RESEARCH ON SPEECH PERCEPTION

Progress Report No. 3

August 1975 - November 1976

David B. Pisoni
Principal Investigator

Department of Psychology

Indiana University

Bloomington, Indiana 47401

Supported by:

Department of Health, Education, and Welfare
U.S. Public Health Service

National Institute of Mental Health
Grant No. MH-24027-03

and

National Institutes of Health
Grant No. NS-12179-02

CONTENTS

INTRODUCTION

This is the third report of research activities on speech processing conducted in the Department of Psychology at Indiana University. As with our previous progress reports, our main goal has been to summarize our various research activities over the past year and make them available to interested colleagues in the field. Some of the papers contained in this report are extended manuscripts that have been prepared for formal publication as journal articles or book chapters. Other papers are simply short reports of research that has been presented at professional meetings during the past year or brief progress reports on the status of on-going research projects in the laboratory. We also have included new information on instrumentation developments and software support when we think this information would be of interest or help to others.

In addition to the progress report, we have also begun a technical report this past year for much longer manuscripts and theses that were carried out in connection with the overall objectives of our research on speech. Copies of these technical reports are made available to interested colleagues at the time they are issued. Additional copies may be obtained by request as long as the supply lasts.

We decided to issue a yearly progress report of our research activities primarily because of the lag in journal publications and the resulting delay in the dissemination of new information and research findings. We are, of course, very interested in following the work of other colleagues who are carrying out research on speech and, therefore, would be most grateful if you would send us copies of your own recent reprints, preprints and progress reports as they become available so that we can keep up with your latest findings. Please address all correspondence to:

Professor David B. Pisoni
Department of Psychology
Indiana University
Bloomington, Indiana 47401
U.S.A.

EXTENDED MANUSCRIPTS

Identification and Discrimination of the Relative Onset Time of Two Component Tones: Implications for Voicing Perception in Stops*

David B. Pisoni

Research Laboratory of Electronics

Massachusetts Institute of Technology

Cambridge, Massachusetts 02139

Abstract

Experiments on the voiced-voiceless distinction in stop consonants have shown sharp and consistent labeling functions and categorical-like discrimination functions for synthetically produced speech stimuli differing in voice-onset time (VOT). Other research has found somewhat comparable results for young infants and chinchillas as well as cross-language differences in the perception of these same synthetic stimuli. In the present paper, four experiments were carried out to investigate a possible underlying basis of these seemingly diverse results. All of the experiments employed a set of non-speech tonal stimuli that differed in the relative onset time of their components. In the first experiment identification and discrimination functions were obtained with these signals which showed strong evidence for categorical perception: the labeling functions were sharp and consistent, the discrimination functions showed peaks and troughs which were correlated with the labeling probabilities. Other experiments provided evidence for the presence of three distinct categories along this non-speech stimulus continuum which were separated by narrow regions of high discriminability. Based on these findings a general account of voicing perception for stops in initial position is proposed in terms of the discriminability of differences in the temporal order of the component events at onset.

* This paper was written while the author was on leave at M.I.T. during the academic year 1975-1976.

Identification and Discrimination of the Relative Onset Time of Two Component Tones: Implications for Voicing Perception in Stops*

David B. Pisoni

Research Laboratory of Electronics

Massachusetts Institute of Technology

Cambridge, Massachusetts 02139

Introduction

Within the last few years considerable attention has been devoted to the study of the voicing feature in stop consonants, particularly in terms of the dimension of voice onset time (VOT). The important work of Lisker and Abramson (1964, 1970) has shown that the voicing and aspiration differences among stop consonants in a wide diversity of languages can be characterized by changes in VOT, which, in turn, reflect differences in the timing of glottal activity relative to supralaryngeal events. According to Lisker and Abramson (1964) it appears that there are three primary modes of voicing in stops: (1) prevoiced stops in which voicing onset precedes the release burst, (2) shortlag voiced stops in which voicing onset is simultaneous or briefly lags behind the release burst, and (3) long lag voiceless stops in which the voicing onset lags behind the release burst. From acoustic measurements, Lisker and Abramson (1964) found relatively little overlap in the modal values of VOT for the voicing distinctions that occurred in the eleven languages that they studied. Moreover, in perceptual experiments with synthetic stimuli they found that subjects identify and discriminate differences in VOT in a categorical-like manner that reflects the phonological categories of their language (Abramson & Lisker, 1965, 1970; Lisker & Abramson, 1970). That is, subjects show consistent labeling functions with a sharp cross-over point from one phonological category to another and discontinuities in discrimination that are correlated with the abrupt changes in the labeling functions. Subjects can discriminate two synthetic stimuli drawn from different phonological categories better than two stimuli selected from the same phonological category (Liberman, Harris, Hoffman & Griffith, 1957; Liberman, Harris, Kinney & Lane, 1961).

The categorical perception of these synthetic stimuli has been interpreted as evidence for the operation of a special mode of perception, a speech mode, that is unique to the processing of speech signals (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967; Liberman, 1970; Studdert-Kennedy, Liberman, Cooper & Harris, 1970). The argument for the presence of a specialized speech mode is based primarily on three empirical findings. First, non-speech signals are typically perceived in a continuous mode; discrimination is monotonic with the physical scale. It is well known that subjects can discriminate many more differences than they can reliably label on an absolute basis. Second, until recently, no convincing demonstrations of categorical perception had been obtained with non-speech signals. Third, it has generally been assumed that the non-monotonic discrimination functions are entirely the result of labeling processes associated with phonetic categorization. Indeed, the non-speech control experiments carried out by Liberman et al., (1961) and Mattingly, Liberman, Syrdal and Halwes (1971) were designed specifically to determine whether the discontinuities in the speech discrimination functions are due to the acoustic or psychophysical attributes of the signals themselves rather than some speech-related labeling process. Since both of these studies failed to find peaks in the nonspeech discrimination functions at phoneme boundaries, it was concluded that the discrimination functions for the speech stimuli were attributable to phonetic categorization resulting from the stimuli being perceived as speech.

Additional support for the existence of a specialized speech perception mode has come from the results of Eimas and his associates who found that two and three month old infants could discriminate synthetic speech sounds varying in VOT in a manner comparable to that of English-speaking adults (Eimas, Siqueland, Jusczyk & Vigorito, 1971). The infants could discriminate between two speech sounds selected from across an adult phoneme boundary but failed to discriminate two stimuli selected from within an adult phonological category even though the acoustic differences between the pairs of stimuli were apparently constant. The implication of these findings is that the infants have access to mechanisms of phonetic categorization at an extremely early age. Furthermore, it has been suggested that these mechanisms are in some way innately determined or develop very rapidly after birth. The important point is that it has been assumed that infants are responding to differences in VOT in a "linguistically relevant" manner which is a consequence of phonetic coding of these signals rather than responding to psychophysical differences prior to phonetic categorization (however, see Stevens & Klatt, 1974). If this claim is true, or even partly true, it would provide very strong support for an account of phonological perception based on a set of universal phonetic features which are innately determined.

It would also suggest that the environment plays a secondary role in phonological development.

Several recent studies, however, have provided some strong evidence for reevaluating this interpretation of the infant data as well as the more general claims associated with a specialized mode of speech perception. These results are based on perceptual experiments with chinchillas (Kuhl & Miller, 1975), two cross-language experiments with young infants (Lasky, Syrdal-Lasky & Klein, 1975; Streeter, 1976) and a study involving more complex non-speech signals (Miller, Wier, Pastore, Kelly & Dooling, 1976). The common property of these seemingly diverse studies is that they have focused on the voicing distinction in stop consonants, specifically VOT.

Kuhl & Miller (1975) showed that chinchillas could be trained to respond differentially to the consonants /d/ and /t/ in syllables produced by four talkers in three vowel contexts. More importantly, however, was the finding that the training generalized to a continuum of synthetically produced stimuli varying in VOT. The identification functions for chinchilla were quite similar to human data: the synthetic stimuli were partitioned into two discrete categories with a sharp cross-over point. The phoneme boundary for chinchilla occurred at almost precisely the same place as for humans which suggests a psychophysical rather than a phonetic basis for the labeling behavior. Since chinchillas presumably have no spoken language and consequently have no phonological coding system, Kuhl & Miller assumed that the labeling behavior in response to synthetic stimuli would be determined exclusively by the acoustic attributes and psychophysical properties of these signals. The results of this study indicate that the boundary between voiced and voiceless labial stops that occurs at about +25 msec is probably a "natural" region of high sensitivity along the VOT continuum and, at least in the case of the chinchilla, has little to do with phonetic coding.

Following Eimas' work with infants from English speaking environments, two cross-language studies were conducted recently using similar methodology and comparable synthetic stimuli differing in VOT. Lasky et al. (1975) studied four to six and one-half-month-old infants born to Spanish-speaking parents and found evidence for the presence of three categories in their discrimination functions. One boundary occurred in the region of +20 to +60 msec which corresponds to the English voiced-voiceless distinction, whereas the other boundary occurred in the region between roughly -20 and -60 msec. These results are interesting because Spanish has only one phoneme boundary separating its voiced-voiceless stops and this boundary does not coincide with either of the two boundaries found in the infant data of Lasky et al. (see for example, Abramson & Lisker, 1973, for relevant adult Spanish data). One conclusion that can be drawn from these findings is that the environment probably plays only a minor role in phonological development at this age and that the infants are more likely to be responding to some set of acoustic attributes independently of their phonetic status.

In another related study Streeter (1976) found that Kikuyu infants also show evidence of three categories of voicing for labial stops. This result is important because in Kikuyu there are no voicing contrasts for labial stops although there are contrasts at other places of articulation. Since this particular distinction is not phonemic in the adult language, it probably never occurred in the language environment of these infants. As a consequence, the infants' discrimination performance must be due entirely to the acoustic and psychophysical attributes of the stimuli. This conclusion is supported by the fact that the categories and boundaries found in this study were quite comparable to those obtained in the Lasky et al. study.

The results of both cross-language investigations of voicing perception are quite similar and indicate that infants can discriminate differences in VOT. Moreover, the pattern of results suggests that infants have the ability to deal with at least three modes of voicing. The basis of these distinctions, however, may be the result of naturally defined regions of high discriminability along the VOT continuum rather than processes of phonetic categorization. Thus, the infants may not be responding to these signals linguistically as suggested by the earlier interpretation of Eimas, but instead may be responding to some complex psychophysical relation that occurs between the components of the stimulus at each of these modes of voicing. In anticipation, one such relation is strongly suggested by the results of the present series of non-speech experiments in terms of changes in sensitivity to differences in temporal order between two components of a stimulus complex. The infants may be responding simply to differences between simultaneous and successive events.

In another study, Miller et al. (1976) generated a set of non-speech control signals that were purported to be analogous to VOT stimuli. The stimuli differed in the duration of a noise burst preceding a buzz. Identification and discrimination functions were obtained with adults in a manner comparable to those collected in the earlier adult speech perception experiments.

For discrimination, the stimuli were presented in an oddity paradigm, whereas for labeling the subjects responded with two choices, either "no noise" or "noise" present before the onset of the buzz. The results of this study revealed identification and discrimination functions that were similar to those found with stop consonants differing in VOT. Discrimination was excellent for stimuli selected from between categories and quite poor for stimuli from within a category. The labeling functions were sharp and consistent; the peak in discrimination occurred at roughly the boundary between the two categories.

Miller et al. (1976) offered a psychophysical account of these categorical-like results in terms of the presence of a perceptual threshold at the boundary between two perceptually distinctive categories. According to Miller et al., in the case of noise-buzz stimuli, there is a certain value of noise-lead time below which subjects can no longer detect the presence of the noise preceding a buzz. At values below this duration the stimuli are perceived as members of one category and subjects cannot discriminate differences in duration between stimuli because they are below threshold. At noise durations slightly above this value there are marked changes in sensitivity and response bias as a threshold is crossed and a new perceptual quality emerges from the stimulus complex. Miller et al. suggest that discrimination of differences above this threshold value follow Weber's law and, consequently, constant ratios are needed rather than constant differences in order to maintain the same level of discriminability. The boundary between these categories separates distinct sets of perceptual attributes and results in the partitioning of the stimulus continuum into equivalence classes. These equivalence classes for most purposes are categorical: the relation defining membership in a class is symmetrical, reflexive and transitive (Bruner, Goodnow & Austin, 1956).

The account of the labeling and discrimination data offered by Miller et al. suggests the presence of naturally determined boundaries at specific regions along the VOT continuum. These boundaries occur at places where a new perceptual attribute emerges in the course of continuous variations in one or more parameters of a complex signal. In the Miller et al. study the experimental variable was the duration of a noise burst preceding a buzz which was varied over a relatively smaller range of values. Based on their suggestion we generated a set of non-speech signals that differed in the temporal order of the onsets of two component tones of different frequencies. The stimuli varied over a range from -50 msec where the lower tone leads the higher tone, through simultaneity, to +50 msec where the lower tone lags behind the higher tone. Our goal in producing these stimuli was to have a set of non-speech stimuli that differed on a variable known to play an important role in the perception of voicing, namely the relative timing between two events. A well-known and important cue to voicing in stops is laryngeal timing. One of several cues to laryngeal timing is the onset of the first formant relative to the second, the "cutback cue" (Liberman, Delattre & Cooper, 1958). Thus, in using non-speech signals such as these, we hoped to learn something about how the timing relations in stop consonants are perceived. Moreover, we hoped that these results would provide the basis for a more general account of the diverse findings obtained with adults, infants and chinchillas on VOT stimuli as well as an account of the results obtained with the non-speech stimuli.

Experiment I

In this experiment, subjects were trained to identify stimuli selected from a non-speech auditory continuum by means of a disjunctive conditioning procedure (Lane, 1965). The results of this study serve as the baseline for our subsequent experiments.

Method

Subjects. Eight paid volunteers served as subjects. They were recruited by means of an advertisement in a student newspaper and were paid at a base rate of $2.00 per hour plus whatever they could earn during the course of the experiment. All were right-handed native speakers of English.

Stimuli. The stimuli consisted of eleven two-tone sequences that were generated digitally with a computer program that permits the user to specify the amplitude and frequency of two sinusoids at successive moments in time.+ Schematic representations of three of the signals are shown in Figure 1.

Insert Figure 1 about here

The lower tone was set at 500 Hz, the higher tone at 1500 Hz. The amplitude of the 1500 Hz tone was 12 dB lower than the 500 Hz tone to preserve the amplitude relations found for a neutral vowel. The experimental variable under consideration was the onset time of the lower tone relative to the higher tone. For the -50 msec stimulus, the lower tone leads the higher by 50 msec, for the 0 msec condition both tones were simultaneous, and for the +50 msec condition the lower tone lags the higher tone by 50 msec. All the remaining intermediate values, which differed in 10 msec steps from -50 msec through +50 msec, were also generated. Both tones were terminated together. In all cases, the duration of the 1500 Hz tone was held constant at 230 msec and only the duration of the 500 Hz tone was varied to produce these stimuli. The eleven stimuli were recorded on audio tape and later digitized via an A-D converter and stored in digital form.

Procedure. All experimental events involving the presentation of stimuli, collection of responses, and feedback were controlled by a small laboratory computer. The digitized waveforms of the test signals were reconverted to analog form via a D-A converter and presented to subjects binaurally through Telephonics (TDH-39) matched and calibrated headphones. The stimuli were presented at a comfortable listening level of about 80 dB (re 0.0002 dyne/cm^2) which was maintained consistently throughout all the experiments to be reported here.

The present experiment consisted of two 1-hour sessions which were conducted on separate days. All subjects were run in small groups. The order of presentation of the test sequences is given in Table 1. On Day 1 subjects received identification training sequences; on Day 2 they were tested for identification and ABX discrimination.

Insert Table 1 about here

In the initial training sessions, subjects were presented with the end point stimuli, -50 and +50, in a random order and were told to learn which one of two buttons was associated with each sound. Immediate feedback for the correct response was provided by turning on a light above the response button although no explicit coding or labeling instructions were given. Subjects were free to adopt their own strategies. After 320 trials, two additional intermediate stimuli (-30 and +30) were included as training stimuli. Immediate feedback was maintained throughout the training conditions.

For identification, subjects were presented with all 11 stimuli in random order and told to respond in this condition exactly as before. However, no feedback was provided in this condition. In ABX discrimination all nine two-step pairs along the continuum were arranged in the four ABX permutations and presented to subjects with feedback for the correct response. Subjects were told to determine whether the third sound was most like the first sound or the second sound. Timing and sequencing in the experiment were self-paced to the slowest subject in a given session.

Results and Discussion

All eight subjects learned to respond to the endpoint stimuli with a probability of greater than .90 during the training sessions. The results of the identificaiton and ABX discrimination tests are shown in Figure 2.

Insert Figure 2 about here

The labeling functions are shown by the filled circles and triangles connected by solid lines. For five of the eight subjects (S1, S2, S4, S7, S8), the labeling functions are extremely sharp and consistent and show only a very small region of ambiguity between the two response categories. For the remaining three subjects (S3, S5, S6), the labeling functions are less consistent although with additional training these functions would probably have leveled out. For the most part, however, the labeling data for these non-speech signals are quite good, given the modest number of training trials (560) over the two day experiment.

The cross-over points for the category boundary for six of the eight subjects do not occur precisely at the 0 onset time value but are displaced towards the category containing lagging stimuli. These are shown separately for each subject in Figure 2. This asymmetry might be due to either the relatively greater masking of high frequencies by low frequencies or to some limitation on processing temporal order information. In order to test the masking interpretation, we ran a pilot study in which the amplitude of the 1500 Hz tone was varied over a 24 dB range from -12 dB through +12 dB relative to the amplitude of the 500 Hz tone. If the asymmetry in the labeling function is caused by masking of the higher tone by the lower tone, we would expect increases in the amplitude of the higher tone to produce systematic shifts in the locus of the category boundary towards progressively shorter onset-time values. No such shift was observed in the pilot experiment, which suggests that the temporal order account is the more likely cause of the asymmetry in the placement of the category boundary. The results of the subsequent experiments reported below also support this conclusion.

The observed two-step ABX discrimination functions are shown by open circles and broken lines and are plotted over the corresponding labeling functions for comparison. Most subjects show evidence of categorical-like discrimination: there is a peak in the discrimination function at the category boundary and there are troughs within both categories. Subject S2 is the most extreme example in the group, showing very nearly the idealized form of categorical perception (Studdert-Kennedy et al., 1970).

The labeling data and the discrimination functions indicate that categorical perception can be obtained with these non-speech signals. To test the strength of these results against the categorical perception model, the ABX predictions from the labeling probabilities were compared with the observed discrimination functions (Liberman et al., 1957). A chi-square test was used to assess the goodness of fit between the expected discrimination functions and the observed functions (Pisoni, 1971). The observed and predicted discrimination scores, as well as the individual chi-square values for each subject are given in Table 2.

Insert Table 2 about here

The fit of the observed and prediction functions is quite good in several cases such as S2 and S6 as shown by the low chi-squares. In other cases, however, the fits are poor and the chi-squares reach a very conservative level of significance (i.e., S1, S4). In the case of S4 the discrimination function is the right shape and level but is just shifted slightly from the discrimination functions predicted from the labeling probabilities.

In general, however, the data from the present experiment show categorical perception effects that are at least as comparable as those obtained with speech sounds, particularly stop consonants. Thus, the results of this study serve as another demonstration of categorical perception with non-speech signals and suggest that this form of perception is not unique to speech stimuli (see Cutting & Rosner, 1974). But what is the basis for the present categorical perception results? Are these results due to some labeling process brought about by the training procedures as Lane (1965) has argued or is there a simpler psychophysical explanation? In order to rule out the labeling explanation, it is necessary to obtain ABX discrimination functions before any training experience. If peaks in discrimination still remain in the absence of any labeling experience we will have reason to suspect some psychophysical basis to the observed discrimination functions. The next experiment was carried out to test this hypothesis.

Experiment II

Method

Subjects. Twelve volunteers served as subjects. They were recruited in the same way as the subjects from the previous experiment and met the same selection requirements.

Stimuli. The eleven stimuli of Experiment I were also used in the present experiment.

Procedure. The procedures for the ABX discrimination tests were identical to those used in the previous experiment. The experiment consisted of two one-hour sessions held on separate days. On each day the subjects received 360 ABX trials with immediate feedback provided for the correct response. In the course of the experiment each of the nine two-step stimulus comparisons was responded to 80 times by each subject.

Results and Discussion

The ABX discrimination functions for all twelve subjects are shown in Figure 3. Except for S1 whose performance is close to chance, all of the

----

Insert Figure 3 about here

----

other subjects show one of two patterns of discrimination performance. Four of the subjects show evidence of a single peak in the discrimination function at approximately +20 msec, whereas the rest of the subjects show discrimination functions with two peaks. For this group one peak occurs at approximately +20 msec whereas a second peak occurs at approximately -20 msec. Broken vertical lines have been drawn through the discrimination functions at values of -20 msec and +20 msec to facilitate these comparisons.

The peak in discrimination at +20 msec is comparable to that found in the previous experiment. A re-examination of Figure 2 also shows some evidence of a smaller peak in the -20 msec range for several subjects in Experiment I, although the major peak occurs at +20 msec and is correlated with changes in the labeling function.

It is clear from the results of the present experiment that the peaks in discrimination do not arise solely from the training procedures employed in Experiment I and the associated labels. Rather, it appears that natural categories are present at places along the stimulus continuum that are marked by narrow regions of high sensitivity. Based on these results, it is possible to describe three categories within the -50 msec through +50 msec region. Going from left to right, the first category contains stimuli with the lower tone leading by 20 msec or more; the second category contains stimuli in which both tones occur more-or-less simultaneously within the -20 msec to +20 msec region, whereas the third category contains stimuli in which the lower tone lags behind the higher tone by 20 msec or more. Within the context of this experiment, the three regions correspond, respectively, to leading, simultaneous and lagging temporal events.

The presence of peaks in the ABX discrimination functions for these non-speech stimuli is in sharp contrast to the results obtained previously by Liberman et al., (1961) and Mattingly et al., (1971) who found marked differences in discrimination between speech and non-speech signals. In these experiments, non-speech control stimuli were created that nominally contained the same acoustic properties of speech but, nevertheless, did not sound like speech. For example, in the Liberman et al., (1961) study, the synthetic spectrograms of the /do/-/to/ stimuli were inverted before being converted to sound on the pattern playback. In the Mattingly et al. (1971) study, the second formant transitions (i.e., chirps) were isolated from the rest of the stimulus pattern since it was assumed that these acoustic cues carry the essential information for place of articulation. When these non-speech stimuli were presented to subjects in a discrimination task the discrimination functions that were obtained failed to show peaks and troughs that corresponded to those found with the parallel set of speech stimuli from which they were derived. The discrimination functions were flat and very nearly close to chance in most of the cases, especially in the earlier study by Liberman and his co-workers.

The failure to find peaks and troughs in the discrimination functions of the non-speech control stimuli may have been due to the lack of familiarity with these stimuli and the absence of any feedback during the discrimination task. With complex multidimensional signals it may be difficult for subjects to attend to the relevant attributes that distinguish these stimuli. For example, if the subject is not specifically attending to the initial portion of the stimulus but focusses instead on other properties, his discrimination performance may be no better than chance. Indeed, the Liberman et al. results indicate precisely this. Moreover, without feedback in experiments such as this the subject may focus on one aspect or set of attributes on a given trial and a different aspect of the stimulus on the next trial. As a result, the subject may respond to the same stimulus quite differently at different times during the course of the experiment. The results of the present experiment strongly indicate that non-speech signals can be responded to consistently and reliably from trial to trial when the subject is provided with information about the relevant stimulus parameters that control his response.

The argument for the presence of three natural categories and our interpretation of the previous non-speech control experiments would be strengthened if it could be demonstrated that subjects can classify these same stimuli into three distinct categories whose boundaries occur at precisely these regions on the continuum. We addressed this question in the next experiment.

Experiment III

In this experiment we used the same training procedures as in the first experiment except that subjects were now required to use three response categories instead of two. Our aim was to determine whether subjects would partition the stimulus continuum consistently into three distinct categories and whether the boundaries would lie at the same points of high discriminability identified in the previous experiment.

Method

Subjects. Eight additional subjects were recruited for the present experiment. They were obtained from the same source and met the same requirements as the subjects used in the previous experiments.

Stimuli. The same basic set of eleven tonal stimuli were also used in the present experiment.

Procedure. The experiment took place on two separate days. The first day was devoted to shaping and identification training with three stimuli; on the second day the labeling tests were conducted. Subjects were not given any explicit labels to use in the task and, as in the previous experiment, were free to adopt their own coding strategies. The procedure used in this experiment was similar to that used in Experiment I. Subjects were presented with three training stimuli, -50, 0 and +50 msec and were told to learn to respond differentially to these signals by pressing one of three buttons located on a response box. The order of presentation of the test sequences is given in Table 3. Immediate feedback was provided for the correct response in each case.

Insert Table 3 about here

Results and Discussion

The identification functions for the eight subjects are shown in Figure 4. As shown here, all subjects partitioned the stimulus continuum

----

Insert Figure 4 about here

----

into three well-defined categories. As anticipated, the boundaries between categories occur at approximately -20 msec and +20 msec. While there is some noise in the data when compared to the results of the first experiment, it is clear that subjects could reliably and consistently use the three responses and associate them with three distinct sets of attributes along the stimulus continuum. There is very little confusion or overlap between the three response categories although the results are not nearly as consistent as those obtained with the stop consonants by Liberman and others (Liberman et al., 1957; Mattingly et al., 1971; Pisoni, 1971).

The identification data from this experiment would probably have been more consistent if additional members of each category were used during training as in the first experiment and if the range of stimuli was expanded slightly. Because of time constraints we used only one exemplar of each category during training. Further experiments are currently underway to resolve these issues.

In this experiment we did not explicitly provide subjects with an appropriate set of labels to use in encoding these sounds, although it is likely that they invented ones of their own. We assumed that by training subjects on representative members of a category we could reveal some aspects of the underlying categorization process, and therefore gain some insight into the basis for defining category membership. The results of these experiments have revealed the presence of three natural categories that can be defined by the presence of certain distinct perceptual attributes at onset. For many subjects these categories are separated by regions of high discriminability corresponding more or less to what might be called a perceptual threshold. We suggested earlier that the three categories observed along this continuum could be characterized by the subject's ability to discriminate differences in temporal order among the components of a stimulus complex. Thus, the middle category corresponds to stimuli that listeners judge to be more or less simultaneous at onset whereas both of the two other categories contain stimuli that are judged to contain two distinct events at onset, separated by a discriminable temporal interval (see Hirsh, 1959).

In order to provide additional support for this account, we carried out another experiment in which subjects were required to determine whether there were one or two distinct events at stimulus onset. The results of this study should provide information bearing on the potential range of attributes that define the perceptual qualities which result from continuous variations in the relative onset of the two tonal components.

Experiment IV

Method

Subjects. Eight additional volunteers were recruited as subjects. None had participated in the previous experiments nor had any of them taken part in a previous psychophysical experiment. Thus, they were experimentally naive observers.

Stimuli. The same eleven tonal stimuli were also used in the present experiment.

Procedure. The experiment was conducted in a single one-hour experimental session. Each of the eleven stimuli was presented singly, in a random order. There were forty replications of each stimulus, which gave a total of 440 trials. Subjects were told to listen to each sound carefully and then to determine whether they could hear one or two events at stimulus onset. They were told that on some trials the two tones would be simultaneous at onset whereas on other trials they would be successive. Subjects were provided with a response box and told to press the button labeled "1" for one event at onset or the button "2" for two events at onset. No feedback was given at any time during the experiment. There was a short break after the first 220 trials.

Results and Discussion

The results for each of the eight subjects are shown in Figure 5 where

---

Insert Figure 5 about here

---

the per cent judgments of two events are displayed as a function of the stimulus value. All subjects showed similar U-shaped functions with fairly sharp cross-over points between categories. There is a region in the center of the continuum, bounded by -20 and +20 msec, which was judged by every subject to contain stimuli whose components are predominantly simultaneous. On the other hand, there are two distinct regions at either end of the continuum in which subjects can reliably judge the presence of two distinct temporal events at stimulus onset, one leading and one lagging. Thus, the results of this experiment, as well as the findings of the other experiments, indicate the presence of three natural categories that may be distinguished by the relative discriminability of the temporal order of the component events. These judgments appear to be relatively easy to make and are consistent from subject to subject. The findings suggest the presence of a fairly robust perceptual effect for processing temporal order information which may also underlie the perception of voicing distinctions in stop consonants in initial position.

General Discussion

The results of the present series of experiments are consistent with the findings of Hirsh (1959), Hirsh & Sherrick (1961), and more recently Stevens and Klatt (1974) who found that 20 msec is about the minimal difference in onset time needed to identify the temporal order of two distinct events. Stimuli with onset times greater than about 20 msec are perceived as successive events; stimuli with onset times less than about 20 msec are perceived as simultaneous events.

Based on the results of the present set of experiments with non-speech stimuli differing in relative onset time, we would like to offer a general account of the labeling and discrimination data that can handle the four seemingly diverse sets of findings that have been previously reported in the literature. To review briefly, these four sets of findings are the perceptual results obtained for: (1) infants, (2) adults, (3) chinchillas with synthetic speech sounds differing in VOT, and (4) the recent findings obtained for adults with non-speech control stimuli differing in noise-lead time. Although specific accounts have been proposed to handle these findings individually, in our view a more general account of voicing perception is preferable.

We suggest that the four sets of findings may simply reflect a basic limitation on the ability to process temporal order information. In the case of the voicing dimension, the time of occurrence of an event (i.e., onset of voicing) must be judged in relation to the temporal attributes of other events (i.e., release from closure). The fact that these events, as well as others involved in VOT, are ordered in time implies that highly distinctive and discriminable changes will be produced at various regions along the temporal continuum. Although continuous variations in the temporal relations may nominally be present in these stimuli, at least according to the experimenter's operational criteria, the only perceptual change to which the listener is sensitive appears to be the presence or absence of a discrete attribute rather than the magnitude of difference between events. Thus, the discrimination of small temporal differences in tasks such as these is relatively poor whereas the discrimination of discrete attributes is excellent. This, of course, is the implication of the previous categorical perception experiments. Phonological systems apparently have exploited this principle during the evolution of language. As Stevens and Klatt (1974) have remarked, the inventory of phonetic features used in natural languages is not a continuous variable but rather consists of the presence or absence of sets of attributes or cues. This seems also to be the case with non-speech stimuli having temporal properties similar to speech.

The account of voicing perception proposed here does not minimize the importance of the F1 transition cue (Stevens & Klatt, 1974; Lisker, 1975) or of the duration of aspiration noise preceding voicing onset (Miller, et al., (1976) as well as the numerous other cues to the voiced-voiceless distinction (Lisker & Abramson, 1964). We would suggest that these cues are simply special cases of the more general process underlying voicing distinctions, namely, whether the events at onset are perceived as simultaneous or successive and if successive what their temporal order is.

It should be pointed out, however, that while the line of argument in this paper has emphasized the temporal domain, there is also strong evidence for some temporal-spectral interaction in voicing perception (Summerfield, 1975). A complete account of voicing perception in stops will, of necessity, have to deal with the findings that the voiced-voiceless boundary varies as a function of the place of articulation of the consonant and the following vowel context. Additional experiments with comparable non-speech signals are currently in progress to see whether these differences are a consequence of phonetic categorization or some more general psychophysical process.

The range of values found in the present experiments between -20 msec and +20 msec probably represents the lower limits on the region of perceived simultaneity. We assume that experience in the environment probably serves to tune and align the voicing boundaries in different languages and, accordingly, there will be some slight modification of the precise values associated with different regions along a temporal continuum such as VOT. It is also possible, as in the case of English voicing contrasts, that if appropriate experience is not forthcoming with the particular distinction, its discriminability will be substantially reduced. The exact mechanism underlying these processes, as well as their developmental course, is under extensive investigation (see Eimas, 1975, 1976).

In summary, the results of these four experiments suggest a general explanation for the perception of voicing contrasts in initial position in terms of the relative discriminability of the temporal order between two or more events. These findings may be thought of as still another example of how languages have exploited the general properties of sensory systems to represent phonetic distinctions. As Stevens (1972) has suggested, all phonetic features of language probably have their roots in acoustic attributes with well-defined properties. We suggest that one of these properties corresponds to simultaneity at stimulus onset as reflected in the perception of voicing contrasts in stops.

Acknowledgements

This research was supported, in part, by NIH grants NS-07040 and NS-04332 to the Research Laboratory of Electronics, Massachusetts Institute of Technology and, in part, by NIMH grant MH-24027 to Indiana University. I thank Jerry Forshee and Judy Hupp for their assistance in carrying out these experiments at Indiana while I was on leave at M.I.T. I am also grateful to Professor Kenneth Stevens and Dr. Dennis Klatt for their interest and advice at various stages of this project.

References

Abramson, A. S. & Lisker, L. (1965) Voice onset time in stop consonants: Acoustic analysis and synthesis. *Proceedings of the 5th International Congress of Acoustics*, Liege, September.

Abramson, A. S. & Lisker, L. (1970) Discriminability along the voicing continuum: Cross-language tests. In *Proceedings of the 6th International Congress of Phonetic Sciences*. Prague: Academic, pp. 569-573.

Abramson, A. S. & Lisker, L. (1973) Voice-timing perception in Spanish word-initial stops. *Journal of Phonetics*, *1*, 1-8.

Bruner, J. S., Goodnow, J. J. & Austin, G. A. (1956) *A study of thinking*. New York: John Wiley.

Cutting, J. E. & Rosner, B. S. (1974) Categories and boundaries in speech and music. *Perception & Psychophysics*, *16*, 564-570.

Eimas, P. D. (1975) Auditory and phonetic coding of the cues for speech: Discrimination of the r-l distinction by young infants. *Perception & Psychophysics*, *18*, 341-347.

Eimas, P. D. (1976) Developmental aspects of speech perception. In R. Held, H. Leibowitz & H. L. Teuber (Eds.) *Handbook of Sensory Physiology: Perception*. New York: Springer-Verlag.

Eimas, P. D., Siqueland, E. R., Jusczyk, P. & Vigorito, J. (1971) Speech Perception in infants. *Science*, *171*, 303-306.

Hirsh, I. J. (1959) Auditory perception of temporal order. *Journal of the Acoustical Society of America*, *31*, 759-767.

Hirsh, I. J. & Sherrick, C. E. (1961) Perceived order in different sense modalities. *Journal of Experimental Psychology*, *62*, 423-432.

Kuhl, P. K. & Miller, J. D. (1975) Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. Science, 190, 69-72.

Lane, H. L. (1965) The motor theory of speech perception: A critical review. Psychological Review, 72, 275-309.

Lasky, R. E., Syrdal-Lasky, A. & Klein, R. E. (1975) VOT discrimination by four to six and a half month old infants from Spanish environments. Journal of Experimental Child Psychology, 20, 213-225.

Liberman, A. M. (1970) Some characteristics of perception in the speech mode. In D. A. Hamburg (Ed.) Perception and Its Disorders, Proceedings of A. R. N. M. D. Baltimore: Williams & Wilkins Co., pp. 238-254.

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

Liberman, A. M., Delattre, P. C. & Cooper, F. S. (1958) Some cues for the distinction between voiced and voiceless stops in initial position. Language and Speech, 1, 3, 353-167.

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

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

Lisker, L. (1975) Is it VOT or a first-formant transition detector? Journal of the Acoustical Society of America, 57, 1547-1551.

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

Lisker, L., & Abramson, A. S. (1970) The voicing dimension: Some experiments in comparative phonetics. Proceedings of the 6th International Congress of Phonetic Sciences, Prague. Academia, pp. 563-567.

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

Miller, J. D., Wier, C. C., Pastore, R., Kelly, W. J., & Dooling, R. J. (1976) Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. Journal of the Acoustical Society of America, 00, 000-000.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Ph.D. Thesis, University of Michigan. Also: Supplement to Status Report on Speech Research. SR-27 New Haven: Haskins Laboratories, 1971, pp. 1-101.

Stevens, K. N. (1972) The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David, Jr. and P. B. Denes (Eds.) Human Communication: A Unified View. New York: McGraw-Hill, pp. 51-66.

Stevens, K. N. & Klatt, D. H. (1974) The role of formant transitions in the voiced-voiceless distinction for stops. Journal of the Acoustical Society of America, 55, 653-659.

Streeter, L. A. (1976) Language perception of 2-month old infants shows effects of both innate mechanisms and experience. Nature, 259, 39-41.

Studdert-Kennedy, M., Liberman, A. M., Harris, K., & Cooper, F. S. (1970) The motor theory of speech perception: A reply to Lane's critical review. Psychological Review, 77, 3, 234-249.

Summerfield, A. Q. (1975) Information-Processing Analyses of Perceptual Adjustments to Source and Context Variables in Speech. Ph.D. Thesis, Queen's University of Belfast.

Footnotes

[*] Reprints may be obtained from the author who has now returned to the Department of Psychology, Indiana University, Bloomington, Indiana 47401.

[+] I am indebted to Dr. Dennis Klatt for his help with the program used to generate these stimuli.

Table 1

Order of Presentation of Training and Test Sequences for Experiment I

Day	Type of Session	Sequence Description	Feedback	Number of Trials
1	Training	Initial Shaping Sequence (-50, +50)	YES	160
1	Training	Identification Training (-50, +50)	YES	160
1	Training	Identification Training (-50, -30; +30, +50)	YES	160
2	Training	Warmup Sequence (-50, +50)	YES	80
2	Labeling	Identification Sequence (all 11 stimuli)	NO	165
2	Discrimination	ABX Discrimination (9 two-step comparisons)	YES	252

Table 2

Observed and Predicted ABX Discrimination Scores and Chi-square Values for Goodness of Fit

Subject		Stimulus comparison (msec)									
		-50/-30	-40/-20	-30/-10	-20/0	-10/+10	0/+20	+10/+30	+20/+40	+30/+50	SUM
1	Observed (O)	.50	.71	.54	.46	.68	.82	.79	.79	.71	
	Predicted (P)	.50	.50	.50	.50	.50	.61	.88	.56	.51	
	O-P	0	.21	.04	.04	.18	.21	.09	.23	.20	
	Chi-square	0	5.04	0	0	3.64	5.32	1.96	5.88	4.76	26.6*
2	Observed	.50	.50	.50	.50	.46	.86	.79	.50	.50	
	Predicted	.50	.50	.50	.50	.50	.88	1.00	.51	.50	
	O-P	0	0	0	0	-.04	.02	.21	.01	0	
	Chi-square	0	0	0	0	0	0	0	0	0	0
3	Observed	.46	.54	.46	.50	.43	.43	.54	.54	.46	
	Predicted	.50	.50	.50	.50	.55	.51	.61	.68	.50	
	O-P	-.04	.04	-.04	0	.12	.08	.07	.14	-.04	
	Chi-square	0	0	0	0	1.68	.56	.56	2.52	0	5.4

Table 2, continued

Observed and Predicted ABX Discrimination Scores and Chi-square Values for Goodness of Fit

		Stimulus comparison (msec)									
Subject		-50/-30	-40/-20	-30/-10	-20/0	-10/+10	0/+20	+10/+30	+20/+40	+30/+50	SUM
4	Observed	.43	.39	.46	.50	.68	.96	.79	.82	.43	
	Predicted	.50	.50	.50	.50	.50	.68	.94	.58	.50	
	O-P	-.07	-.11	-.04	0	.18	.28	-.15	.24	-.07	
	Chi-square	.56	1.12	0	0	3.64	10.36	10.36	6.72	.56	33.3*
5	Observed	.57	.64	.71	.54	.61	.82	.64	.71	.50	
	Predicted	.51	.50	.50	.58	.68	.52	.50	.54	.51	
	O-P	.06	.14	.21	-.04	-.07	.30	.14	.17	.01	
	Chi-square	.56	2.24	5.04	.28	.56	10.80	2.24	3.64	0	24.6
6	Observed	.50	.43	.50	.57	.82	.61	.43	.68	.50	
	Predicted	.51	.50	.52	.58	.56	.56	.52	.51	.50	
	O-P	-.01	-.07	-.02	-.01	.26	.05	-.09	.17	0	
	Chi-square	0	.56	0	0	7.84	.28	.84	.28	0	9.8

Table 2, continued

Observed and Predicted ABX Discrimination Scores and Chi-square Values for Goodness of Fit

		Stimulus comparison (msec)									
Subject		-50/-30	-40/-20	-30/-10	-20/0	-10/+10	0/+20	+ 0/+30	+20/+40	+30/+50	SUM
7	Observed	.50	.61	.54	.43	.71	.93	.96	.50	.57	
	Predicted	.50	.50	.50	.50	.50	.77	1.00	.54	.50	
	O-P	0	.11	.04	-.07	.21	.16	-.04	-.04	.07	
	Chi-square	0	1.12	0	.56	5.04	3.92	0	0	.56	11.2
8	Observed	.46	.50	.68	.57	.68	.79	.96	.75	.61	
	Predicted	.50	.50	.50	.50	.52	.68	.72	.58	.51	
	O-P	-.04	0	.18	.07	.16	.11	.24	.17	.10	
	Chi-square	0	0	3.58	.56	2.80	1.40	8.12	3.36	1.12	20.9

* p < .001

df = 8

Table 3

Order of Presentation of Training and Test Sequences for Experiment III

Day	Type of Session	Sequence Description	Feedback	Number of Trials
1	Training	Initial Shaping Sequence (-50, 0, +50)	YES	180
1	Training	Identification Training (-50, 0, +50)	YES	300
2	Training	Warmup Sequence (-50, 0, +50)	YES	90
2	Labeling	Identification Sequence (all 11 stimuli)	NO	165

Figure Captions

Figure 1. Schematic representations of three stimuli differing in relative onset time: leading (-50 msec), simultaneous (0 msec) and lagging (+50 msec).

Figure 2. Labeling functions are shown by the filled circles and triangles (left ordinate) and ABX discrimination functions by open circles (right ordinate) for individual subjects in Experiment I.

Figure 3. ABX discrimination functions for individual subjects in Experiment II. Broken lines have been drawn through -20 msec and +20 msec to illustrate the three regions of onset time.

Figure 4. Labeling functions for individual subjects after training on -50 msec, 0 msec and +50 msec stimuli as representative of each of the three categories, R1, R2, and R3.

Figure 5. Per cent judgment of two events for individual subjects as a function of relative onset time of the two components. Broken lines have been drawn through -20 msec and +20 msec to permit comparisons.

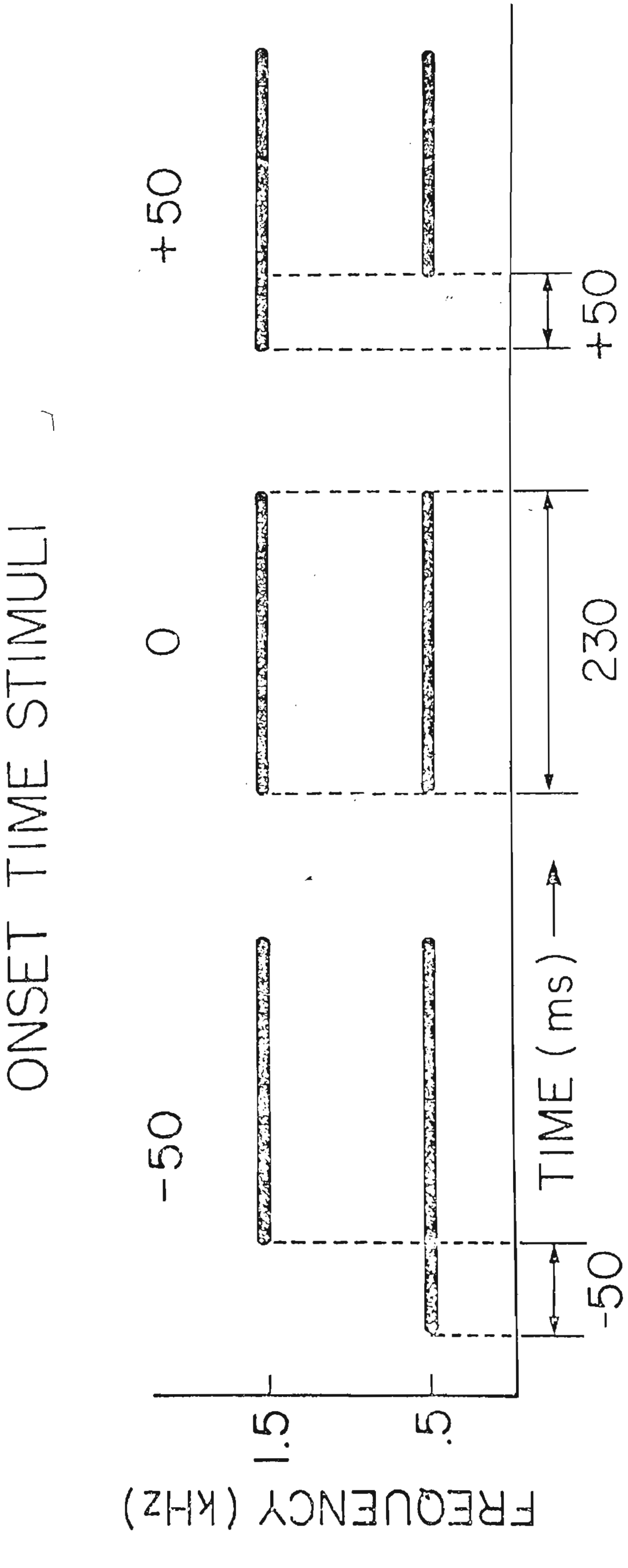


Figure 1.

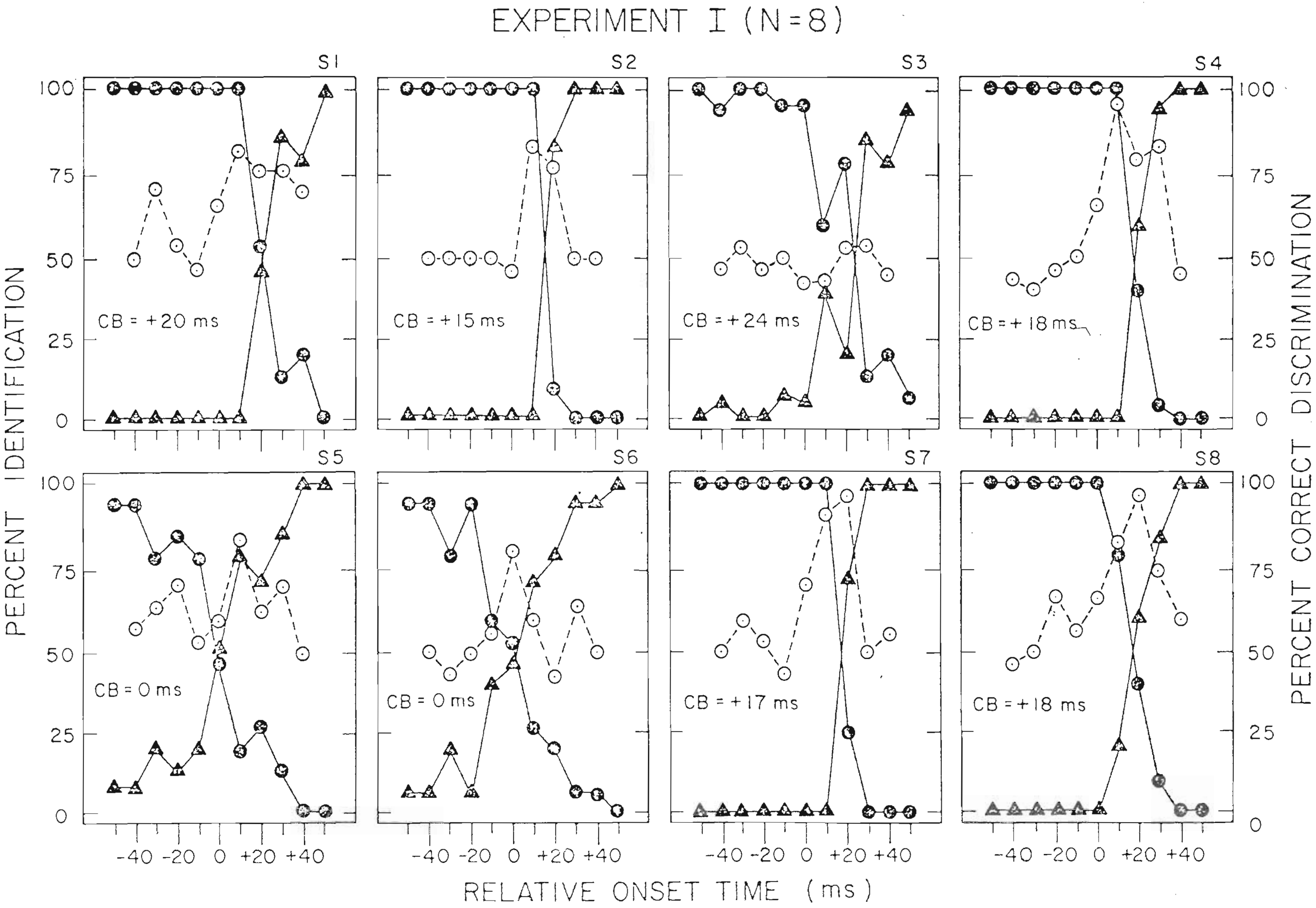

EXPERIMENT I (N=8)
S1
S2
S3
S4
S5
S6
S7
S8
CB = +20 ms
CB = +15 ms
CB = +24 ms
CB = +18 ms
CB = 0 ms
CB = 0 ms
CB = +17 ms
CB = +18 ms
PERCENT IDENTIFICATION
PERCENT CORRECT DISCRIMINATION
100
75
50
25
0
-40 -20 0 +20 +40
RELATIVE ONSET TIME (ms)

Figure 2.

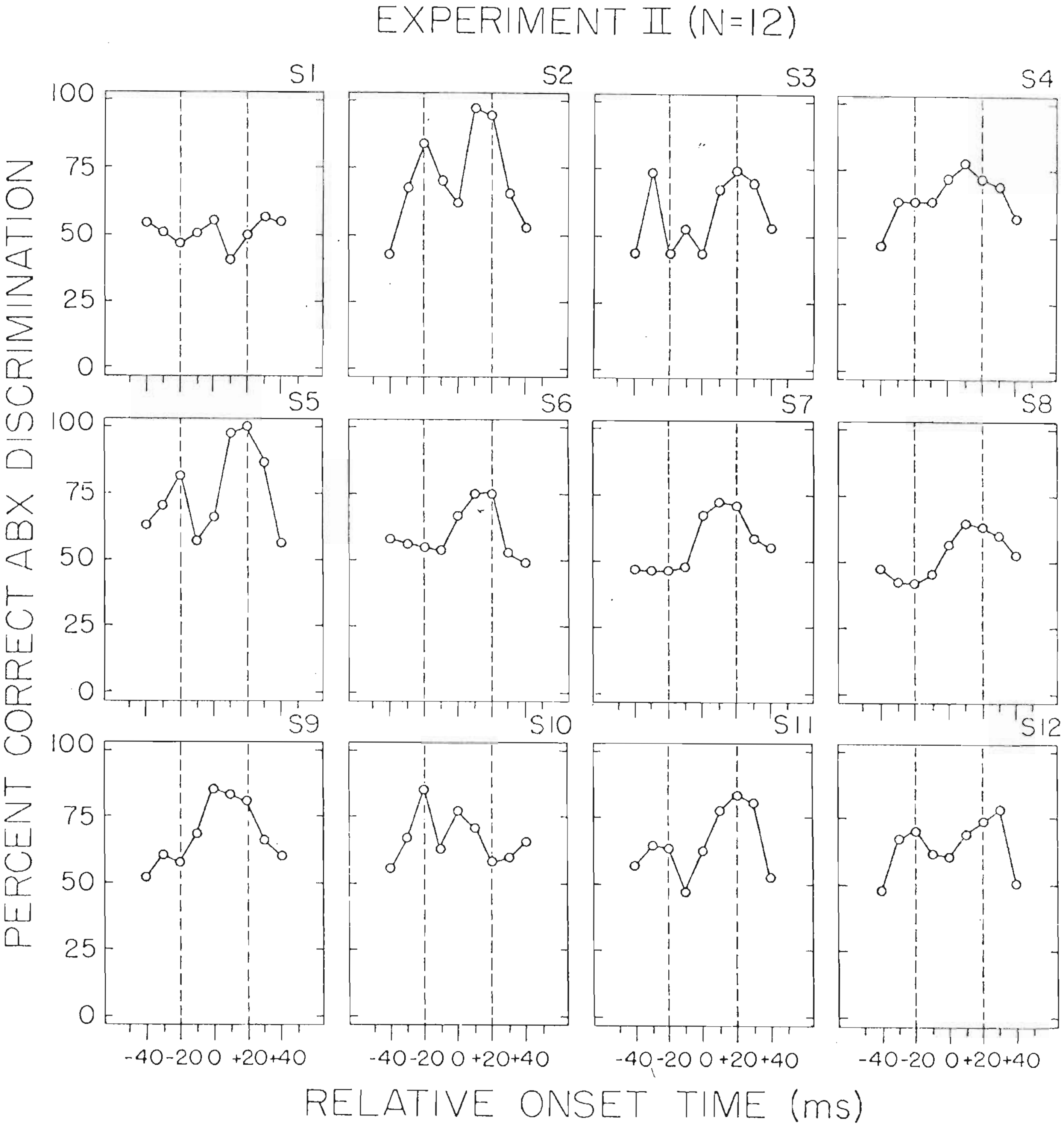

EXPERIMENT II (N=12)
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
100
75
50
25
0
PERCENT CORRECT ABX DISCRIMINATION
-40-20 0 +20+40
RELATIVE ONSET TIME (ms)

Figure 3.

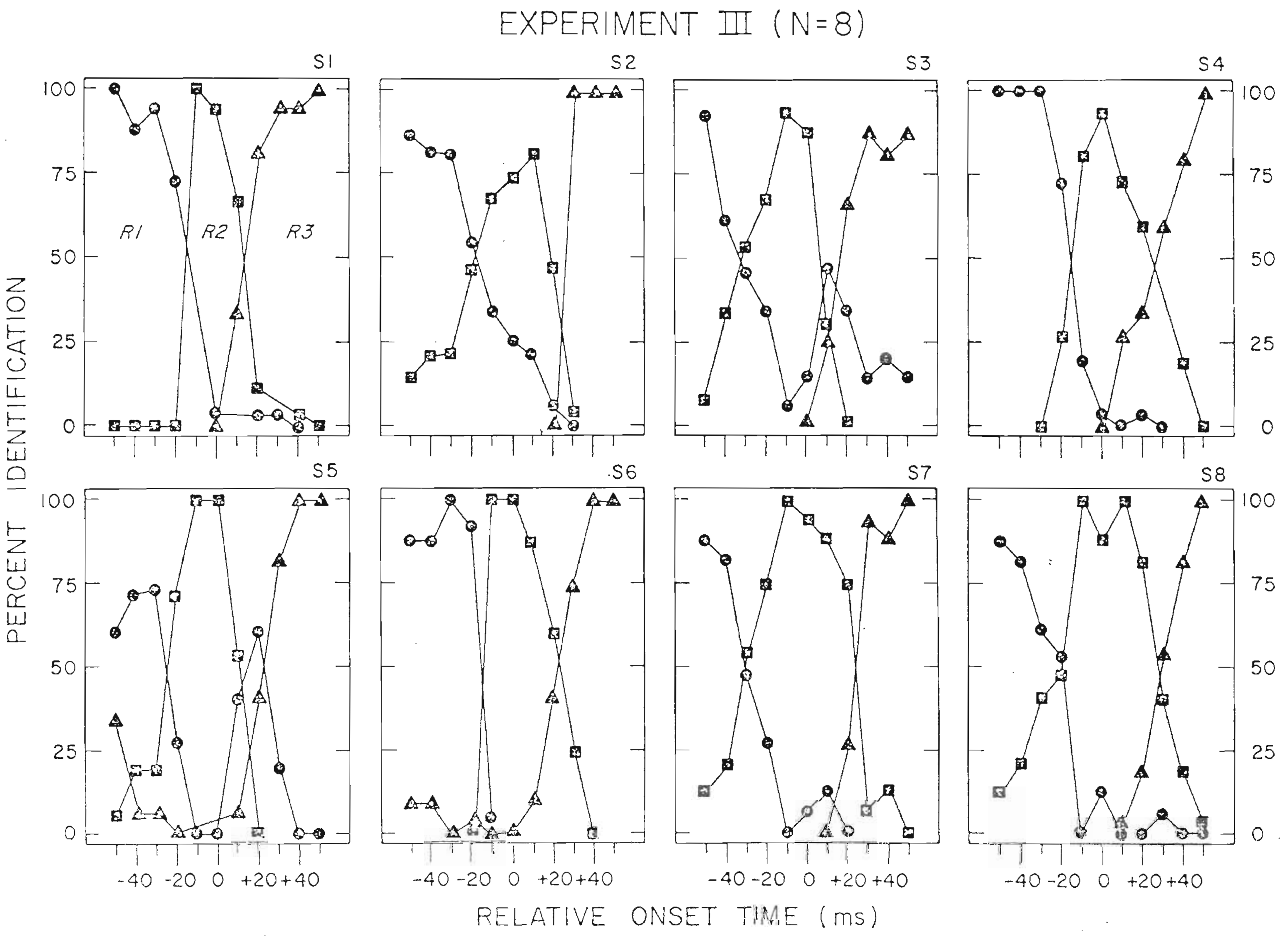

EXPERIMENT III (N=8)
S1
S2
S3
S4
S5
S6
S7
S8
R1
R2
R3
PERCENT IDENTIFICATION
100
75
50
25
0
-40 -20 0 +20 +40
RELATIVE ONSET TIME (ms)

Figure 4.

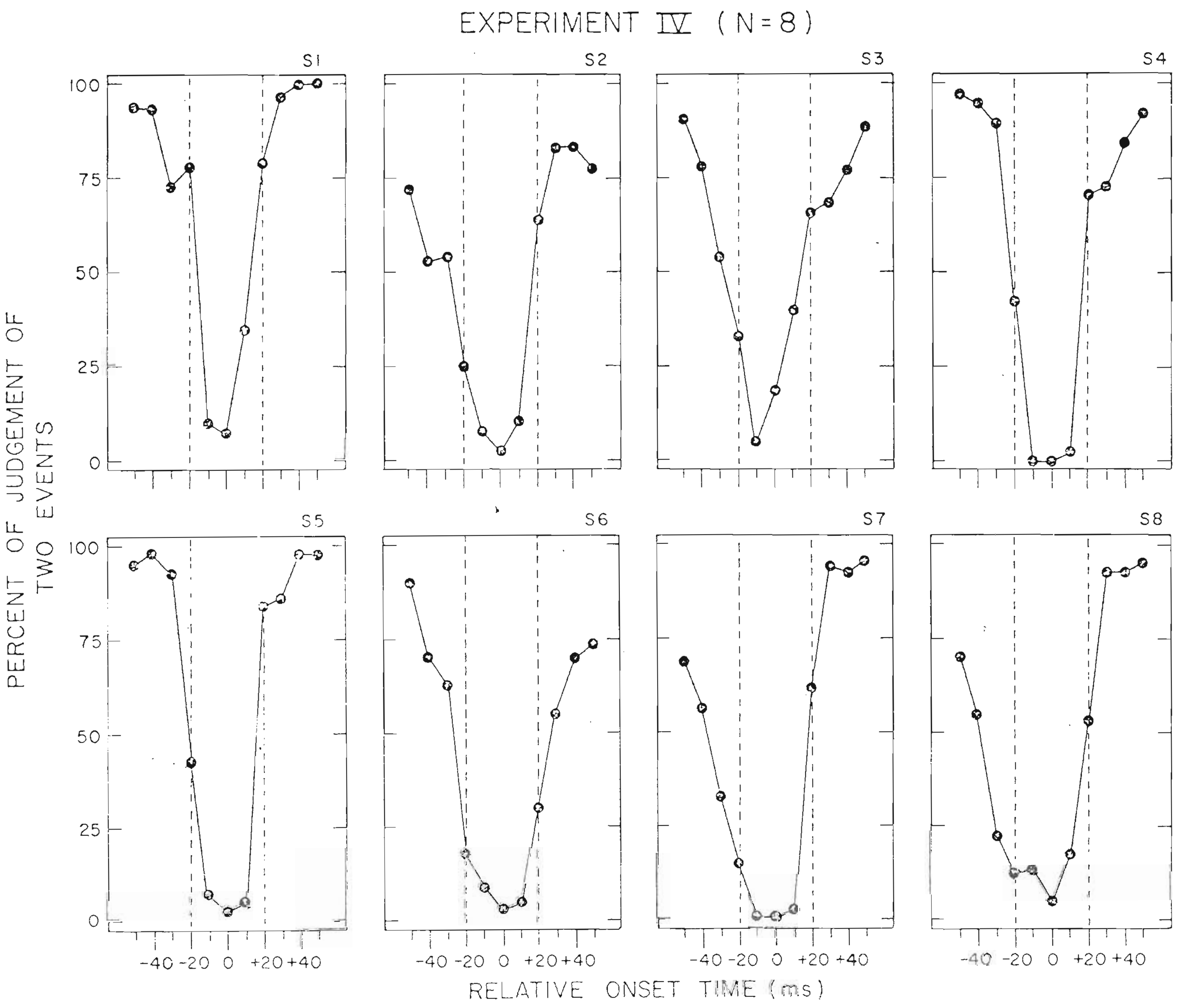

EXPERIMENT IV (N=8)
S1
S2
S3
S4
S5
S6
S7
S8
100
75
50
25
0
PERCENT OF JUDGEMENT OF TWO EVENTS
-40 -20 0 +20 +40
RELATIVE ONSET TIME (ms)

Figure 5.

On the Perception of Speech Sounds as Biologically Significant Signals*

David B. Pisoni

Indiana University

Bloomington, Indiana 47401

* This paper is a substantially revised and expanded version of a talk given as part of a symposium entitled "Perception of Species-specific Vocalizations: From Primates to Human Infants" at the Midwest Regional Meeting of the Animal Behavior Society in Chicago on November 13, 1976. It is to be published in a monograph edited by Philip Morse.

On the Perception of Speech Sounds as Biologically Significant Signals[1]

Introduction

The purpose of this paper is to briefly review some of the evidence currently available in support of the view that the perception of speech sounds may require specialized biological mechanisms. Evidence has accumulated rapidly over the last thirty years to suggest that speech is a special type of acoustic signal that has species-specific properties unique to humans. While the bulk of the evidence for this view has come from work on the acoustic analysis of speech and perceptual experiments with adults using synthetic speech sounds, more recent findings have been obtained with pre-linguistic infants and animals. This work has expanded our knowledge of speech perception in a number of directions and has shown the value of comparative experimentation for the evaluation of species-specific behaviors, particularly those associated with acoustic signaling systems.

Before turning to the main arguments of this paper and the evidence to be discussed, it would be useful first to briefly review the role of speech in language. One of the essential "design features" of all spoken languages is what Hockett (1958) refers to as "duality of patterning." That is, at the message level all languages have two levels of representation, one consisting of the arrangement of meaningless elements (phonemes) and the other consisting of the arrangement of meaningful elements (morphemes or words). Differences between morphemes and hence differences in meanings are realized by variations in the sequencing and arrangement of the constituent phonemes and their features. The phonology of a language represents the sound patterning or structural arrangement of the phonemes whereas the syntax represents the rules for arranging and ordering morphemes and words.

Although the formal analysis of natural languages has been carried out for quite some time and a fairly good understanding of the message units exists, the same cannot be said for the analyses of animal communication systems. Indeed, knowledge of the message level is still a somewhat problematical question. The distinction between message units and sound units or segments should be emphasized here because much of the work to be discussed below tacitly assumes the existence of the phoneme as the smallest unit of linguistic analysis. And while the phoneme may not have physical reality as represented in current analyses of the acoustic waveform, it does have psychological reality in terms of the functional sound systems employed by natural languages. In studying animal communication systems that use vocal signals, an investigator is unfortunately faced with the problem of not having access to knowledge of the units and code appropriate to the message level and therefore must focus his attention entirely on the physically realized sound segments of the signal and their possible correlation to some observable response by the organism. This preoccupation with the physical realization of the message creates the obvious problem, in many cases, of failing to recognize the functional role that a signal plays in the communication system of an organism. With these preliminary remarks completed we can now turn our attention to relevant issues in speech perception.

Human speech production imposes certain well-defined constraints on the resulting acoustic waveform. These constraints are derived from a consideration of the anatomy and physiology of the production mechanism and the associated resonant properties of the vocal tract (Stevens & House, 1955; Fant, 1960; Flanagan, 1972). Although a good deal of work still remains to be done in understanding speech perception, there are a number of empirical findings which define the basic issues in the field of speech perception and set it apart from other related areas such as auditory psychophysics and general auditory perception. The nature of these problems form the basis, in part, for suspecting that speech perception may require the use of specialized mechanisms for perceptual analysis.

The research in speech perception may be described in terms of two general lines of investigation: (1) studies aimed at establishing the acoustic cues to the perception of speech sound segments and, (2) studies aimed primarily at demonstrating the effects of manipulating syntactic and semantic variables on speech perception. This contribution will be concerned primarily with research that falls into the first category. Three sets of findings will be discussed: (1) early research on the acoustic cues for stop consonants, (2) experiments on the identification and discrimination of speech and speech-like sounds, and (3) research on developmental aspects of speech perception in young infants. The material to be reviewed in these sections is quite selective in order to highlight some of the major problems and theoretical issues that have been studied in speech perception over the last few years and to see how these have influenced the kinds of theoretical approaches that are currently prominent in the field.

The study of speech perception differs in several important ways from the study of auditory perception and psychophysics. First, the signals typically used to study the functioning of the auditory system are simple, discrete and usually well defined mathematically. Moreover, they typically vary along only a single dimension. In contrast, speech sounds involve complex spectral relations that vary as a function of time; changes that occur in a single parameter often affect the perception of other attributes of the stimulus. Secondly, most of the research in auditory psychophysics over the last two decades has been concerned with the discriminative capacities or resolving power of the transducer and the peripheral auditory mechanism. In the perception of speech, the relevant mechanisms are, for the most part, centrally located. Moreover, experiments in auditory psychophysics have commonly focused on experimental tasks involving discrimination rather than absolute identification. This is rarely the situation of listeners when they perceive and understand speech. In fact, the listener must almost always attempt to identify, on an absolute basis, a particular stretch of speech. As a consequence, it is generally believed that a good deal of what we have learned from traditional auditory psychophysics is unfortunately only marginally relevant to the study of the complex cognitive processes that are involved in speech perception.

In addition to differences in the signal, there are also marked differences in the way speech sounds are processed by listeners. For the most part, when people are presented with speech signals they respond to them as linguistic entities rather than simply as auditory events in the environment. Speech signals are categorized and labeled almost immediately with reference to the listener's linguistic background. Moreover, as we shall see, a listener's ability to discriminate certain speech sounds is often a function of the extent to which the particular acoustic distinction under study plays a functional role in the listener's linguistic system. That is, speech sounds are interpreted by listeners as biologically significant acoustic signals which have an important functional role in the communicative process.

Perceptual Constancy and Acoustic Cues for Stop Consonants

One of the most firmly established findings in the speech perception literature is that the acoustic correlates of a number of consonantal features are highly dependent on context. This is especially true for the stop consonants (b,d,g; p,t,k) which show the greatest amount of contextual variability. As a result of co-articulation effects, one sound segment often carried information about two or more successive phonemes in an utterance. And, conversely, a single phoneme often exerts an influence on several successive sound segments in the acoustic signal. Although the physical signal often varies continuously, listeners perceive speech as consisting of a sequence of discrete segments arrayed in time.

The earliest experimental studies of speech perception were aimed at uncovering the relation or mapping between attributes of the acoustic signal and the linguistic units derived from perceptual analysis. The outcome of these initial studies indicated that there were very few discrete isolatable and invariant sound segments in the physical signal that correspond uniquely to perceived phonemes. This lack of correspondence between attributes of the acoustic signal and units of linguistic analysis has been, and still currently is, one of the most important and controversial issues in speech perception. Because of the prominence of this single issue, it is appropriate to briefly review some of the early findings.

The initial work on the acoustic cues for phonemes involved two related procedures. First, spectrographic analyses were carried out on minimal pairs to identify the potentially important acoustic attributes that distinguished these utterances. Then synthesis experiments were conducted to verify the significance of these acoustic cues in perception. In the first study to use these combined methods, Liberman, Delattre and Cooper (1952) examined the relation between the frequency of a noise burst and the perception of the voiceless consonants p, t and k. An examination of sound spectrograms of real speech showed that the voiceless stops could potentially be distinguished by the frequency of a brief burst of noise, the acoustic counter-part of the articulatory explosion at the release of stop closure. Liberman et al., (1952) systematically varied the frequency of a synthetic noise burst before a number of different two formant vowels and observed its effect on perception. Their results showed that subjects' identification of a particular stop consonant varied not only according to the frequency of the noise burst but also in terms of the relation of the burst to the vowel with which it was paired.

The results of this initial study with synthetic speech were replicated with natural speech stimuli in a study by Schatz (1954). She found that release bursts spliced from /ki/, /ka/ and /ku/ were perceived as different stops depending upon the following vowel context. For example, the release burst excised from /ki/ is perceived as /t/ when spliced before /a/ but is perceived as /p/ before /u/. The findings from both of these experiments indicate that identification of the consonant does not depend exclusively on the absolute frequency of the burst but rather depends on the attributes of the burst in relation to the vowel which follows.

In another study, Cooper, Delattra, Liberman, Borst and Gerstman (1952) studied the role of formant transitions in the perception of stop consonants. Variations in the first formant transition were found to provide cues to voicing and manner whereas variations in the second formant transition were found to provide cues to place of articulation among the stops /b,d,g/ and /p,t,k/. Cooper et al., (1952) also studied a range of second formant transitions with the same vowels that were used in the previous burst experiment. Subjects were required to identify the stimuli at b, d or g. The results of this study indicated that while most subjects heard rising transitions as /b/ in almost all vowel contexts, perception of the falling transitions as either /d/ or /g/ varied as a function of the following vowel. Thus, as in the previous burst experiments, the perception of the formant transitions as a particular phoneme also depended on the following vowel.

Three important conclusions have been drawn from the results of these early perceptual experiments with synthetic speech. First, with regard to stop consonants, no invariant acoustic cues could be identified which corresponded uniquely to the same phoneme in all environments. Second, because burst and transition cues depend, to a large extent, on properties of the following vowel, the minimal acoustic unit seemed more likely to be about the size of a consonant-vowel syllable than an isolated phoneme. Indeed, as Cooper et al., (1952) remark, "one may not always be able to find the phoneme in the speech wave, because it may not exist there in free form." Finally, within the context of these experiments which have employed relatively simple synthetic stimuli, it was apparent that the acoustic information for a particular phoneme was encoded into the sound stream in a complex way; no one-to-one correspondence could be found between perceived phoneme and acoustic segment. These findings made it very unlikely at the time that the recognition of speech, particularly its phonological structure, could be carried out passively by simple template matching or filtering of the acoustic waveform. Indeed, these methods have yet to be successful as candidates for recognition schemes even some twenty-five years later. The earliest experiments on speech perception as well as those which followed made it very clear that the linguistic information in the message was restructured and encoded in the waveform in a complex and non-trivial way. Understanding the process by which the human listener carries out this task has been one of the fundamental problems in the field of speech perception. At the present time there is still no adequate account of the processes by which this is accomplished or the perceptual mechanisms involved.

Although these early findings failed to uncover the invariant acoustic attributes for phonemes (e.g., stop consonants) and emphasized a complex relation between acoustic attribute and phonetic unit, numerous investigators have continued, nevertheless, to search for a description of the acoustic signal which would reveal invariant attributes for phonemes (Fant, 1960, 1962, 1973; Stevens, 1967, 1973). The experimental literature on this topic during the 1950's, 1960's and even the 1970's is extensive and cannot be reviewed here (see Pisoni, 1977).

It is worth re-emphasizing here that the invariance problem in speech perception has, by no means, been resolved yet by anyone. The work reviewed in this section has been limited to only the stop consonants in initial position in stressed CV syllables. But some idea of the magnitude of the invariance problem in speech perception can be obtained by considering the contextual effects that appear for stop consonants in other phonetic environments such as medial and final position of syllables as well as consonant clusters. Moreover, the problem becomes enormous when we add to it the contextual variability found for other classes of speech sounds such as fricatives, liquids, nasals and vowels as well as the inherent variability associated with phonetic context, speaking rate and individual talker differences.

The Speech Mode and Categorical Perception of Speech Signals

The earliest perceptual experiments with speech stimuli showed that listeners respond to these sounds quite differently from other auditory signals. Liberman and his colleagues at Haskins Laboratories found that listeners perceived synthetic speech stimuli varying between /b/, /d/ and /g/ as members of distinct categories (Liberman, Harris, Hoffman & Griffith, 1957). When these same listeners were required to discriminate pairs of these sounds, they could discriminate stimuli drawn from different phonetic categories but could not discriminate stimuli drawn from the same phonetic category. The obtained discrimination functions showed marked discontinuities at places along the stimulus continuum that were correlated with changes in identification.

The ideal case of this form of perception, "categorical perception," is illustrated in Figure 1. In an experiment such as this, two or more phonetic segments are selected to represent end points and a continuum of synthetic stimuli is generated spanning this range. Subjects are required to carry out two tasks: identification and discrimination. In the identification task, stimuli are selected from the continuum and presented one-at-a-time in random order for labeling into categories defined by the experimenter. In the discrimination task, pairs of stimuli are selected from the continuum and presented to listeners for some discriminative response.

Insert Figure 1 about here

The basic finding of the categorical perception experiments is that listeners can discriminate between two speech sounds which have been identified as different phonemes much better than two stimuli which have been identified as the same phoneme even though the acoustic differences are comparable. At the time, the categorical perception results were considered by Liberman and others to be quite unusual when compared with the results typically obtained in most psychophysical experiments with non-speech stimuli. In general, stimuli that lie along a single continuum are perceived continuously resulting in discrimination functions that are monotonic with the physical scale. As is well known, there are capacity limitations on information transmission in terms of absolute identification (Miller, 1956). Listeners can discriminate many more acoustic stimuli than they can identify in absolute terms (Pollack, 1952, 1953). However, in the case of categorical perception the situation is quite different. The listener's differential discrimination appears to be no better than his absolute identification. In the extreme case of categorical perception, a listener's discrimination performance can be predicted from his identification function under the strong assumption that the listener can discriminate between two stimuli only to the extent that these stimuli are identified as different on an absolute basis (Liberman et al., 1957).

These initial findings with stop consonants led to a similar experiment with synthetic vowels which varied in acoustically equal steps through the range /I/, /ɛ/ and /æ/. Fry, Abramson, Eimas and Liberman (1962) reported that these stimuli were perceived continuously, much like non-speech stimuli. That is, the discrimination functions did not yield discontinuities along the stimulus continuum which were related to changes in identification but were relatively flat across the whole continuum. Moreover, it was observed that vowels were, in general, more discriminable than stop consonants indicating that listeners could perceive many more intra-phonemic differences.

The differences in perception between stop consonants and steady-state vowels have been assumed to reflect two basic modes of perception, a categorical mode and a continuous mode. Categorical perception reflects a mode of perception in which each acoustic pattern is always and only, perceived as a token of a particular phonetic type (Studdert-Kennedy, 1974).

Listeners can discriminate between two different acoustic patterns if the stimuli have been categorized into different phonetic categories, but they cannot discriminate two different acoustic patterns that have been categorized into the same phonetic category. Information about the acoustic properties of these stimuli appears to be unavailable for purely auditory judgments as a consequence of phonetic classification. What remains available to the decision process is a more abstract and permanent code based on the listener's interpretation of the stimulus event (see Pisoni, 1971; Pisoni & Tash, 1974).

Although the stimulus generalization reflected in categorical perception might seem at first glance to be a more primitive form of stimulus control, it may provide, on the other hand, a more efficient mode of response for absolute and rapid decisions concerning the presence or absence of particular attributes such as those required in the processing of connected speech. Indeed, the great interest expressed in categorical perception of speech presumably derives from the assumption that listeners do indeed make categorical decisions when listening to continuous speech.

Continuous perception, on the other hand, may be thought of as reflecting an auditory mode of perception where discrimination is independent of category assignment. Although listeners can assign acoustically different stimuli to the same category, they may still discriminate between tokens selected from the same category. Thus, an auditory, non-phonetic basis for discrimination is available to the listener.

For a number of years the categorical perception results were assumed to be a unique aspect of speech perception and primarily a consequence of phonetic or linguistic categorization. Indeed, the differences in perception between consonants and vowels lead Liberman (1970) to argue strongly for a specialized mode of perception, a "speech mode," to characterize the way these acoustic stimuli are perceived by humans. Other findings have suggested that a specialized perceptual mechanism--a "special speech decoder" may exist for processing speech sounds (Studdert-Kennedy & Shankweiler, 1970).

The interpretation that categorical perception reflects a specialized mode of perception unique to speech has come under strong criticism from a number of directions in the last few years. Several investigators have argued that the differences in perception between consonants and vowels reflect differences in the psychophysical properties of the acoustic cues which distinguish these two classes of speech sounds (Lane, 1965; Pisoni, 1971; Studdert-Kennedy, 1974). For the stop consonants there is a relatively complex relation between phoneme and its representation as sound; the essential acoustic cues are contained in the rapidly changing spectrum at onset (i.e., release burst and formant transitions) which is weak, relatively brief in duration (30-50 msec) and transient in nature. On the other hand, the cues to the vowels, at least the ones used in these experiments, involve changes in the steady-state frequencies of the first three formants which have a relatively long duration and more uniform spectral properties as well as greater intensity. In support for this, Fujisaki and Kawashima (1969, 1970) and Pisoni (1971, 1975) have shown that the differences in perception between consonants and vowels are due, in part, to the duration of their respective acoustic cues. Vowels of very short duration (i.e., 40-50 msec) are perceived more categorically than identical stimuli having longer durations.

Other findings have shown that categorical perception is also due, in part, to encoding processes in short-term memory that are a consequence of the particular type of discrimination task used in these experiments (Pisoni, 1971, 1973, 1975). The ABX procedure has been used in almost all of the speech perception experiments demonstrating categorical perception. In this task the subject is presented with three sounds successively, ABA or ABB. A and B are always acoustically different and the subject has to indicate whether the third sound is identical to the first or second sound. This is basically a recognition memory paradigm. In order to solve the discrimination task, the subject is forced to encode the individual stimuli in temporal succession and then base his decision on the encoded representations that have been maintained in short-term memory rather than to respond to the magnitudes of difference between stimuli within an ABX triad. In a number of experiments Pisoni (1971, 1973) has shown that differences between categorical and continuous modes of perception are crucially dependent on the memory requirements of the particular discrimination procedure and the level of encoding required to solve the task (Pisoni, 1971, 1973).

Several recent experiments employing non-speech stimuli have also suggested that categorical perception may not be peculiar to speech sounds or a specialized "speech" mode of perception as once supposed, but may be a more general property relevant to processing complex signals that involves categorization and coding of the stimulus input and later storage in memory. For example, Cutting and Rosner (1974) recently demonstrated categorical perception for non-speech musical sounds varying in rise-time that could be labeled easily by listeners as a "pluck" or a "bow." Miller, Wier, Pastore, Kelly and Dooling (1976) have shown comparable categorical perception effects for non-speech stimuli varying in the onset of a noise preceding a buzz. And in another study Pisoni (1976) has reported similar results for stimuli differing in the relative onset-time of two component tones. In each case, these non-speech experiments showed that discrimination was better for pairs of stimuli selected from different perceptual categories than pairs of stimuli selected from the same category. Moreover, discrimination of stimuli selected from within a category was very nearly close to chance performance as predicted by the categorical perception assumption.

The results obtained with non-speech stimuli have provided some important insights into the underlying basis of categorical perception for speech stimuli. These non-speech experiments have succeeded in demonstrating categorical perception when previous attempts have failed primarily for three reasons. First, the investigators employed relatively complex acoustic stimuli in which only a single component was varied relative to the remainder of the stimulus complex. In most of the previous non-speech experiments only simple stimuli were used. Second, while these complex stimuli may be characterized as varying in linear steps along some nominally physical continuum, on both psychophysical and perceptual grounds, the stimulus continuum that is generated results in several distinctive perceptual attributes or qualities that are present for some stimuli but not others. These perceptual attributes, in turn, define quantal regions along the stimulus continuum that are separated by natural psychophysical boundaries: within these regions sensitivity is low whereas between these regions it is high. Finally, because the stimulus continuum can be partitioned into several perceptually distinctive classes, subjects can easily employ a set of labels or codes to represent these attributes in short-term memory. These codes can then be assigned to stimuli presented in the subsequent ABX discrimination task.

Categorical perception of both speech and complex non-speech signals, therefore, can be explained, in part, by the presence of well-defined psychophysical boundaries which separate stimuli into distinctive perceptual categories and by the use of verbal labels which can be used to encode these attributes in short-term memory. Thus, this account of categorical perception involves two distinct components, a sensory component and a memory or labeling component involving some interpretative process.

But what is the basis for the distinctive perceptual attributes of speech sounds and what biological significance would they have? One approach to this problem can be found in the recent work of Stevens (1972) on the Quantal Theory of Speech. According to Stevens, phonetic features are grounded in a close "match" between articulatory and auditory capacities of human speech perception and production. The acoustic attributes common to a phonetic category are determined, in part, by articulatory constraints on the speech production mechanism and, in part, by the distinctiveness of the resulting acoustic signals in perception. One piece of evidence for a perceptual match between speech perception and production according to Stevens, is the categorical perception results. The acoustic correlates of certain phonetic features that show quantal properties in production are also precisely those features that show categorical-like discrimination in perceptual experiments. Thus, there is some very close relation between attributes of production and distinctions made in perception.

It should be pointed out, however, that the category boundaries separating phonetic features cannot be inherently fixed since their precise location and alignment on a stimulus continuum shifts as a function of linguistic experience. Indeed, as Popper (1972) has suggested, "people who speak different languages may tune their auditory systems differently (p. 218)." Cross-language research has shown, in fact, that the categorizations imposed on synthetic speech stimuli are based on both the particular acoustic attributes of the stimuli and the linguistic experience of the listener. To take one example, Abramson and Lisker (1965) generated a continuum of synthetic stimuli varying in voice onset time between /da/ and /ta/ and presented them for labeling to listeners of three different language backgrounds. The labeling functions for English, Spanish and Thai subjects are shown in Figure 2. The results indicate that these listeners categorized the same stimuli in quite different ways depending on the phonological structure of their language. The phoneme boundaries are not only placed somewhat differently along the continuum in each case, but the Thai subjects show an additional category in their labeling behavior. This result was expected of course since in Thai a phonological distinction exists between the voiceless aspirated stop [t^h] and the

----

Insert Figure 2 about here

----

voiceless unaspirated stop [t].[2] This phonetic difference is not realized in either English or Spanish and consequently fails to play a role in the listener's identification and discrimination. The importance of these findings is that the phonological systems of different languages make use of the acoustic and phonetic distinctions that exist between different speech sounds in different ways and these distinctions must be learned from the local environment.

In summary, several implications can be drawn from the categorical perception research reviewed in this section. First, the perception of speech sounds appears to have certain quantal properties because listeners treat acoustically different sounds as functionally the same. Second, the categorical perception results can be thought of as representing a phonetic mode of perception in which the listener responds to speech signals in terms of the auditory features deployed in his own linguistic system. Thus, categorical perception of speech by humans may represent a species-specific form of coding acoustic signals for subsequent linguistic processes. The extent to which the mechanisms underlying these perceptual processes are innately determined or modified by the environment has been a topic of recent interest in speech perception as shown below.

Speech Perception in Infants

Much of what we currently know about the development of language and the acquisition of phonology is based on data obtained in studies of speech production (see McNeill, 1970; Menyuk, 1971; Jakobson, 1968). One conclusion that has been drawn by a number of investigators is that the developmental process proceeds from the general to the specific and gradually involves a greater and greater differentiation of language skills. Within the last five years a number of pioneering studies by Eimas and others have demonstrated that infants as young as one month of age are capable of making fine discriminations among a number of the distinctive attributes of speech sounds (see Eimas, Siqueland, Jusczuk, & Vigorito, 1971; Eimas, 1974, 1975). These results obviously call into question the validity of the differentiation assumption. While increasing differentiation may very well be true of the development of productive language skills, it may not be true of the perception of speech. The recent evidence from studies of infant speech perception points to a loss of discriminative abilities over time for certain speech sounds if specific experience with these distinctions fails to take place in the local environment.

The procedure used in these speech perception experiments involved a discrimination paradigm in which the infant was first familiarized with a particular stimulus and then shifted to another stimulus. If the infant showed an increased response rate after the shift, it was assumed that the infant could discriminate the difference between the two stimuli (see Morse, this volume). In the first experiment using this procedure, Eimas et al. (1971) studied the voicing feature which distinguishes /b/ from /p/. The stimuli were synthetically produced CV syllables that varied in acoustically equal steps of voice-onset time (VOT) between [ba] and [p^ha]. The results revealed two important findings. First, infants could discriminate between two speech sounds selected from different phonetic categories. Second, infants could <u>not</u> reliably discriminate between two acoustically different stimuli selected from within the same phonetic category. The latter finding is particularly important since it permitted Eimas and his collaborators to argue that their infants perceived the voicing distinction in a more nearly categorical manner and, therefore, in a linguistic mode comparable to that found with adults.

Two cross-language studies of infant speech perception have also been carried out recently on the voicing distinction using similar synthetic stimuli and methodology. In one study, Lasky, Syrdal-Lasky and Klein (1975) found that infants from Spanish-speaking environments could discriminate three categories along the voicing continuum. One boundary occurred between +20 and +60 msec and another between -20 and -60 msec. The first boundary is consistent with the discrimination findings of Lisker and Abramson (1970) for English-speaking adults and the previous results of Eimas et al. with infants and suggests a possible innate or sensory basis to the voicing distinction. However, the presence of a second boundary in the discrimination data was of particular interest. Spanish-speaking adults distinguish between only two categories of voicing and, based on the adult discrimination data of Lisker and Abramson (1970), their phoneme boundary does not correspond to either of the two VOT boundaries found with these infants. The implication of these findings is that infants are capable of perceiving three major voicing distinctions in the absence of any specific experience with these particular voicing contrasts in the environment.

In another related study, Streeter (1976) found that Kikuyu infants are capable of discriminating voicing differences between labial stops which are not used phonologically by the adults in their language environment. In Kikuyu, a Bantu language spoken in Kenya, there is only one labial stop with a VOT value in the range of -60 msec (i.e., a pre-voiced stop). However, the Kikuyu infants could discriminate differences between three voicing categories corresponding roughly to the same ones found in the Lasky et al. study. Thus, the discrimination of these voicing contrasts can also be made in the absence of relevant linguistic experience with the specific sound contrasts. The results of the cross-language experiments suggest that infants may be predisposed, in some sense, to deal with these acoustic attributes with only a very limited exposure to the specific sounds and obviously well before any experience in producing these distinctions.

The developmental course of speech perception seems to be somewhat different from other forms of perceptual development in which it is assumed that environmental experience serves primarily to sharpen the discriminative capacities of an organism (Gibson, 1969). Since the child is capable of making relevant discriminations between the important distinctive acoustic attributes of speech sounds at a very early age, the effects of linguistic experience may be restricted primarily to learning that particular distinctions are <u>not</u> functional within a child's language environment. Thus, the course of development may not involve learning to make finer and finer discriminations among stimulus attributes but may be more analogous to the effects of acquired similarity or equivalence where there is a loss of discriminative abilities (Gibson & Gibson, 1955; Liberman, Harris, Kinney & Lane, 1961). As Eimas (1976) has recently suggested "the course of development of phonetic competence is one characterized by a loss of abilities over time if specific experience is not forthcoming." Like the adult, if the phonetic distinctions are not used functionally in the language, sensitivity to the relevant acoustic attributes is lowered or even lost and the child will fail to process or interpret them. One proposal to account for the infant results is considered below in terms of a detector system with specialized feature detectors each sensitive to a restricted range of acoustic information in the speech signal. These detectors are assumed to be available innately to the infant for processing the relevant acoustic attributes of speech although they can be modified and tuned substantially by specific linguistic experience in the environment.

To explain the results of the categorical-like discrimination found in infants, Eimas (1974) proposed an approach to speech perception based on the idea of specialized feature detectors which are finely tuned to restricted ranges of acoustic information in the speech signal. This particular idea did not originate with Eimas as a number of other investigators have remarked on the possibility of some sort of feature detecting mechanism in speech perception (see for example, Whitfield, 1965; Liberman et al., 1967; Abbs and Sussman, 1971; Lieberman, 1970; Stevens, 1972). However, it was left to Eimas and Corbit (1973) to introduce an experimental paradigm--selective adaptation, to speech perception that could reveal the workings of these hypothesized detectors in some detail (see Cooper, 1975 for an extensive review). In selective adaptation, repetitive presentation of a stimulus alters the perception of a set of test stimuli. For example, in the initial study, Eimas and Corbit (1973) investigated the voicing feature and showed that adaptation with the syllable [ba] caused the locus of the phonetic category boundary between [ba] and [p^{h}a] to shift towards the [ba] end of the continuum. Stimuli near the boundary which were identified as [ba] when the listener was in an unadapted state were subsequently labeled as [p^{h}a] after adaptation with [ba]. Similar findings were obtained when [p^{h}a] was used as an adaptor; the locus of the phonetic boundary shifted toward the [p^{h}a] end of the stimulus continuum.

Eimas and Corbit (1973) interpreted the selective adaptation findings as support for the hypothesis that the perception of voicing in stop consonants involves two distinct types of feature detectors organized as opponent pairs, a voiced detector (+V) and a voiceless detector (-V). Each detector is assumed to be selectively tuned to a range of partially overlapping voice onset-time values. When a stimulus containing a particular VOT value is presented repetitively, it fatigues the detector most sensitive to that range of the feature and, accordingly, its sensitivity is reduced. After adaptation, the opponent or unadapted detector provides a greater output to the decision process in identification than the adapted detector and, accordingly, produces a shift in the locus of the phonetic category boundary.

One of the intriguing questions that the infant speech perception work has raised is the extent to which environmental input determines the development and sensitivity of these hypothesized feature detectors. There is now an extensive literature on the role of early experience in the development of the visual system which indicates that early environmental experience can modify the selectivity of cortical cells in kittens (e.g., Hirsch & Spinelli, 1970; Blakemore & Cooper, 1970). The analogy to this developmental work has already been drawn by Eimas (1976) who argues that the lack of experience with specific phonetic distinctions in the local environment during language acquisition has the effect of modifying the appropriate detectors by reducing their sensitivity. Some detectors originally designed to process certain phonetic features may be captured or subsumed by other detectors after exposure to specific acoustic stimuli from the linguistic environment. The nonspecific detectors might, therefore, assume the specificity for only those attributes present in the stimuli to which they are exposed. The poor within category discrimination of speech sounds found with adults and infants in the categorical perception experiments discussed earlier may not only be due to phonetic coding of these signals but may also be a consequence of the modification of the needed discrimination mechanism. In considering the recent infant findings it may well be the case that the general program of development of speech perception is genetically determined with experience in the environment playing only a role in the tuning and alignment of the system. Moreover, the presence of linguistic universals, particularly in terms of the relatively small number of phonetic features found across many different languages, lends some support to this contention and implies that at least some of the structural mechanisms underlying speech perception are part of the biological endowment of the organism. This, of course, would not be a surprising conclusion for a biologist or ethologist but it is one that psychologists are only recently coming to appreciate and take seriously.

References

Abbs, J. H. & Sussman, H. M. Neurophysiological feature detectors and speech perception: A discussion of theoretical implications. Journal of Speech and Hearing Research, 1971, 14, 23-36.

Abramson, A. S. & Lisker, L. Voice onset time in stop consonants: Acoustic analysis and synthesis. Proceedings of the 5th International Congress of Acoustics, Liege, 1965.

Blakemore, C. & Cooper, G. F. Development of the brain depends on the visual environment. Nature, 1970, 228, 477-478.

Cooper, F. S., Delattre, P. C. Liberman, A. M., Borst, J. M. & Gerstman, L. J. Some experiments on the perception of synthetic speech sounds. Journal of the Acoustical Society of America, 1952, 24, 597-606.

Cooper, W. E. Selective adaptation to speech. In F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindman & D. B. Pisoni (Eds.) Cognitive theory: Vol. 1. Potomac, Maryland: Erlbaum Assoc., 1975.

Cutting, J. E. & Rosner, B. S. Categories and boundaries in speech and music. Perception & Psychophysics, 1974, 16, 564-570.

Eimas, P. D. Auditory and linguistic processing of cues for place of articulation by infants. Perception & Psychophysics, 1974, 16, 513-521.

Eimas, P. D. Speech perception in early infancy. In L. B. Cohen and P. Salapatek (Eds.), Infant perception. New York: Academic Press, 1975.

Eimas, P. D. Developmental aspects of speech perception. In R. Held, H. Leibowitz & H. L. Teuber (Eds.) *Handbook of sensory physiology: Perception*. New York: Springer-Verlag, 1976.

Eimas, P. D. & Corbit, J. D. Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, 1973, *4*, 99-109.

Eimas, P. D., Siqueland, E. R., Jusczyk, P. & Vigorito, J. Speech perception in infants. *Science*, 1971, *171*, 303-306.

Fant, G. *Acoustic theory of speech production*. The Hague: Mouton, 1960.

Fant, C. G. M. Descriptive analysis of the acoustic aspects of speech. *Logos*, 1962, *5*, 3-17.

Fant, C. G. M. *Speech sounds and features*. Cambridge: M.I.T. Press, 1973.

Flanagan, J. L. *Speech analysis, synthesis & perception*. (Second edition) New York: Academic Press, 1972.

Fry, D. B., Abramson, A. S., Eimas, P. D. & Liberman, A. M. The identification and discrimination of synthetic vowels. *Language and Speech*, 1962, *5*, 4, 171-189.

Fujisaki, H. & Kawashima, T. On the modes and mechanisms of speech perception. *Annual Report of the Engineering Research Institute*, *Vol. 28*, Faculty of Engineering, University of Tokyo, Tokyo, 1969, 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*, *Vol*. *29*, Faculty of Engineering, University of Tokyo, Tokyo, 1970, 207-214.

Gibson, E. J. *Principles of perceptual learning and development*. New York: Appleton Century Crofts, 1969.

Gibson, J. J. & Gibson, E. J. Perceptual learning: Differentiation or enrichment? *Psychological Review*, 1955, *62*, 32-41.

Hirsch, H. V. B. & Spinelli, D. N. Visual experience modifies distribution of horizontally and vertically oriented receptive fields in cats. *Science*, 1970, *168*, 869-871.

Hockett, C. F. *A course in modern linguistics*. New York: The MacMillan Company, 1958.

Jakobson, R. *Child language, aphasia and phonological universals*. The Hague: Mouton, 1968.

Lane, H. L. The motor theory of speech perception: A critical review. *Psychological Review*, 1965, *72*, 275-309.

Lasky, R. E., Syrdal-Lasky, A. & Klein, R. E. VOT discrimination by four to six and a half month old infants from Spanish environments. *Journal of Experimental Child Psychology*, 1975, *20*, 215-225.

Lisker, L. & Abramson, A. S. The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the Sixth International Congress of Phonetic Sciences*, Prague, 1967. Prague: Academia, 1970, Pp. 563-567.

Liberman, A. M. Some characteristics of perception in the speech mode. In D. A. Hamburg (Ed.), *Perception and Its Disorders*, *Proceedings of A. R. N. M. D.* Baltimore: Williams and Wilkins Co., 1970. Pp. 238-254.

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. The role of selected stimulus variables in the perception of the unvoiced stop consonants. American Journal of Psychology, 1952, 65, 497-516.

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, 1956, 54, 358-368.

Lieberman, P. Towards a unified phonetic theory. Linguistic Inquiry, 1970, 1, 3, 307-322.

McNeill, D. The acquisition of language. New York: Harper & Row, 1970.

Menyuk, P. The acquisition and development of language. Englewood Cliffs: Prentice-Hall, 1971.

Miller, J. D., Wier, C. C., Pastore, R., Kelly, W. J. & Dooling, R. J. Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. Journal of the Acoustical Society of America, 1976, 60, 2, 410-417.

Pisoni, D. B. On the nature of categorical perception of speech sounds. Supplement to Status Report on Speech Research. Haskins Laboratories, New Haven, SR-27, November, 1971.

Pisoni, D. B. Auditory and phonetic memory codes in the discrimination of consonants and vowels. Perception and Psychophysics, 1973, 13, 2, 253-260.

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

Pisoni, D. B. Identification and discrimination of the relative onset of two component tones: Implications for the perception of voicing in stops. Progress Report No. 118 Research Laboratory of Electronics, M.I.T., June, 1976, Pp. 212-230. Also accepted for publication in *The Journal of the Acoustical Society of America*, 1977, *00*, 000-000.

Pisoni, D. B. Speech Perception. In W. K. Estes (Ed.) *Handbook of learning and cognitive processes: Volume 6*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977. (In press)

Pisoni, D. B. & Tash, J. B. Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, 1974, *15*, 285-290.

Pollack, I. The information in elementary auditory displays. *Journal of the Acoustical Society of America*, 1952, *24*, 745-749.

Pollack, I. The information in elementary auditory displays II. *Journal of the Acoustical Society of America*, 1953, *25*, 765-769.

Popper, R. D. Pair discrimination for a continuum of synthetic voiced stops with and without first and third formants. *Journal of Psycholinguistic Research*, 1972, *1*, 205-219.

Schatz, C. The role of context in the perception of stops. *Language*, 1954, *30*, 47-56.

Stevens, K. N. Acoustic correlates of certain consonantal features. Paper presented at Conference on Speech Communication and Processing, M.I.T. Cambridge, Massachusetts, November 6-8, 1967.

Stevens, K. N. The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David, Jr. and P. B. Denes (Eds.), *Human communication: A unified view*. New York: McGraw-Hill, 1972.

Stevens, K. N. Further theoretical and experimental bases for quantal places of articulation for consonants. _Quarterly Progress Report No. 108_, Research Laboratory of Electronics, M.I.T., 1973, 247-252.

Stevens, K. N. & House, A. S. Development of a quantitative description of vowel articulation. _Journal of the Acoustical Society of America_, 1955, _27_, 484-493.

Streeter, L. A. Language perception of 2-month-old infants shows effects of both innate mechanisms and experience. _Nature_, 1976, _259_, 39-41.

Studdert-Kennedy, M. The perception of speech. In T. A. Sebeok (Ed.), _Current trends in linguistics_, Vol. XII, The Hague: Mouton, 1974.

Studdert-Kennedy, M., Liberman, A. M., Harris, K. S. & Cooper, F. S. Motor theory of speech perception: A reply to Lane's critical review. _Psychological Review_, 1970, _77_, 234-249.

Studdert-Kennedy, M. & Shankweiler, D. Hemispheric specialization for speech perception. _Journal of the Acoustical Society of America_, 1970, _48_, 2, 579-594.

Whitfield, I. C. "Edges" in auditory information processing. In _Proceedings of the XXIII International Congress of Physiological Sciences_. Tokyo, September, 1965, 245-247.

Footnotes

[1] Preparation of this paper was supported in part by NIMH Grant MH-24027-03 and NIH Grant NS-12179-02.

[2] Phonetic differences are represented by brackets whereas phonological or phonemic differences are represented by slashes. This notation is used throughout the remainder of the paper. Differences in the notation are used to call attention to the differences between phonetic and phonological levels of linguistic structure. The phonetic level corresponds to a level at which the articulatory gestures and their acoustic attributes can be represented as static or discrete states each corresponding to a set of distinctive features. In contrast, the phonological level corresponds to a somewhat more abstract level of perceptual analysis where only linguistically significant information would be represented. The phonetic level is assumed to represent universal properties of the vocal tract common to all speakers whereas the phonological level is necessarily by definition confined to language specific distinctions.

Figure Captions

Figure 1. Idealized form of categorical perception showing the identification function (left ordinate) and the discrimination function (right ordinate). (From Studdert-Kennedy, Liberman, Harris & Cooper, 1970, with permission).

Figure 2. Labeling functions for a set of synthetic speech stimuli varying in voice onset time (VOT) that were presented to speakers of English, Spanish and Thai. (From Abramson & Lisker, 1965, with permission).

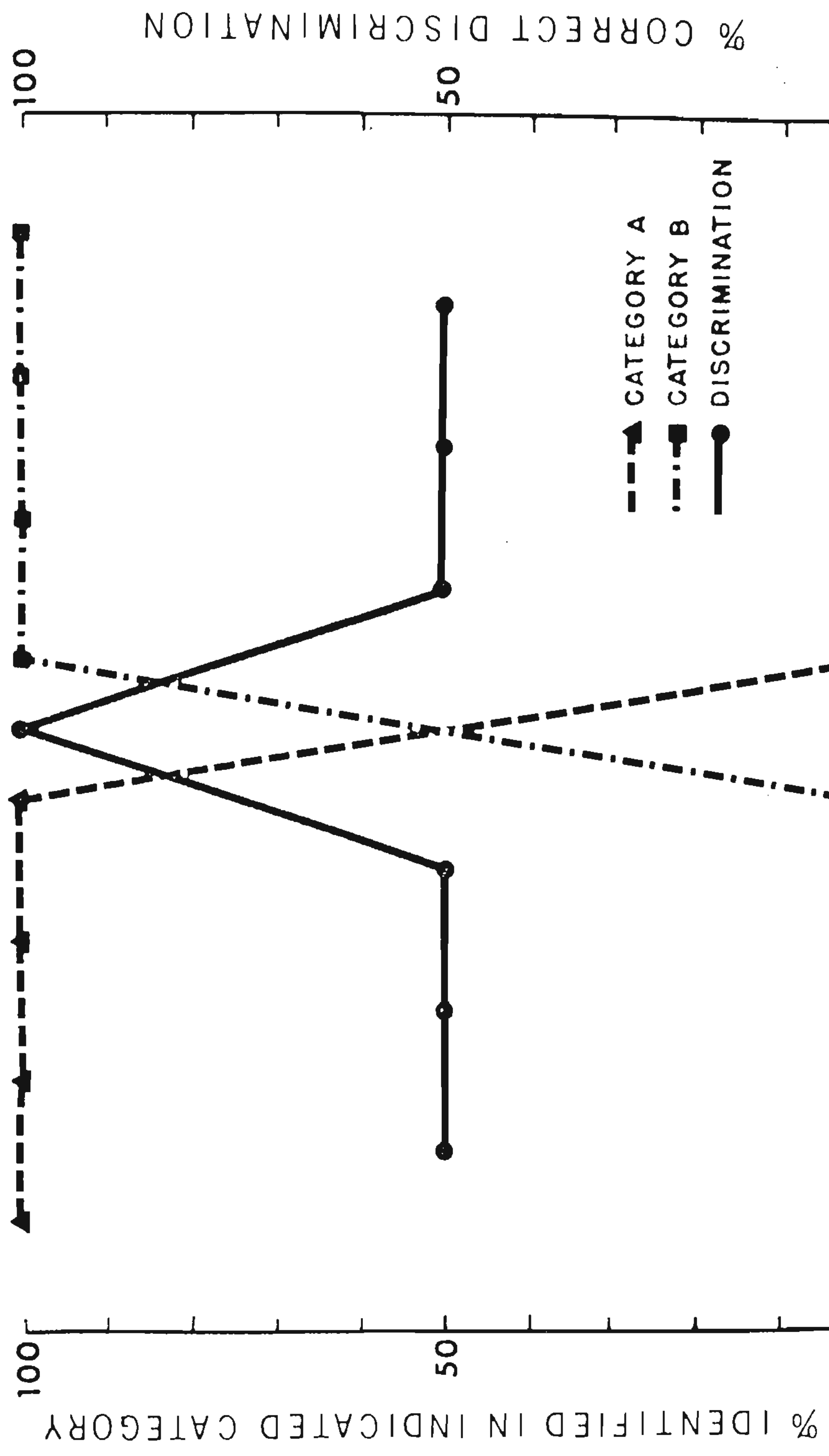

% IDENTIFIED IN INDICATED CATEGORY
100
50
0
% CORRECT DISCRIMINATION
100
50
0
STIMULUS NUMBER ALONG CONTINUUM
1
2
3
4
5
6
7
8
CATEGORY A
CATEGORY B
DISCRIMINATION

Figure 1.

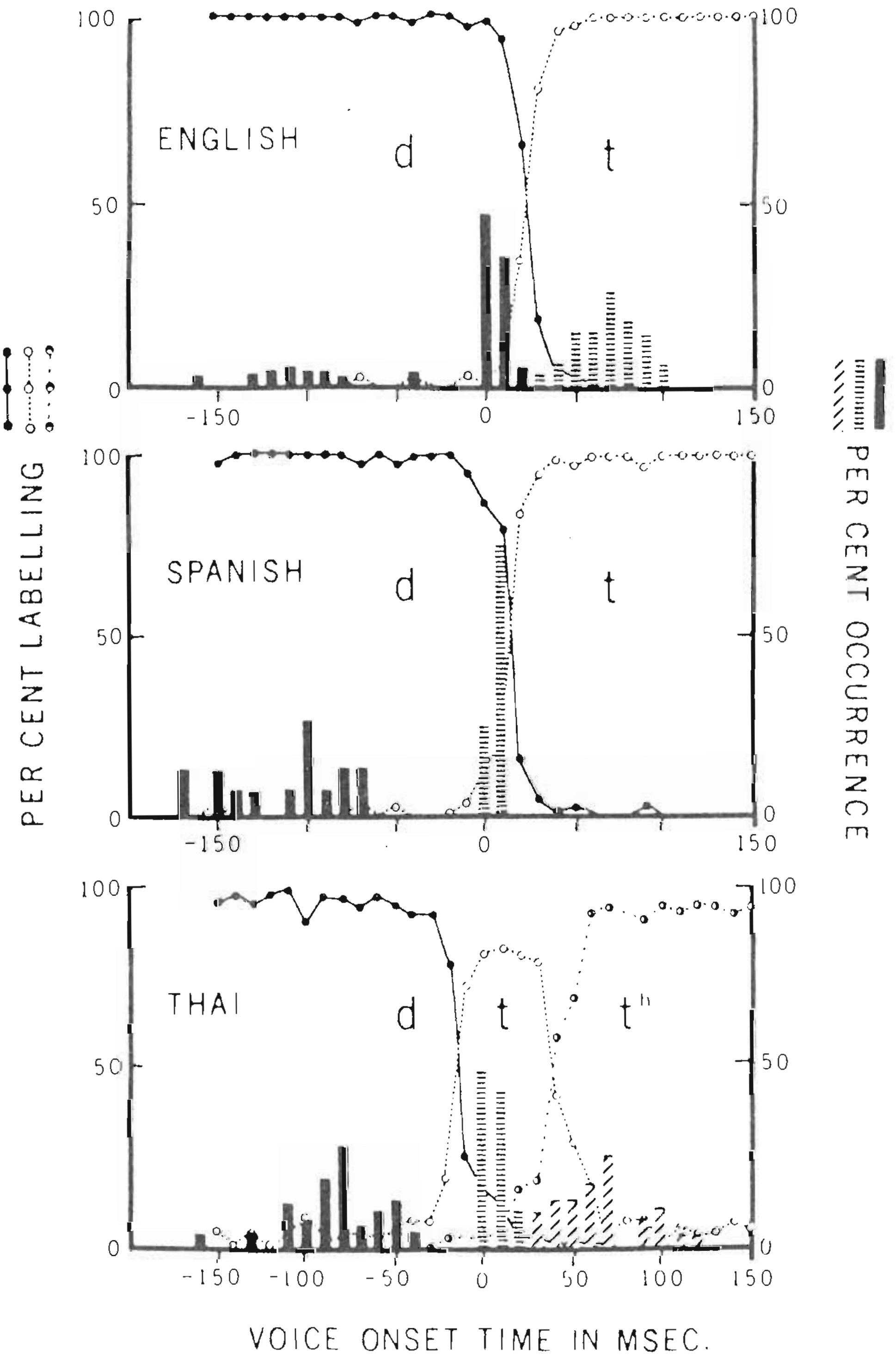

ENGLISH
d
t
SPANISH
d
t
THAI
d
t
tʰ
100
50
0
-150
0
150
-100
-50
50
100
PER CENT LABELLING
PER CENT OCCURRENCE
VOICE ONSET TIME IN MSEC.

Figure 2.

Vowel Mimicry and the Quantal Theory of Speech: A First Report*

David B. Pisoni

Indiana University

Bloomington, Indiana 47401

* This paper is a report of research carried out while the author was on leave at the Research Laboratory of Electronics, Massachusetts Institute of Technology during the academic year 1975-76. I am grateful to Dr. Victor Zue for his kind help and cooperation with regard to the spectral analysis of the data and Professor Kenneth N. Stevens for his interest in this work.

Abstract

This paper reports the results of a study in which variability of formant frequencies for different vowels was examined with special regard to several predictions derived from the Quantal Theory of Speech. Two subjects were required to mimic eight different steady-state synthetic vowels which were presented repeatedly in a randomized order. Spectral analysis was carried out on the vowel responses in order to obtain means and standard deviations of the formant frequencies. In the spirit of the Quantal Theory, it was predicted that the point vowels, /i/, /a/ and /u/ would show lower standard deviations than the non-point vowels because these vowels are assumed to be produced at places in the vocal tract where small perturbations in articulation produce only minimal changes in the resulting formant frequencies. That is, these vowels are assumed to be quantal vowels. The results of this study provided no support for the hypothesis under consideration. A discussion of the outcome of the results as well as some speculation as to its failure to find support for the Quantal approach is provided in the report. Some comments are also provided about computer simulation studies of speech production and the need for additional empirical studies on the problem of vowel production.

Vowel Mimicry and the Quantal Theory of Speech: A First Report

David B. Pisoni

Indiana University

Within the last few years considerable effort has been devoted to establishing well-defined anatomical, physiological and perceptual correlates for distinctive phonetic features as well as attempting to rationalize various phonological processes that occur with some regularity in natural languages. Particular emphasis has been placed on the articulatory space used to form phonetic segments and features and the resulting acoustic attributes of speech sounds generated by these configurations (see Lindblom, 1972).

Although the vocal apparatus can theoretically assume a relatively large number of articulatory positions, only a small number of preferred or "natural regions" of the articulatory space are actually used in phonological systems of languages. Stevens (1972) has recently proposed that speech sounds produced in these so-called natural regions have certain "quantal" properties or attributes that seem to be good candidates for the inventory of phonetic features used in language. According to Stevens, all phonetic features that occur in languages probably have their basis in acoustic attributes that have such quantal properties: perturbations in articulation at one of these natural regions produce only small changes in the acoustic output. Stevens argues that these acoustic attributes are well-matched to the properties of the auditory system and therefore have some special status perceptually.

The basic line of reasoning behind Stevens' Quantal Theory of Speech Production can be illustrated by reference to Figure 1. Assume that one could manipulate some articulatory parameter continuously along a particular dimension. And further assume that one could obtain a measure of some specific acoustic parameter of the speech signal that would be controlled by changes in this articulatory parameter. As the articulatory parameter

----

Insert Figure 1 about here

----

is varied continuously one might expect to find continuous changes in the output of this acoustic parameter. However, what seems to happen is shown schematically in Figure 1. There are places in this space where very small changes in the articulatory parameter, as in Region II, produce large variations in the acoustic parameter whereas in other places, as Region III, large changes in the articulatory parameter produce only small changes in the acoustic output. Relations such as these between vocal-tract shape and sound output have been studied in some detail by Stevens (1972, 1973) and others with the use of computer simulation models of the vocal tract (see also Lindblom and Sundberg, 1969; Liljencrants and Lindblom, 1972). Small changes have been made systematically in one or more articulatory parameters and the resultant effects on the properties of sound output calculated quite precisely.

To see how Stevens carried out the simulation and arrived at these views it will be helpful to briefly examine the resonance characteristics of the vocal tract. Figure 2 shows an approximation of the vocal tract shape for the vowel [a] in terms of a two tube resonator (Fant, 1960).

Insert Figure 2 about here

The spectrum envelope produced by this configuration is shown below in the lower panel. The size of the pharyngeal cavity is constricted when compared with the size of the oral cavity. Using a vocal tract simulation which calculates formant frequencies from area functions, Stevens found that over a wide range, variations in the lengths of the cavities, (i.e., d_1 and d_2) and the cross-sectional areas (i.e., A_1 and A_2) did not affect the formant frequencies of the vowel to any large degree. The results of this simulation are shown in Figure 3 where the calculated

----

Insert Figure 3 about here

----

formant frequencies are plotted as a function of the length of d_1. Notice that there is an area in the middle of this figure where variations in d_1 produce only small changes in the formant frequencies of F1 and F2. On the other hand, there are regions at the extremes where a small variation in d_1 produces a much larger shift in the formant frequencies, especially the values of the second formant.

Stevens (1972, 1973) has employed these simulation techniques to study other vowels as well as several types of consonants and concludes that there are certain places in the vocal tract which show quantal relations between articulation and the attributes of the sound output.

Perturbations in articulation at these places apparently have only small effects on the acoustic signal and therefore, according to Stevens, would presumably affect perception only minimally.

One of the proposals that Stevens has made about these findings is that the sound segments used in languages are selected from a relatively small range of features or distinctive attributes. A careful examination of these acoustic attributes has revealed that they have a natural basis in terms of certain vocal tract configurations. These configurations occur in precisely those regions where variability in articulation can be tolerated without the sound output being appreciably affected.

Using the same types of simulation procedures, Liljencrants and Lindblom (1972) have studied the vowel systems of a number of languages in order to predict their phonetic structure. Since human vowels represent only a small subset of the possible combinations of formant frequencies, it was of interest to see how the vowels of different languages were distributed in acoustic space. Liljencrants and Lindblom found that the vowel systems of quite diverse languages can be described in terms of a simple principle of "maximal perceptual contrast" as defined by the linear distance in mel units between the points representative of two vowels. The whole vowel space for a particular language seems to be organized in terms of sound contrasts that are highly discriminable. This is true of languages having as few as three vowels to as many as twelve. These two sets of findings are important because they suggest that the sound systems of natural languages may be matched, in some sense, to attributes of both perception and production.

It should be noted here that not all of the potential distinctive features that exist at the phonetic level are used distinctively in the phonological system of a particular language. Similarly, no natural language has as many phonemes in its phonological system as there are logically possible combinations of utilized distinctive features. The sound systems of languages may have evolved the way they are for the following two reasons. First, the distinctions used in language are easily pronounceable for talkers thus permitting some variability in production. Secondly, these same articulatory distinctions are used to generate acoustic attributes that have highly distinctive properties resulting in signals that can be identified and discriminated under very poor listening conditions.

The basic assumption of the quantal theory of speech is that there are nonlinear relationships between articulation and acoustic output. According to Stevens, as a result of these nonlinearities a speaker "manipulates his speech generating mechanism to select sounds with well-defined acoustic attributes that are relatively insensitive to small perturbations in articulation" (Stevens, 1972, p. 65). While the quantal theory has a certain amount of face validity to it and obviously may have important implications for a number of areas such as phonological development in young infants, the articulatory and acoustic basis of distinctive features, and physiological constraints on coarticulation phenomena, very little empirical work has been directed towards testing its basic assumptions. This is also not too surprising at the present time because Stevens has not offered any specific predictions from the theory. Moreover, much of the work on the theory has been done under very idealized conditions involving numerous assumptions in order to implement the simulation procedures.

One obvious set of predictions that would seem to follow from the spirit of the quantal approach to speech deals with the stability or precision of the articulatory gestures underlying the production of different speech sounds. The precision of these gestures presumably would be reflected in the variability of the measured formant frequencies in the acoustic domain (Fant, 1960). In connection with vowels, there has been some discussion in recent years that the point vowels /i/, /a/ and /u/ may be thought of as stable reference points in both articulatory and acoustic space (see Gerstman, 1968; Lieberman, 1970, 1976). Indeed, the present investigation was undertaken in the hope of learning something about the variability associated with the production of a number of steady-state vowels and the possible relations that might be found between the variances of the measured formant frequencies.

The general hypothesis under consideration in this report was that some vowels, such as the point vowels, may be produced by inherently more stable articulatory configurations than other vowels and this articulatory stability or precision may be reflected in the pattern of variances of the formant frequencies over successive repetitions of a set of vowels by the same speaker. The study to be reported below used a mimicry paradigm in which the subject was presented with one of eight synthetically generated vowel targets and then was required to mimic the sound during a specific response interval. Reaction time measures were obtained as well as formant frequencies for each response.

Before proceeding to the experiment proper it would be appropriate to review briefly some of the previous work dealing specifically with the mimicry paradigm and the studies which lead up to the present inquiry. Several investigations have been carried out with this procedure in the past, although they were not concerned with the same questions under consideration here.

Some years ago Chistovich, Fant, de Serpa-Leitao and Tjernlund (1966) carried out a study on the imitation of a set of synthetic vowels in order to determine whether the internal representation of a vowel was a continuous motor representation or a discrete or finite set representation. A single subject was presented with 12 synthetic vowels that followed a continuous trajectory in the F_1, F_2, F_3 pathway from /a/ to /ɛ/ to /i/ and was required to perform three different experimental tasks. In the first task the subject was required to mimic the synthetic vowel as accurately as possible after listening to the whole stimulus. In the second task the subject had to shadow the vowel with the shortest delay possible. In the third task the subject had to identify the vowel by writing down an appropriate Russian letter for the vowel phoneme. The latter task was carried out with a metal plate and a special writing pen so that latencies would be obtained. Latencies were also obtained for the first two tasks by examination of Minograph records. Analyses of the spectra of the response vowels produced in the first two tasks were carried out with a 51-channel filter bank spectrum analyzer so that estimates of the frequencies of the first two formants could be obtained.

In order to determine whether the internal representation for vowels was continuous or discrete, Christovich et al. (1966) examined three characteristics of the spectral data obtained in the first two tasks. First, they sought to determine the relationship between the mean formant frequencies of the response vowels and the formant frequencies of the stimulus or target vowels. Evidence of any non-linearities in the form of a "step-like quantizing function" could be taken as support for the discrete or categorization hypothesis concerning the internal representation of these vowels. The second analysis of the spectral data involved an examination of the standard deviations of the formant frequencies of the response vowels. Under the categorization hypothesis, the standard deviations of the formant frequencies should be minimal at the center of the category as defined by the presence of the step-like functions observed earlier and should increase toward the extremes of the category. That is, there should be greater precision or stability of the responses within rather than between response categories. Finally, Chistovich and her colleagues examined the topography or distribution of the formant frequencies of the response vowels. The presence of pronounced peaks in the distribution reflecting modal response values at particular formant frequencies could also be interpreted as support for the discrete transformation view of the internal representation of vowels.

The results obtained by Chistovich et al. in the mimicry task for all three criteria favored the view that the internal representation of vowels is discrete and involves categorization whereby the instructions to the articulators have undergone some form of transformation resulting in a reduction of information. While there were some differences in the spectral data for the mimicry and shadowing tasks, the overall results showed that the observed perceptual categories do not appear to reflect the phonemes of the speaker's language but correspond more nearly to the allophones in her repertoire. The conclusion was based on the observation that there were more categories observed in the shadowing and mimicry tasks than in the writing task in which letter symbols were used as responses to represent the phonemes present in the subject's native language. Thus, the mimicry paradigm as used by Chistovich et al. appears to provide much more detailed information about the phonetic knowledge that the speaker has access to as well as the means used to implement these distinctions in the articulatory domain.

In a number of more recent studies, Kent (1973, 1974, 1976) has also used the mimicry paradigm to study the imitation of synthetic steady-state vowels as well as the formant motions present in time-varying vocalic stimuli. Kent (1973) found in one study that English speakers showed evidence of a continuous representation for synthetic vowels with formant trajectories ranging from /u/ to /i/ but a more nearly discrete or categorical representation for vowels with trajectories ranging from /æ/ through /i/. The /u/ to /i/ stimulus continuum contains only two vowels that are phonologically distinctive in English whereas the /æ/ through /i/ continuum contains several English vowels including: /i/, /I/, /e/, /ɛ/ and /æ/. While there was a tendency in this study for the standard deviations of the formant frequencies of the response vowels to show the expected pattern in terms of the categorization hypothesis described earlier, Kent reported relatively large individual differences across his four subjects. As expected, however, there were well-defined peaks in the distributions of the responses corresponding to phonological categories along the /æ / to /i/ continuum.

In another study which compared the imitation of English vowels with non-English vowels by adults and children, Kent (1976) found that familiarity with the vowels had a large influence on the reliability of reproduction as indexed by the standard deviations of the formant frequencies of the response vowels. Moreover, Kent also reported that the children had standard deviations roughly twice as large as those obtained with adults. While Kent attributed this finding to the possible measurement errors associated with the higher fundamental frequency of children, these results may also reveal something about the relative precision of the articulatory control children have in producing vowels under these conditions.

Although these previous investigations have used a mimicry paradigm, the interest for the most part has been limited to questions concerning the internal representation of vowels and possible familiarity effects. In the present study, we were concerned with a somewhat different issue concerning the relative precision with which different vowels could be produced. More specifically, we were interested in determining whether we could find some support for the idea that some vowels are inherently more stable than others and consequently we would anticipate that their formant frequencies should show lower variances. If such a pattern of results could be observed in a vowel mimicry paradigm, they could be taken as some empirical support for the Quantal Theory of Speech which assumes quantal regions for some vowel articulations.

The independent variable of most interest in this study was the particular vowel phoneme itself. The vowel target stimuli that we used were selected to scan the range of the vowel space of General American English encompassing the so-called point vowels, /i/, /a/ and /u/, as well as the non-point vowels, /I/, /ε/, /æ/, /Λ/, and /ɔ/. Several dependent measures were obtained from computer analyses of the responses. These included the mean frequencies of the first three formants and their respective standard deviations, the duration of the vowel response, and the latency of the response. Some estimates of the reliability of these measurements as well as the reliability of the overall pattern of results were obtained by means of correlations between data obtained in two separate test sessions.

Method

Stimuli

The stimuli used in this experiment consisted of the following eight synthetically generated steady-state vowels: /i/, /I/, /ɛ/, /æ/, /Λ/, /ɔ/, and /u/. All of the vowels had a duration of 300 msec with a pitch contour that fell linearly from 125 Hz to 80 Hz over the duration of the stimulus. The formant frequencies of the stimuli were originally taken from the averaged measurements provided by Peterson and Barney (1952) for their male talkers although some slight changes were made in these values to improve the vowel quality at the time of synthesis. These values are given in Table 1. The stimuli were originally synthesized on

----

Insert Table 1 about here

----

a digitally controlled software speech synthesizer at the Research Laboratory of Electronics, M.I.T. (Klatt, 1972). Since this is a serial synthesizer, formant amplitudes were determined automatically. The eight stimuli were converted to analog form, recorded on audio tape and then subsequently transferred to digital form on the PDP-11 computer system in the Psychology Department at Indiana University where the experiment was carried out.

Subjects

Two subjects were used in the present experiment. Both were right-handed, native speakers of English and reported no history of a speech or hearing disorder. One subject (B.S.) was a laboratory technician employed in the Psychology Department, the other subject (M.W.) was a graduate student in psychology. Subject B.S. was born in Northwestern Indiana and lived most of his life there. Subject M.W. spent most of his life in Southern California. Both subjects were naive with respect to the hypothesis under consideration in this study.

Procedure

The subjects were run one at a time in a single-walled sound attenuated room (Industrial Acoustics Model 401). There were two sessions lasting about one hour held on two successive days. The subject was seated comfortably in front of a high-quality microphone (Electro-Voice

Model 664) and was told to keep his head positioned so that a distance of 10 inches could be maintained between his lips and the microphone. The stimuli were presented through earphones (Telephonics TDH-39) at a comfortable listening level which was the same for each subject. All of the target stimuli were stored in digital form on disk and presented in real-time during the course of the experimental sessions. The subject's response as well as the target vowels output from the computer were recorded at 7½ ips on separate channels of a two-channel tape deck (Ampex AG-500) for later spectral analysis. Response latency was recorded by means of a voice operated relay interfaced appropriately to the computer. All stimulus timing, warning and response lights were controlled automatically by the computer.

There were three parts to the experiment, familiarization, practice and test. During the familiarization phase, subjects were presented with the set of eight target vowels in order six times in a row with a 4 sec interval between successive stimuli. They were told to listen carefully to each vowel and to study the following list which was present in front on them since it would help them to learn to distinguish each vowel sound:

	Vowel Sound	Example of Word with Vowel Sound
1.	"EE"	"heed"
2.	"IH"	"hid"
3.	"EH"	"head"
4.	"AE"	"had"
5.	"UH"	"hud"
6.	"AH"	"hod"
7.	"AW"	"hawed"
8.	"UU"	"who'd"

The second phase of the experiment consisted of having subjects practice producing the appropriate vowel as a response to the target vowels which were presented randomly for forth trials. Data were not collected during this phase. Subjects received the following set of detailed instructions describing the mimicry task and how their response was to be produced:

> This is an experiment dealing with the way people imitate vowel-like sounds. On each trial you will hear through your earphones a synthetic vowel like the ones you heard earlier. They will be presented in a random order one at a time. Your task is to try to mimick the sound as best as you can. That is, say out loud the same vowel as you would produce it. Remember, the sounds you will hear are synthetic and are approximations to the real vowels we want you to produce. There will be a "warning" light on the console in front of you signaling when the trial will begin. When you see the light go on you should get ready to utter the vowel that will come up next. Approximately 1 second after the light goes on the target vowel sound will come through your earphones. As soon as you know which vowel sound it is you should get ready to initiate your utterance. When the "go" light comes on you should make your response. After each vowel is presented you will have 1 second in which to initiate your mimic of the vowel. If you take longer than this the red light on the panel in front of you will flash and we will proceed to the next trial. You should consider each trial independently of previous trials. The stimuli are arranged in a completely random order, so you should not try to predict what stimuli will be coming up next. There is a microphone in front of you and you should be sure that your lips are always the same distance from the microphone. It is important that you produce your vowel responses as naturally as you can. It is not necessary for you to exert any extra effort since the microphone is very sensitive and we want you to do this as naturally as possible.

The last part of the experiment, the test phase, was identical to the practice session except that subjects received four blocks of 80 trials with a one-minute rest period between successive blocks. When subjects returned for the second test session on the next day, they received only a repetition of the test phase consisting of an additional four blocks of 80 trials. Within each block of 80 trials each of the eight stimulus vowels appeared randomly ten times each with the only restriction that no stimulus ever followed itself in the test sequence. Thus, each block of 80 trials could be considered a completely independent replication of the basic experiment with ten trials per stimulus vowel. Stimuli within a block were separated by 3.5 sec.

Spectral Analysis

Spectral data from the responses were obtained by means of linear prediction analysis (see McCandless, 1974). The audio tapes for each subject were taken to M.I.T.'s Lincoln Laboratories where a highly interactive computer facility for studies of the acoustic properties of speech could be used for the analysis (for a description of this facility see Zue, 1976). With the kind help and collaboration of Dr. Victor Zue we were able to obtain formant tracks of the first three formants for each response and a computer print-out of the numerical values of these formant frequencies sampled every five msec. Armed with these printouts we could then obtain values of the first three formants for each response vowel and its duration as well as estimates of the fundamental frequency. The measurements contained in these print-outs formed the basis for the subsequent data analysis reported below.

Results

Because of the large number of responses obtained in the present investigation and the amount of time required to perform the spectral analysis, only a small portion of the available data from this experiment have been processed to date. For analysis purposes we choose the first block of 80 trials in each test session. That is, only analyses of the data obtained on trials 1 through 80 and trials 321 through 400 will be reported below. At some point in the future we hope to carry out the analyses of the remaining data.

Formant frequencies for the first three formants of each response were obtained by selecting a point 50 msec from the onset of the vowel and recording the numerical values for F_1, F_2 and F_3 provided in the computer printout. Some care was taken at the time these values were recorded to insure that they did not represent spurious peaks in the formant tracks or marked deviations from the immediately surrounding values of the formant frequencies. In all cases, the values selected at 50 msec seemed to be chosen from fairly stable regions located centrally in each vowel. Additional analyses of these formant tracks is currently underway at the present time and will be presented in another report.

The mean formant frequencies for the first and second formants averaged over both experimental sessions is shown in Figure 4 separately for

---

Insert Figure 4 about here

---

each subject. The filled circles in this figure show the formant frequencies for the synthetic target vowels whereas the open triangles and squares show the response vowels for each subject respectively. The

F_1-F_2 vowel space for both subjects shows a reasonably good approximation to the space of the target vowels although there are some noticeable differences that can be observed in this figure. For example, both subjects show some deviation for the high front vowel /i/ and the vowels /I/ and /u/. In each of these cases, the frequency of F_1 is higher than in the target vowels. The mean formant frequencies of F_1, F_2 and F_3 for each vowel broken down separately by test session are given in Table 2.

Insert Table 2 about here

Product-moment correlations were carried out on the means from the two test sessions to determine whether the observed values were reliable and whether the pattern of the mean values was stable. In each case, extremely high and statistically significant correlations were obtained for both subjects on the mean formant frequencies for F_1, F_2 and F_3 across sessions. These correlations are shown in Table 3.

Insert Table 3 about here

In order to examine the relative stability or precision with which the different vowels were produced in this experiment, we also calculated the standard deviations of the formant frequencies for each vowel. Figure 5 shows the standard deviations of the first two formants averaged

----

Insert Figure 5 about here

----

over the two test sessions for each subject as a function of the target vowel. While the variability of F_1 is somewhat lower than F_2 this would be anticipated since the range over which the frequency of F_1 can vary is much smaller than F_2. Nevertheless, by inspection of the standard deviations of F_2, it can be observed that some vowels show somewhat more variability than other vowels. This is particularly true for subject B.S., shown in the left hand panel, who shows more variability for the back vowels than front vowels. Unfortunately the same pattern of results is not observed in the data from subject M.W., shown on the right. The standard deviations of the formant frequencies of F_1, F_2 and F_3 for each vowel are also shown in Table 2 separately for each test session.

To assess the reliability of the pattern of variances associated with the production of different vowels across test sessions, product-moment correlations were also calculated. These values are shown separately for each subject in Table 3. In contrast to the correlations obtained for the means, the correlations for the standard deviations were, in each case, very low and non-significant suggesting that the pattern of variability across vowels is not stable from session to session. Correlations were also carried out across subjects for each formant frequency taken separately. These correlations were, not too surprisingly, also quite low and non-significant suggesting that both subjects in this experiment were apparently employing quite different patterns of articulatory control in producing the same vowel responses. We did find, however, significant correlations, at least for B.S., between the frequency of F1 and its standard deviation ($r = +.753$,

$p < .05$) and the frequency of F2 and its standard deviation ($r = +.752$, $p < .05$). Both effects may be seen in Figure 5 in the left hand panel.

We also collected latency data during the production of these vowels with the hope that some interesting pattern of results might emerge. Unfortunately, because of programming problems and limitations on the computer's cycle time, we had to instruct subjects to wait until a "go" light was present before executing their response. There is no doubt that this gives an unrealistic estimate of the response latency and this will be corrected in future work. Nevertheless, Table 4 shows the means and standard deviations for response latency in each of the two sessions and the average over both sessions. While there were some differences between different vowels, they do not appear to be reliable across test sessions nor even across subjects which is probably to be expected based on the procedure used. However, it is clear that there is a substantial difference in the latencies for both subjects. The mean response latency across all vowels for subject B.S. was 256 msec whereas the mean latency for M.W. was 670 msec.

Insert Table 4 about here

Although the synthetic vowels that were used as target stimuli were all synthesized with the same overall duration of 300 msec, it was of some interest to examine the durations of the vowel responses produced by the two subjects. The means and standard deviations for duration of the vowel response are shown in Table 5. Inspection of this table shows that while there are some slight differences in vowel length across the

----

Insert Table 5 about here

----

different vowel targets, for the most part, no systematic pattern appears to emerge, at least according to traditional criteria on vowel duration in the phonetic literature. However, the pattern of durations observed here does seem to be reliable across both test sessions since very high correlations were observed for both subjects. It is also of some interest to note here that in all cases the two subjects produced responses of greater duration than the target stimuli, often by as much as 100 msec. It may be recalled from the instructions described earlier that we did not explicitly provide any information to the subjects concerning the natural inherent durations of these vowels and simply left this as a free variable.

Discussion

We began this report with a brief discussion of the Quantal Theory of Speech and its basic assumption that non-linearities can be found between articulation and acoustic output. In keeping with the spirit of the quantal theory as it might be applied to the production of vowels, we predicted that some vowels, particularly the point vowels, would show greater stability or precision in articulation which in turn would be reflected in the variabilities of the measured formant frequencies of vowels. To a first approximation, the results obtained for two subjects in a mimicry task do not provide very convincing support for this hypothesis. No clear pattern of variances appeared to emerge for the eight different vowels studied. While the mean values of the formant frequencies for the responses were highly reliable over the two test sessions analyzed so far, no such reliability was observed for the pattern of variances associated with these means. Moreover, comparisons between the two subjects revealed quite different patterns of variances for the formant frequencies suggesting that at least these subjects appeared to carry out the mimicry task with somewhat different production strategies. Indeed, these two subjects showed rather substantial differences in their overall latencies in initiating their vowel response.

What can be concluded from the results obtained in this study with regard to the predictions of the quantal theory? First, there is probably some reason to be suspicious of the pattern of variances obtained from the data analyzed so far primarily because this result does not appear to be very reliable across test sessions. We are currently analyzing substantially more of the data in the hopes of finding a more stable and reliable pattern of variances over test sessions. If these subsequent analyses are successful we will then be able to more accurately assess the original hypothesis under consideration.

Secondly, it may very well be the case that the outcome of this experiment is dependent to a large extent on the experimental procedure used, especially requiring our subjects to wait for the "go light" before initiating their response. Some modifications have already been made in the hardware in our laboratory which will permit a replication of this experiment without requiring a delay in the vocal response. The results of this study should then provide some information about whether the vowel responses should be produced spontaneously without delay. We also plan to place the vowel targets in various phonetic environments since these are known to exert rather strong effects, not only on the formant frequencies of the vowels themselves but also on the pattern of variabilities associated with the formant frequencies (Stevens and House, 1963).

Finally, it is quite possible, though unlikely, that the predictions derived from the quantal theory may be simply wrong in the context of this experiment or ones similar to it. It should be pointed out here that much of the evidence used to support the quantal theory and its predictions with regard to the formant frequencies of vowels and their presumed stability has been obtained under highly idealized conditions in the context of simulation experiments (see however, Stevens, 1975). As far as we know, no empirical data have been collected with real talkers under conditions approximating those assumed by the model that Stevens has used. Obviously, for the theory to have any claim to reality some empirically motivated set of criteria would have to be specified and subsequent data collected on the issue.

Some support for the predictions of the quantal theory, at least with regard to vowels, has already been reported in the literature although in a somewhat different context. For example, in an extensive investigation on the effects of consonantal context on vowel formant frequencies, Stevens and House (1963) reported standard deviations for

F_1 and F_2 for eight different vowels averaged over 14 different consonantal contexts for three talkers. Examination of these results indicated that tense vowels have substantially lower variances than lax vowels. Moreover, /i/ and /a/, two of the vowels under special consideration in this study, showed substantially smaller variances for F_2 than the other vowels. The variance of F_2 for the vowel /u/ was the highest, a fact easily accounted for by the addition of pronounced lip rounding in most consonantal contexts. After reporting these data, Stevens and House (1963) even point out at the time that "on the basis of these data it is evident that some vowel articulations are more stable than others" (p. 120). It may well be the case that the quantal theory is a reasonable account of some aspects of vowel articulation, but only when more constraints are placed on the surrounding phonetic environment of the vowel as it might appear in spontaneously produced utterances or continuous speech. Thus, the theory may be correct but not necessarily because of the evidence obtained in the simulation work carried out by Stevens.

In summary, based on the results obtained in the present mimicry experiment there would appear to be little support for the prediction that point vowels are inherently more stable than non-point vowels, at least with regard to measured formant frequencies. The pattern of results observed in this study as well as the procedures used in the mimicry paradigm suggested several modifications for future work which employs the pattern of variances of formant frequencies as an index of the precision or stability of vowel production.

References

Chistovich, L. Fant, G., de Serpa-Leitao, A., & Tjernlund, P. Mimicking of synthetic vowels. _Quarterly Progress and Status Report, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm_, 1966, No. 2, 1-18.

Fant, G. _Acoustic theory of speech production_. The Hague: Mouton, 1960.

Gerstman, L. J. Classification of self-normalized vowels. _IEEE Transactions on Audio and Electroacoustics_, 1968, _AU-16_, 1, 78-80.

Kent, R. D. The imitation of synthetic vowels and some implications for speech memory. _Phonetica_, 1973, _28_, 1-25.

Kent, R. D. Auditory-motor formant tracking: A study of speech imitation. _Journal of Speech and Hearing Research_, 1974, _17_, 203-222.

Kent, R. D. Imitation of synthesized vowels by children and adults. Paper presented at the 92nd meeting of the Acoustical Society of America, San Diego, California, November 15-19, 1976.

Klatt, D. H. Acoustic theory of terminal analog speech synthesis. _Proceedings of the 1972 International Conference on Speech Communication and Processing_. Boston, Mass. _IEEE_ 1972, No. 72, CHO, 567-7 AE, 131-135.

Lieberman, P. Towards a unified phonetic theory. _Linguistic Inquiry_, 1970, _1_, 3, 307-322.

Lieberman, P. Phonetic features and physiology: A reappraisal. _Journal of Phonetics_, 1976, _4_, 91-112.

Liljencrants, J., & Lindblom, B. Numerical simulation of vowel quality systems: The role of perceptual contrast. _Language_, 1972, _48_, 839-862.

Lindblom, B. E. F. Phonetics and the description of language. In A. Rigault and R. Charbonneau (Eds.) *Proceedings of the Seventh International Congress of Phonetic Sciences*. The Hague: Mouton, 1972, 63-97.

Lindblom, B. & Sundberg J. A quantitative model of vowel production on the distinctive features of Swedish vowels. *Speech Transmission Laboratory*, *Quarterly Progress and Status Report No*. *1*, 1969, pp. 14-30.

McCandless, S. S. An algorithm for automatic formant extraction using linear prediction spectra. *IEEE Transactions on Acoustics*, *Speech and Signal Processing*, 1974, *22*, 2, 135-141.

Peterson, G. E. & Barney, H. L. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 1952, *24*, 175-184.

Stevens, K. N. The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David, Jr. and P. B. Denes (Eds.), *Human communication*: *A unified view*. New York: McGraw-Hill, 1972.

Stevens, K. N. Further theoretical and experimental bases for quantal places of articulation for consonants. *Quarterly Progress Report No*. *108*, Research Laboratory of Electronics, M.I.T. 1973, 247-252.

Stevens, K. N. Quantal configurations for vowels. Paper presented at the 89th meeting of the Acoustical Society of America, Austin, Texas, April, 1975.

Stevens, K. N. & House, A. S. Perturbation of vowel articulations by consonantal context: An acoustical study. *Journal of Speech and Hearing Research*, 1963, *6*, 2, 111-128.

Zue, V. W. Acoustic characteristics of stop consonants: A controlled study. Unpublished doctoral thesis, Massachusetts Institute of Technology, 1976.

Table 1

Formant Frequencies of Synthetic Target Vowel in Hz

	i	I	ɛ	æ	ʌ	a	ɔ	u
F1	270	374	530	660	640	730	570	300
F2	2290	2070	1840	1720	1190	1090	840	870
F3	3010	2666	2480	2410	2390	2440	2410	2240

Table 2

Means and Standard Deviations in Hz for F_1, F_2 and F_3 by Session for Two Subjects

		Subject B.S.								Subject M.W.							
	Vowel Target	i	I	ɛ	æ	ʌ	a	ɔ	u	i	I	ɛ	æ	ʌ	a	ɔ	u
Means	$\overline{X}F_1$ (1)	304	476	560	658	600	775	590	322	320	434	501	610	560	742	576	336
	(2)	317	468	552	650	600	778	608	352	342	421	484	580	593	769	558	348
	Mean	311	472	556	654	600	777	600	337	331	428	493	595	577	756	567	342
	$\overline{X}F_2$ (1)	2259	1923	1861	1775	1201	1141	933	889	2039	1907	1791	1784	1237	1040	871	979
	(2)	2246	1996	1839	1713	1250	1179	999	998	2140	1920	1824	1664	1235	1086	897	855
	Mean	2253	1960	1850	1744	1226	1160	966	945	2090	1914	1808	1724	1236	1063	884	917
	$\overline{X}F_3$ (1)	2954	2561	2550	2466	2427	2502	2431	2349	3122	2594	2486	2427	2153	2345	2234	1998
	(2)	2900	2537	2570	2474	2451	2404	2332	2292	2715	2512	2438	2395	2329	2353	2294	2060
	Mean	2927	2549	2560	2470	2439	2453	2382	2321	2919	2553	2462	2411	2241	2349	2264	2029
Standard Deviations	σF_1 (1)	23	31	17	32	18	49	9	16	15	21	35	53	77	26	40	20
	(2)	12	8	22	38	19	44	35	2	16	30	18	28	19	31	19	9
	Mean	18	20	20	35	18	47	22	9	16	26	27	40	48	29	25	15
	σF_1 (1)	86	63	45	74	33	43	57	35	48	63	51	91	61	27	32	57
	(2)	67	144	107	66	33	39	53	75	108	35	45	71	33	80	30	91
	Mean	77	104	76	70	33	41	55	56	78	49	48	81	47	54	31	74
	σF_3 (1)	146	76	56	113	60	83	79	127	51	48	43	55	129	28	61	50
	(2)	91	53	148	23	193	30	43	97	183	39	30	44	54	53	32	88
	Mean	119	65	102	68	127	57	61	112	117	44	37	50	92	41	47	69

Table 3

Product-Moment Correlations for Means and

Standard Deviations for F_1, F_2 and F_3 Across Test Sessions

	Subject B.S.	Subject M.W.
Means		
F_1	+.997*	+.986*
F_2	+.996*	+.883*
F_3	+.965*	+.891*
Standard Deviations		
F_1	+.427	+.080
F_2	+.229	-.008
F_3	-.344	-.063

* With = 6 these correlations are significant beyond the .005 level.

Table 4

Response Latency in Msec by Test Session for Two Subjects

Vowel Target	i	I	ɛ	æ	ʌ	a	ɔ	u	Means
Subject B.S.									
Session 1									
Mean	239	259	259	243	246	252	234	213	243
σ	102	63	50	64	75	57	53	45	63
Session 2									
Mean	234	305	243	226	244	232	219	222	240
σ	93	71	32	31	26	28	44	43	46
All Sessions*									
Mean	228	296	255	250	270	274	243	238	256
σ	60	68	45	48	71	69	45	54	58
Subject M.W.									
Session 1									
Mean	676	652	635	584	608	552	568	598	609
σ	74	74	96	83	44	70	58	59	70
Session 2									
Mean	711	727	658	656	709	677	651	714	687
σ	75	61	78	76	54	57	97	91	74
All Sessions*									
Mean	724	698	657	657	672	645	621	692	671
σ	85	75	81	83	78	79	109	94	86

* Each mean is based on 80 responses per vowel obtained over all sessions including trials 1-640.

Table 5

Vowel Response Duration in Msec by Test Session for Two Subjects

Vowel Target	i	I	ɛ	æ	ʌ	a	ɔ	u
Subject B.S.								
Session 1								
Mean	415	416	404	380	418	415	446	475
σ	37	24	30	35	25	58	27	29
Session 2								
Mean	401	401	397	385	400	412	410	437
σ	36	35	35	20	17	21	23	32
Overall								
Mean	408	409	401	383	409	414	428	456
σ	37	30	33	28	21	40	25	31
Subject M.W.								
Session 1								
Mean	387	403	406	422	427	433	426	387
σ	23	21	37	24	27	25	35	24
Session 2								
Mean	378	408	406	412	411	431	420	370
σ	17	19	31	18	21	25	24	41
Overall								
Mean	83	406	406	417	419	432	423	379
σ	20	20	34	21	24	25	30	33

Figure Captions

Figure 1. Hypothetical quantal relations between a parameter that describes some aspect of articulation and the resulting acoustic parameter of speech. (Adapted from Stevens, 1972).

Figure 2. Panel (a) shows a two-tube resonator approximating the vocal tract configuration for the vowel [a]. A_1 and A_2 represent the cross-sectional areas of the pharyngeal and oral cavities, respectively. Panel (b) shows an approximation of the spectrum envelope for the vowel [a] produced by the idealized configuration above. The peaks in the function represent the center frequencies of the formants. (Adapted from Stevens, 1972).

Figure 3. Results of a vocal tract simulation showing the relation between the frequencies of the first and second formants (F_1 and F_2) when the length of the back tube, d_1, is varied between 5 and 15 cm. (Adapted from Stevens, 1972).

Figure 4. Mean values of the first- and second-formant frequencies in Hz for two subjects, B.S. and M.W., obtained in the mimicry task. The filled circles show the formant frequencies for the synthetic target vowels, whereas the solid lines in each panel are drawn through the vowel space of the observed responses.

Figure 5. Standard deviations in Hz of the mean formant frequencies for F_1 and F_2 for the eight target vowels shown separately for each subject.

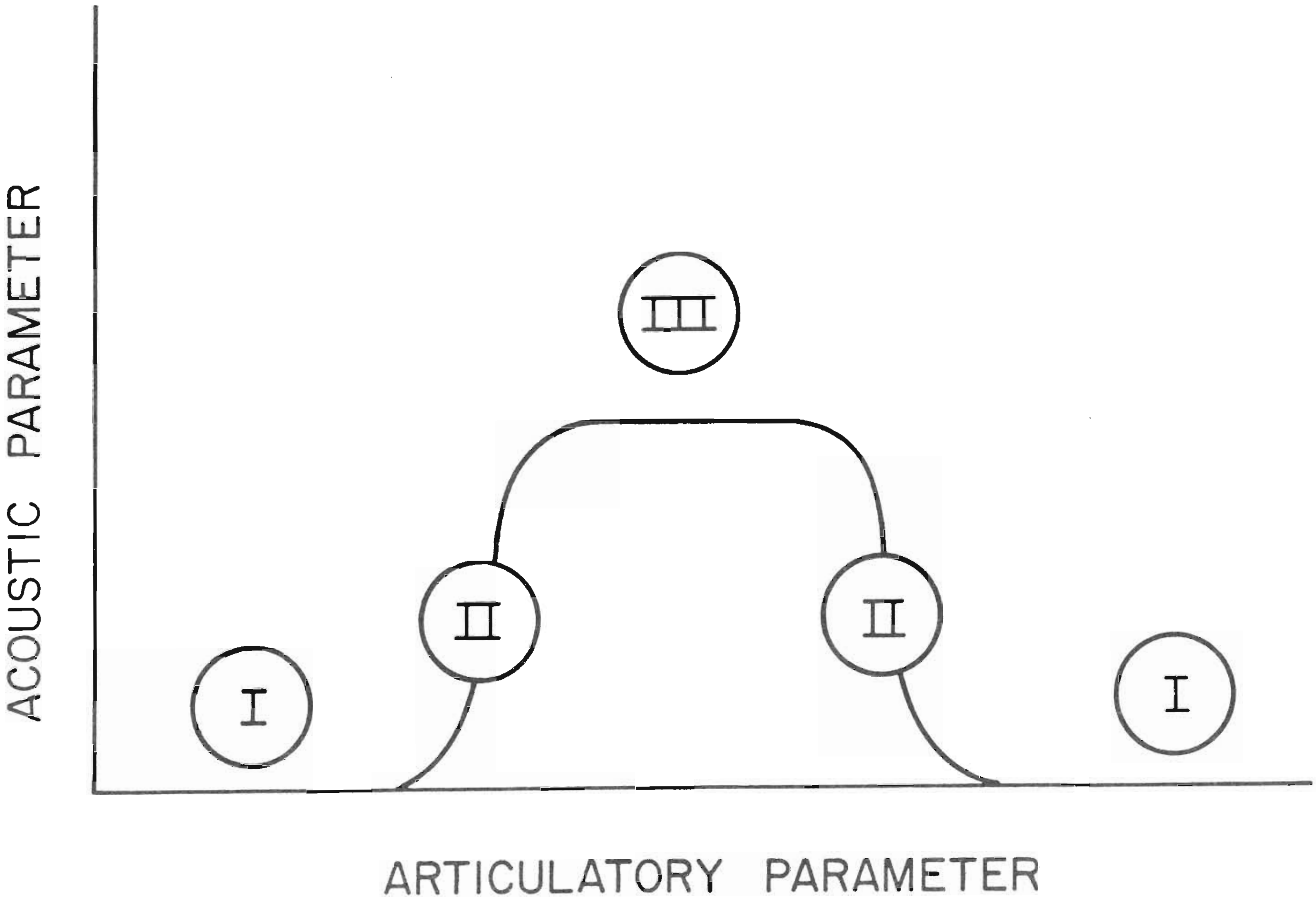


Figure 1.

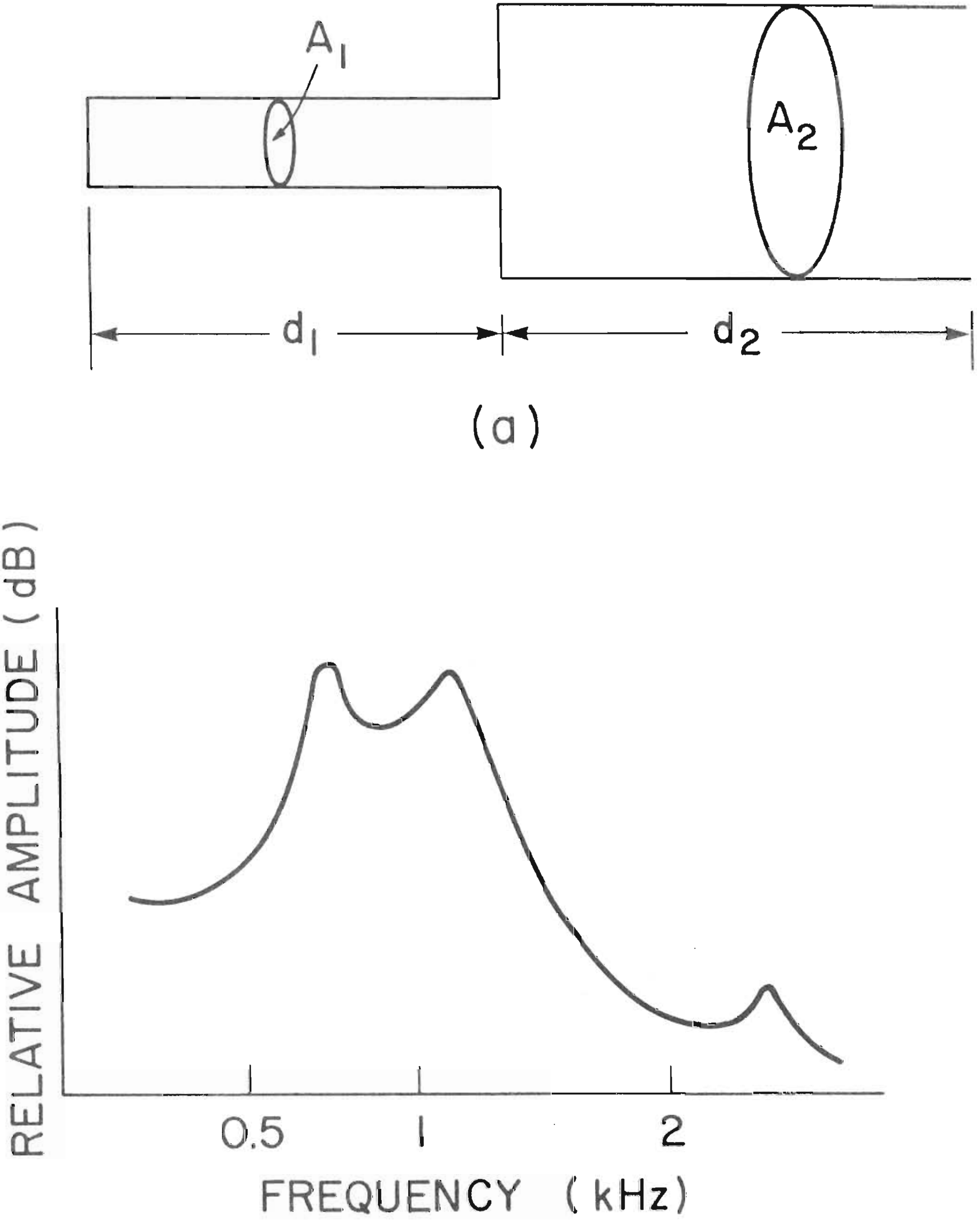


Figure 2.

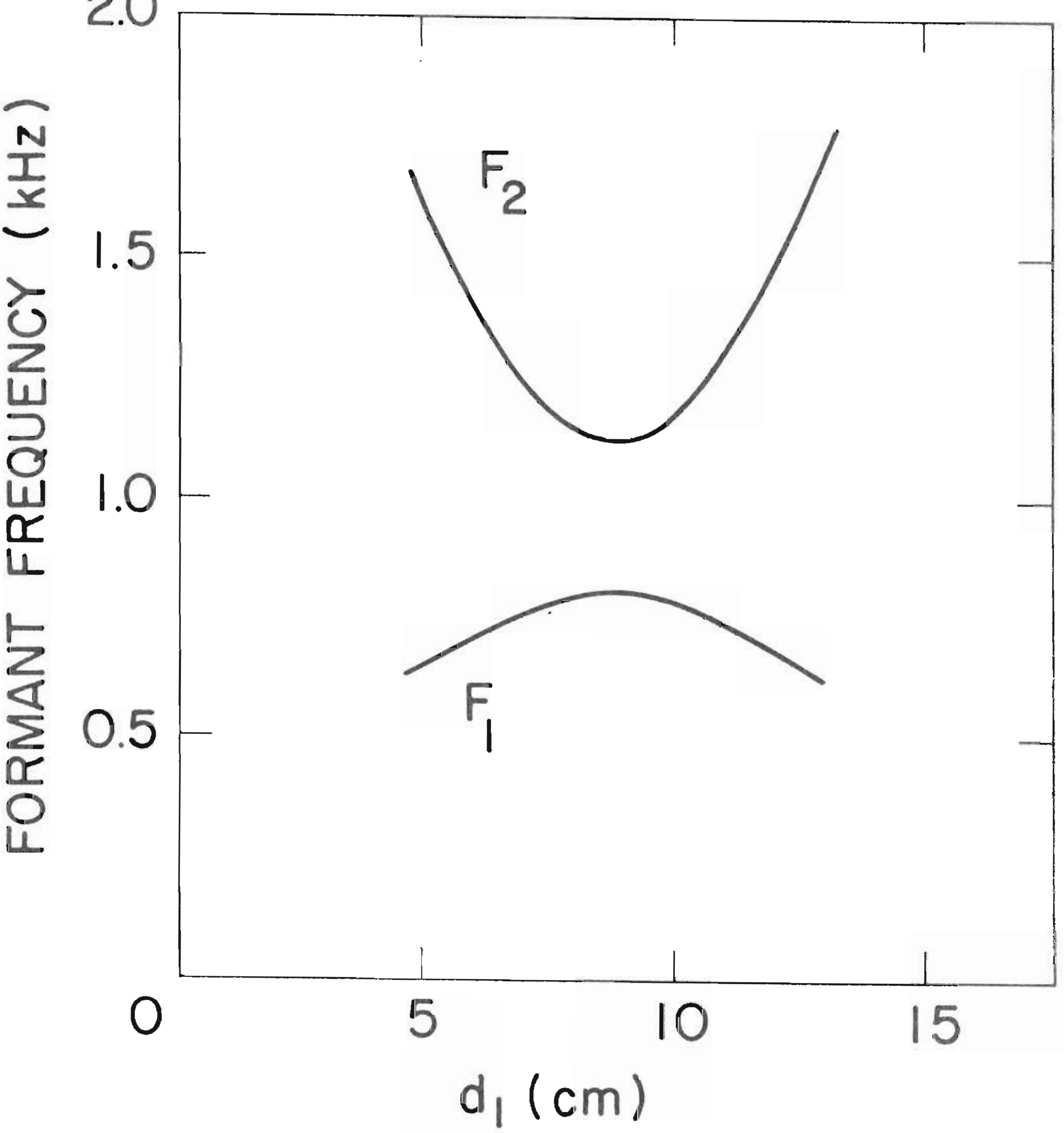


Figure 3.

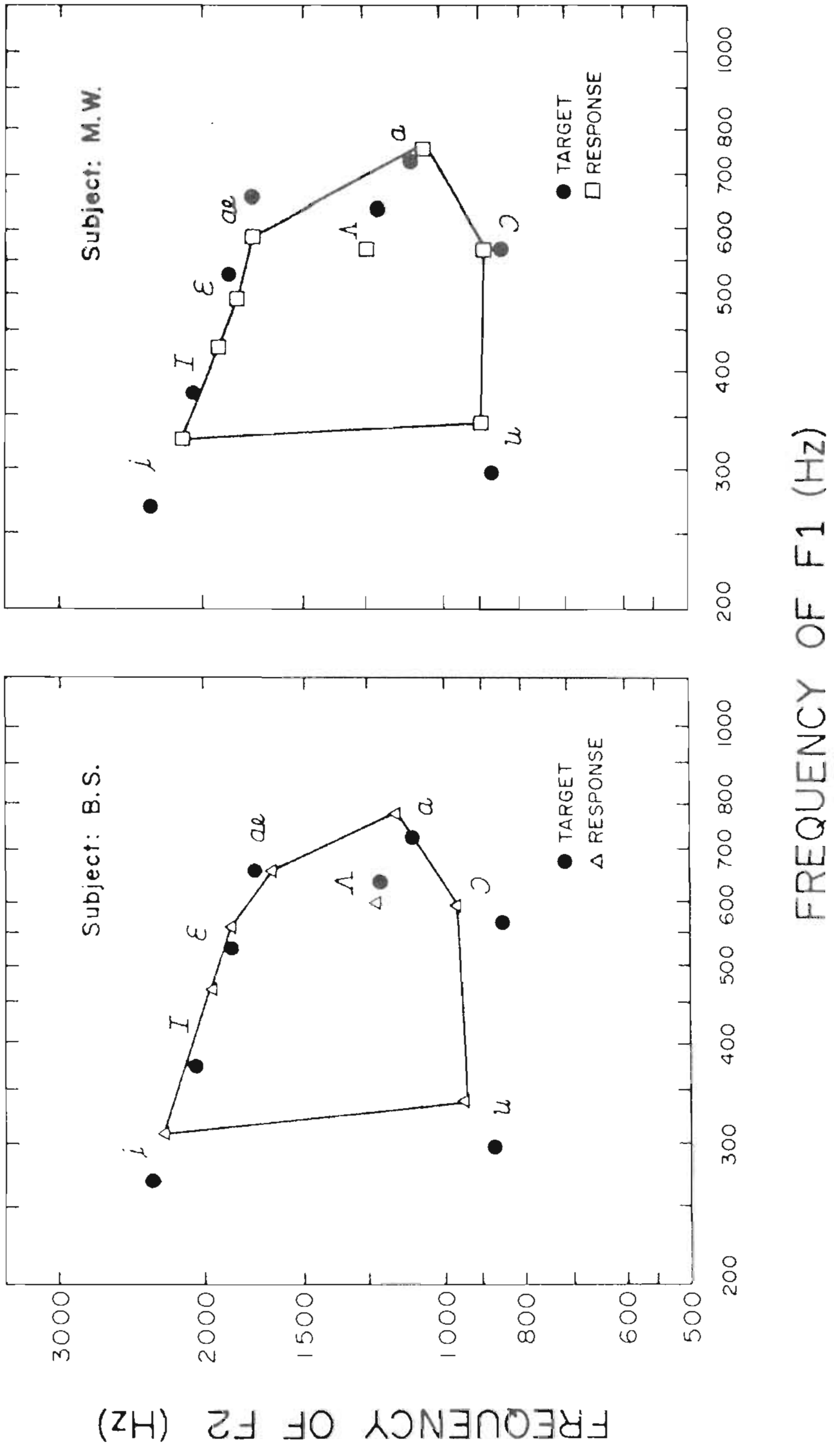

Subject: B.S.
Subject: M.W.
● TARGET
Δ RESPONSE
□ RESPONSE
i
I
ε
æ
a
ʌ
ɔ
u
FREQUENCY OF F2 (Hz)
3000
2000
1500
1000
800
600
500
FREQUENCY OF F1 (Hz)
200
300
400
500
600
700
800
1000

Figure 4.

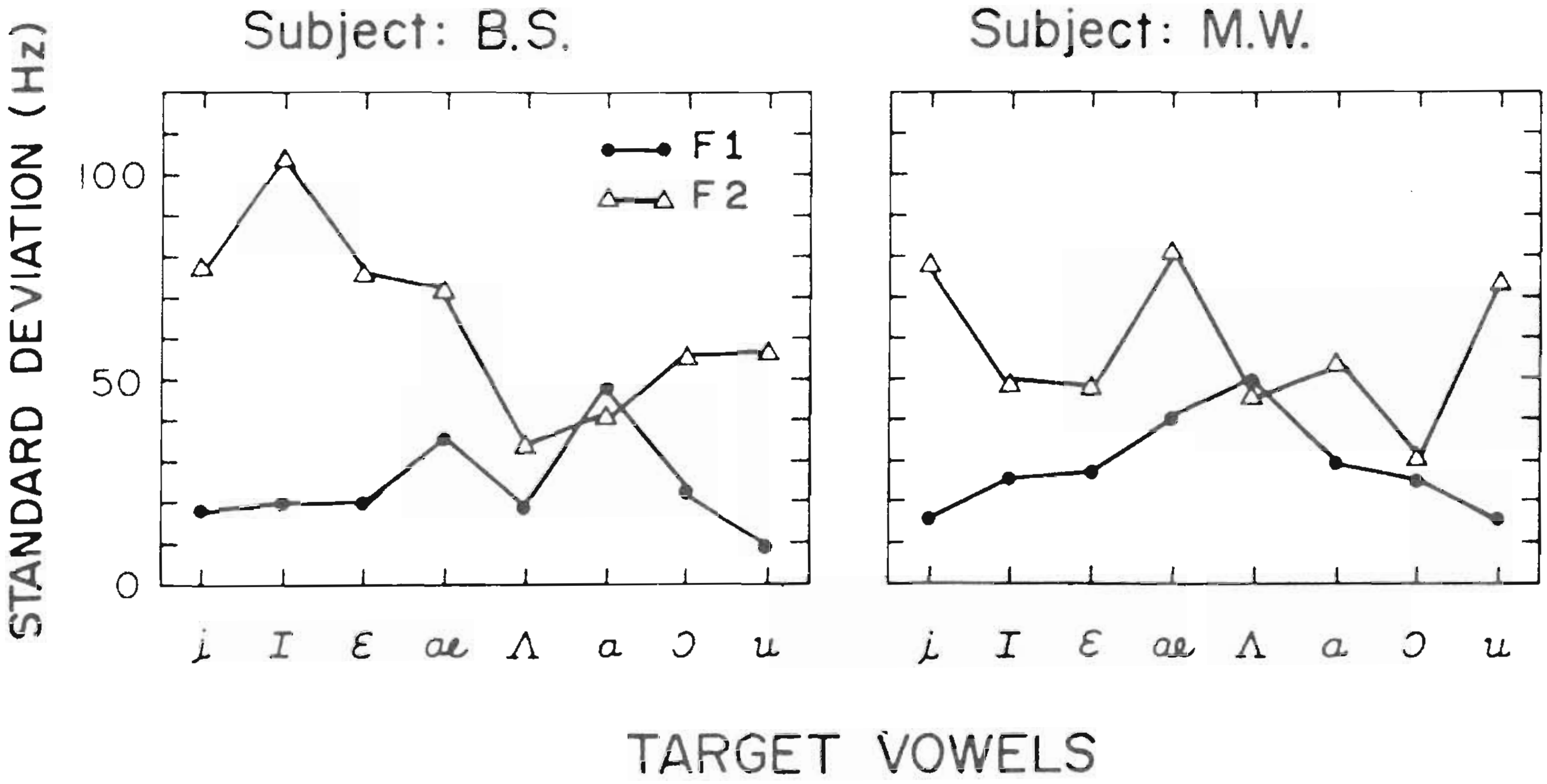

Subject: B.S.
Subject: M.W.
F1
F2
STANDARD DEVIATION (Hz)
100
50
0
i I ɛ æ ʌ a ɔ u
i I ɛ æ ʌ a ɔ u
TARGET VOWELS

Figure 5.

Some Effects of Discrimination Training on the Identification and Discrimination of Rapid Spectral Changes*

David B. Pisoni

Indiana University

Bloomington, Indiana 47401

Abstract

This paper reports the results of two experiments in which subjects were trained to identify stimuli selected from a nonspeech auditory continuum by means of a disjunctive conditioning procedure. The stimuli varied in the direction and extent of a very brief rapid spectral change. These changes were comparable to those typically used as F_2 formant transitions in synthetic speech sounds. Following the training procedure, the subject's identification responses generalized to the intermediate stimuli such that the stimulus continuum was partitioned into two equivalent response classes. Discrimination functions were also obtained by means of an ABX procedure. The discrimination results for these nonspeech stimuli showed peaks at the boundary between the categories and troughs within the categories which were correlated with changes in the identification functions. The results were discussed within the context of recent work on categorical perception of nonspeech sounds, the role of environmental modification of perceptual processing and the influence of early experience on the development and tuning of perceptual analyzers for speech signals.

* This work was carried out at Indiana University while the author was on leave at the Research Laboratory of Electronics, Massachusetts Institute of Technology during the academic year 1975-1976. I thank Jerry Forshee for his help in all phases of this work and Judy Hupp for her able assistance in running subjects.

Some Effects of Discrimination Training on the Identification and Discrimination of Rapid Spectral Changes

David B. Pisoni

Indiana University

Within the last few years there has been a renewed and quite vigorous interest in categorical perception of speech sounds in adults and infants as well as extensions of this work to other organisms such as the chinchilla and monkey. Several investigators have also recently reported very strong evidence for categorical-like perception of nonspeech signals having properties similar to speech (Cutting & Rosner, 1974; Miller, Wier, Pastore, Kelly & Dooling, 1976; Pisoni, 1977). These results with nonspeech signals suggest the possibility that categorical perception may not be unique to speech perception or a linguistic mode of processing as once thought but may be a more general property of all sensory systems. As Stevens (1972) has suggested, it may be that speech perception has simply exploited certain properties of the auditory system in selecting the sound attributes used for signaling phonetic distinctions.

Some years ago, Lane (1965) attempted to put forth a somewhat similar argument with regard to the experimental evidence cited at the time in support of the motor theory of speech perception. Most of the evidence for the motor theory concerned the differences found in perception between speech and nonspeech sounds, particularly with regard to the way listeners identified and subsequently discriminated between stimuli selected from continua representing equal physical differences along some complex acoustic dimension.

In several papers, Lane (1965, 1967) sought to demonstrate categorical perception effects with non-speech stimuli. Using a variety of visual and auditory stimuli, he tried to show that if two stimuli sampled from the same continuum evoke different identification responses, they will be more readily discriminated than two stimuli, separated by equal physical units, that evoke the same identification responses. Lane maintained that the observed relation between identification and discrimination found in categorical perception studies is not peculiar to the speech sounds, but rather is the result of very general procedures for discrimination training and testing with discrete response repertoires. That is, categorical perception is due to labeling responses that subjects have learned in the course dealing with speech sounds.

Support for this account of categorical perception came from a series of experiments reviewed by Lane (1965). The primary experiment cited by Lane was an otherwise unpublished study carried out by Lane and Schneider (1963). In this experiment, subjects were trained to identify the extreme members of a non-speech auditory continuum, the inverted spectrograms of the /do/-/to/ control continuum of Liberman et al., 1961, by a disjunctive conditioning procedure (see Cross and Lane, 1962). Lane and Schneider trained subjects to identify the extreme members of this non-speech continuum as /do/ or /to/. Following this training, the responses generalized to the intermediate stimuli such that the stimulus continuum was partitioned into two equivalent response classes. Before initiating the training procedure with the extreme stimuli, the ABX discrimination functions for the non-speech continuum showed chance performance. However, after training, the discrimination functions showed peaks at the boundary between the two response categories.

Lane's contention at the time was that categorical perception could be obtained "quickly and easily in the laboratory after a few minutes of conditioning." If this view is correct, there would be little justification for considering the perception of speech as special in some sense and possibly involving perceptual operations and mechanisms basically different from those used in the processing of other auditory signals.

Lane's account of categorical perception was examined in great detail by the Haskins group in their reply some five years later (Studdert-Kennedy et al., 1970). Studdert-Kennedy et al. focused most of their attention on the Lane and Schneider experiment in which Lane purported to show categorical perception with nonspeech sounds. After reconstructing the data for the subjects which Lane presented, they questioned the effectiveness of his discrimination training procedures in producing categorical perception with nonspeech stimuli. Their major criticism of the Lane and Schneider experiment dealt with the ABX discrimination data for the three subjects presented by Lane in his 1965 paper. Studdert-Kennedy et al. argued that two of the conditions for categorical perception were not reliably established in this experiment. First, the discrimination functions were in every case less consistent than the data shown by Lane for the one "representative" subject. Second, the discrimination functions after training could not be accurately predicted from the corresponding identification functions for any of the three subjects.

The Haskins group undertook to replicate the Lane and Schneider experiment. Their attempt to produce categorical perception by discrimination was summarized in only one paragraph in their reply to Lane. From this description it appeared that they were no more successful than Lane in establishing categorical perception by discrimination training procedures. Of the five subjects in the Haskins replication study, only one achieved better than 75% accuracy in identifying the two extreme stimuli although no indication of the number of training trials was provided. The effect of discrimination training for this one subject was to raise his overall discrimination performance rather than producing the expected categorical discrimination function.

The conflicting results of the Lane and Studdert-Kennedy experiments with nonspeech sounds indicated a real need for an additional attempt at establishing categorical perception by discrimination training. In 1971 we (Pisoni, 1971) carried out an experiment that was designed specifically to replicate the original Lane and Schneider training procedures with a stimulus variable that was previously shown to be perceived categorically in speech context but continuously in nonspeech contexts (Mattingly, Liberman, Syrdal and Halwes, 1971). The experimental variable under consideration was the direction and extent of a second formant transition removed from speech context. The results of that study were not entirely satisfactory for a number of reasons and, as a consequence, the present experiments were carried out to determine what effects discrimination training procedures might have on the identification and discrimination of nonspeech signals. This study was undertaken with full knowledge of the work of Miller et al. (1976) and Cutting and Posner (1974) on nonspeech signals. In those nonspeech studies as well as the one carried out recently by Pisoni (1977) the stimuli could be easily labeled by observers and therefore the question of training in these studies becomes only a minor issue. In this study the focus is on the training procedure itself.

Experiment I

Method

Stimuli

A set of seven single formant patterns was designed as nonspeech control stimuli for use in discrimination training. These patterns consisted of a steady-state second formant at 1386 Hz and seven initial second formant transitions. Figure 1 shows schematized spectrograms of stimuli 1 and 7. The F_2 transition was 35 msec in duration while the steady-state portion was 165 msec. The starting frequencies of the transitions were varied systematically from 921 Hz to 1845 Hz in approximately equal steps. The transitions for stimuli 1, 2 and 3 had positive slopes, beginning below the steady-state frequency and rising while stimuli 5, 6 and 7 had negative slopes, beginning above the steady-state frequency and falling. Stimulus 4 had no transition at all since it began at the same frequency as the steady-state portion of the formant. The fundamental frequency was held steady at 114 Hz for the first 100 msec and then dropped linearly to 70 Hz over the remainder of the stimulus. The seven stimuli were originally produced on the Haskins parallel-resonance synthesizer and recorded on magnetic tape. The stimuli were later digitized and the wave forms stored on the PDP-11/10 system in the Psychology Department at Indiana University.

Insert Figure 1 about here

Subjects

Eight undergraduate students at Indiana University served as subjects in the present experiment. The subjects were obtained from an advertisement in the student newspaper. All subjects were right-handed native speakers of English and reported no past history of a hearing disorder or speech impediment. The subjects were paid at a base rate of $2.00 per hour. The opportunity to earn additional money in each session was determined by their performance during the training conditions. None of the subjects had ever heard any synthetic speech and were naive to the experimental hypothesis under test.

Procedure

Subjects were tested in small groups in an experimental room containing six test booths. Each subject served for approximately an hour a day on two consecutive days. Each test booth was equipped with a response box containing appropriate response buttons and feedback lights as well as a set of high quality headphones (TDH-39). Stimulus presentation, timing of experimental events, feedbacks and monitoring was carried out on line in real-time under the control of the PDP-11/10 computer.

The experiment was divided into three major phases: discrimination training, identification testing, and discrimination testing. The first day of the experiment consisted entirely of discrimination training with the endpoint stimuli of the test series, Stimulus 1 and Stimulus 7. Subjects received three blocks of 160 trials. In the first block the two stimuli were arranged in an order appropriate for initial shaping of the desired response. In the other two blocks subjects received Stimuli 1 and 7 in a random order. Feedback was provided during all training sessions by flashing a light above the correct response after each trial. Subjects received the following set of instructions concerning the discrimination training phase of the experiment:

Discrimination Training Instructions

In this part of the experiment we are going to train you to identify some complex sounds. These sounds are actually the speech sounds /bae/ and /dae/ which have been modified by a computer to make them sound somewhat different from their original form. In front of you is a response box with several buttons. The button on the left is marked with the syllable /bae/ and the one on the right is marked with the syllable /dae/. What we would like you to do is to identify each sound by pressing either the /bae/ button or the /dae/ button. You can earn additional money by simply responding with a /bae/ or a /dae/ at the appropriate time. We cannot tell you now when or how you should use these two responses. That is obviously for you to figure out and learn for yourself. Indeed, that is part of the experiment. But all you have to do is wear the headset in front of you and listen to the sounds that will be presented to you. We also want you to watch the lights on the display unit in front of you since you will be receiving feedback after each trial indicating which was the correct response. During the course of the experiment you will only hear two sounds, one of them will sound something like a /bae/ and the other will sound something like a /dae/. Each time you respond appropriately you will receive a point which will be added to your score on the computer in the next room. If you do not respond correctly you will not get a point and the computer will subtract two points from your score. The total amount of money that you can earn in this experiment will depend on how many points you can accumulate by the end of the session. The beginning of each trial will be signaled when all of the lights on the display panel go on. About a half second later a stimulus will come over the earphones. Try to identify whether the sound was a /bae/ or /dae/. Then press the appropriate button and wait to see if you got the correct response. Remember, the light will always indicate what was the correct response on a trial. You will obviously want to earn as high a score as possible because the amount of money we will pay you at the end of the experiment will be determined by your final score. That is, the total amount of money you earn is dependent on the number of correct responses you make during the session. Please be sure to respond on every trial if you have to guess. And be sure to watch the feedback lights which will let you know which was the correct response after each trial.

If you have any questions, the experimenter will be happy to help you if he/she can. Please raise your hand.

The second day of the experiment consisted of three separate phases. First, subjects received a brief warmup period (80 trials) in which Stimuli 1 and 7 were presented in a random order as on the previous day and feedback was provided for the correct response. In the second phase, identification testing, all seven stimuli were presented twenty times each in a random order for 140 trials. However, no feedback was provided after each response. Finally, in the last phase, two-step ABX discrimination was measured. This consisted of 100 ABX trials without feedback. Each of the five two-step discrimination comparisons was responded to 20 times by each subject.

Results and Discussion

The identification and ABX discrimination functions for each of the eight subjects are shown in Figure 2. As can be observed, almost all of the eight subjects show very good labeling functions for the nonspeech stimuli even after only 560 training trials. These results are substantially better than those obtained in Pisoni (1971).

Insert Figure 2 about here

The category boundaries (CBs) were determined by a computer program that finds the 50 per cent point by linear interpolation. These are shown in Figure 2 separately for each subject. In each case, the category boundary occurs earlier than Stimulus 4 which has no transitions and therefore might be considered the natural dividing point of the stimulus continuum.

The ABX discrimination functions are also shown in Figure 2 superimposed on the identification functions. For some subjects such as S4, S5, S7, and S8 the peak in discrimination occurs at the category boundary as expected, whereas for others, the discrimination functions appear to show less correspondence with the identification data. These results are also substantially better than those obtained in the earlier study by Pisoni (1971) suggesting that discrimination training can indeed produce categorical-like perception for nonspeech signals even with only a very modest amount of training. While there is substantial variability among subjects, the shape of the discrimination functions do, in some cases, parallel those that would be predicted from the categorical perception hypothesis. The fact that the locus of the category boundary occurs at a lower stimulus value than the middle stimulus is of some interest since it might indicate some possible psychophysical landmark or perceptual discontinuity. To pursue this possibility a little further we ran an additional experiment in which we tried to modify the location of the boundary and therefore specify the criterial attributes of the response categories to our subjects.

Experiment II

Method

Stimuli

The same set of seven stimuli were used in this experiment.

Subjects

Eight additional subjects were recruited for this experiment. They met the same requirements as the subjects used earlier and were naive with respect to the experimental hypothesis under test.

Procedure

The procedure was quite similar to that used in the previous experiment except for several changes in the training procedure and an increase in the number of trials run in identification and ABX discrimination. Each subject served for one hour a day but this experiment lasted three days instead of two. On Day 1 subjects received the same 160 trial shaping procedure using Stimuli 1 and 7 as in the previous experiment. This was followed by three separate blocks of 160 trials of training stimuli. In the first block, Stimuli 1 and 7 were presented randomly. In the second block Stimuli 2 and 6 were added to the set of stimuli. Finally, in the last block Stimuli 3 and 5 were also added to the set. During the whole discrimination training session feedback was provided as before. The instructions to subjects were identical to those used earlier.

On Days 2 and 3 subjects returned and received a 160 trial warmup sequence containing only Stimuli 2 and 6. This was followed by the identification test using all seven stimuli in the set. In this condition, subjects received each stimulus 20 times a day for two days with no feedback. Finally, after the identification sequence, the subjects received 200 ABX discrimination trials a day with feedback for the correct response.

Results and Discussion

The identification and ABX discrimination functions are shown in Figure 3 for individual subjects. As with the results from the previous

---

Insert Figure 3 about here

---

experiment, there are clearly large individual differences in both the labeling performance and discrimination functions for these subjects. There is some evidence for a correspondence between identification and discrimination for S1, S2, S6, and S7 but there is also a good deal of variability in the data for the other subjects. Because of this we ran another six subjects through the same experimental procedures to provide additional data. The results from this replication are shown in Figure 4.

Insert Figure 4 about here

Both sets of data indicate that the category boundaries are systematically shifted to progressively higher stimulus values along the continuum as compared with the results obtained in Experiment I. This would appear to be a direct consequence of the training procedure employed in which other exemplars of the particular category were used to define the criterial attributes of the category. From these results, it may be inferred that the boundaries can be modified selectively by discrimination training and that the alignment and precise location of the boundaries may not be rigidly fixed as if there were a "perceptual threshold" or psychophysical notch in the stimulus continuum. Thus, training procedures such as these do appear to be effective in producing categorical-like perception for nonspeech signals as well as playing an important role in defining the acoustic attributes that specify category membership for a class of acoustic signals having properties like speech.

The results of the present experiment are, surprisingly enough, quite different from those obtained by Pisoni (1971) in an earlier study using the same stimuli. While the data from the present study show the expected labeling and discrimination functions that would be predicted from Lane's account of the role of discrimination training in categorical perception, there are several cases in which the results are not as entirely convincing as might be anticipated. Nevertheless, one is compelled to conclude that there is something to what Lane claimed in his 1965 paper although this may not be the whole account. Discrimination training procedures of the kind used here can produce selective effects on both identification and discrimination of nonspeech signals. It is also clear that labeling and discrimination functions such as those obtained in the present experiments can be obtained quickly and easily but only after several hundred trials.

The question remains, of course, as to the particular kind of perceptual learning that is going on in experiments of this sort. As we noted in some previous work on the identification and discrimination of the relative onset of two tones, the training procedures used here may simply provide the observer with information that helps to specify the important criterial attributes of the stimulus (Pisoni, 1977). The observed labeling and discrimination functions may simply reflect the basic underlying psychophysical properties of the stimuli modified by experience with these signals. In the case of speech and speech-like stimuli, the important cues are often very brief or transient relative to the acoustic attributes that subjects may be initially set to listen for at the onset of the experiment. By employing some form of training procedure we may be helping subjects to use their basic auditory capacities to a greater or lesser extent depending on the particular cue to be discriminated and the properties of the stimulus configuration itself.

Additional work on the use of discrimination training procedures applied to the problem of "re-acquisition" of phonological contrasts in several languages is currently underway in our laboratory. One project is concerned with training Arabic speakers to re-acquire the /b/-/p/ distinction which is not present in their phonological system. Another project will train English speakers to perceive distinctions in voicing as well as differences in tone that do not occur in English but which have well-defined acoustic properties which can be made salient for listeners in the context of a discrimination training paradigm such as the one used here. Our interest in this line of work derives, of course, from theoretical issues concerning the role of early experience in perceptual development and the extent to which the perceptual apparatus can be modified or re-aligned by experimental intervention.

References

Cross, D. V. & Lane, H. L. On the discriminative control of concurrent responses: The relations among response frequency, latency and topography in auditory generalization. Journal of the Experimental Analysis of Behavior, 1962, 5, 487-496.

Cutting, J. E. & Rosner, B. S. Categories and boundaries in speech and music. Perception & Psychophysics, 1974, 16, 564-570.

Lane, H. L. The motor theory of speech perception: A critical review. Psychological Review, 1965, 72, 275-309.

Lane, H. L. A behavioral basis for the polarity principle in linguistics. In K. Salzinger and S. Salzinger (Eds.), Research in verbal behavior and some neurophysiological implications. New York: Academic Press, 1967. Pp. 79-96.

Lane, H. L. & Schneider, B. A. Discriminative control of concurrent responses by the intensity, duration and relative onset time of auditory stimuli. Unpublished report, Behavior Analysis Laboratory, University of Michigan, Ann Arbor, 1963.

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

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

Miller, J. D., Wier, C. C., Pastore, R., Kelly, W. J. & Dooling, R. J. Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. _Journal of the Acoustical Society of America_, 1976, _60_, 2, 410-417.

Pisoni, D. B. On the nature of categorical perception of speech sounds. Ph.D. Thesis, University of Michigan, August, 1971. Also: _Supplement to Status Report on Speech Research_, SR-27. New Haven: Haskins Laboratories, 1971, pp. 1-101.

Pisoni, D. B. Identification and discrimination of the relative onset of two component tones: Implications for the perception of voicing in stops. Progress Report No. 118, Research Laboratory of Electronics, M.I.T., June, 1976, Pp. 212-230. Also accepted for publication in the _Journal of the Acoustical Society of America_, 1977, _00_, 000-000.

Stevens, K. N. The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David, Jr. and P. B. Denes (Eds.), _Human communication: A unified view_. New York: McGraw-Hill, 1972.

Studdert-Kennedy, M., Liberman, A. M., Harris, K., & Cooper, F. S. The motor theory of speech perception: A reply to Lane's critical review. _Psychological Review_, 1970, _77_, 3, 234-249.

Figure Captions

Figure 1. Schematized spectrographic patterns representing the extreme members of the test continuum, Stimulus 1 and Stimulus 7.

Figure 2. Identification and ABX discrimination functions for eight subjects in Experiment I who were trained only on Stimulus 1 and Stimulus 7. The values in each panel show the locus of the category boundary.

Figure 3. Identification and ABX discrimination functions for eight subjects in Experiment II who were trained on Stimuli 1, 2, 3 for one response and Stimuli 5, 6, 7 for the other response.

Figure 4. Identification and ABX discrimination functions for an additional six subjects run under the same conditions as those used in Experiment II.

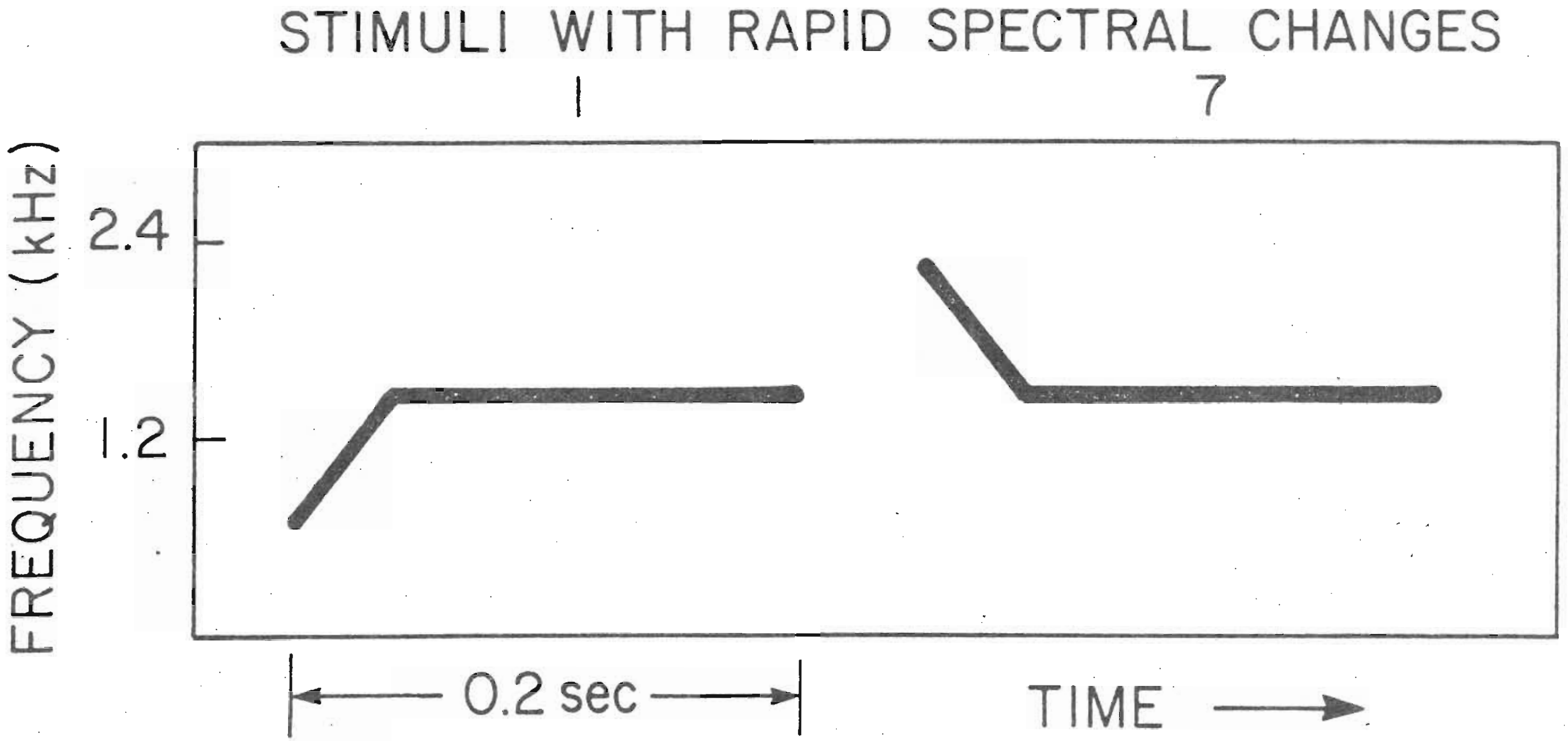

STIMULI WITH RAPID SPECTRAL CHANGES
1
7
FREQUENCY (kHz)
2.4
1.2
0.2 sec
TIME

Figure 1.

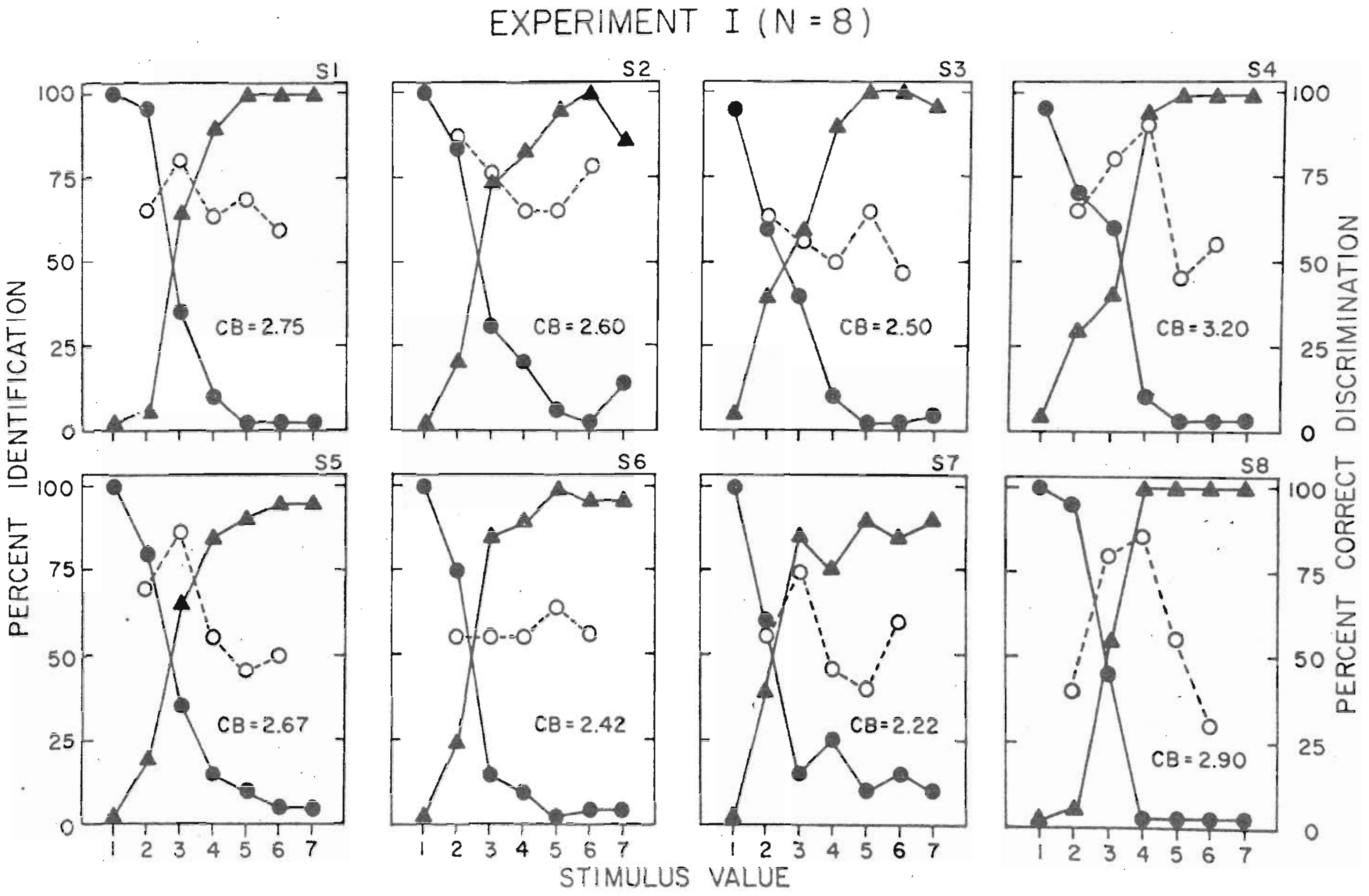

EXPERIMENT I (N = 8)
S1
S2
S3
S4
S5
S6
S7
S8
CB = 2.75
CB = 2.60
CB = 2.50
CB = 3.20
CB = 2.67
CB = 2.42
CB = 2.22
CB = 2.90
PERCENT IDENTIFICATION
PERCENT CORRECT DISCRIMINATION
STIMULUS VALUE
100
75
50
25
0
1
2
3
4
5
6
7

Figure 2.

EXPERIMENT II (N=8)

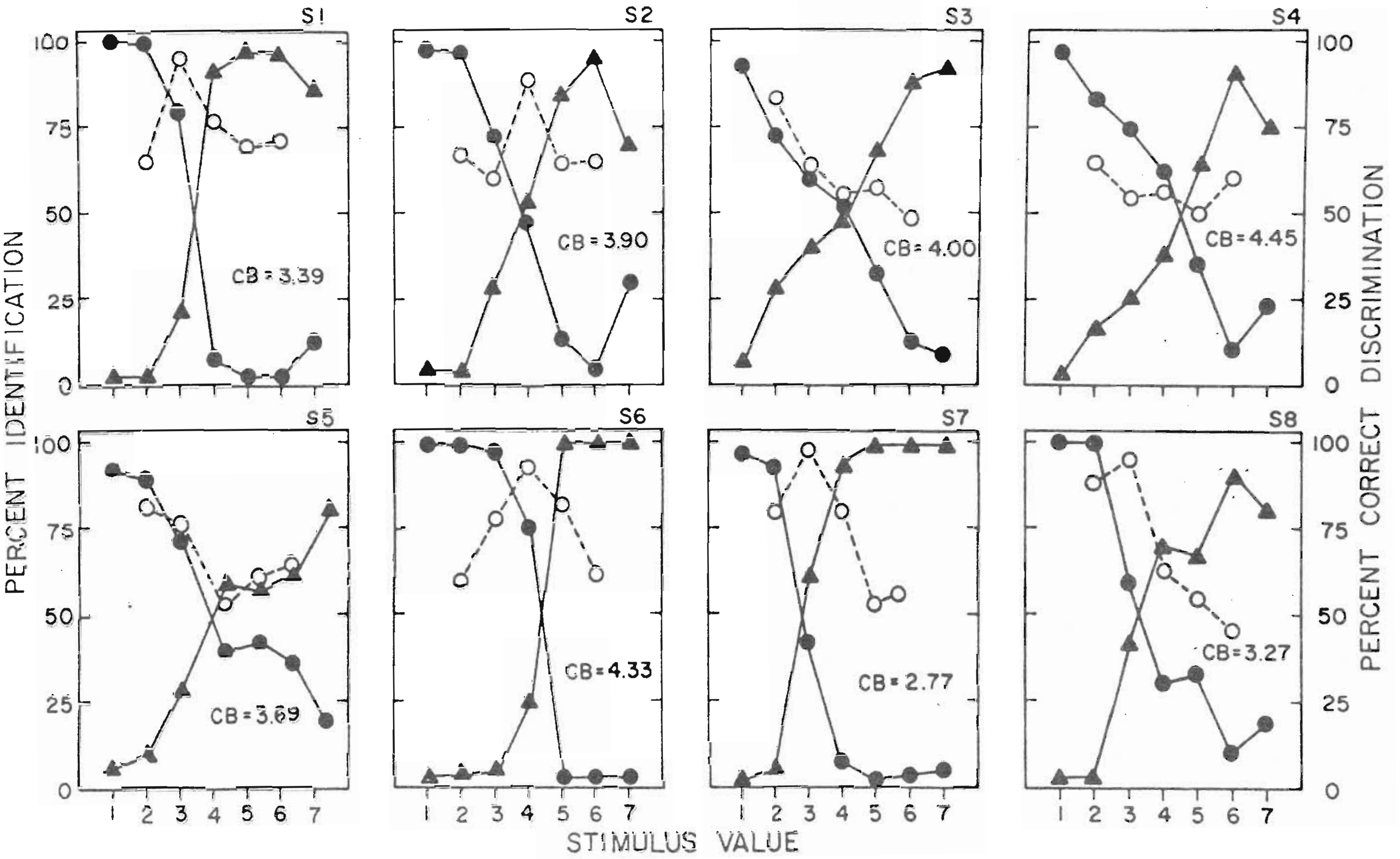


Figure 3.

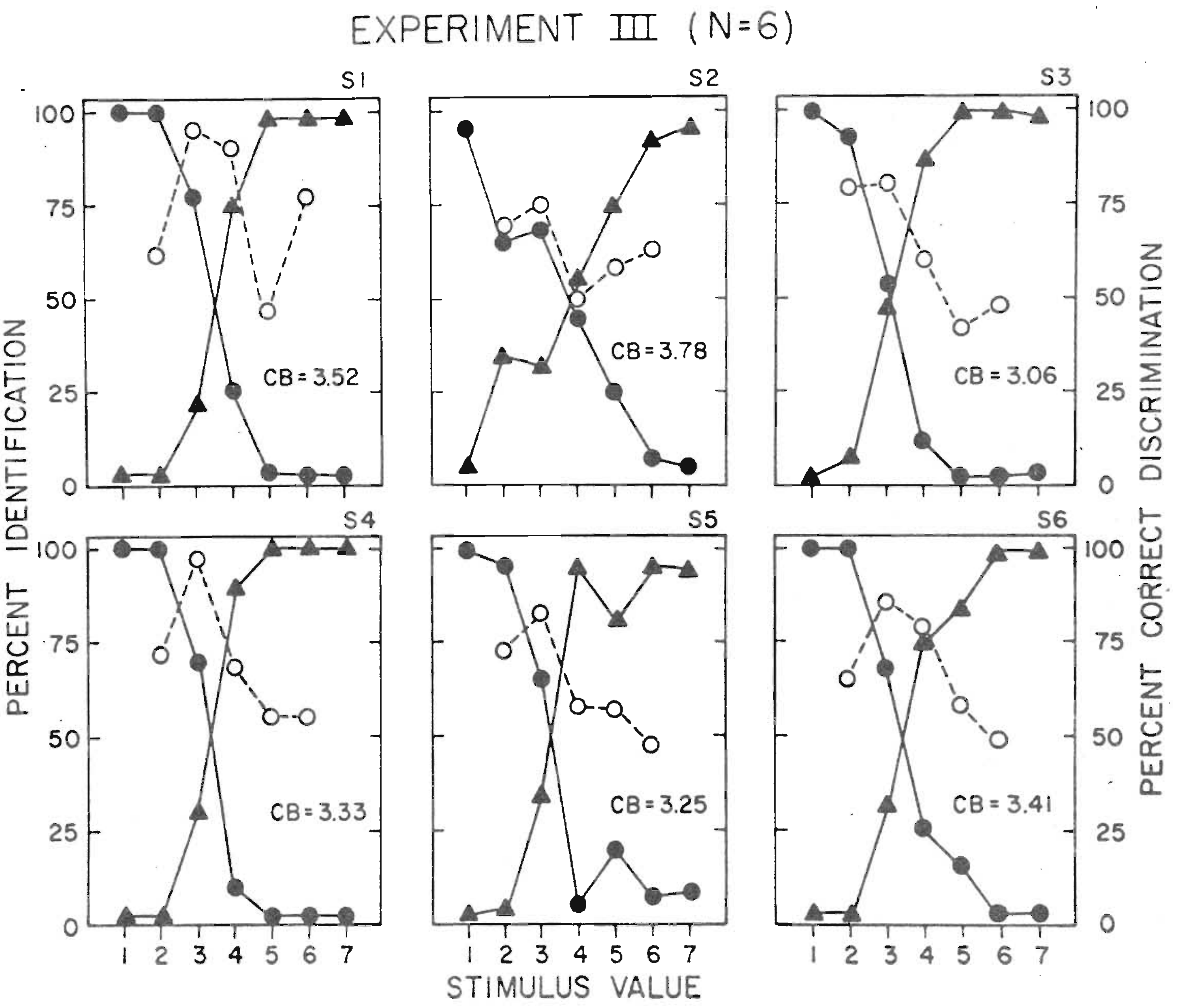

EXPERIMENT III (N=6)
S1
S2
S3
S4
S5
S6
CB = 3.52
CB = 3.78
CB = 3.06
CB = 3.33
CB = 3.25
CB = 3.41
PERCENT IDENTIFICATION
PERCENT CORRECT DISCRIMINATION
0
25
50
75
100
1
2
3
4
5
6
7
STIMULUS VALUE

Figure 4.

SHORT REPORTS AND WORK-IN-PROGRESS

Fundamental Frequency and Perceived Vowel Duration*

David B. Pisoni

Research Laboratory of Electronics

Massachusetts Institute of Technology

Cambridge, Massachusetts 02139

Abstract

What is the relationship between variations in fundamental frequency and perceived vowel duration? At the last meeting of the Society, Professor Lehiste presented some data dealing with this question. Basically, she found that when listeners are asked to judge the duration of pairs of vowels of equal duration they typically judge the first member of the pair as longer than the second. However, when the second member of the pair contained some variation in F_o, listeners perceived it as longer than the first stimulus. This paper presents the results of an experiment that bears on the general questions raised by Professor Lehiste's findings. In the present study, subjects were presented with all possible pairs of synthetic vowels which varied in duration (160, 200, 240 msec) and fundamental frequency (falling: 200-90 Hz; level: 145 Hz; or rising: 90-200 Hz). They were asked to judge which vowel was longer. The results showed that, in general, vowels containing either a rise or a fall in fundamental frequency were perceived as longer than the corresponding vowels of the order of presentation of the pair of vowels. The results are discussed in terms of the interaction between segmental and suprasegmental cues in speech perception.

* This paper was presented at the 91st Meeting of the Acoustical Society of America, Washington, D.C., April 7, 1976. The research was supported by PHS grants NS-12179-01 to Indiana University and NS-07040-01 to the Research Laboratory of Electronics, M.I.T. This work was carried out while the author was on leave at M.I.T.

Fundamental Frequency and Perceived Vowel Duration

Much of the past research in speech perception has been concerned with identifying the acoustic cues or attributes that underlie various segmental distinctions. For the most part, little interest has been focused on the temporal features of speech such as duration, rate of change or variations in fundamental frequency. Moreover, there has been very little interest in the way these variables influence the processing of segmental information. Fortunately this situation has changed somewhat in the past few years. There are now a number of studies that have found substantial effects of suprasegmental variations on the perception of segmental features. Obviously one major reason for interest in the relation between suprasegmental attributes on the one hand, and the processing of segmental features on the other, is the fact that speech is a multi-dimensional signal containing numerous complex changes that vary as a function of time. In order to gain a greater understanding of the perceptual process it is necessary to study as many aspects of the signal as possible that may be employed by the listener in his decision process.

At the last meeting of the Society, Professor Lehiste presented the results of a brief but interesting experiment that examined the effects of changes in F_0 on the perception of vowel duration. She reported that a rising-falling pattern (i.e., a "cap" pattern) or a falling-rising pattern (i.e., a "cup" pattern) is perceived to be longer than a level F_0 pattern. She also found a large "time-order" error in the judgments that interacted with the effects of F_0. As a result, she could arrive at only a tentative conclusion about the effects of F_0 on judgments of duration. In the present study, we examined the effects of only a single change in F_0 either a rise or a fall, on the perception of vowel duration.

The stimuli used in this experiment consisted of synthetically produced vowels that varied in both F_0 and overall duration. Figure 1 shows the arrangement of the stimuli and the variables that were manipulated in this study.

Insert Figure 1 about here

On each trial a subject was presented with a pair of vowels and was required to determine which vowel was longer, the first or the second. The stimuli within a pair were separated by 250 msec and both had identical formant frequencies appropriate for a schewa-like vowel (F1 = 500 Hz, F2 = 1500 Hz, and F3 = 2500 Hz). As shown in this figure, each vowel in the pair could independently take on any one of three durations (i.e., 160, 200, 240 msec) and any one of three F_0 contours (i.e., level, falling or rising). Eighty-one different stimulus pairs were obtained by combining all possible values of each of the three variables for each stimulus position in the pair.

Pairs of stimuli were presented to subjects through headphones in real-time under the control of a PDP-11 computer. The computer arranged the timing sequences and recorded the subjects responses automatically. Six subjects were run simultaneously in parallel. A total of thirty undergraduate students at Indiana University served as subjects in sessions that lasted about an hour each. In a given session, five different randomizations of the basic 81 trials were presented for judgment.

The results are shown in Figure 2. The dependent measure employed here is the probability that the second stimulus in the pair is judged longer. This is plotted on the ordinate in Figure 2. The panel on the left shows the effects of variations in duration and F_0 contour on the first stimulus whereas the panel on the right shows the effects of these variables on the second stimulus. Let us first consider the effect of stimulus duration. From an examination of Stimulus 1, on the left, it can be seen very clearly that as the duration of the first stimulus increases from 160 to 200 to 240 msec, the probability of judging the second stimulus as longer decreases. This effect holds for each of the F_0 contours taken separately. Precisely the opposite result is obtained when the duration of the second stimulus is increased in the same way. Now the probability that the second stimulus is longer increases as its duration increases. Both findings are, of course, not too surprising. However, consider the effects of changes in F_0 contour on the two stimuli. In each case, the rising F_0, as shown by the filled circles in this figure, is interpreted as a lengthening cue by the subjects. In the case of stimulus 1, a rising F_0 pattern shifts the judgments towards lower probabilities relative to the falling and level patterns which are shown by the open squares and circles respectively. This finding indicates that the first stimulus is judged longer than the second stimulus in the pair when there is a change in F_0. The opposite result is obtained for the same variations in stimulus 2 as shown in the right-hand panel. A rising or falling F_0 pattern shifts the

----

Insert Figure 2 about here

----

judgments towards increasingly higher values indicating that the second stimulus is perceived as longer than the first.

A careful examination of the data shown in both panels of this slide also indicates that within the context of these experimental conditions, there appears to be a trade-off relation between stimulus duration and F_0 contour. For example, as shown in the left panel, the duration of a vowel can be effectively reduced by about 1/3 from 240 msec to 160 msec and still be judged to be about equal to a level pattern if the F_0 contour of the shorter vowel is rising. This same relation also appears to be roughly comparable for the data given in the panel on the right of this figure.

A rough idea of the magnitude and direction of these effects as well as the size of the time order error can be seen in Figure 3. Here we plot on

---

Insert Figure 3 about here

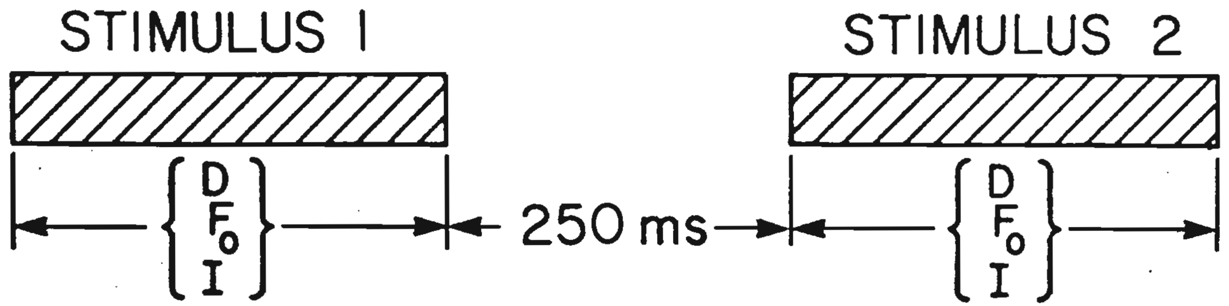
---

the ordinate the mean difference between the obtained values and .50 which we can take to be the equi-probable value assuming no time-order error in judgment. Positive values above the zero line in this figure indicate that the second stimulus is judged longer; negative values below the zero line indicate that the first stimulus is judged longer.

Note first that most of the data fall in the lower half of the figure below zero indicating an overall bias to judge the first stimulus as longer--this is the time-order error. But also notice how the judgments shift systematically as the F_0 patterns change from level to falling to rising in each panel. In the case of stimulus 1, there is a progressive shift in the judgments from positive values to negative values as duration is changed within each F_0 contour. The direction of the shift is just the reverse for stimulus 2. The judgments also change progressively but this time they move from negative values to positive values as the F_0 contour varies from level to falling to rising. Thus, although there is a time-order error pressed in the data, the influence of the two variables on judgments of duration can be clearly observed in the present experiment.

The overall results of the present study, therefore, indicate that changes in F_0 are sufficient to influence the judgment of vowel duration in quite systematic ways. One general conclusion that can be drawn from these findings is that there would appear to be a rather intimate relationship between changes in F_0 and the segmental attributes associated with vowel duration. These types of relations may play an important role in linking the level of segmental features with the higher levels of syntactic analysis. For example, the fact that vowel lengthening appears to be, in part, syntactically determined as shown by the work of Klatt and Cooper, would lead one to expect some relation between F_0 and constituent structure. Indeed, precisely such findings have been reported by Wayne Lea in some of his analysis work related to speech understanding. Since the human listener also makes use of this sort of information in rather subtle ways, it may well be the case that temporal features of speech such as duration, rate of change, and differences in F_0 play a much greater role in the perception of segmental attributes than has been assumed in the past. Obviously, additional experimental work is needed on this neglected and potentially important topic in speech perception.

To summarize, we have found that, in general, vowels containing either a falling or rising pitch contour are perceived as longer than vowels of the same physical duration with level F_0 contours. These findings were obtained regardless of the order of presentation of the stimuli in a pair thus replicating and then extending the findings previously reported by Professor Lehiste at the last meeting of the Society.



(a) DURATION : (1) 160 ms
(2) 200 ms
(3) 240 ms

(b) F_0 : (1) LEVEL : 145 Hz
(2) FALLING : 200-90 Hz
(3) RISING: 90-200 Hz

(c) ΔI : (1) 0 dB
(2) -3 dB
(3) -6 dB

Figure 1.

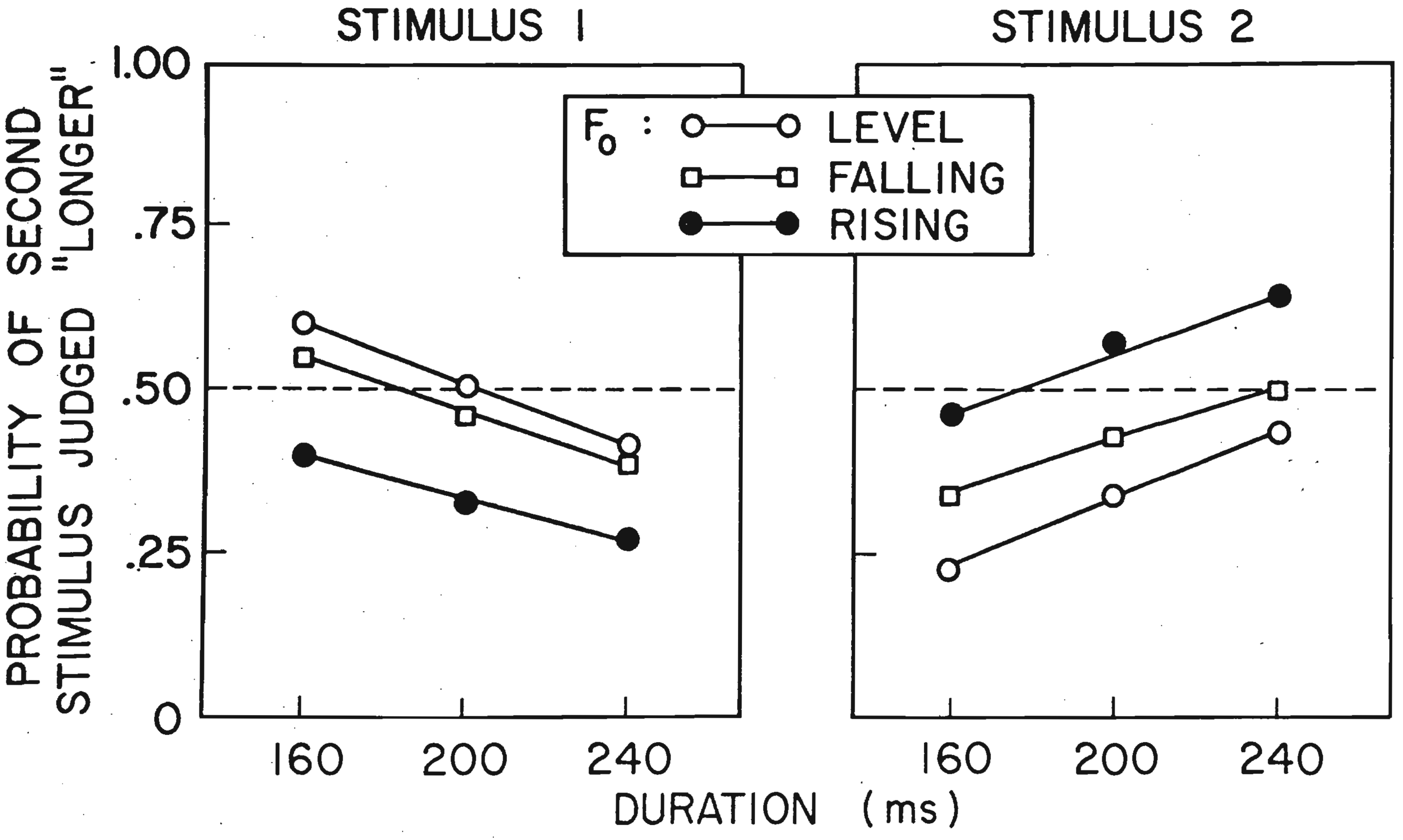

STIMULUS 1
STIMULUS 2
PROBABILITY OF SECOND STIMULUS JUDGED "LONGER"
1.00
.75
.50
.25
0
160
200
240
160
200
240
DURATION (ms)
F_0 :
○—○ LEVEL
□—□ FALLING
●—● RISING

Figure 2.

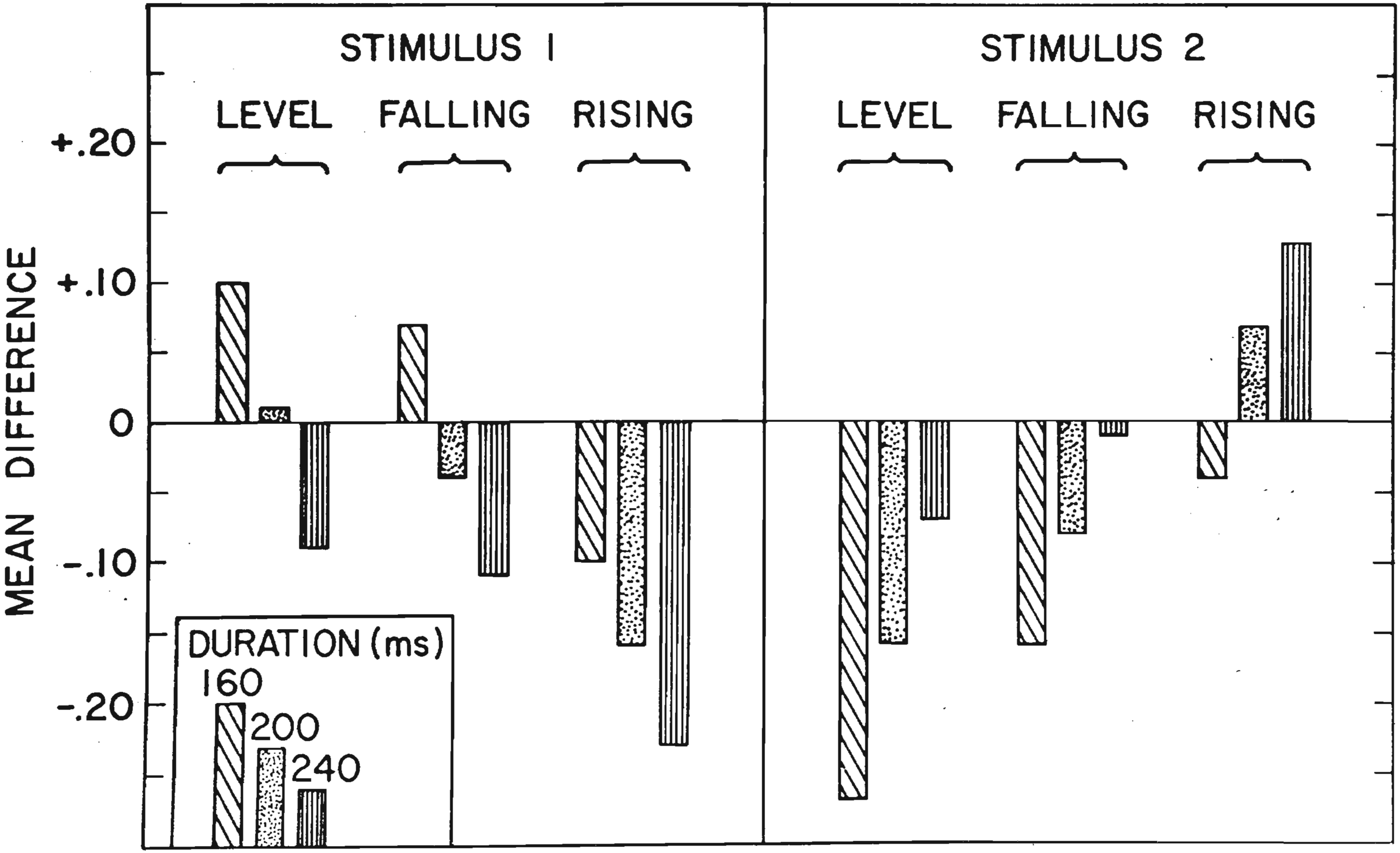

STIMULUS 1
STIMULUS 2
LEVEL
FALLING
RISING
LEVEL
FALLING
RISING
MEAN DIFFERENCE
+.20
+.10
0
-.10
-.20
DURATION (ms)
160
200
240

Figure 3.

Discrimination of Brief Frequency Glissandos*

David B. Pisoni

Indiana University

Bloomington, Indiana 47401

Abstract

The results of a brief experiment on the discrimination of frequency glissandos having properties similar to formant transitions in stop consonants are reported. A stimulus continuum containing variations in the duration and extent of the onset frequency of two tones was generated and presented to three observers in an ABX format. Discrimination functions were obtained for comparisons selected along the continuum in the tradition of previous speech perception studies. The results showed marked discontinuities analogous to those found with speech sounds varying in the acoustic cues underlying place of articulation in stop consonants. Some implications of the role of stimulus context in speech-nonspeech comparisons is discussed.

* This work was carried out while the author was on leave at the Research Laboratory of Electronics, M.I.T. during 1975-1976. Special thanks are due Dr. Dennis H. Klatt for his interest and advice on this project and his help in generating the signals at M.I.T.

Discrimination of Brief Frequency Glissandos

David B. Pisoni

Indiana University

For a number of years, primarily because of their interest in distinctive feature theory and its emphasis on relational invariants, Stevens (1967, 1971, 1973) and Fant (1960, 1973) have sought to find acoustic invariants for distinctive features. The basic argument of their approach is that well-defined acoustic correlates can be found for the principal places of articulation for consonants, particularly the stops, and that previous research with synthetic speech has tended to obscure some of the important cues found in natural speech. According to Fant and Stevens, these invariant acoustic attributes are contained in the rapid changes in spectral energy that occur during the first 10-30 msec after the release of the consonant. Both investigators emphasize that burst and formant transitions, which previously have been assumed to be independent cues, should now be regarded as a single integrated stimulus event or cue. Thus, burst and transitions may be thought of as an overall spectral change at onset.

Based primarily on acoustic analyses of CV syllables, Fant (1969) has claimed invariant spectral relations for labials, post-dentals and velars. For labials, spectral energy is weak and spread with a major concentration at low frequencies; for post-dentals, spectral energy is stronger and spread although the major concentration occurs at high frequencies; for velars, the spectral energy is compact and concentrated at mid-frequencies.

Focusing on the rapid spectrum changes that accompany the release of a stop consonant, Stevens (1967, 1973) has also argued for invariant spectral patterns for place of articulation. According to Stevens, labials can be characterized by low frequency onsets followed by upward or rising changes in spectral energy whereas post-dentals have high frequency onsets followed by downward or falling changes in spectral energy. Velars have spectral energy concentrated narrowly in the mid-frequency range at onset followed by spreading of energy to frequencies above and below this region. Some examples of these idealized spectral relations are shown schematically in Figure 1. Stevens (1975) believes that these rapid spectrum changes at onset "identify features of place of articulation without reference to acoustic events remote from this point in time" and that "these cues are absolute properties of the speech signal and are context-independent" (p. 319).

Insert Figure 1 about here

More recently, however, Stevens and Blumstein (1976) have modified this earlier account somewhat and now claim that the invariant properties for stops in initial position involve simply the location and diffuseness of spectral energy at stimulus onset. This position is now closer to Fant's views summarized above. The main direction of both Stevens' and Fant's position has been to focus on somewhat more complex integrated acoustic attributes of consonants rather than simple isolated cues. By following this strategy it is assumed that an integrated acoustic pattern will show some form of invariance when each of its components fails to do so when considered separately in isolation.

The strategy of isolating a particular acoustic cue in synthetic speech has been a common practice in a number of recent perceptual experiments (Liberman, 1970; Mattingly, Liberman, Syrdal, & Halwes, 1971; Miyawaki, Strange, Verbrugge, Liberman, Jenkins & Fujimura, 1975; Eimas, 1974, 1975). These experiments have compared the perception of an acoustic cue in speech context to perception of the same acoustic cue when it was presented in isolation, removed from its original speech context. The general finding of these studies was that there are substantial differences in perception between the two context conditions. In the speech context, the acoustic cue is typically perceived categorically whereas in isolation discrimination performance is substantially lower overall and the typical peaks in the discrimination function are absent. The explanations typically invoked to account for these differences have concentrated primarily on differences between speech and nonspeech modes of perception and the need for specialized perceptual analyzers for speech signals. As far as we know, little attention has been paid to the possibility that the differences might be due to the shift in stimulus context, particularly in the case where the acoustic cues for a particular phonetic distinction have been isolated from the other frequency components in the original short-term spectrum. The present study was carried out in the tradition of the previous studies comparing speech and nonspeech signals. Our aim was simply to determine whether presumed continuous variations in frequency glissandos much like those found in formant transitions would show discontinuities in discrimination. If this result is obtained, it would suggest that the discontinuities found in discrimination for the acoustic cues in speech context may well be due to the presence of well-defined acoustic attributes in the signal rather than a shift in mode of processing. It would also provide some basis for explaining why the isolated cues appear to be perceived so differently from the cues in context.

Method

Stimuli

To avoid some of the potential difficulties of using natural or synthetic speech signals which elicit strong linguistic biases, we generated a set of thirteen very brief nonspeech stimuli that resemble the formant transitions in stop consonants (see Shattuck & Klatt, 1976; Klatt & Shattuck-Hufnagel, 1976). Each stimulus consisted of three sinusoids of rapidly changing frequency. All stimuli had an overall duration of 50 msec. Figure 2 shows schematized examples of Stimuli 1, 7 and 13.

Insert Figure 2 about here

The lowest tone, T1, remained constant throughout the series. Its onset frequency began at 350 Hz and varied linearily to 750 Hz over its total duration. The starting point of the onset frequencies of T2 and T3 was the experimental variable and followed trajectories such as those shown in Figure 3. The endpoint of T2 was 1695 Hz whereas the endpoint of T3 was

----

Insert Figure 3 about here

3195 Hz. The frequency of T2 and T3 varied linearly over the duration of the stimulus. T1 and T2 were of equal amplitude, T3 was set 6 dB lower.

The thirteen test signals were generated digitally on the PDP-9 computer at the Research Laboratory of Electronics, M.I.T. and then recorded on audio tape. The stimuli were later played back and digitized on the PDP-11/10 in the Psychology Department at Indiana University where the present experiment was carried out.

Subjects

Three subjects were used in the present experiment. All were experienced listeners associated with the Psychology Department and were familiar with perceptual tests of this kind. None of the subjects had any hearing disorder. One subject was the present experimenter.

Procedure

Discrimination data were collected by means of an ABX procedure. Each of the eleven two-step comparisons along the stimulus continuum was presented sixteen times in a random order for a total of 176 trials. Stimuli in a trial were separated by 500 msec. Trials were self-paced to the slowest subject in the group. The stimuli were presented in real-time under computer control via TDH-39 headphones at a comfortable listening level of approximately 80 dB SPL. Immediate feedback was provided after each trial. Subjects ran simultaneously in a small group in individual listening booths each of which was equipped with response buttons and feedback lights. The experiment took approximately twenty minutes.

Results and Discussion

The ABX discrimination data are shown in Figure 4 for each subject

----

Insert Figure 4 about here

----

separately. As can be seen, there are clear discontinuities in discrimination performance as a function of the test comparison along the stimulus continuum. Both S1 and S2 show evidence for peaks in their discrimination functions at roughly comparable stimulus values. S3 shows a gradual increase in discrimination in the center of the continuum although his performance on several comparisons is at ceiling thus precluding any firm conclusions about the possible shape of his discrimination function.

The results of this brief study support the idea that the discontinuities found in comparable speech stimuli varying in the direction and extent of the F2 and F3 transitions might be a consequence of some complex interaction between the frequency components of the stimulus complex. While other work on this problem is currently underway in our laboratory, it is clear, at least from these findings, that the context in which an acoustic cue occurs does exert some combined effect on perceptual analysis. Precisely what these effects are remains to be determined in future work. It may well be that these interactions contribute collectively to processing of rapid spectrum changes at stimulus onset in stop consonants as suggested by Stevens (1975).

Studies which have selectively isolated these acoustic cues may have modified the context so substantially that the perceptual results no longer reflect the processing of the same short-term spectral composition of the original speech signals after which they were modeled. The results obtained in perception between speech and nonspeech signals in studies of both adults and prelinguistic infants might, therefore, have a principled explanation in the acoustic structure and subsequent psychophysical properties of the test signals themselves. Invariant properties or attributes for stop consonants in CV syllables may be contained in the relations observed between components of the stimulus rather than in terms of absolute context independent properties per se.

References

Eimas, P. D. Auditory and linguistic processing of cues for place of articulation by infants. Perception & Psychophysics, 1974, 16, 513-521.

Eimas, P. D. Auditory and phonetic coding of the cues for speech: Discrimination of the r-l distinction by young infants. Perception & Psychophysics, 1975, 18, 341-347.

Fant, G. Acoustic theory of speech production. The Hague: Mouton, 1960.

Fant, C. G. M. Stops in CV-syllables. Speech Transmission Laboratory Quarterly Progress and Status Report No. 4, 1969. pp. 1-25.

Fant, C. G. M. Speech sounds and features. Cambridge: M.I.T. Press, 1973.

Klatt, D. H. & Shattuck-Hufnagel, S. R. Perceptual importance of the second formant during rapid spectrum changes. RLE Progress Report No. 117, M.I.T., Cambridge, (1976), pp. 291-304.

Liberman, A. M. Some characteristics of perception in the speech mode. In D. A. Hamburg (Ed.), Perception and its disorders, Proceedings of A. R. N. M. D. Baltimore: Williams and Wilkins Co., 1970. Pp. 238-254.

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

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J. & Fujimura, O. An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. Perception & Psychophysics, 1975, 18, 331-340.

Shattuck, S. R. & Klatt, D. H. The perceptual similarity of mirror-image acoustic patterns in speech. Perception & Psychophysics, 1976, 20, 6, 470-474.

Stevens, K. N. Acoustic correlates of certain consonantal features. Paper presented at the Conference on Speech Communication and Processing, M.I.T., Cambridge, Mass., November 6-8, 1967.

Stevens, K. N. Airflow and turbulence noise for fricative and stop consonants: Static considerations. Journal of the Acoustical Society of America, 1971, 50, 1180-1192.

Stevens, K. N. Further theoretical and experimental bases for quantal places of articulation for consonants. Quarterly Progress Report No. 108, Research Laboratory of Electronics, M.I.T., 1973, 247-252.

Stevens, K. N. The potential role of property detectors in the perception of consonants. In G. Fant & M. A. A. Tatham (Eds.) Auditory Analysis and Perception of Speech. New York: Academic Press, 1975, pp. 303-330.

Stevens, K. N. & Blumstein, S. E. Context-independent properties for place of articulation in stop consonants. Paper presented at the 91st meeting of the Acoustical Society of America, Washington, D. C., April, 1976.

Figure Captions

Figure 1. Idealized invariant patterns for labial, post-dental and velar stop consonants in prevocalic position. (Adapted from Stevens, 1975).

Figure 2. Schematized spectrograms of three nonspeech glissando stimuli.

Figure 3. Trajectories of the starting frequencies for T2 and T3 for each stimulus value along the continuum.

Figure 4. ABX discrimination functions for three observers.

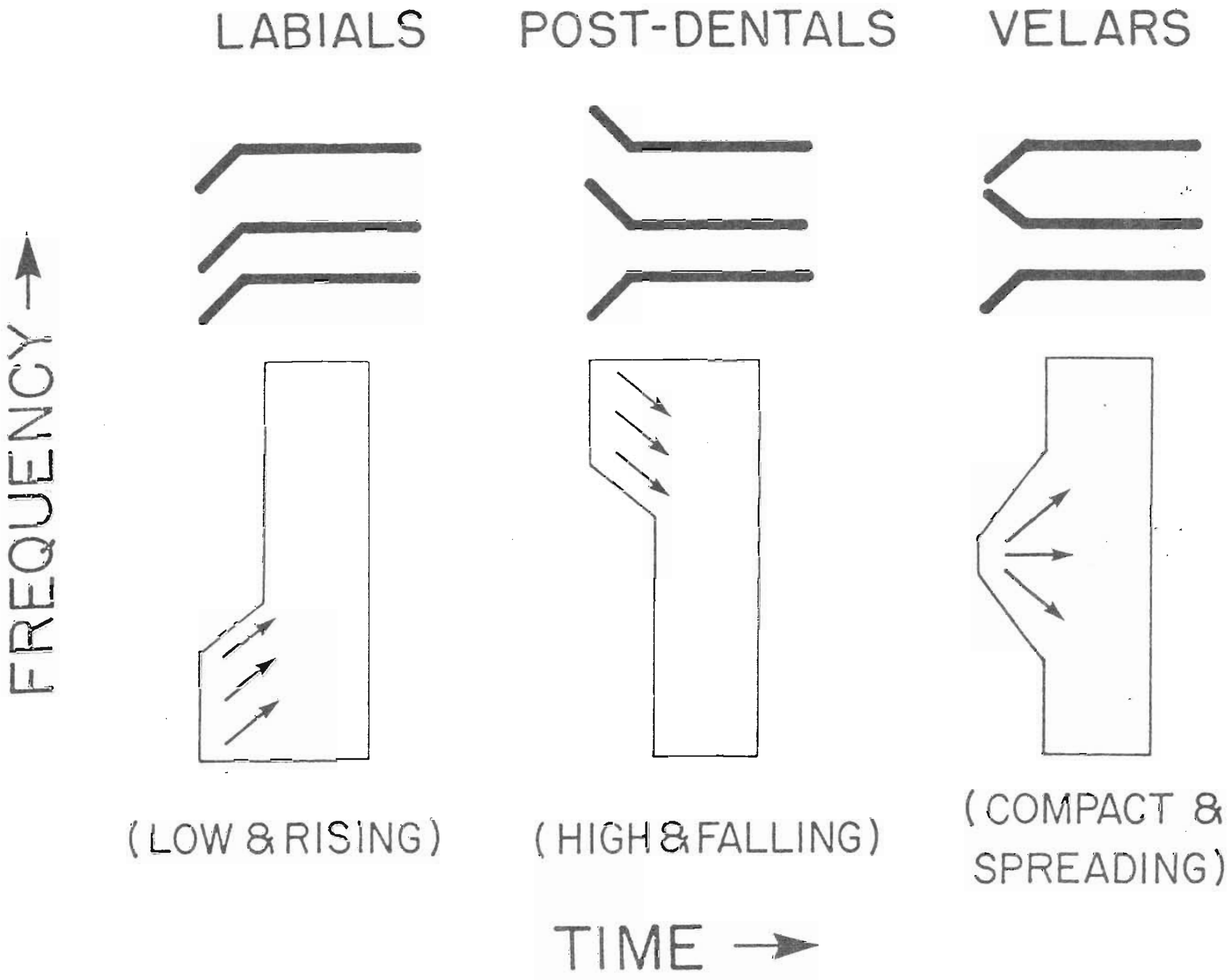

LABIALS
POST-DENTALS
VELARS
FREQUENCY →
(LOW & RISING)
(HIGH & FALLING)
(COMPACT & SPREADING)
TIME →

Figure 1.

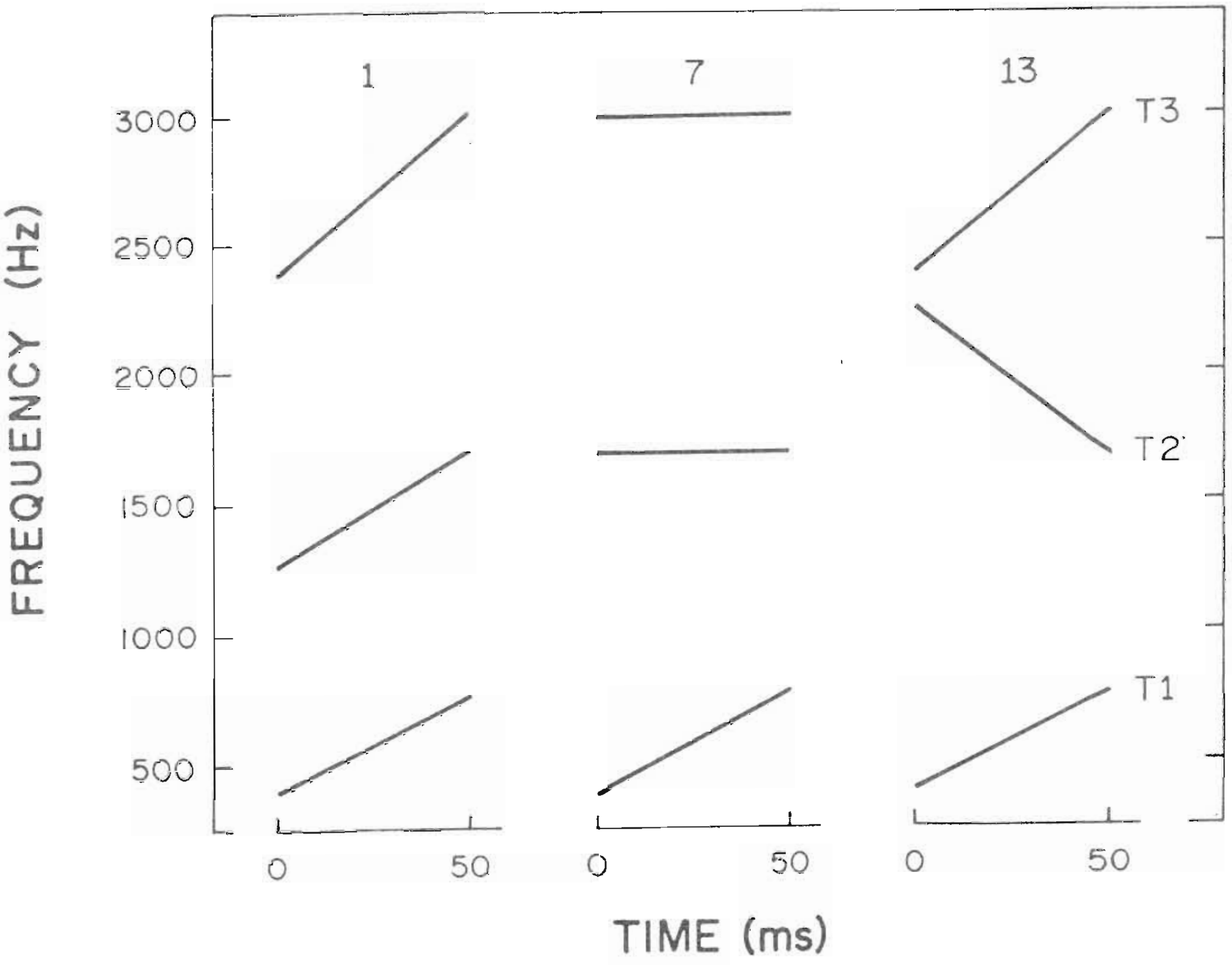

1
7
13
3000
2500
2000
1500
1000
500
FREQUENCY (Hz)
T3
T2
T1
0
50
0
50
0
50
TIME (ms)

Figure 2.

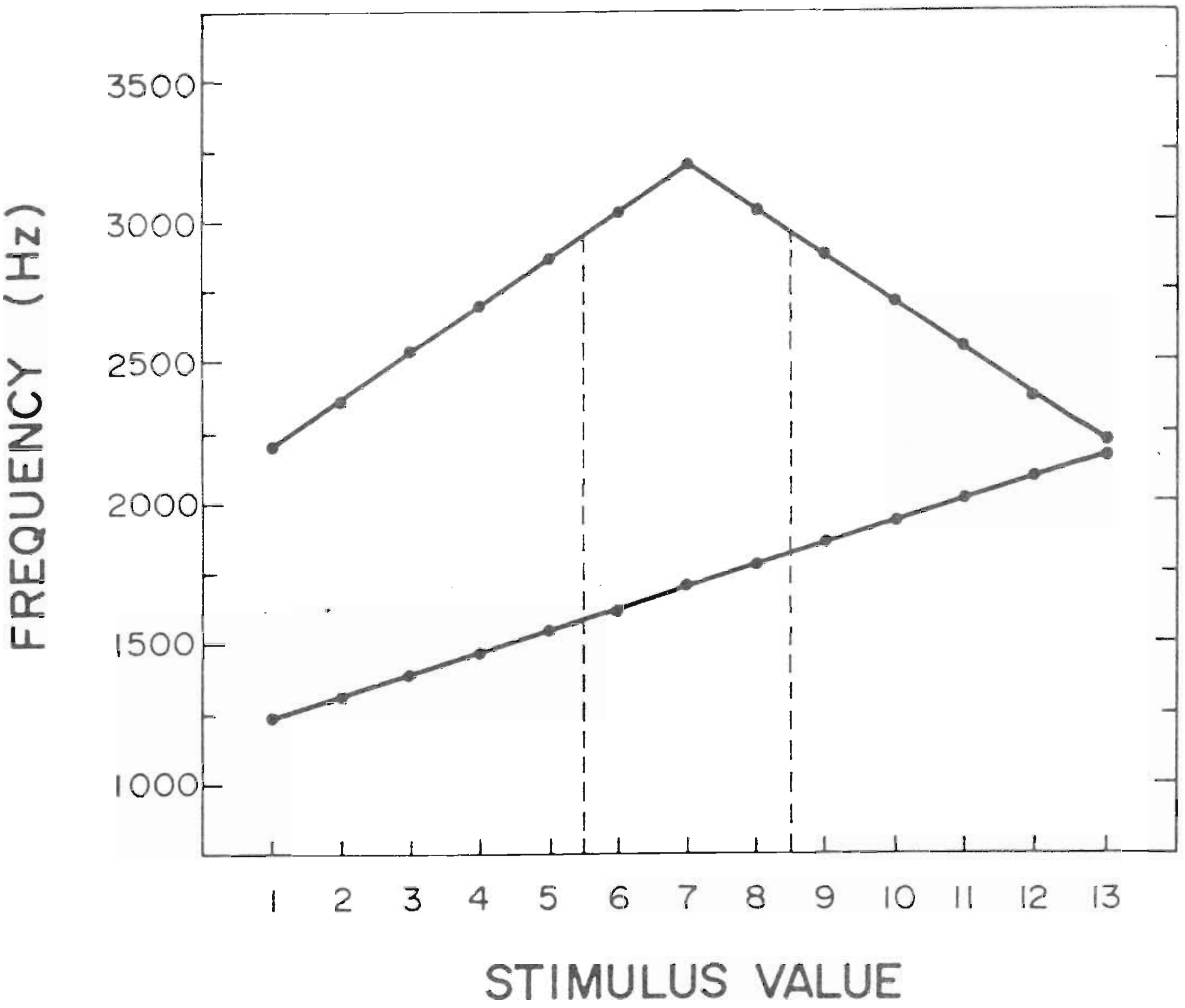

3500
3000
2500
2000
1500
1000
FREQUENCY (Hz)
1
2
3
4
5
6
7
8
9
10
11
12
13
STIMULUS VALUE

Figure 3.

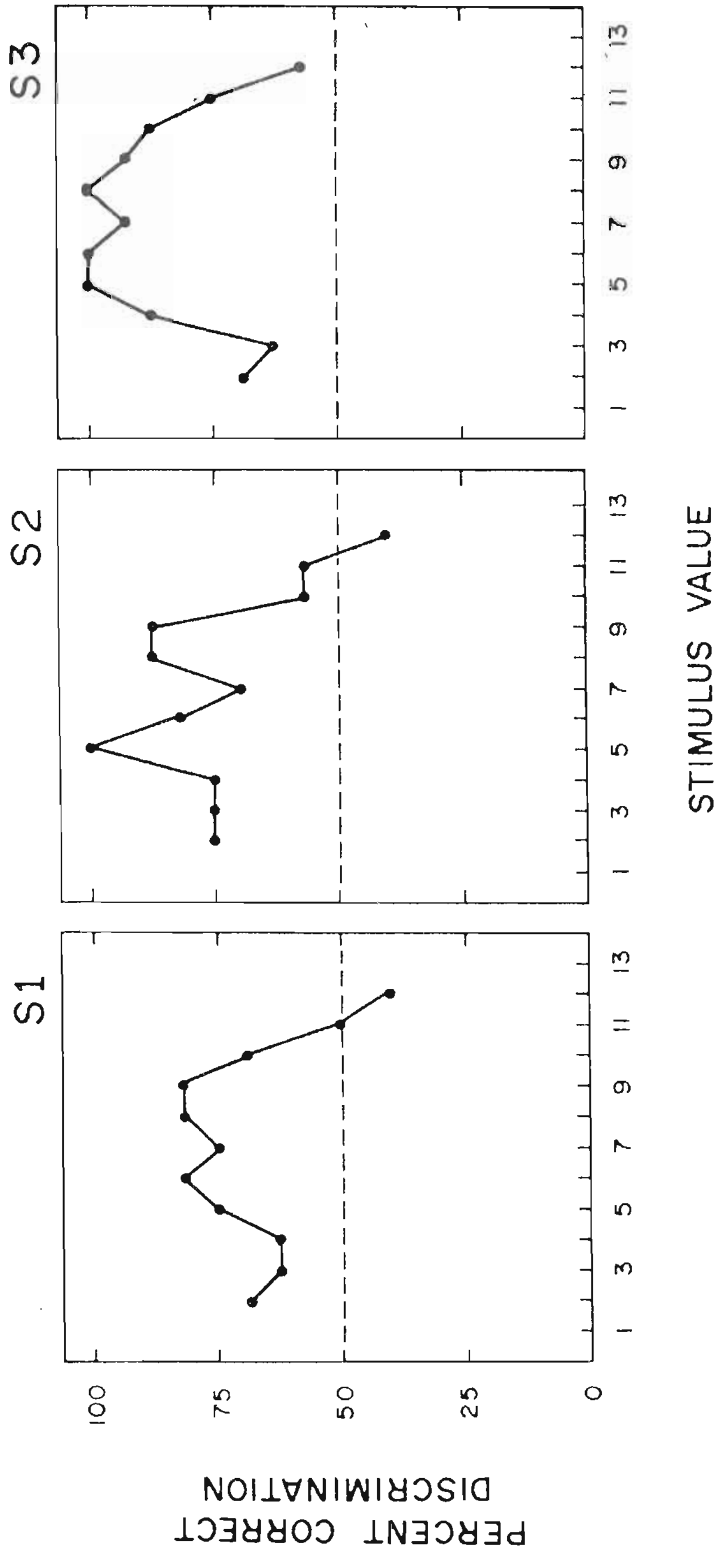

S1
S2
S3
PERCENT CORRECT DISCRIMINATION
100
75
50
25
0
1
3
5
7
9
11
13
STIMULUS VALUE

Figure 4.

A Head-turning Paradigm for Research on Speech Perception in Infants

Richard N. Aslin and David B. Pisoni

Indiana University

An operant head-turning procedure to study speech discrimination in infants is described and an overview of the apparatus is provided in this report. The results of a pilot investigation using the apparatus and methodology with 5 and 6 month old infants are reported and discussed briefly. Some comments are offered about the implications of these results for future work using behavioral measures for discrimination and generalization of speech and speech-like signals in young infants.

In the past few years, three basic methodological procedures have been used to study infant speech perception: (1) high amplitude sucking (HAS), (2) heart-rate change (HR), and (3) operant-conditioned head-turning (OHT). The HAS procedure is based on the infant's tendency to increase the rate of non-nutritive sucking when that sucking leads to the presentation of visual or auditory stimulation (Siqueland & DeLucia, 1969). The infant typically increases the rate of sucking from a non-contingent baseline level and then decreases that rate after several minutes presumably due to satiation with the reinforcing or novel properties of the speech stimulus. After this decline in sucking rate some infants (i.e., the experimental group) are presented with a new speech stimulus which is contingent upon sucking whereas the other groups receive no change (control). If the rate of sucking in the experimental group increases after the decline in sucking rate and subsequent stimulus change, then discrimination between the two stimuli can be inferred provided that the post-shift increase is greater than any spontaneous increase in the control condition. This general HAS procedure appears to be particularly suited to testing very young infants (under three months), is clearly sensitive enough to detect small changes in speech stimuli as shown by Eimas (1975, 1976) and Morse (1972, 1974) in their well known work, but unfortunately has the drawback of a relatively large attrition rate (60-70%) and the inability to be used to assess individual infants' discrimination performance. Almost all past studies have been based on between subject comparisons.

The HR procedure is based on an organism's tendency to exhibit a generalized orienting reaction when presented with a novel stimulus (see Graham & Jackson, 1970). One component of the orienting reaction is a phasic heart-rate deceleration which habituates as a stimulus is repeated and therefore presumably becomes less novel (Jeffrey & Cohen, 1971). Studies of infant speech perception using the HR procedure have repeated one speech sound several times followed by presentation of either a new or the same speech sound. Unfortunately, most studies have analyzed all HR changes after the termination of the experiment, thus eliminating control over possible between subject differences in rate of habituation. In addition, motor movements, known to affect HR, are not easily controlled nor have they been typically reported. These two problems have again limited the HR procedure to group data.

The OHT procedure consists of shaping a directional head-turn response toward a visual stimulus, so that eventually the infant is cued to anticipate the presentation of a visual stimulus when a change from one speech sound to another is made. Fodor, Garrett & Brill (1975) and Moore, Wilson & Thompson (1977) have used this procedure employing an animated mechanical toy as the visual stimulus which acts as a reinforcer for a stimulus contingent head-turn response. Both studies placed the speaker delivering the speech sound at the same azimuth as the reinforcer, which was only visible to the infant if a direction-specific head-turn was made after the speech stimulus was changed. The OHT procedure has the advantages of allowing the discrimination ability of individual infants to be assessed in the course of one or two sessions as well as a relatively low attrition rate (10-20%).

In this paper we wish to report the progress we have made in the last six months with a variation of the OHT procedure as well as some preliminary results on discrimination of vowels with a few young infants. This work has convinced us of the advantages of the OHT procedure over some of the other procedures used with infants in the past few years.

Method

Apparatus

The apparatus which we constructed is shown in Figure 1. The

Insert Figure 1 about here

---

infant is positioned in a padded seat facing a visual display panel (a). At eye level is a series of multicolored lights which flash on and off to attract the infants gaze. The rate at which the lights go on and off

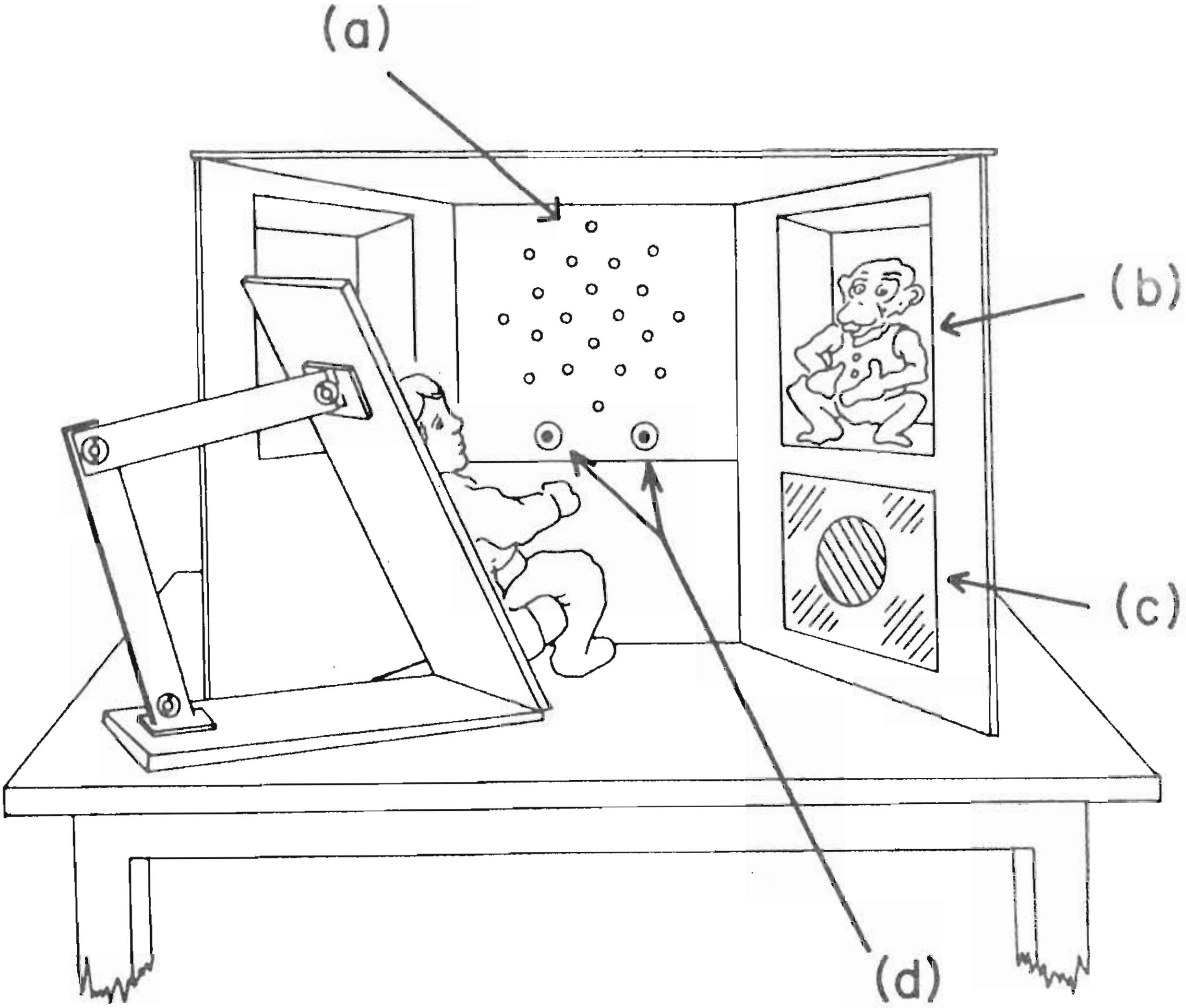


Figure 1. Apparatus for measuring sound discrimination using the operant head-turning procedure. The infant is seated facing a central visual display (a). The visual reinforcer is an animated toy monkey (b) which becomes visible to the infant if a head-turn response occurs contingent upon a sound change delivered through a loudspeaker (c). Two observers are present behind peepholes (d) to record the head-turn response on each trial.

as well as the visual patterns produced by the display can be controlled by the experimenter. Approximately 45° to the right and left of the display panel at eye level is a window-screen (b) (1' square) behind which is located a sound proof box containing a mechanically operated toy monkey which serves as a visual reinforcer. Below this is a loudspeaker (c). The monkey is only visible through the display screen when the two observers, who are positioned behind the central light pattern viewing the infant through peepholes (d), both depress buttons indicating that the infant turned to look right. A chart recorder provides a permanent record of the number and relative timing of the trials and the observer's judgments of the direction of the infant's head-turn response.

Procedure

Thus far we have used the apparatus to study discrimination of two synthetic vowels. Before discrimination testing can be started, however, some initial control over the head-turning response is desirable and as a result we have had to begin each session with some shaping trials. In the data described below, each infant received a variable number of shaping trials ranging from 10-25 trials in a single session.

Shaping consisted of conditioning the infant to anticipate the presence of an animated visual reinforcer (i.e., the toy monkey) at one location (i.e., to the right) with respect to the central orienting visual display. Shaping trials consisted of initially getting the infant to look toward the source of a sound. While the infant was fixating the central visual display, a speech sound (S_1) was used as a discriminative stimulus for the appropriate directional head-turn response. The visual reinforcer was turned on to elicit the appropriate directional head-turn response. As shown below, our pilot testing indicated that this stimulus-response contingency is learned by infants very rapidly, typically within 5-25 shaping trials. If the infant was still cooperative, an initial attempt at discrimination testing was carried out.

If shaping was successful we then moved to the discrimination testing phase of the procedure. After the infant learned to anticipate the appearance of the visual reinforcer by making an appropriate directional head-turn when a speech sound was introduced, a background contrast stimuli was then turned on. This background stimulus (S_2) was presented repeatedly at a relatively low intensity with a 500 msec interstimulus interval and remained on for at least 10 sec. When the infant fixates the central visual display after this 10 sec interval one observer pushes a button which switches channels on the tape recorder (Crown CX822) to present four repetitions of the original shaping stimulus (S_1). The background stimulus was then attenuated 10 dB relative to the discriminative stimulus (S_1) and finally no difference in intensity was present for testing. Additional shaping trials were sometimes necessary for some infants given the added complexity of the second discrimination task relative to the initial no-sound vs. sound discrimination. For some infants, however, we went directly to the discrimination testing phase of the experiment without shaping the head-turn response at all.

Results and Discussion

Table 1 shows the various experimental conditions and results we

---

Insert Table 1 about here

Table 1

Shaping and Discrimination Data for Pilot Subjects

Subject	Age (Months)	Sex	Day	Stimulus Conditions	No. trials of Shaping	No. trials of discrimination testing	Probability Correct Response	Total No. Trials
KL	6.5	F	D1	1. Ø vs /a/	16			
				2. Ø vs /a/		15	1.00	
				3. /i/ vs Ø		5	1.00	36
			D2	1. /i/ vs /a/		16	.70	16
SU	6.25	M	D1	1. Ø vs /a/	14			
				2. Ø vs /a/		10	.90	
				3. /i/ vs /a/		9	.00	
				4. Ø vs /a/		7	1.00	
				5. /i/ vs /a/		6	.17	
				6. Ø vs /a/		5	1.00	37
			D2	1. /a/ vs /i/		23	.00	
				2. Ø vs /i/		8	.00	
				3. Ø vs /a/		12	.00	43
JN	5.5	F	D1	1. /a/ vs /i/	15			15
AJ	5.5	M	D1	1. /a/ vs /i/	6			
				2. /a/ vs /i/		10	.60	
				3. /a/ vs /i/		26	.54	
				4. /a/ vs /i/		3	.66	45
MM	6.0	M	D1	1. Ø vs /i/	11			11
			D2	1. Ø vs /i/	23			
				2. Ø vs /i/		27	.26	23
MC	6.5	F	D1	1. Ø vs /a/	26			
				2. Ø vs /a/		23	.61	49
			D2	1. /i/ vs /a/	5			
				2. /i/ vs /a/		16	.13	21

Ø = Silence
/i/ = Vowel, F1=270 F2=2290 F3=3010
/a/ = Vowel, F1=730 F2=1090 F3=2440

obtained with the last six subjects run in the apparatus shown in Figure 1. These subjects all received some initial form of shaping and were then transferred to the discrimination test phase of the experiment. Two subjects (KL & AJ) showed discrimination of the desired contrast between the two synthetic vowels /i/ and /a/. One subject (JN) could be shaped successfully but did not transfer to the discrimination situation. Another subject (MM) could not be shaped at all even after two separate sessions on successive days. We have no motivated explanation for this failure at present. Two other subjects (SU & MC) showed relatively high levels of discrimination for the silence vs. /a/ condition but failed to transfer to the harder discrimination of the vowels /i/ vs. /a/.

Table 1 also shows the total number of trials carried out during each session which lasted about 20 minutes. This ranged from 11 trials for subject MM to 49 trials for subject MC. The mean number of trials run including shaping was 29.6.

Although these results are based on a small number of subjects we are very encouraged by the fact that five out of six infants could be shaped and of these four out of five showed strong evidence for transfer of the response in the discrimination paradigm. We feel that, in contrast to the other procedures such as heart-rate and high amplitude sucking, the head-turning methodology offers many advantages for future work on infants. Perhaps the most important one is the relatively small dropout rate we have observed with this behavioral technique. Being in a community such as Bloomington, we are very much aware of the small subject population available to us for this work.

We anticipate extending our work with the head-turning response to a generalization paradigm in the next few months in which the infant makes two discrete responses each of which is reinforced appropriately with a visual display of a toy animal. A new apparatus is being constructed which contains three loudspeakers and two toy monkeys. One speaker will be placed in the center of the apparatus under the light display. The other two speakers will each be placed at 45° from center below windows containing the toy animals, one on the left and one on the right.

After shaping the infant to respond to the no-sound vs. sound situation as in our pilot work, the infant will then be trained to turn toward one sound, S_1, when that sound is introduced through the right speaker or to turn in the other direction toward S_2 when it is introduced through the left speaker. Some shaping will probably be needed in this paradigm for the infant to learn that a direction-specific head-turn is needed to gain access to visual reinforcement. After the infant has attained a 75% correct response rate for each of the two stimuli, the final testing phase will begin. This consists of presenting S_1 and S_2 from the centrally located speaker rather than from separate speakers. After the infant has attained a correct response criterion of 75%, generalization testing will begin. Testing will consist of presenting S_1, S_2 as well as other generalization stimuli from the central speaker and then observing the infant's response. In this condition all responses will be reinforced during testing under a 50% schedule of partial reinforcement. The partial reinforcement procedure during testing is used to ensure that the infant continues to make head-turns in both directions. The entire generalization procedure is expected to take from two to three sessions for each infant. A short re-training phase will probably be necessary at the beginning of each follow-up session.

Over the next few months we hope to obtain more precise measures of response latency and head-turning automatically by recording EOGs during the course of an experimental session. We also hope to interface the infant auditory laboratory and associated apparatus to the PDP-11 so that the experiments can be run more efficiently in real-time and under computer control.

In final summary, we feel that substantial progress has been made over the last few months in developing the appropriate behavioral methodology using a head-turning response to study various aspects of infant speech perception, particularly involving discrimination. We plan to continue and expand this line of research in a number of directions involving speech and speech-like signals.

References

Eimas, P. D. Speech perception in early infancy. In L. B. Cohen & P. Salapatek (Eds.) *Infant perception*. New York: Academic Press, 1975.

Eimas, P. D. Developmental aspects of speech perception. In R. Held, H. Leibowitz & H. L. Teuber (Eds.) *Handbook of sensory physiology: Perception*. New York: Springer-Verlag, 1976.

Fodor, J. A., Garrett, M. F. & Brill, S. L. Pi ka pu: The perception of speech sounds by prelinguistic infants. *Perception & Psychophysics*, 1975, *18*, 74-78.

Graham, F. K. & Jackson, J. C. Arousal systems and infant heart rate responses. In H. W. Reese and L. P. Lipsitt (Eds.) *Advances in child development and behavior*, Volume 5. New York: Academic Press, 1970.

Jeffrey, W. E. & Cohen, L. P. Habituation in the human infant. In H. W. Reese (Ed.) *Advances in child development and behavior*, Volume 6. New York: Academic Press, 1971.

Moore, J. M., Wilson, W. R. & Thompson, G. Visual reinforcement of head-turn responses in infants under twelve months of age. *Journal of Speech and Hearing Disorders*, 1977, in press.

Morse, P. A. The discrimination of speech and non-speech stimuli in early infancy. *Journal of Experimental Child Psychology*, 1972, *14*, 477-492.

Morse, P. A. Infant speech perception: A preliminary model and review of the literature. In R. L. Schiefelbusch & L. L. Lloyd (Eds.)

Language perspectives--acquisition, retardation, and intervention. Baltimore: University Park Press, 1974.

Siqueland, E. R. & DeLucia, C. A. Visual reinforcement of non-nutritive sucking in infants. Science, 1969, 165, 1144-1146.

INSTRUMENTATION AND SOFTWARE DEVELOPMENT

Computer Resources in the Speech Perception Laboratory[1]

Jerry C. Forshee

Indiana University

This paper describes the laboratory computer system as it has evolved to date. First, the physical components of the computer systems configuration are presented, indicating the major functional use of each one. Next the major components of the software system are described and finally, some future plans for the computer system are briefly presented.

Computer Configuration

The configuration of the PDP11 computer system began as a packaged OEM (original equipment manufacture) system, the PDP11e/05. This packaged system offered the basic configuration that was needed, and being purchased through an OEM offered the prospect for substantial savings. The original system consisted of the processor (KD11B), with three integral options: 16K of 900 μsec core memory (2 MM11-L's), a line frequency clock (KW11L), and a console terminal port (Serial Communications Line). This packaged system also included several peripheral options: a 30 character/sec hard copy terminal (LA30), a dual drive cassette tape system (TA11) which has a storage capacity of 92K bytes/drive, and a random access moving head disk system (RK11) with a capacity of 1.2 M words/drive. Also included in the system was a ROM bootstrap for the disk (BM792YB). Figure 1

Insert Figure 1 about here

----

graphically presents the computer system configuration which is composed of this initial system and the subsequent expansion.

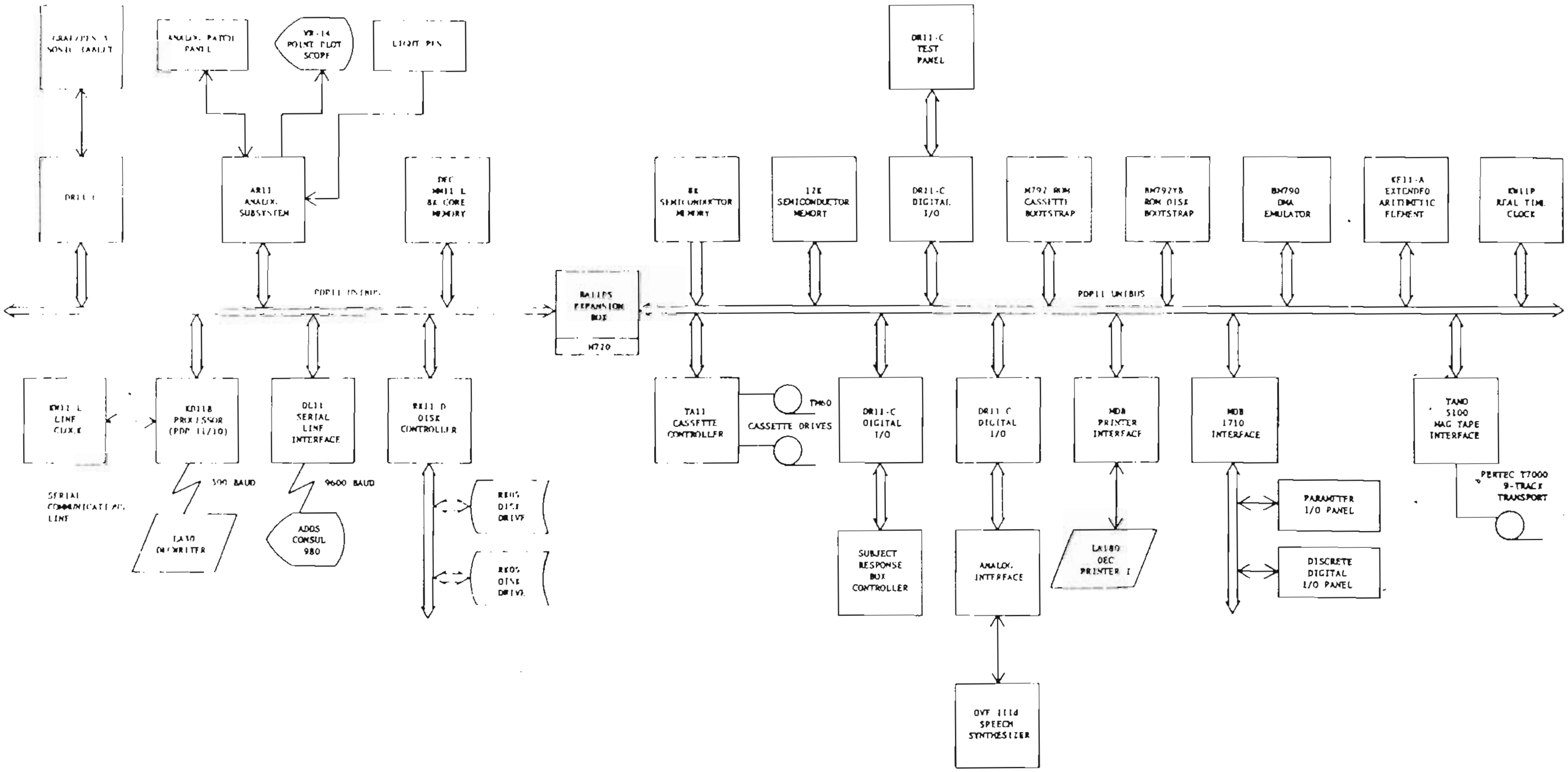


Figure 1. PDP-11 Computer Configuration

The PDP11e/05 was chosen as the starting nucleus of the laboratory computer system because it offered some form of support for each of the essential functions. The RK11 disk system offered the high speed random access secondary storage needed for the operating system (RT-11) and for storing and retrieving speech stimuli. The TA11 provided the low cost off-line storage medium needed for limited system back-up, and for storage of stimulus and response data. Operator interaction with the system for such functions as text editing, program execution, program debugging, system management and maintenance as well as hardcopy output such as program listings, data analysis and simulation output could be supported by the LA30 DECwriter.

As the development of the laboratory progressed and the applications to which the computer system was applied grew, the physical capacity of the original computer system quickly became inadequate. To accommodate the larger peripherals and interface systems, two additional equipment racks were added adjacent to the one which housed the original system. To accommodate the several SPC's (Small Peripheral Controller) a BA11ES expansion box and a H720 power supply were added to the system increasing the available SPC slots by 24.

To increase the throughput of the system, primarily for text editing and secondarily for increasing the speed of all dialogue between the console terminal (CPU) and the user, a high speed CRT terminal (ADDS Consul 980) replaced the LA30 as the console terminal. The new function of the LA30 became a secondary device for slow speed hardcopy output, or for a separate terminal for dialogue with the experimenter from the foreground experiment control program. The addition of an asynchronous serial line interface (DL11) allowed the CRT terminal to be connected to the computer. All user computer dialogue and text editing is currently done on the high speed (9600 baud) CRT terminal. Trial listings from FORTRAN and assembly programs are also done on this terminal. When hardcopy is desired, output can be directed to the serial line printer, the DEC LA180; which has a printing speed of 180 characters/sec.

The large number of speech stimuli used on the system, their large memory size, and the many software components on the system soon made a one disk system very cumbersome. A second RK05 disk drive was added to allow one disk to be used as a system disk which contained the operating system, and all the laboratory and real time support software while the other disk could be used for stimulus and data files for a particular experiment, project or user. This organization is very convenient since all users have access to the entire software system on the system disk (which they use but never modify). Each user is kept separate from all others by having his own experiment disk pack, during the term of his project, on which the stimulus and other required files reside. With each user having his own disk pack for his particular project, the problem of one user's error disrupting another's research project is reduced substantially.

For reascns of both convenience and speed a ROM (M792), which we programmed with a cassette bootstrap, was added to the system. All of the diagnostic programs originally supplied by DEC were delivered to us on cassette tapes, and the addition of a cassette bootstrap helped speed up the procedure whenever it was necessary to run diagnostic programs. With the addition of the second disk drive, all diagnostics were moved to disk storage. The primary use of the cassettes now is for storage of user programs, stimuli and data on a low cost off-line medium.

We found that the size of our real time experimental control programs with the necessary buffers for storing speech waveforms was greater than the available 16K core memory. To alleviate this problem and expand the system to the maximum memory size of 28K, a 12K semiconductor memory system (supplied by Monolithic Systems) was added. Not only has this been a very cost effective solution, but we have found 28K of memory to be sufficient for all of our activities thus far.

Several of the software projects that were begun on the original system taxed the computing power of the PDP11/05 CPU. The real time hardware synthesizer driver program required that many words of data be shifted a variable number of multiple bits. The program to generate complex waveforms consisting of several sinusoids had to perform thousands of sine calculations in generating each stimulus. The software synthesizer programs, data analysis programs, and other simulation programs have all placed heavy demands on the arighmetic processing capabilities of the CPU. To alleviate this problem as much as possible, at a reasonable cost, a KE11-A Extended Arithmetic Element (EAE) was added to the system. The EAE performs five different arithmetic operations: fixed point multiplication and division, logical and arithmetic multiple bit shifts, and normalization.

An industry compatible 9-track magnetic tape system was also added to the system primarily as a secondary bulk storage device and for use during experiments requiring on-line real-time digitization of speech signals. The addition of another peripheral device to support this function was necessary because the length of such a signal, in digital form, would exceed the capacity of the disk system. The addition of this mag tape system has had several very useful secondary functions including high speed local back-up and off-line storage of all types of data (operating system, source programs, stimulus and data files). Also, it provides a compatible medium by which data can be conveniently transferred to the campus computing center. For several current projects large volumes of real-time subject data are taken, via mag tape, to the campus computing center where a large number of "canned" programs and our own programs are used for analysis, simulation and prediction.

The current relatively high data rate, 20 kHz, used in processing speech signals requires that the real time A/D and D/A driver programs be extremely efficient. The lack of equipment funds to allow us to obtain a DMA (direct memory access) interface for stimulus I/O required that direct program I/O techniques be used. This required that the stimulus I/O driver programs be capable of filling or dumping a buffer at the rate of one word every 50 μsec, and never fall behind. These programs were very carefully coded and tightly constructed and were just fast enough; but they did not allow for other real-time processing operations simultaneous with stimulus I/O. Since some experimental paradigms require the ability to record subject response latencies from the beginning of stimulus onset, a new stimulus I/O driver scheme had to be developed. A used ROM bootstrap was purchased and reprogrammed with a short program that emulates DMA operation. Substantial program execution time is saved over the old direct program I/O method as the access time for the ROM memory is 2½ to 3 times faster than regular memory. This means that the action taken by the old direct program I/O driver, which took just under 50 μsec to execute, could be done by the DMA emulation program in the ROM in something under 20 μsec. We expect, however, to add the DMA to the system before long.

The collection of subject response latencies to the resolution of milliseconds necessitated a clock with greater temporal resolution than the KW11L, line frequency clock. A programmable real-time clock (KW11P) was added to the system for this function. In addition to timing the latencies for up to 16 subject stations, this clock is also used to precisely synchronize the presentation of stimuli in dichotic presentation where there is a temporal lag between stimuli presented to the two ears. Several other miscellaneous timing functions are served by this clock such as inter-trial intervals, inter-stimulus intervals, and other program sequencing tasks.

With the addition of an AR11, Analog Real-time Subsystem, which consists of A/D inputs, D/A outputs, and another programmable clock support for these additional functions could be provided. The 16 channel multiplexed A/D inputs, to which the user has access from a patch panel mounted on the computer rack, are currently used for some pilot studies measuring certain physiological reactions by subjects to the presentation of speech stimuli. We also anticipate collecting some physiological data from the subjects used in our infant lab. These A/D inputs can be used as general purpose A/D channels available to the user to implement as he desires. Four of the A/D input channels have been assigned to X-Y joysticks used in the graphics system to position cursors on the display screen. The graphics display, which employs a VR-14 point plot display scope is driven by the two D/A channels of the AR11. A light pen is also used with this system and is connected to the system via the erase return input on the AR11. The third component of the AR11, the programmable clock, is used in conjunction with the KW11P clock in complex experimental paradigms where two events are being timed simultaneously and the onset and offset of the two events are asynchronous.

The graphics system is further enhanced with the use of a digitizer tablet. This device is used to input parameter data for synthesis and analysis programs that use the graphics display mode in interacting with the user. This system allows the computer to read the hardcopy output of such devices as sound spectrographs or polygraphs and is much faster than transcribing such data point-by-point and entering it via a terminal keyboard.

A general purpose interface system has been constructed to support discrete digital I/O devices such as single push buttons, switches, lights, relays, beepers, etc. This interface is built around the MDB Systems 1710 General Purpose Interface foundation module. This interface system connects two panels to the computer system. The Parameter I/O panel allows the user to input to a program several parameter values without stopping the program to read from the terminal keyboard. Program options in effect can also be displayed as well as selected parameter values. The Parameter

I/O panel consists of a LED register of 16 individual LEDs, a switch register consisting of 16 option switches, four display readouts each consisting of four digits of seven-segment display, and four parameter value selectors each consisting of a four-digit thumbwheel switch which allows values from 0 to 9999 to be selected. The Discrete Digital I/O panel provides a number of single point input and output bits. The panel consists of DPDT outputs from a bank of 16 relays, 16 output bits suitable for driving devices of moderate current load ($\leq$ 330 ma) such as small relays, indicator lamps, etc., 16 interrupting push button inputs, and 16 non-interrupting inputs for TTL or switch closure.

Three parallel digital I/O modules (DR11-Cs), each consisting of 16 bits of digital input and output plus several control and status bits, were added to support the special purpose hardware that we developed. One module was added to connect the PDP11's Unibus with the subject Response Box Controller (RBC). The RBC serves to interface the human subjects to the computer. It provides two essential services. First, it provides output of cue and feedback information to subjects. The RBC has the capacity to control from one to 16 subject stations. Each of the several stations may be equipped, for output, with from one to 16 discrete output event lights and, for input, with from one to 16 discrete pushbutton switches or an ASCII keyboard. The response manipulanda option can be varied easily by choosing one of several different response boxes. Each of the several different boxes is designed with a particular experimental paradigm in mind, i.e., two button for two choice discrimination, six button for confidence rating, ASCII keyboard for recall, etc. A more detailed discussion of this interface system appeared in our last Progress Report (see Forshee, 1975).

A second DR11-C was added to the system to support the Analog Subsystem Controller, which manages the speech processing functions on the system. The analog interface implements several distinct functions combined physically as a single integral unit. The analog Subsystem Controller consists of the following functional units: (a) a two-channel 12-bit digital-to-analog (D/A) converter, (b) a one-channel analog-to-digital (A/D) converter, (c) a buffered interface for the OVE IIId Speech Synthesizer, (d) a two-channel audio mixer and amplifier, and (3) two programmable attenuators. A detailed description of the analog Subsystem Controller was also given in the last Progress Report.

A third DR11-C was added to support a special digital I/O test panel which is used in testing the special purpose hardware and interfaces as they are developed.

Program Supported Research Activities

We have been successful in implementing a full range of activities that allow for the on-line generation of synthetic speech, the digitization of natural and synthetic speech signals, the presentation of these signals in real time to human observers and subsequent collection and analysis of the responses of these observers in a variety of experimental paradigms.

The first step in the process of conducting experiments in speech perception is to generate the experimental signals to be presented to subjects. Currently we have several ways of generating acoustic signals on our system. Synthetic speech can be generated by the hardware OVE IIId speech synthesizer or a software driven speech synthesizer. In both cases elaborate control programs are available for the user to employ in creating parameter files that drive the synthesis process. The waveforms generated by either the OVE IIId or the software synthesizer are, at any instant, the result of parameter values input for each of a dozen or so control parameters. These control programs allow the user to manipulate parameter files, create, modify and examine sets of parameter values used to specify the synthesis output, playout the parameter values and produce a synthetic waveform, and to generate a series of synthetic or natural signals for off-line audio recordings on magnetic tape.

Non-speech "control" signals for a project investigating the characteristics of voice onset time (VOT) can also be generated by software. For this project special stimuli are constructed as a complex sine wave composed of from one to three different input sine waves (see Port, this report, pp.   ). The program that generates these signals interacts with the user and allows him complete control over the several input sine waves, allowing each to vary in frequency, phase, and amplitude. In addition to generating the waveforms, the program executes over the entire stimulus generating process permitting the user to perform other necessary functions, such as: inputting old stimulus files, creation and deletion of stimulus files, listing a directory of stimulus files, outputting any stimulus through the D/A for listening and verification.

Additional stimuli can be obtained from an off-line source. This would include all forms of analogue waveform generators (i.e., tone and noise), microphone input, audio tape recorder input, including audio tapes of special stimuli prepared at other laboratories, and other sources of acoustic waveforms of interest to the investigator which can be connected to the system via the interface and A-D subsystem.

The second step necessary to run an on-line perceptual experiment in the laboratory is to make the stimuli available on the system in digital form. This involves a two-step process of digital quantification of the stimulus and then refinement or conditioning of the digital representation through digital filtering and editing. For software generated stimuli, digitization is not necessary because the output form of the synthesis software is already a digital waveform. However, for other stimulus sources the analog signal must be digitized. As we mentioned earlier, the source of the waveform to be digitized can be an audio tape, human speaker, hardware synthesizer, tone generator, noise generator, or any other analog signal that can be input to the Analog Subsystem Controller.

Once a stimulus has been digitized, it is represented on the system as a stimulus disk file which can be digitally processed. Future software developments of the system include plans for implementing digital filtering algorithms. At the present time some editing of the waveform can be accomplished with the graphic waveform display system. This editing system with graphic waveform display output allows the user to perform simple operations on the waveform such as waveform subsection or windowed display, insert, display and clear fixed time cursors; waveform or subsection insertion, deletion, appending, store on disk file, retrieve from disk file; time the duration display between two fixed cursors, and outputting the waveform or a subsection between two fixed cursors through the D/A simultaneous with graphic display of the waveform.

Once the stimuli have been properly prepared they can be used in an actual experiment. Most of our experimentation is conducted on-line, but several programs are available to the user to produce stimulus sequences to be recorded on audio tape and used in stimulus presentation in off-line experiments. The audio tape generating programs allow the user to define a set of stimulus items, a presentation sequence and all the necessary timing parameters for stimulus sequencing.

For the experiments to be conducted on-line in real-time the third step in the experimental process is to write the FORTRAN control program that actually controls the entire experiment. The experimenter for this task has available to him a large library of assembly language subroutines which perform all the necessary real-time functions to present stimuli to subjects, time stimulus intervals, and collect subject responses and their latencies and store the responses for subsequent analysis.

With the complicated program sequencing and sophisticated interrupt servicing (for D/A transfers and subject responses) provided by assembly language subroutines, we have found FORTRAN to be the most suitable language for controlling experiments. FORTRAN has the advantage of being almost universally known by most psychologists and once the new laboratory user becomes familiar with the subroutine library it becomes quite easy to program any particular experiment.

For most experiment projects this third step is not required for each user since we have developed a set of standard experiment control programs for several of the most common experimental paradigms currently used in speech perception research. The standard paradigms which we have implemented and are currently available to a user include identification or labeling, selective adaptation, ABX, AX, 4IAX discrimination, same-different reaction time, recognition masking, dichotic listening and numerous discriminatory training procedures for use with speech and non-speech signals.

The last stage of the experimental process on the computer system involves the recovery and summarizing of subject response and latency data which is carried out by a standard set of programs. A common data file structure has been adopted throughout the laboratory so that if the standard programs do not provide the user with the necessary summary statistic he requires, he can quickly adopt one of these standard programs or derive his own. Once the data have been summarized and the desired statistics have been output to hard copy, the raw trial-by-trial data can be transferred to mag tape for storage. Often the tapes are taken to the campus computing center where further statistical analysis can be performed with the available BMD and SPSS statistical program packages.

Future System Developments

Continual effort is planned in the gradual evolution of all our major software systems including: additional experimental paradigms available in our library of developed FORTRAN experiment control programs; enhancements to the digital waveform editing and graphics display system; improvements to the I/O drivers for stimulus I/O; expansion of statistical procedures available in our standard real time data recovery and analysis programs; improvement in the quality of synthetic speech produced by the software synthesizer; synthesis by rule; and more extensive support for the generation of a wider range of non-speech acoustic signals.

One new major software project for the coming year will be to bring up a Linear Predictive Coding (LPC) analysis program at the campus computing center. This analysis program package will be used to analyze the digital representations of the vocal responses produced by subjects in a vowel mimicry project (see Pisoni, this report, pp.

Three other new projects will be initiated soon which will enhance the software and hardware of the system as well as increase efficiency in data processing. A three-fold telecommunications facility is planned for the laboratory computer system. With the addition of an auto-answer, send-receive modem and a DL11 interface, and the appropriate software modules, we plan to be able to support the following three new activities. First, we want to use the LA30 DECWriter in the laboratory as a terminal for the Wrubel Computing Center timesharing system. This feature will be used for working on large simulation programs, data reduction and data analysis using many available "canned" programs available at the computing center. Secondly, we eventually want to be able to communicate directly between the PDP11/05 and the campus computing center in order to initiate frequent job sequences at the center, transfer limited amounts of data and to check the status of batch jobs in progress at the center. Finally, we hope to use the PDP11/05 to communicate via a dial-up telephone line which will allow the computer to be controlled from other laboratories where it can be used to transfer data or make batch processing available from other locations if needed.

We also plan to add an LSI-11 micro-computer system to enhance our graphics capabilities significantly from what they are currently. Not only will it be possible to display larger waveform files, but the PDP11/05 CPU will be freed from the task of graphics screen refresh and will be better able to perform other processing operations such as those described above concerning waveform editing. These operations can then be carried out several times faster with the graphics display running concurrently.

One additional project is planned which only involves a change in the hardware. The addition of a DMA (direct memory interface) to the system will allow all A/D and D/A transfers involving the Analog Subsystem Controller to be handled much faster and with much less programming overhead. This will free-up processor time and will allow maximum concurrent activity with stimulus I/O that will be needed in several upcoming projects which require more precise timing of external events which are simultaneous with stimulus I/O.

At the present time we feel that the computer system is adequate for our own main research activity which involves primarily real-time perceptual experimentation. The system has been expanded almost to capacity now and seems to be quite efficient for most applications.

References

Forshee, J. C. Speech perception research laboratory: the state of the computer system. _Research on Speech Perception Progress Report_, Indiana University, Bloomington, Indiana, 1975, _2_,

Sawusch, J. R. A description of the OVE IIId Control Program: OVEXEC. _Research on Speech Perception Progress Report_, Indiana University, Bloomington, Indiana, 1975, _2_,

Footnote

[1] Work on the development and implementation of the laboratory facilities and the computer system described in this report was supported by NIMH Grant MH-24027 and NIH Grant NS-12179 to Indiana University.

A Complex-Tone Generating Program

Diane Kewley-Port

Indiana University

This paper describes the development of TONE, a program written to generate complex waveforms consisting of three sinusoidal components. The program permits the user to vary the frequency, amplitude and relative onset of each of three sinusoids independently. The output is a digital waveform which can be used for presentation to observers in various experimental paradigms.

TONE is a new program developed to synthesize non-speech signals having properties similar to speech sounds. The purpose is to generate digitally non-speech stimuli consisting of the sum of three sine waves which can vary over time in frequency and amplitude under user control. The sine waves can be specified to approximate the amplitude and frequency variations of three formants of a speech stimulus. In particular, the parameters used to generate a three-formant stimulus on a speech synthesizer can be matched very closely for a three-tone complex stimulus.

TONE was developed on the Speech Perception Laboratory PDP-11/05 computer (see Forshee, 1975, 1976). Parameter values for each of the three tones are typed on the ADDS consul and may be printed out for the user. Each of the three tones can be specified independently. A tone may begin at any point in time referenced to t = 0 msec. The tone parameters are specified in 5 msec intervals. For each point in time that the tone changes, three parameter values are chosen; the time relative to t = 0, the relative amplitude, 0 to 66 dB, and the frequency in

Hz. Changes in frequency or amplitude between any two points are interpolated linearly. The maximum length of the complex-tone waveform is 500 msec at the present time.

The output of TONE is a 12 bit digital waveform in memory. The user may store and retrieve the waveform from disk using his own six character name. The user may listen to the complex-tone repeatedly through earphones or on a loudspeaker. These stimuli can later be presented to subjects directly from disk via the D-A using the software designed for experimentation by Forshee (1975).

The algorithms used in TONE were developed from suggestions and extensive consultation with Dr. Dennis Klatt at the Research Laboratory of Electronics, M.I.T. to whom we are very grateful. The algorithms were based on the standard equation:

$$T = a\mathrm{SIN}(2\pi ft + \phi)$$

where: a = amplitude

f = frequency

t = time

ϕ = phase

To compute the desired tone or complex tone, three separate strategies were used in implementing the above equation, one for amplitude, one for frequency and one for the digital computation of the SIN function. Each will be discussed separately below.

Frequency Algorithm

It was decided that the phase between the three component tones was not an important parameter for the experimentation proposed, so ϕ was dropped from the frequency calculations. A tone T is referenced to the output sample rate which is currently 20,000 samples/second. For each sample k, k = 0 to 20,000 (for 1 sec), a value T(k) is calculated as:

$$T(k) = A(k) \cdot \mathrm{SIN}\left(2\pi f(k)\frac{k}{20000}\right).$$

If the frequency were steady state, values for one period (or a half-period) of the waveform could be calculated, stored and appropriately concatenated for the duration of the tone. Since the purpose of the TONE program is to allow for continuously varying changes in frequency such as those that occur in real speech, it is necessary to calculate T(k) separately for each k.

Care must be taken in specifying an algorithm for changing frequency if the associated problems of phase shift and boundary discontinuities at the 5 msec time intervals are to be avoided. In particular, the SIN(x) function is computed for a parameter x in radians. To produce a smoothly varying SIN function, the following sample algorithm in FORTRAN can be used to calculate the correct radian value, R(k), which is summed over the entire tone length. F1 and F2 are the frequencies of the endpoints of a given time interval, TIME.

```
      .
      .
      .
      DF = (F2 - F1)/TIME     ! DELTA FREQUENCY IN HERTZ
      DFR = (2*PI/20000)*DF   ! DF IN RADIANS
      FR = (2*PI/20000)*F1    ! INITIAL FREQUENCY IN RADIANS
      DO 100 J=1, TIME
      K = K + 1
      R(K) = R(K) + FR        ! RUNNING VALUE OF RADIANS
      T(K) = A(K)*SIN(R(K))
100   FR = FR + DFR
      CONTINUE
      .
      .
      .
```

Amplitude Algorithm

The amplitude of the complex-tone waveform provides a maximum of 12 bits or signed 11 bits to the D/A converters. This is equivalent to a 66.226 dB range which was rounded to 66 dB. The relative amplitude of each tone is therefore specified as 0 to 66 dB on input to the program. In order to use the maximum range of the D/A converters, the absolute maximum value of the complex-tone waveform is obtained, and then the waveform is scaled up or down to the full 12 bits. Overall amplitude of the stimuli may be attenuated under computer control 0 to 63 dB during a particular experiment.

To calculate A(k) for each tone, the input parameters in dB are converted to absolute values as (in FORTRAN):

```
A(K) = 10.**(DB/20.).
```

It was arbitrarily decided that the amplitude variation would be piecewise linear, interpolated in absolute (not dB) values.

SIN Function

The greatest difficulty we encountered in getting TONE to run rapidly on the PDP-11/05 involved the SIN function and therefore we had to develop an appropriate SIN function algorithm. For a sample rate of 20K we needed a fast SIN algorithm with accuracy of only 12 bits for output. The FORTRAN SIN function was totally inadequate for this purpose. A formula was taken from Hastings (1955, sheet 14) with 14 bit accuracy. An algorithm was then developed for fixed point arithmetic, with a resulting accuracy of 12 bits. The machine language version utilizes the KE11-A Extended Arithmetic Element. As an example of the improvement in speed of our own SIN function, a 200 msec, three-tone stimulus took 16 seconds with our routine and 90 sec with the FORTRAN routine. More information about the SIN algorithm will be sent on request to interested colleagues.

Future Extensions of TONE

The TONE program will be expanded and modified to incorporate some new hardware acquisitions in the coming year. A 10K sample rate option for output will be available allowing the duration of a complex tone to be extended to 1 sec, or with double buffering to even several seconds if the need arises. When the graphic tablet arrives input parameters to the program can be made from diagrams of the stimuli as well as sound spectrograms.

In summary, we now have available a software program for generating complex tones which have speech-like properties. The amplitude and frequency of three sinusoids can be manipulated independently under digital control by the user. The relative onset of these three components can also be varied independently thus providing a way of studying some of the spectral and temporal properties found in speech.

References

Forshee, J. C. Speech Perception Research Laboratory: The State of the Computer System. *Research on Speech Perception Progress Report No. 2*, 1975, 202-220.

Forshee, J. C. Computer Resources in the Speech Perception Laboratory. *Research on Speech Perception Progress Report No. 3*, 1976,

Hastings, C. *Approximations for digital computers*. Princeton, N.J.: Princeton University Press, 1955.

Publications:

Castellan, N. J., Pisoni, D. B., & Potts, G. (Eds.) *Cognitive theory: Volume II*. Hillsdale, N.J.: Erlbaum Associates, 1977.

Pisoni, D. B. Information processing and speech perception. In G. Fant (Ed.), *Speech communication: Volume 3*. New York: John Wiley, 1975, pp. 331-337.

Pisoni, D. B. & Sawusch, J. R. Some stages of processing in speech perception. In A. Cohen & S. Nooteboom (Eds.) *Structure and process in speech perception*. Heidelberg: Springer-Verlag, 1975, pp. 16-34.

Pisoni, D. B. & Tash, J. Auditory property detectors and processing place features in stop consonants. *Perception & Psychophysics*, 1975, *18*, 401-408.

Pisoni, D. B. Review of "The Psychology of Language" by J. A. Fodor, T. G. Bever and M. F. Garrett. *Language*, 1976, *52*, 3, 682-689.

Pisoni, D. B. Mechanisms of auditory discrimination and coding of linguistic information. In J. V. Irwin (Ed.) *The Second Auditory Processing and Learning Disabilities Symposium*. Memphis: Memphis State University Press, 1976, pp. 73-119.

Sawusch, J. R. Selective adaptation effects on end-point stimuli in a speech series. *Perception & Psychophysics*, 1976, *20*, 61-65.

Sawusch, J. R. & Pisoni, D. B. Response organization in selective adaptation to speech sounds. *Perception & Psychophysics*, 1976, *20*, 413-418.

Technical Reports:

1. Pisoni, D. B. Speech perception. *Technical Report No. 1*, June, 15, 1976, Pp. 119.

2. Sawusch, J. R. The structure and flow of information in speech perception: Evidence from selective adaptation of stop consonants. *Technical Report No. 2*, August 15, 1976, Pp. 217.

Manuscripts to be published:

Cutting, J. E. & Pisoni, D. B. An information processing approach to speech perception. In J. F. Kavanagh & W. Strange (Eds.) *Implications of basic speech and language research to the school and clinic*. Cambridge: The M.I.T. Press, 1977. (In Press).

Liberman, A. M. & Pisoni, D. B. Evidence for a special speech perception subsystem in the human. In T. H. Bullock (Ed.) *Recognition of complex acoustic signals*. Berlin: Dahlem Konferenzen, 1977. (In Press).

Pisoni, D. B. On the perception of speech sounds as biologically significant signals. To appear in a special issue of *Brain, Behavior, and Evolution*, 1977. (In Press).

Pisoni, D. B. Speech perception. In W. K. Estes (Ed.) *Handbook of learning and cognitive processes*: *Volume 6*. Hillsdale, N.J.: Erlbaum Associates, 1977. (In Press).

Pisoni, D. B. Stages of processing in speech perception: Feature analysis. *In Proceedings of the Eighth International Congress of Phonetic Sciences*, Leeds, England. (In Press).

Pisoni, D. B. Identification and discrimination of the relative onset of two component tones: Implications for voicing perception in stop consonants. *Journal of the Acoustical Society of America*, 1977. (In Press).