlled PCM system for investigation of dichotic speech perception. *Journal of the Acoustical Society of America,* **46**, 115 (Abstract).

COOPER, W. E. (1974*a*). Adaptation of phonetic feature anlyzers for place of articulation. *Journal of the Acostical Society of America,* **56**, 617–27.

COOPER, W. E. (1974*b*). Contingent feature analysis in speech perception. *Perception and Psychophysics,* **16**, 201–4.

COOPER, W. E. (1975). Selective adaptation to speech. In RESTLE, F., SHIFFRIN, R. M., CASTELLAN, N. J., LINDMAN, H. and PISONI, D. B. (Eds), *Cognitive Theory.* Pp. 23–54, Hillsdale, N. J.: Erlbaum.

CUTTING, J. E. (1976). The magical number two and the natural categories of speech and music. In SUTHERLAND, N. S. (Ed.), *Tutorial Essays in Psychology.* Hillsdale. N .J.: Erlbaum. In press.

CUTTING, J. E. and EIMAS, P. D. (1975). Phonetic feature analyzers in the processing of speech by infants. In KAVANAGH, J. F. and CUTTING, J. E. (Eds), *The Role of Speech in Language.* Pp. 127–48. Cambridge, Mass.: M.I.T. Press.

CUTTING, J. E. and ROSNER, B. S. (1974). Categories and boundaries in speech and music. *Perception and Psychophysics,* **16**, 564–70.

DARWIN, C. J. (1971). Ear differences in the recall of fricatives and vowels, *Quarterly Journal of Experimental Psychology,* **23**, 46–62.

DIEHL, R. L. (1975). The effect of selective adaptation on the identification of speech sounds. *Perception and Psychophysics,* **17**, 48–52.

DUIFHUIS, H. (1973). Consequences of peripheral frequency selectivity for nonsimultaneous masking. *Journal of the Acoustic Society of America,* **54**, 1470–88.

EIMAS, P. D. (1974). Auditory and linguistic processing of cues for place of articulation by infants. *Perception and Psychophysics,* **16**, 513–21.

EIMAS, P. D., COOPER, W. E. and CORBIT, J. D. (1973). Some properties of linguistics feature detectors. *Perception and Psychophysics,* **13**, 247–52.

EIMAS, P. D. and CORBIT, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology,* **4**, 99–109.

EIMAS, P. D., SIQUELAND, E. R., JUSCZYK, P. and VIGORITO, J. M. (1971). Speech perception in infants. *Science,* **171**, 303–6.

FUJISAKI, H., and KAWASHIMA, T. (1968). The influence of various factors on the identification and discrimination of synthetic speech sounds. *Reports on the Sixth International Congress on Acoustics, Tokyo,* B95–8.

FUJISAKI, H. and KAWASHIMA, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. *Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo,* **29**, 207–14.

JUSCZYK, P. W., ROSNER, B. S., CUTTING, J. E., FOARD, C. F. and SMITH, L. (1975). Categorical perception of nonspeech sounds in the two-month-old infant. Paper presented to the bienniel meeting of the Society for Research in Child Development, April, Denver, Colorado.

KAY, R. H. and MATTHEWS, D. R. (1972). On the existence in human auditory pathways of channels selectively tuned to the modulation present in frequency-modulated tones. *Journal of Physiology*, **225**, 657–77.

LANE, H. (1965). Motor theory of speech perception: a critical review. *Psychological Review*, **72**, 275–309.

LIBERMAN, A. M. (1970). Some characteristics of perception in the speech mode. *Perception and its Disorders*, **48**, 238–54.

LIBERMAN, A. M., COOPER, F. S., SHANKWEILER, D. P. and STUDDERT-KENNEDY, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 631–61.

LIBERMAN, A. M., HARRIS, K. S., HOFFMAN, H. S. and GRIFFITH, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, **54**, 358–68.

LIBERMAN, A. M., MATTINGLY, I. G. and TURVEY, M. T. (1972). Language codes and memory codes. In MELTON, A. W. and MARTIN, E. (Eds), *Coding Processes in Human Memory*. Pp. 307–34. Washington, D.C.: Winston.

LIEBERMAN, P., CRELIN, E. S. and KLATT, D. (1972). Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. *American Anthropologist*, **74**, 287–307.

LOCKE, S. and KELLAR, L. (1973). Categorical perception in a nonlinguistic mode. *Cortex*, **9**, 355–69.

MATTINGLY, I. G., LIBERMAN, A. M., SYRDAL, A. K. and HALWES, T. (1971). Discrimination in speech and nonspeech modes. *Cognitive Psychology*, **2**, 131–57.

MILLER, J. D., WIER, C. C., PASTORE, R. E., KELLY, W. M. and DOOLING, R. M. (1976). Discrimination and labelling of noise-buzz sequences with varying noise-lead times: an example of categorical perception. *Journal of the Acoustical Society of America*, **60**. In press.

PALERMO, D. (1975). Developmental aspects of speech perception: problems for a motor theory. In KAVANAGH, J. F. and CUTTING, J. E. (Eds), *The Role of Speech in Language*. Pp. 149–54. Cambridge, Mass.: M.I.T. Press.

PISONI, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, **13**, 253–60.

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

POSNER, M. I. (1976). The temporal course of pattern recognition in the human brain. In INBAR, G. F. (Ed.), *Signal Analysis and Pattern Recognition in Biomedical Engineering*. Tel Aviv: Israel Universities Press. In press.

STEVENS, K. N. (1960). Toward a model for speech recognition. *Journal of the Acoustical Society of America*, **32**, 47–55.

STEVENS, K. N. (1972). Segments, features and analysis by synthesis. In KAVANAGH, J. F. and MATTINGLY, I. G. (Eds), *Language by Ear and by Eye*. Pp. 47–52. Cambridge, Mass.: M.I.T. Press.

STEVENS, K. N. and HOUSE, A. S. (1972). Speech perception. In TOBIAS, J. (Ed.), *Foundations of Modern Auditory Theory*. Vol. 2, pp. 1–62. New York: Academic Press.

STUDDERT-KENNEDY, M. (1975). From continuous signal to discrete message: syllable to phoneme. In KAVANAGH, J. F. and CUTTING, J. E. (Eds), *The Role of Speech in Language*. Pp. 113–26. Cambridge, Mass.: M.I.T. Press.

STUDDERT-KENNEDY, M., LIBERMAN, A. M., HARRIS, K. S. and COOPER, F. S. (1970). Motor theory of speech perception: a reply to Lane's critical review. *Psychological Review*, **77**, 234–49.

WOOD, C. C. (1974). Parallel processing of auditory and phonetic information in speech perception. *Perception and Psychophysics*, **15**, 501–8.

WOOD, C. C. (1975). Auditory and phonetic levels of processing in speech perception: Neurophysiological and information-processing analyses. *Journal of Experimental Psychology: Human Perception and Performance*, **1**, 3–20.

WOOD, C. C. and DAY, R. S. (1975). Failure of selective attention to phonetic segments in consonant–vowel syllables. *Perception and Psychophysics*, **17**, 346–50.

Received 14 *July* 1975