RESEARCH ON SPEECH PERCEPTION

Progress Report No. 6

September 1979 - December 1980

David B. Pisoni

Principal Investigator

Department of Psychology

Indiana University

Bloomington, Indiana 47405

Supported by:

Department of Health and Human Services
U.S. Public Health Service

National Institute of Mental Health
Research Grant No. MH-24027-06

National Institutes of Health
Research Grant No. NS-12179-05

and

National Institutes of Health
Training Grant No. NS-07134-01

CONTENTS

INTRODUCTION

This is the sixth annual progress and status report of research activities on speech perception, analysis and synthesis 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 presented at professional meetings during the past year or brief summaries of 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 other colleagues.

We are distributing progress reports 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 in the field. We are, of course, very interested in following the work of other colleagues who are carrying out research on speech perception, production, analysis and synthesis and, therefore, we would be 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 47405
U.S.A.

Copies of this report are being sent primarily to libraries and research institutions rather than individual scientists. Because of the rising costs of publication and printing it is not possible to provide multiple copies of this report or issue copies to individuals.

I. EXTENDED MANUSCRIPTS

Discrimination of Voice-onset-time by Human Infants: New Findings Concerning Phonetic Development

Richard N. Aslin

David B. Pisoni

Beth L. Hennessy *

Alan J. Perey

Indiana University

Running Head: Discrimination of VOT

Abstract

Two experiments examined the ability of 6- to 12-month-old infants to discriminate differences in voice-onset-time (VOT) at several regions along a synthetic VOT continuum ranging from -70 msec to +70 msec. An operant headturning technique, combined with an adaptive staircase testing algorithm, was used to determine whether individual infants could discriminate fixed VOT contrasts and to estimate thresholds for resolving small differences in VOT. Infants from an English-speaking environment provided reliable within-subject evidence for discrimination of VOT contrasts located at both the plus and minus regions of the VOT continuum. Threshold ΔVOT values indicated that the infants were more sensitive to VOT differences in the plus region of the VOT continuum than in the minus region. Similar ΔVOT results were obtained from English-speaking adults tested under identical experimental conditions. However, the adults were more sensitive than the infants at every location along the VOT continuum. In addition, the adults showed a peak sensitivity to VOT differences surrounding the voiced-voiceless boundary in the plus region of the VOT continuum, a peak that was not evident in the infants' data. These results provide strong evidence that infants from an English-speaking environment are capable of discriminating VOT contrasts that are not phonemic in English. Previous claims that the ability to discriminate VOT contrasts in the minus region of the VOT continuum is acquired from early linguistic inputs, as well as claims that infants perceive VOT differences in a phonetic mode, are critically evaluated.

Discrimination of Voice-onset-time by Human Infants: New Findings Concerning Phonetic Development

The past decade has witnessed a remarkable upsurge of interest in the sensory and perceptual capabilities of young infants. Of particular importance to this trend was the pioneering study of speech sound discrimination by Eimas, Siqueland, Jusczyk and Vigorito (1971). The Eimas, et al. study has been extremely influential not only because of the infant discriminative abilities demonstrated but also because of the interpretation offered concerning the developmental mechanisms underlying speech perception. It is to these broad interpretive and theoretical issues that the present paper is addressed.

The findings and interpretations reported by Eimas, et al. in their study of infants' perception of synthetic speech sounds consisted of four major points. First, infants as young as one-month of age were shown to be capable of discriminating small differences in an acoustic parameter, voice-onset-time (VOT),[1] a dimension that is sufficient to differentiate the phonetic categories of voiced (b, d, g) from voiceless (p, t, k) stop consonants. Second, only VOT differences that straddled the voiced-voiceless boundary for English-speaking adults (approximately +25 msec) were discriminated by young infants from an English-speaking environment. Comparable differences in VOT selected from within the phonemic categories of English were not discriminated. Third, the differences in infants'

sensitivity to VOT contrasts within and between categories were interpreted as evidence of categorical perception. And fourth, based on a traditional view that categorical perception is unique to speech sounds, and the fact that young infants have little opportunity to acquire phonetic categories prior to testing, their discrimination performance was interpreted as evidence of "perception in a linguistic mode (that) may well be part of the biological makeup of the organism (p. 306)." In recent years, however, new data from several areas of auditory perception have raised many questions concerning these initial interpretations, despite overwhelming support for the conclusion that infants appear to discriminate a wide variety of phonetic contrasts (see reviews by Aslin and Pisoni, 1980b; Eilers, 1980; Eimas, 1975, 1978; Eimas and Tartter, 1979; Jusczyk, 1980; Kuhl, 1976, 1978; 1980; Morse, 1978, 1979; Trehub, 1979).

The interpretation that young infants perceive VOT differences categorically was premature for two reasons. First, categorical perception has been characterized by three properties: (1) a sharp cross-over in the VOT identification function at the category boundary, (2) a non-monotonic discrimination function with peak VOT sensitivity at the category boundary, and (3) an accurate prediction of the discrimination function from the identification function (Studdert-Kennedy, Liberman, Harris and Cooper, 1970). Only data on VOT discriminability were obtained in the Eimas, et al. study.[2] Moreover, the correspondence between discrimination

and identification functions must be verified with individual subject data, and only group data were available from the Eimas, et al. study because of the high variability associated with the high amplitude sucking technique. Since there are always many potential reasons for an infant's failure to perform at above chance levels in any perceptual study (e.g., insensitive methods, poor attention, etc.), the discovery of a discontinuity in VOT discrimination was only suggestive of categorical discrimination. For example, in adults there are peak regions of discriminability between adjacent vowel sounds despite the non-categorical nature of vowel discrimination (Stevens, Liberman, Studdert-Kennedy and Ohman 1969; Pisoni, 1975).

A second reason that the Eimas, et al. interpretation of categorical perception in infants was premature stems from the correspondence between the infants' presumed peak in VOT discriminability and the location of the category boundary for English-speaking adults. If Eimas, et al. had found that infants only discriminated a VOT contrast that did not correspond to the category boundary of English-speaking adults, then their interpretation would have been somewhat different. In fact, as shown in Figure 1, the precise location of adults'

Insert Figure 1 About Here

---

category boundary for voicing differences varies considerably from language to language (Lisker and Abramson, 1964). Moreover, adults appear to be most sensitive to VOT differences

located at or near the category boundary used in their native language (Abramson and Lisker, 1970). These differences in VOT sensitivity among adults are clearly the result of experiential influences exerted during the language learning process, although some underlying sensory or psychophysical mechanism may also operate to limit the number and general location of the category boundaries employed by various phonological systems.[3] Thus, for example, if the results obtained by Eimas, et al. had been gathered by researchers in Guatemala (Spanish has a VOT boundary of +4 msec) or Thailand (Thai has two VOT boundaries near -30 msec and +25 msec), they might not have been interpreted as evidence for categorical perception, but simply as evidence for infants' differential sensitivity to VOT differences. In other words, it may have been fortuitous that the infants in the Eimas, et al. study discriminated a VOT contrast that corresponds closely to the peak in VOT discriminability for the voicing boundary in English-speaking adults.

Given the interpretation by Eimas, et al. that infants perceive stop consonants categorically, the further inference that infants perceive these sounds in a "linguistic mode" according to innate phonetic categories was undeniably a seductive conclusion. Unfortunately, a definitive answer to the question of whether infants perceive speech sounds as <u>linguistic</u> units rather than <u>acoustic</u> entities demands more than simple discrimination data. If, however, one is limited to discrimination data, then there are two key tests of the

hypothesis that infants perceive stop consonants as phonetic segments (the phonetic-coding hypothesis).[4] First, since a phonetically-based perceptual system presumably performs some type of special analysis on speech sounds, then phenomena that are unique to speech, such as categorical perception, should not be present for nonspeech control signals. However, several recent studies with adults (Cutting and Rosner, 1974; Cutting, Rosner and Foard, 1976; Miller, Wier, Pastore, Kelly and Dooling, 1976; Pisoni, 1977) have conclusively demonstrated that categorical perception is not unique to speech sounds. In addition, two studies (Jusczyk, Rosner, Cutting, Foard and Smith, 1977; and Jusczyk, Pisoni, Walley and Murray, 1980) have clearly demonstrated that nonspeech sounds are discriminated in a categorical-like manner by human infants. Thus, the seemingly plausible interpretation that infants who exhibit categorical-like discrimination of stop consonants are employing a phonetic level of analysis must be reexamined in light of these recent findings.

The second key test of the phonetic-coding hypothesis involves comparative studies of the perception of human speech signals by non-human species. Since non-humans do not have access to the phonology of a natural language, their perceptual behavior should not exhibit any property assumed to be unique to human speech perception, such as categorical perception. Several recent studies (Kuhl and Miller, 1975, 1978; Morse and Snowden, 1975; Waters and Wilson, 1976) have demonstrated that in chinchillas and monkeys some of the characteristics of

categorical perception (either non-monotonic discrimination functions or the presence of sharp identification functions) are present. Thus, if categorical perception can be demonstrated by non-humans presented with human speech sounds, and conversely by human adults presented with nonspeech signals, then the conclusion that categorical-like discrimination of VOT by human infants is mediated by a phonetic level of analysis is equivocal at best.

Despite awareness of the foregoing evidence against a phonetically-based speech perception system in human infants, numerous researchers have continued to interpret infant speech discrimination data according to the phonetic-coding hypothesis (Eimas, Note 1; Eimas and Miller, 1980; Eimas and Tartter, 1979; Eilers, Gavin and Wilson, 1979; Miller and Eimas, 1979; Morse, 1978). One approach to resolving this controversy consists of studying young infants' ability to discriminate a wide range of VOT contrasts, particularly those contrasts that cross a category boundary that is not used in the phonological system of the infant's native language. If, during the course of early infancy, the peak in VOT discriminability begins to correspond precisely with the VOT category boundary of the infant's native language, then the process of phonetic tuning may be inferred.[5] This process of phonetic tuning could involve the acquisition of sensitivity to only those VOT regions surrounding category boundaries used in the infant's language environment, or a selective attenuation and/or realignment of any peak(s) in VOT sensitivity that is (are) already present at

birth (see Aslin and Pisoni, 1980b, for a discussion of these possible experiential modifications).

Within the last few years, four studies have been carried out to examine the abilities of infants to discriminate VOT contrasts that do not occur in the infant's native language. Eimas (1975) found only weak evidence, despite employing a very large (80 msec) difference in VOT, that infants from an English-speaking environment discriminated a VOT difference that contrasts a voiced /d/ from a voiceless unaspirated /t/, a difference which is used phonemically in Thai. Lasky, Syrdal-Lasky and Klein (1975) found that Guatemalan infants from a Spanish-speaking environment discriminated a voiced /b/ from a voiceless unaspirated /p/, as well as a voiceless unaspirated /p/ from a voiceless aspirated /p^h/. However, neither of these contrasts crossed the Guatemalan adult category boundary of +4 msec (Williams, 1977). Interestingly, a third group of Guatemalan infants failed to discriminate the VOT contrast that straddled the adult boundary. Streeter (1976) found that infants from a Kikuyu language environment discriminated both the voiced /b/ from the voiceless unaspirated /p/ and the voiceless unaspirated /p/ from the voiceless aspirated /p^h/. It should be noted here that Kikuyu adults do <u>not</u> employ contrasts in voicing at the labial place of articulation in their native language, although the voiced-voiceless unaspirated distinction is used phonemically at the velar place of articulation. Finally, Eilers, Gavin and Wilson (1979) have recently reported that infants from a Spanish-speaking environment discriminated a

voiced /b/ from a voiceless unaspirated /p/, as well as a voiceless unaspirated /p/ from a voiceless aspirated /p^h/. Recall that Spanish does not employ the voiceless aspirated labial stop phonemically (see Figure 1). Eilers, et al. also reported that infants from an English-speaking environment discriminated the VOT contrast that is phonemic in English (voiceless unaspirated /p/ from voiceless aspirated /p^h/), but failed to discriminate the voiced-voiceless unaspirated contrast that is phonemic in Spanish (see, however, Aslin and Pisoni, 1980a, for a critique of these conclusions).

To a first approximation, the results of these four discrimination studies suggest that infants from all language environments studied to date are able to discriminate VOT differences in the plus (i.e., short lag) region of the VOT continuum that signal differences in aspiration between the voiceless stops in a number of languages. However, only infants from a language environment that uses the voiced-voiceless unaspirated distinction phonemically are apparently able to discriminate VOT differences in the minus or lead region of the VOT continuum. Despite the lack of precise correspondence between the adult category boundary and the presumed peak in VOT discriminability for infants in the Lasky, et al. and Streeter studies, the absence of any evidence for VOT discrimination in the minus VOT region by infants from English-speaking environments has suggested to several researchers that the voiced-voiceless unaspirated distinction must be "learned" through experience in a specific linguistic environment (e.g.,

Eilers, et al., 1979). However, several important methodological and interpretive issues have been raised recently concerning the Eilers, et al. study in particular, and for that matter, any study that claims to demonstrate that infants fail to discriminate specific VOT contrasts (see critique by Aslin and Pisoni, 1980a, and reply by Eilers, Gavin and Wilson, 1980). Thus, it is of particular theoretical importance at this time to determine whether infants from an English-speaking environment are, in fact, capable of discriminating VOT contrasts in the minus region of the VOT continuum. Positive evidence of such discriminative abilities would demand a careful reinterpretation of the phonetic-coding hypothesis as it has been applied to infants' discrimination of speech signals.

In the present study, the capacities of infants to discriminate VOT differences were examined along a broad range of a synthetic VOT continuum encompassing voiced, voiceless unaspirated and voiceless aspirated bilabial (b, p, p^h) stop consonants. In addition, the limits of the infants' ability to detect small differences in VOT were assessed in a psychophysical testing procedure using an adaptive staircase technique. The discrimination procedure employed an operant headturning paradigm that enabled us to collect data over multiple sessions from the same infant. If reliable evidence for the ability of infants to discriminate VOT contrasts in the minus region of the VOT continuum could be obtained, then the phonetic-coding hypothesis regarding infant speech perception would be further weakened, as would be claims that the

voiced-voiceless unaspirated distinction must be learned from early experiences in the language- learning environment. In addition, such results would also imply that previous failures to demonstrate that infants can discriminate phonologically-irrelevant VOT contrasts may well have been subject to subtle methodological artifacts. Finally, positive evidence of discrimination of both lead and lag VOT contrasts would imply that both voicing distinctions (voiced-voiceless unaspirated and voiceless unaspirated- voiceless aspirated) are discriminated on the basis of an underlying psychophysical mechanism that is language-independent.

Experiment 1

Method

Subjects. A total of 92 infants ranging in age from 5.5 to 11.5 months participated in Experiment 1. Forty-two of these infants failed to show any reliable evidence of acquiring the headturn response after two sessions and were dropped from the study. [6] Infants were solicited on a volunteer basis by letter and phone from the birth announcements in the local newspaper. All infants were presumed to have normal hearing, although intermittent illnesses (and poor performances) may have been accompanied by middle ear infections (otitis media). The parents were paid $3 for each testing session.

Stimuli. The stimuli consisted of a set of 15 synthetic labial stop consonant-vowel syllables that were generated on the

cascade-parallel software synthesizer originally designed by Klatt (1980) and subsequently modified and expanded in our laboratory by Kewley-Port (1978). The 15 stimuli differed in 10 msec steps of VOT from -70 msec to +70 msec. Spectrograms of the -70 msec, 0 msec and +70 msec stimuli are illustrated in Figure 2. The values used for synthesis of these stimuli were

---

Insert Figure 2 About Here

---

chosen from measurements of natural speech originally made by Klatt (Note 2) as well as our own measurements from spectrograms of the speech of a native English-speaker (RFP), a trained phonetician, producing various voicing contrasts. The stimuli consisted of a 255 msec steady-state pattern with formant values appropriate for the vowel /a/ (F1=700Hz, BW1=90 Hz; F2=1200 Hz, BW2=90 Hz; F3=2600 Hz, BW3=130 Hz; F4=3300 Hz, BW4=400 Hz; F5=3700 Hz, BW5=500 Hz). The formant transitions into the vowel were 40 msec in duration and had starting frequencies appropriate for the bilabial stop in stressed syllable initial position (F1=438 Hz, F2=1025 Hz, F3=2425 Hz). Voicing lead was simulated by passing a combination of the sinusoidal voicing source (ASV) and normal voicing (AV) through F1 which was set at 180 Hz with a bandwidth of 150 Hz. The amplitude values of the voicing source were chosen to match natural productions measured from broad-band spectrograms and average amplitude contours. A 10 msec release burst was generated by passing a turbulent noise source (AF) through the bypass channel (AB) of the parallel

branch of the synthesizer which has a broadband (5 KHz) flat spectrum. The amplitude of the release burst was chosen on both theoretical and empirical grounds after a long series of listening tests. Finally, the aspiration associated with voiceless stops was generated by passing a noise source (AH) through the cascade branch of the synthesizer to simulate the turbulence produced at the glottis. The amplitude of aspiration was again chosen to match measurements of spectrograms of natural speech. During the period of aspiration, the bandwidth of F1 was also widened to 300 Hz. To simulate breathiness at the end of the syllable, the bandwidth of F1 was widened linearly from 90 Hz to 180 Hz and some aspiration noise was added to the final 35 msec of the stimulus. The pitch contour had a slight rise at the onset of the release of the consonant from 120 Hz to 125 Hz and then fell linearly to 100 Hz over the remaining steady-state portion of the vowel.

Apparatus. The testing situation is illustrated in Figures 3 and 4. Infants were tested individually in a single-wall sound attenuated booth (IAC Model 402, internal dimensions 2 X 2 m) while seated on their parent's lap. An assistant was seated facing the infant and parent. The assistant had access to a group of small toys that were used to attract the infant's gaze during the testing procedure. The booth was equipped with a single loudspeaker (Radio Shack MC-1000), a cassette tapedeck that presented masking music over headphones to the assistant and parent, and a visual reinforcer. The visual reinforcer consisted of a smoked plexiglas enclosure (30 cm cube)

containing a motorized toy animal (bear, monkey or dog) that could be activated by the experimenter who was located outside the testing booth. When the reinforcer was activated, the inside of the plexiglas enclosure was illuminated by two fluorescent bulbs. Thus, the toy animal was not visible until the lights were illuminated, at which time the toy became animated (drummed, clapped or jumped).

Insert Figures 3 and 4 About Here

The entire experimental procedure was controlled on-line in real-time by a PDP-11/34 computer. The experimenter viewed the infant on a closed-circuit TV monitor during the entire testing procedure. The experimenter initiated trials and coded the infant's behavior with a button box that was interfaced to the computer. The computer program presented the speech sounds at a 10 KHz sampling rate through a 12 bit D/A converter to the loudspeaker in the booth and controlled the onset and offset of the visual reinforcer. The program was designed to code the experimenter's responses, present all stimuli, and deliver reinforcement in accordance with a set of constraints dictated by the experimental procedure. The coding of all responses and stimulus contingencies on each trial of the experiment were stored by the computer on disk for later analysis.

Procedure. The basic testing procedure employed a variant of the operant headturning technique originally developed by Wilson (1978), with the addition of more rigorous control

procedures to eliminate experimenter bias and an adaptive staircase testing algorithm to estimate the limits of the infant's discriminative abilities. The basic operant headturning procedure involves the repetitive presentation of a background signal that is interrupted by several repetitions of a novel target signal. The discriminative response is a unidirectional headturn toward the direction of the visual reinforcer and loudspeaker, which in our laboratory are located to the infant's left as shown in Figures 3 and 4. Initially, the infant is naive to the contingency between the change in the background signal and the headturn response. Thus, the infant must be shaped to orient toward the location of the change in the stimulus which signals the appearance of the visual reinforcer. Evidence of positive discrimination in this procedure consists of an anticipatory headturn on stimulus-change (experimental) trials and the absence of an anticipatory headturn on no-change (control) trials.

Our modification of this basic headturning procedure involved three distinct phases: (1) shaping, (2) criterion testing and (3) staircase testing. Throughout all three phases the speech stimuli were presented with a constant interstimulus interval of 1200 msec, thus eliminating any potential temporal cue that might occur during the shift from background to target. During the shaping phase the experimenter viewed the infant on a TV monitor and, when the infant's gaze was steadily directed toward the toys being manipulated by the assistant, initiated an experimental trial consisting of three repetitions of the target

stimulus. In the first testing session for each infant, shaping trials began with the background at 70 db and the target at 80 db. This additional redundant intensity cue was used to facilitate the infant's acquisition of the basic orienting response toward the sound change. On subsequent shaping trials, the intensity of the target was reduced in 5 db steps until the target and background were equated at 70 db. The experimenter typically reduced the target intensity after 2 or 3 trials on which a correct anticipatory headturn response was observed. Throughout the shaping phase the experimenter could hear the stimuli being presented to the infant, thus enabling her to judge how well the infant was performing and ensuring that the visual reinforcer was presented immediately after a headturn. In addition, the experimenter sometimes presented the visual reinforcer simultaneously with the onset of the sound change to establish the contingency between sound change and visual reinforcer for the infant. When the experimenter judged that the infant was consistently anticipating the appearance of the visual reinforcer on trials in which the target and background were equal in intensity, the computer program was signalled to proceed to the next phase of the experiment, the criterion testing phase.

During the criterion testing phase the experimenter wore headphones through which a tone was presented in synchrony with the background and target stimuli. The experimenter continued to initiate trials whenever the infant's gaze was consistently directed toward the toys being manipulated by the assistant in

the booth. Trials now consisted of two types: experimental (change) or control (no-change). These two trial types were presented according to one of eight pseudo-random orders (Fellows, 1967) with the constraint that no more than two control trials be presented successively and that each type of trial occur with equal frequency in each block of ten trials. The observation interval for scoring the infant's headturn response began with the onset of the first target repetition and ended two seconds after the third (and final) target repetition (total scoring interval = 5.285 sec). If the experimenter judged that a headturn had occurred during the scoring interval on an experimental trial, a button was pressed which signalled the computer to immediately present the visual reinforcer for three seconds. If the headturn button was pressed on a control trial, the computer did not present the visual reinforcer. The computer program also kept a running tally of the infant's percent correct responding on the previous five experimental trials and the previous five control trials. When this percent correct met an 80% correct criterion for both experimental and control trials, the computer automatically terminated the criterion testing phase of the experiment and immediately proceeded to the staircase testing phase. This 80% criterion, applied to the five experimental and five control trials, results in a probability of .055 (based on the binomial expansion, p= 0.5, q= 0.5) of falsely accepting the hypothesis that the infant was responding above chance. Note that the experimenter was "blind" as to the type of trial (experimental

or control) presented to the infant on any trial during the criterion testing phase. The parent and the assistant in the booth were also "blind" as to the time of onset and type of trial presented. Thus, there was no potential for experimenter bias to influence the infant's discriminative behavior in this testing procedure (see Aslin and Pisoni, 1980a).

The staircase testing phase was identical to the criterion testing phase except that the VOT value of the target stimulus was varied as a function of the accuracy of the infant's performance on preceding trials. A staircase algorithm was used to measure the smallest VOT difference between the background and target that the infant could reliably discriminate. The algorithm followed a 2 up and 1 down rule in which two correct responses at a given target VOT value resulted in a change in the VOT value of the target that decreased the background-target VOT difference. A single incorrect response at a given target VOT value resulted in an increase in the background-target VOT difference. This staircase algorithm generates a probability of correct response equal to 70.7% (see Levitt, 1970).

Figure 5 illustrates the response sequence of a particular infant tested in the staircase phase on a VOT series with 0 msec as the background stimulus and -70 msec as the initial target stimulus. The VOT step size in this series was 10 msec, so that two correct responses (hits) or one incorrect response (miss) resulted in a 10 msec change in the target VOT value, either up or down, relative to the constant background signal. This infant consistently responded to the shift from background to

target at target VOT values of -70, -60, -50, -40 and -30 msec before failing to respond on the first trial at -20 msec. The "miss" at -20 msec resulted in a return of the target VOT value to -30 msec on the subsequent trial. As shown in Figure 5, the 2 up and 1 down algorithm generates a series of reversal points which can be used to estimate the target VOT value that is discriminated from the background at the 70.7% level. At least five reversals were required in a given staircase testing session for the infant's data to be considered acceptable. The computer program terminated the staircase phase when six reversal points had been collected. The estimate of the VOT threshold value was computed as the mean of the midpoints of the five or six reversal points. Finally, in carrying out this procedure, we also included probe trials which occurred whenever the infant failed to respond correctly on two consecutive trials. Probe trials consisted of presenting the VOT value of the initial target (the largest VOT difference). These probe trials not only helped to maintain the infant's interest in the task during brief periods of inattention, but also provided us with an objective behavioral criterion that could be used to terminate a session prior to reaching the five reversal point requirement. If an infant failed to respond correctly on three consecutive probe trials, the session was terminated and the data were judged to be unacceptable on the grounds that the infant had not maintained selective attention to the criterial properties of the stimuli.

Insert Figure 5 About Here

Design. All infants were initially tested on the full VOT series, with assignment of target (-70 or +70 msec) and background (+70 or -70 msec) randomized across subjects. Of the 50 infants who progressed beyond the shaping phase, 44 completed the criterion testing phase on one of the two full VOT series. Twenty-seven of these infants who met criterion in the testing phase completed the staircase testing phase on the full VOT series. The step size used in the staircase testing phase with the full VOT series was 20 msec. Those infants who failed to meet the 80% correct criterion (6 of 50) and those who failed to complete the staircase phase (17 of 44) either were unable to return for additional testing sessions or showed little further interest in the visual reinforcer.

After completion of the staircase testing phase on the full VOT series, each infant returned for further testing sessions on several truncated versions of the full VOT series. Each of these truncated VOT series spanned a range of 70 msec from background to target and employed a step size of 10 msec. If the infant's background VOT value on the full series was +70 msec, then the background value on subsequent truncated series was +70, +50, +20 and 0 msec. If the initial background VOT value was -70 msec, then on the truncated series it was 0, -20, -50 and -70 msec. Thus, the initial truncated series was located in the plus region of the VOT continuum, and subsequent

truncated series were gradually shifted to the minus region of the VOT continuum.

Finally, it became clear that many infants became satiated both with the visual reinforcer and the entire testing procedure after several sessions. Thus, within a session on the truncated VOT series, the criterion testing phase was eliminated for approximately half of the infants. These infants, however, had met criterion during the full VOT series. For these infants, testing on the truncated series proceeded directly from the shaping phase to the staircase testing phase.

Adult Controls. Two groups of 10 adults were also tested with the same VOT stimuli in a procedure identical to that used with the infants. Each adult was tested individually in the same sound-attenuated booth and listened to the stimuli through the same speaker. The adults were instructed to respond to any difference in the background signal by raising their hand or turning their head to view the reinforcer. The reinforcer was modified by placing a "correct" sign inside the plexiglas enclosure in front of the animated toy. One group of adults was assigned to the -70 msec background and the other group to the +70 msec background for testing on the full VOT series. Subsequent tests on the four truncated series proceeded in the same manner as with the infants. Since the adults had little difficulty in discriminating the five VOT contrasts, they proceeded directly from the shaping phase to the staircase testing phase. Data for all five VOT series were collected in a single one hour testing session.

Results and Discussion

Forty-four of the 50 infants who proceeded from the shaping phase to the criterion testing phase met the 80% criterion on one of the two full VOT series (-70 vs. +70 or +70 vs. -70 msec). Although all of these 44 infants demonstrated reliable evidence of discriminating this large VOT difference between the endpoint stimuli, only 27 infants subsequently completed the staircase testing phase. This additional subject attrition resulted from the high demands of passing through both the criterion and staircase testing phases in a single session (average number of sessions to completion = 3.5). Often, an infant would meet the 80% criterion and begin the staircase phase, but then fail to provide the required five reversals before becoming fussy or inattentive. Nevertheless, the 27 infants who completed the staircase testing phase comprise a relatively large sample, particularly since each infant provided within-subject data rather than simply between-subject group data as in past studies of infant speech discrimination.

Data from each of the 27 infants who completed the staircase phase on one of the two full VOT series are illustrated in Figure 6. Note that the target VOT value at which the infants ceased to respond, i.e., the estimate of their category boundary, was located at or near the VOT value typically associated with the voiced-voiceless boundary in English-speaking adults. The mean boundary value for these 27 infants was 21.6 msec. On the basis of these data alone, it would appear that this boundary location in the voicing lag or

plus region of the VOT continuum is not only very distinctive perceptually, but that infants from an English-speaking environment are incapable of reliably discriminating VOT differences in the minus region of the VOT continuum. However, a stronger test of this hypothesis requires a closer examination of infants' discriminative abilities within more restricted regions of the VOT continuum.

Insert Figure 6 About Here

Eleven of the 27 infants who completed the full VOT series went on to successfully complete the staircase testing phase on one of the two VOT series in the plus region of the VOT continuum (0 vs. +70 or +70 vs. 0). Seven infants subsequently completed another truncated series in the plus region of the VOT continuum (+50 vs. -20 or -20 vs. +50). Boundary values for each infant who completed the staircase phase on VOT contrasts in the plus region of the VOT continuum are shown in Table 1. First, note that on the two +70 series the boundary values were similar to those obtained on the full (-70 vs. +70) series. The mean boundary value for the 11 infants who completed the +70 series was 29.5 msec compared to the 21.6 msec mean on the full VOT series. These data from the truncated +70 series are presumably more accurate estimates of the category boundary because the staircase step size was now set at 10 msec rather than the 20 msec used in the full VOT series. Second, note that the category boundary was shifted

slightly in the +50 series compared to the +70 series. The mean boundary value in the +50 series was 12.4 msec, significantly less than the 29.6 msec in the +70 series (t = 2.59, df = 13, p < .05). Despite the difference in the location of the boundary between the +70 and +50 series, these data provide further evidence that the boundary is located in the plus region of the VOT continuum when the stimulus endpoints are restricted to the -20 msec to +70 msec range.

Insert Table 1 About Here

Although not all infants provided boundary values on all five VOT series because of scheduling conflicts and satiation to the reinforcer, several infants proceeded from the +70 series to the two series in the minus region of the VOT continuum. Eight infants successfully completed the staircase phase in the two -50 series and three infants completed the staircase phase in the two -70 series. These data from individual infants are shown in Table 2. Note that 10 different infants provided boundary values for stimulus contrasts that were in the -70 to +20 msec range. Neither of these stimulus contrasts crossed the category boundary for voicing differences employed by adult speakers of English. Although these boundary values were more variable than the boundaries obtained with contrasts in the plus region of the VOT continuum, these infants nevertheless demonstrated reliable evidence of discriminating these phonologically-irrelevant VOT contrasts.

Insert Table 2 About Here

An important aspect of these data is the fact that they were obtained from individual infants, rather than the between-subject group data typically reported in past studies that have used high-amplitude sucking (HAS) and heart rate (HR) measures. Although there are clearly individual subject differences in the precise location of the boundary values shown in Figure 6 and Tables 1 and 2, there were a number of infants who were remarkably consistent across sessions. Figure 7 illustrates the staircase boundary data from an infant who proceeded through all five VOT series in just seven testing sessions. This infant's boundary values were consistently located near +20 msec for contrasts that crossed this VOT value (+14, +28 and +15 on sessions 2, 3 and 6). However, when the VOT series was shifted to the minus region of the VOT continuum, this infant showed boundary values that were consistently located near -20 msec (-21 and -26 msec on sessions 5 and 7). In addition, this infant was remarkably consistent in providing boundary values based on small differences in the reversal points during the staircase phase. The last four reversal points on all four truncated VOT series were separated by only 20 msec (2 steps). Although one cannot generalize from a single infant to the entire population of infants from English-speaking environments, this particular example illustrates the power of the head-turning technique combined with an adaptive testing

algorithm in obtaining reliable and consistent data on multiple sessions from attentive infants.

Insert Figure 7 About Here

A final comparison between the infants and the adults tested in the same staircase procedure is shown in Table 3. These comparisons must be made somewhat cautiously because of the low sample size for several of the VOT series from the infants. Nevertheless, there are consistencies in the data that suggest a tentative conclusion regarding the ability of infants and adults to discriminate differences in VOT. For adults, mean boundary values in the plus region of the VOT continuum were centered around 24 msec. For infants, the comparable mean boundary values were centered around 21 msec. And, for adults and infants, the mean boundary values in the minus region of the VOT continuum were centered around -23 msec and -16 msec, respectively.[7] These results appear to be remarkably similar. However, if one considers the staircase procedure to be a threshold task, then we have actually obtained estimates of the difference limens (DL) for each background VOT value. From these data one can compute a ΔVOT value for each contrast by calculating the difference between the boundary value and the background. Adults showed a ΔVOT value that was, across all contrasts, 20.1 msec less than the infants' ΔVOT value. Thus, the magnitude of VOT difference between the target and background that was required to discriminate the target at the

70.7% level was consistently greater for infants than for adults.

Insert Table 3 About Here

The ΔVOT values shown in Table 3 illustrate several important points regarding the infants' capacities to discriminate differences in VOT. Note that infants tested on the truncated series in the plus region of the VOT continuum (backgrounds of +70, +50 and +20) showed ΔVOT values that were very similar regardless of the initial target VOT value. The corresponding ΔVOT values on the truncated series in the minus region of the VOT continuum (backgrounds of -20, -50 and -70) were slightly higher. In addition, for the two contrasts with background VOT values of 0 msec, the ΔVOT value in the minus region of the VOT continuum was higher than in the plus region. Finally, at background VOT values of -70 and +70, the ΔVOT values were higher on the full series than on the truncated series. These findings imply that the initial contrast between the background and the target influenced the decision criterion used by the infant throughout the staircase testing session. This context effect was also evident in the data from the adults. Moreover, the adults, like the infants, showed slightly higher ΔVOT values in the minus region of the VOT continuum than in the plus region. Finally, the adults, but <u>not</u> the infants, showed a trough in their ΔVOT values between backgrounds of 0 and +70 msec, indicating greater sensitivity to differences in

this region. This enhanced sensitivity to VOT differences corresponds closely to a region of the VOT continuum utilized by English-speaking adults in making the phonemic distinction between voiced and voiceless stop consonants.

The results from this first experiment strongly suggest that infants from an English-speaking environment are capable of discriminating VOT differences that cross both the English adult voicing boundary and the voiced-voiceless unaspirated boundary employed phonemically in other languages such as Spanish and Thai. In addition, the infants tested in our procedure appeared to show less sensitivity to VOT differences that were located in the minus region of the VOT continuum compared to equivalent VOT differences located in the plus region. This differential sensitivity mirrors the performance of adults tested under identical experimental conditions. Despite the similarities between the adult and infant data, it could be argued that the infant data were potentially confounded by the effects of testing VOT differences in the minus region after having tested VOT differences in the plus region. The order of presentation of VOT contrasts across sessions may have biased the infants to attend primarily to voicing lags, and the repeated testing sessions may have satiated the infants to the testing situation, in general, and the visual reinforcer, in particular.

In an effort to provide a stronger test of the hypothesis that infants from an English-speaking environment are capable of discriminating VOT contrasts in the minus region of the VOT continuum, we conducted a second experiment with two additional

groups of infants. Each infant was tested on only two VOT series: (1) the full -70 vs. +70 msec series and (2) the -70 vs. 0 msec series (both counterbalanced for background VOT value). With the experimental conditions designed in this way, infants should have been less influenced by the order of presentation of the VOT contrasts. Moreover, with substantially fewer VOT contrasts, requiring fewer testing sessions, the infants should have become less satiated to the visual reinforcer.

Experiment 2

Method

Subjects. A total of 33 infants ranging in age from 6 to 11.5 months participated in Experiment 2. Infants were obtained from birth announcements in the local newspaper as in Experiment 1. Seventeen infants were assigned to the +70, -70 msec group (background = -70 msec) and 16 infants to the -70, +70 msec group (background = +70 msec). Twenty-two of the 33 infants completed the criterion testing phase and proceeded to the staircase testing phase.

Stimuli. The stimuli used in Experiment 2 were the same set of 15 synthetic bilabial stop consonant-vowel syllables used in Experiment 1.

Apparatus. The apparatus used in Experiment 2 was identical to that described in Experiment 1 (see Figures 3 and 4).

Procedure. The procedure used in Experiment 2 was identical to that used in Experiment 1 except that only two VOT series were tested with each infant. The infant was initially assigned to one of the two full VOT series (-70 vs. +70 or +70 vs. -70). The infant proceeded through the shaping phase to the criterion testing phase, which also employed an 80% correct criterion as in Experiment 1. The staircase phase immediately followed the criterion testing phase for all infants who had met the 80% criterion. However, if the infant failed to provide 5 or 6 reversals in the staircase phase within the same session as the criterion testing phase, they were not required to repeat the criterion testing phase again before proceeding to the staircase phase in the next session. Those infants who successfully completed 5 or 6 reversals in the staircase phase returned for additional testing on a single truncated VOT series. Infants assigned to the full series with -70 msec as the background were subsequently tested on the 0 vs. -70 msec series, which also had a -70 msec background. Infants assigned to the full series with +70 msec as the background were subsequently tested on the -70 vs. 0 msec series, which had a background of 0 msec. All infants were required to meet the 80% correct response level during the criterion testing phase on the truncated series (-70 vs. 0 or 0 vs. -70) before proceeding to the staircase testing phase.

Results and Discussion

Twenty-two of the 33 infants met the 80% criterion on one of the two full VOT series. The average number of sessions required to meet this criterion was 2.2 (range: 1-4). Although each of the 22 infants who met criterion was performing at the 80% correct level at the end of the criterion testing phase (i.e., the last 10 trials), it could be argued that their overall performance in reaching this criterion may have been below chance. To test this possibility, each infant's percent correct was computed for all trials during the session in which the criterion was met. This mean percent correct across all 22 infants was 79.2%, a value that is significantly above the chance response level of 50% ($t = 16.1$, $df = 21$, $p < .001$). Finally, the mean number of trials within the criterion testing phase required to meet the 80% criterion was 15.9 (S.D. = 6.3).

Fifteen of the 22 infants who met the 80% criterion completed the staircase testing phase on one of the two full VOT series. Individual boundary values for these 15 infants are shown in Table 4. The mean boundary value was 20.4 msec, a value that is nearly identical to the 21.6 msec boundary obtained earlier in Experiment 1. The average number of sessions required to complete the staircase phase was 2.9 (range: 1-4).

Insert Table 4 About Here

Twelve of the 15 infants who completed the staircase phase on one of the full VOT series met the 80% criterion on one of the two truncated series located in the minus region of the VOT continuum (-70 vs. 0 or 0 vs. -70). All three of the infants who did not meet the 80% criterion were unable to return for additional testing sessions, and thus were never actually tested on either of the truncated series. All 12 infants who were tested on the truncated series met the 80% criterion, and they did so within an average of 1.8 additional sessions (range: 1-4). Again, for the session during which the 80% criterion was met, the overall percent correct was 72.7%, a value that is significantly above the chance level of 50% (t = 8.2, df = 11, p < .001). Finally, the mean number of trials within the criterion testing phase required to meet the 80% criterion was 14.7 (S.D. = 5.8).

Six of the 12 infants who met the 80% criterion on one of the truncated VOT series completed the staircase testing phase. These data for individual infants are shown in Table 5. In addition to these boundary values, ΔVOT values for each of the four VOT contrasts are shown in Table 5. For comparison purposes, the ΔVOT values from Experiment 1 are also listed here. Note that these ΔVOT values are very similar to those obtained in Experiment 1, and that relatively large ΔVOT values were evident for background VOT values located in the minus region of the VOT continuum (-70 msec). Despite the small number of infants who completed the staircase phase on contrasts in the minus region of the VOT continuum, the data from

Experiments 1 and 2 provide very strong evidence that infants from an English-speaking environment are capable of discriminating VOT contrasts that do not cross the English adult voicing boundary. Of particular importance is the finding from Experiment 2 that all infants we tested met the 80% criterion on the VOT contrasts in the minus region of the continuum, and that their overall performance across all trials was significantly above chance.

Insert Table 5 About Here

General Discussion

The results of the present set of experiments not only add to the existing literature on infants' discrimination of VOT, particularly with regard to the performance of individual infants, but they also offer the first demonstration that infants from an English-speaking environment can reliably and consistently discriminate a phonologically-irrelevant VOT contrast. What factors are responsible for the failure of other investigators to show that infants from an English-speaking environment can discriminate VOT contrasts in the minus region of the VOT continuum? It appears that the acceptance of categorical perception as a basic mechanism in both adults' and infants' processing of VOT differences has led to an examination of very small VOT constrasts. The idealized version of

categorical perception implies that any VOT difference that crosses the category boundary will be discriminated. However, no study with adults has shown that observed perceptual performance matches this idealized conception of categorical perception. Discrimination is never perfect near the category boundary, and, in fact, it rarely exceeds 90% correct for contrasts as large as 20 msec. Given the rigorous task demands confronting infants in speech perception experiments, as well as the absence of any control over their attentive and motivational states, one would not expect their performance to approach that of adults. Thus, the robust evidence of VOT discrimination in the plus region of the VOT continuum indicates a remarkable sensitivity to this acoustic information even in infants. And, the absence of robust evidence of VOT discrimination in the minus region of the VOT continuum, at least in past studies, indicates that the acoustic information contained in such contrasts is less salient to both the infant and the adult. The present study corroborates these earlier findings, but the application of a more sensitive assessment technique provides strong evidence that contrasts in the minus VOT region <u>can</u> be discriminated reliably by prelinguistic infants.

We strongly suspect that the previous failures to find reliable evidence of discrimination in the minus VOT region were the result of two factors. First, the particular discriminative testing procedures may have been less sensitive than the adaptive testing methods we have used in the present experiments. Second, the reliance on group analyses of

discrimination in HAS, HR and previous operant headturning experiments may have obscured reliable performances by individual subjects. Moreover, the initial use of large VOT differences in our experiments overcomes many of the problems that may have suppressed the infant's performance, since the acoustic information which the infant must selectively attend to is made more explicit. Often in experiments with adults, subtle acoustic differences between complex signals are not initially detected. However, after a brief familiarization period, the acoustic cues may be isolated and excellent perceptual sensitivities can be demonstrated (see Pisoni, Aslin, Perey and Hennessy, Note 3). A similar process most likely occurred for the infants tested in the present study.

The generality of the present findings has been questioned on several grounds (Eilers, Gavin and Wilson, 1980). First, it has been suggested that demonstrations of unusual (i.e., non-phonemic) discrimination by a few selectively attentive infants cannot refute alternative evidence of "typical" (i.e., phonemic) discrimination by an unselected group of infants. However, the results of Experiment 2 leave little doubt that the discrimination of non-phonemic VOT contrasts is in fact quite representative rather than exceptional. Although our results suggest that VOT contrasts in the minus region of the VOT continuum are more difficult to discriminate than comparable differences in the plus region, all infants that we tested were able to provide reliable evidence of discriminating the -70 vs. 0 msec contrast. Moreover, the claim by Eilers and Gavin (in

press) that use of a running criterion, such as our 80% correct criterion, for each infant's preceding 10 trials, is not statistically defensible was mitigated by showing that across all trials in the process of meeting this 80% criterion the performance of the infants was significantly above chance ($p < .001$).

Second, it has also been suggested that our procedures, which require an initial shaping phase and the use of a fairly rigid criterion for accepting the hypothesis that individual infants discriminated a VOT contrast, may have provided the infant with sufficient listening experience to "have learned to perceive non-native VOT contrasts" (Eilers, Gavin and Wilson, 1980, p. 117). Such a suggestion is not only unwarranted given the robust nature of our findings, but it raises important interpretive and methodological questions. First, if infants are so influenced by short-term experiences in a laboratory setting, why are they not influenced by long-term linguistic experience (several months) in a natural setting? One would certainly anticipate that naturally occurring speech sounds presented in a full communicative context would be more influential to a phonetic processing mechanism than synthetic speech presented during a brief laboratory testing session. Moreover, the non-correspondences between infant and adult VOT sensitivities documented by Lasky, et al. (1975) and Streeter (1976) argue against significant experiential influences operating upon the infant's ability to discriminate VOT differences prior to 6 months of age. Second, if one sacrifices

methodological rigor in the testing of infants for purposes of assessing "typicality", then one is in essence abandoning the psychophysical approach to the study of infant sensory and perceptual abilities, an approach that has been extremely productive in areas other than speech perception (see Teller, 1979, for a recent review).

Eilers, Gavin and Wilson (1979) have also criticized the earlier finding by Eimas (1975), who reported that infants provided some evidence of discriminating a +10 vs. -70 contrast between two apical stops. They claimed that "this finding is difficult to interpret since the 80 msec difference between the members of the pair may be discriminable on irrelevant acoustic grounds (e.g., on the basis of perceived loudness)" (p. 17). A similar criticism could be raised concerning the present study since we initially used large VOT contrasts. However, this criticism implies that when infants discriminate phonemically-relevant contrasts, they are employing a phonetic rather than an acoustic mode of analysis. Clearly, such an implication is premature, particularly since there are several instances of non-correspondence between infants' discrimination performance and adults' labeling data for synthetic speech stimuli. A more parsimonious interpretation of all infant VOT studies to date is that all VOT contrasts are discriminated on the basis of acoustic differences rather than according to a phonetic mode of analysis. The phonetic-coding hypothesis simply has not received any unequivocal supporting evidence. Evidence to support this hypothesis in infants would require the

demonstration of a precise correspondence between the discriminative performance of infants and the phonemic categories employed in the infant's native language, a prediction that is not supported by the available findings in the literature.

The present results suggest that conceptions of speech perception in infants should be altered to eliminate simplified views of discrete perceptual categories. The discrimination of VOT differences can be more fruitfully conceptualized according to a sensitivity or ΔVOT function. Infants from an English-speaking environment are less sensitive to VOT differences in the minus region of the VOT continuum. English-speaking adults are also less sensitive in this region of the VOT continuum. However, both infants and adults from an English-speaking environment are capable of discriminating VOT differences in the minus VOT region at above chance levels. Until detailed psychophysical studies are conducted on samples of infants from language environments that do not employ minus VOT contrasts phonemically, it will be impossible to determine if the lower sensitivity in the minus VOT region is the result of selective early linguistic experience (see Aslin and Pisoni, 1980b). Nevertheless, at the present time, the weight of the evidence from infant and adult speech perception studies suggests that the lower sensitivity in the minus region of the VOT continuum is acoustically rather than phonetically based. For example, Pisoni (1977) has shown that adults exhibit categorical perception of pure tone nonspeech signals, and that

those signals which are analogous to minus VOT stimuli are less discriminable than are signals analogous to plus VOT stimuli. In addition, Stevens and Klatt (1974) and Lisker (1978) have correctly noted that syllable-initial stop consonants in the plus region of the VOT continuum contain additional cues (aspiration and absence of the F1 transition) that provide salient acoustic information not present in stops from the minus region of the VOT continuum (see also Summerfield and Haggard, 1977).

The ΔVOT results from the present study further suggest that linguistic experience may operate to enhance the basic sensitivity to VOT differences only at regions along the VOT continuum that are used phonemically (see Aslin and Pisoni, 1980b). Whether the complementary process of attenuation of sensitivity to VOT differences that are not used phonemically also operates during the period of early linguistic experience remains unclear at the present time. However, recent experiments with adults (Pisoni, Aslin, Perey and Hennessy, Note 3) strongly suggest that the basic auditory capacities to perceive phonemically-irrelevant VOT distinctions do not become eliminated by lengthy periods during which the distinction was not used contrastively in the productions heard by the developing child. Rather, such distinctions can be "reacquired", as evidenced by the accurate and reliable discrimination and identification functions obtained from adults after little or no training experience. It is the task of future studies to more accurately delineate the processes by

which sensitivity to phonetic contrasts changes during the period of early linguistic experience.

Reference Notes

1. Eimas, P. D. Infant speech perception: Issues and models. Paper presented at the C.N.R.S. Conference, Paris, June, 1980.

2. Klatt, D. H. Analysis and synthesis of CV syllables in English. Unpublished manuscript, M.I.T., 1978.

3. Pisoni, D. B., Aslin, R. N., Perey, A. J. and Hennessy, B. L. Identification and discrimination of a new linguistic contrast: Some effects of laboratory training on speech perception. Submitted to Language and Speech (1980).

4. Aslin, R. N., Perey, A. J., Hennessy, B. L. and Pisoni, D. B. Perceptual analysis of speech sounds by prelinguistic infants: A first report. Paper presented at the 94th meeting of the Acoustical Society of America, Miami, December, 1977.

References

Abramson, A. S. and Lisker, L. Discriminability along the voicing continuum: Cross language tests. Proceedings of the 6th International Congress of Phonetic Sciences. Prague: Academia, 1970.

Aslin, R. N. and Pisoni, D. B. Effects of early linguistic experience on speech discrimination by infants: A critique of Eilers, Gavin and Wilson (1979). Child Development, 1980, 51, 107-112. (a)

Aslin, R. N. and Pisoni, D. B. Some developmental processes in speech perception. In G. Yeni-Komshian, J. Kavanagh and C. A. Ferguson (Eds.), Child Phonology: Perception and Production. Volume I. New York: Academic Press, 1980. (b)

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

Cutting, J. E., Rosner, B. S. and Foard, C. F. Perceptual categories for music-like sounds: Implications for theories of speech perception. Quarterly Journal of Experimental Psychology, 1976, 28, 361-378.

Eilers, R. E. Infant speech perception: History and mystery. In G. Yeni-Komshian, J. Kavanagh and C. A. Ferguson (Eds.), Child Phonology: Perception and Production. Volume I. New York: Academic Press, 1980.

Eilers, R. E. and Gavin, W. J. The evaluation of infant speech perception skills: Statistical techniques and theory development. In R. Stark (Ed.), Language Behavior in Infancy and Early Childhood. New York: Elsevier, in press.

Eilers, R. E., Gavin, W. J. and Wilson, W. R. Linguistic experience and phonemic perception in infancy: A cross-linguistic study. Child Development, 1979, 50, 14-18.

Eilers, R. E., Gavin, W. J. and Wilson, W. R. Effects of early linguistic experience on speech discrimination by infants: A reply. Child Development, 1980, 51, 113-117.

Eilers, R. E., Wilson, W. R. and Moore, J. M. Speech discrimination in the language-innocent and the language-wise: A study in the perception of voice-onset-time. Journal of Child Language, 1979, 6, 1-18.

Eimas, P. D. Auditory and linguistic processing of cues for place of articulation by infants. Perception and 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 and Psychophysics, 1975, 18, 341-347. (a)

Eimas, P. D. Speech perception in early infancy. In L. B. Cohen and P. Salapatek (Eds.), Infant Perception: From Sensation to Cognition. Volume II. New York: Academic Press, 1975. (b)

Eimas, P. D. Developmental aspects of speech perception. In R. Held, H. Leibowitz and H. L. Teuber (Eds.), Handbook of Sensory Physiology. Volume VIII. New York: Springer-Verlag, 1978.

Eimas, P. D. and Miller, J. L. Contextual effects in infant speech perception. Science, 1980, 209, 1140-1141.

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

Eimas, P. D. and Tartter, V. C. On the development of speech perception: Mechanisms and analogies. In H. W. Reese and L. P. Lipsitt (Eds.), Advances in Child Development and Behavior. Volume 13. New York: Academic Press, 1979.

Fellows, B. J. Chance stimulus sequences for discrimination tasks. Psychological Bulletin, 1967, 67, 87-92.

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

Jusczyk, P. W. Infant speech perception: A critical appraisal. In P. D. Eimas and J. L. Miller (Eds.), Perspectives on the study of speech. Hillsdale, N. J.: Lawrence Erlbaum Associates, 1980.

Jusczyk, P. W., Rosner, B. S., Cutting, J. E., Foard, C. F. and Smith, L. B. Categorical perception of nonspeech sounds by 2-month-old infants. Perception and Psychophysics, 1977, 21, 50-54.

Jusczyk, P. W., Pisoni, D. B., Walley, A. and Murray, J. Discrimination of relative onset time of two-component tones by infants. Journal of the Acoustical Society of America, 1980, 67, 262-270.

Kewley-Port, D. KLTEXC: Executive program to implement the KLATT software speech synthesizer. In Research on Speech Perception: Progress Report No. 4. Bloomington, Indiana: Indiana University, 1978.

Klatt, D. H. Software for a Cascade/Parallel formant synthesizer. Journal of the Acoustical Society of America, 1980, 67, 971-995.

Kuhl, P. K. Speech perception in early infancy: The acquisition of speech-sound categories. In S. K. Hirsh, D. H. Eldredge, I. J. Hirsh and S. R. Silverman (Eds.), Hearing and Davis: Essays Honoring skerHallowell Davis. St. Louis: Washington University Press, 1976.

Kuhl, P. K. Predispositions for the perception of speech-sound categories: A species-specific phenomenon? In F. Minifie and L. Lloyd (Eds.), Communicative and Cognitive Abilities: Early behavioral assessment. Baltimore: University Park Press, 1978.

Kuhl, P. K. Perceptual constancy for speech-sound categories. In G. Yeni-Komshian, J. Kavanagh and C. A. Ferguson (Eds.), Child Phonology: Perception and Production. Volume I. New York: Academic Press, 1980.

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

Kuhl, P. K. and Miller, J. D. Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. Journal of the Acoustical Society of America, 1978, 63, 905-917.

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

Levitt, H. Transformed up-down methods in psychoacoustics. Journal of the Acoustical Society of America, 1970, 49, 467-477.

Lisker, L. In qualified defense of VOT. Language and Speech, 1978, 21, 375-383.

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

Lisker, L. and Abramson, A. S. The voicing dimension: Some experiments in comparative phonetics. Proceedings of the 6th International Congress of Phonetic Sciences. Prague, 1967.

Miller, J. D., Wier, L., Pastore, R., Kelly, W. and Dooling, K. 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, 410-417.

Miller, J. L. and Eimas, P. D. Organization in infant speech perception. Canadian Journal of Psychology, 1979, 33, 353-367.

Morse, P. A. Infant speech perception: Origins, processes and alpha centauri. In F. Minifie and L. Lloyd (Eds.), Communicative and cognitive abilities: Early behavioral assessment. Baltimore: University Park Press, 1978.

Morse, P. A. The infancy of infant speech perception: The first decade of research. Brain, Behavior and Evolution, 1979, 16, 351-373.

Morse, P. A. and Snowden, C. T. An investigation of categorical speech discrimination by rhesus monkeys. Perception and Psychophysics, 1975, 17, 9-16.

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

Pisoni, D. B. Identification and discrimination of the relative onset time of two-component tones: Implications for voicing perception in stops. Journal of the Acoustical Society of America, 1977, 61, 1352-1361.

Stevens, K. N. and Klatt, D. H. Role of formant transitions in the voiced-voiceless distinction for stops. Journal of the Acoustical Society of America, 1974, 35, 653-659.

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

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., Liberman, A. M., Harris, K. S. and Cooper, F. S. Motor theory of speech perception: A reply to Lane's critical review. Psychological Review, 1970, 77, 234-249.

Summerfield, Q. and Haggard, M. On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. Journal of the Acoustical Society of America, 1977, 62, 435-448.

Trehub, S. E. Reflections on the development of speech perception. Canadian Journal of Psychology, 1979, 33, 368-381.

Waters, R. S. and Wilson, W. A. Speech perception by rhesus monkeys: The voicing distinction in synthesized labial and velar stop consonants. Perception and Psychophysics, 1976, 19, 285-289.

Williams, L. The perception of stop consonant voicing by Spanish-English bilinguals. Perception and Psychophysics, 1977, 21, 289-297.

Wilson, W. R. Behavioral assessment of auditory function in infants. In F. Minifie and L. Lloyd (Eds.), Communicative and Cognitive Abilities: Early Behavioral Assessment. Baltimore: University Park Press, 1978.

Acknowledgement

This research was supported by research grants from NIH (HD-11915 and NS-12179) and NIMH (MH-24027 and MH-30424). The first author was also supported by an NICHHD Research Career Development Award (HD-00309). An earlier version of Experiment 1 was presented at the biennial meeting of the Society for Research in Child Development, San Francisco, March, 1979. We thank Jerry C. Forshee and David Link for their technical assistance, Diane Kewley-Port for assisting in the synthesis of the speech stimuli, Wendy Crawford and Kathy Mitchell for organizing and implementing the data collection process, and Nancy Layman for manuscript preparation. Address reprint requests to R. N. Aslin or D. B. Pisoni, Department of Psychology, Indiana University, Bloomington, Indiana 47405. * Beth Hennessy is now at the Institute of Child Development, University of Minnesota.

Footnotes

1. VOT refers to the delay between release from stop closure and the onset of laryngeal pulsing (voicing). In synthetic consonant-vowel (CV) syllables, VOT is realized acoustically by varying the delay between the onset of energy in the first formant and higher formants and the presence of aspiration noise during this interval.

2. Studies of speech sound identification have been attempted by Aslin, Pisoni, Hennessy and Perey (Note 4), Fodor, Garrett and Brill (1975) and Kuhl (1976, 1980), but the tasks developed to date for use with infants have not yet generated data that are directly comparable to adult identification data.

3. Lisker and Abramson (1964) in their cross-language study of voicing in 11 different languages noted that the majority of languages employ one or two contrasts in voicing with boundaries in two regions: a voiced-voiceless unaspirated boundary near -20 msec VOT and a voiceless unaspriated-voiceless aspirated boundary near +20 msec VOT. Thus, every language studied employed at least one voicing contrast characterized by leading, short lag, or long lag VOT values in consonant-vowel syllables.

4. Our use of the term phonetic-coding is based upon the conclusions and interpretations offered by several researchers in the infant and adult speech perception areas. Eimas (1974,

1975a), for example, has claimed that "the sounds of speech would appear to undergo some additional, specialized processing that permits the extraction of distinctive phonetic features" (1974, p. 513), and "There is now considerable experimental evidence indicating that very young, prearticulate infants are able to use a phonetic code in perceiving speech" (1975a, p. 341).

5. Of course, the shift in the region of peak VOT discriminability could simply reflect modifications in some psychophysical aspect of auditory perception that does not involve a subsequent interpretive level of analysis entailing the phonetic coding of these signals as speech sounds.

6. This attrition rate of 45% compares favorably to other methods used in testing speech discrimination with infants such as high amplitude sucking and heart rate. Although other investigators (Eilers, Wilson and Moore, 1978; Eilers, Gavin and Wilson, 1979; Kuhl, 1980) have reported virtually no drop out in their studies using the headturning technique, they have typically employed only a very small number of trials (6 or 8) or "easily" discriminable vowel contrasts. In contrast, our procedure uses a more extensive shaping phase to ensure that the headturning response is under stimulus control before actual data are collected. The reliability of the observer during the shaping phase was indicated by the low (12%) drop out rate for completing the criterion testing phase after proceeding from the

shaping phase.

7. Mean boundary values in the plus VOT region included the six contrasts that straddled the +20 msec VOT value (-70 vs. +70, 0 vs. +70, -20 vs. +50, +70 vs. 0, +50 vs. -20 and +70 vs. -70 msec). Mean boundary values in the minus VOT region included the remaining four contrasts that straddled the -20 msec VOT value (-50 vs. +20, -70 vs. 0, +20 vs. -50 and 0 vs. -70). The overall similarity between infants and adults in their mean boundary values for the plus and minus VOT regions is clearly misleading, since the variance in the boundary values contributing to these means is quite large.

Table 1

Staircase boundary values for individual infants tested on the two truncated series in the plus VOT region.

Target	Background	Subject	Session	Boundary
+70	0	052	5	+26
		065	3	+45
		068	5	+35
		070	4	+26
		072	5	+45
		075	3	+ 9
		078	4	+22
0	+70	020	6	+30
		037	3	+25
		044	3	+35
		071	3	+28
+50	-20	075	4	+14
		101	4	+37
		102	3	+16
-20	+50	020	11	- 7
		037	14	- 7
		071	6	+15
		109	4	+19

Table 2

Staircase boundary values for individual infants tested on the two truncated series in the minus VOT region.

Target	Background	Subject	Session	Boundary
+20	-50	052	7	+ 5
		070	5	+14
		072	6	+12
		078	6	+13
-50	+20	020	10	-42
		037	13	+ 4
		044	9	-38
		071	5	-21
0	-70	075	5	- 5
-70	0	071	7	-26
		081	5	-65

Table 3

Mean boundary values and ΔVOT values for infants and adults.

		Boundary		ΔVOT	
Target	Background	Infant	Adult	Infant	Adult
-70	+70	12.0 (12)	33.6 (10)	58.0	36.4
0	+70	29.5 (4)	42.9 (10)	40.5	27.1
-20	+50	4.9 (4)	33.2 (10)	45.1	16.8
-50	+20	-24.3 (4)	2.8 (10)	44.3	17.2
-70	0	-45.5 (2)	-42.3 (9)	45.5	42.3
+70	0	29.6 (7)	17.2 (10)	29.6	17.2
+50	-20	22.3 (3)	9.6 (10)	42.3	29.6
+20	-50	11.1 (4)	-8.9 (10)	61.1	41.1
0	-70	-5.0 (1)	-44.8 (9)	65.0	25.2
+70	-70	29.3 (15)	7.4 (10)	99.3	77.4

Numbers in parentheses indicate the number of subjects per condition.

Table 4

Staircase boundary values for individual infants tested on the full VOT series.

Target	Background	Subject	Session	Boundary
-70	+70	02	3	+28
		05	2	+40
		08	2	+32
		09	3	- 8
		10	4	+20
		I2	1	+14
		26	4	+18
		28	4	-12
		33	4	+10
+70	-70	11	1	+20
		13	4	+36
		22	4	-14
		24	2	+26
		29	3	+60
		30	3	+36

Table 5

Staircase boundary values for individual infants tested on the truncated VOT series and mean ΔVOT values.

Target	Background	Subject	Session	Boundary
-70	0	08	3	-54
		10	5	-51
		26	6	-63
0	-70	11	5	- 5
		22	6	- 6
		24	4	-12

Target	Background	ΔVOT Exp. 2	ΔVOT Exp. 1
-70	+70	54.2	58.0
-70	0	56.0	45.5
+70	-70	97.3	99.3
0	-70	62.3	65.0

Figure Captions

1. Labeling functions for synthetic bilabial stop consonant-vowel syllables in voice-onset-time for groups of English-, Thai- and Spanish-speaking adults [redrawn from Lisker and Abramson, 1967].

2. Broadband spectrograms of three test stimuli, -70, 0 and +70 msec, from the set of 15 synthetic consonant-vowel syllables varying in voice-onset-time from -70 to +70 msec.

3. Schematic representation of the overall testing environment including the location of the parent, infant and assistant within the sound-attenuated testing booth and the experimenter who controlled the experiment from outside of the booth.

4. Illustration of the testing situation inside the sound-attenuated booth showing the position of the parent, infant and assistant with respect to the location of the speaker and the visual reinforcer.

5. An example of an individual infant's responses during the adaptive staircase testing phase on a specific VOT contrast. Circles indicate test trials and squares indicate probe trials. A plus (+) indicates a correct headturn response and a minus (-) indicates the absence of a headturn.

6. Individual boundary values obtained from the 27 infants tested in the staircase phase on the full VOT series. Infants in Group I were presented with a constant background stimulus with a VOT value of -70 msec and a target stimulus whose VOT value was shifted from +70 msec VOT in 20 msec steps toward the background signal. Infants in Group II were tested under conditions identical to those in Group I except for the reversal of the background and target VOT values.

7. Boundary values obtained from a single infant tested in five sessions of the staircase phase on the full VOT series (background = +70 msec) and the four truncated VOT series (background = +70, +50, +20 and 0 msec).

LISKER & ABRAMSON (1967) CROSS-LANGUAGE LABELING DATA

LABIALS

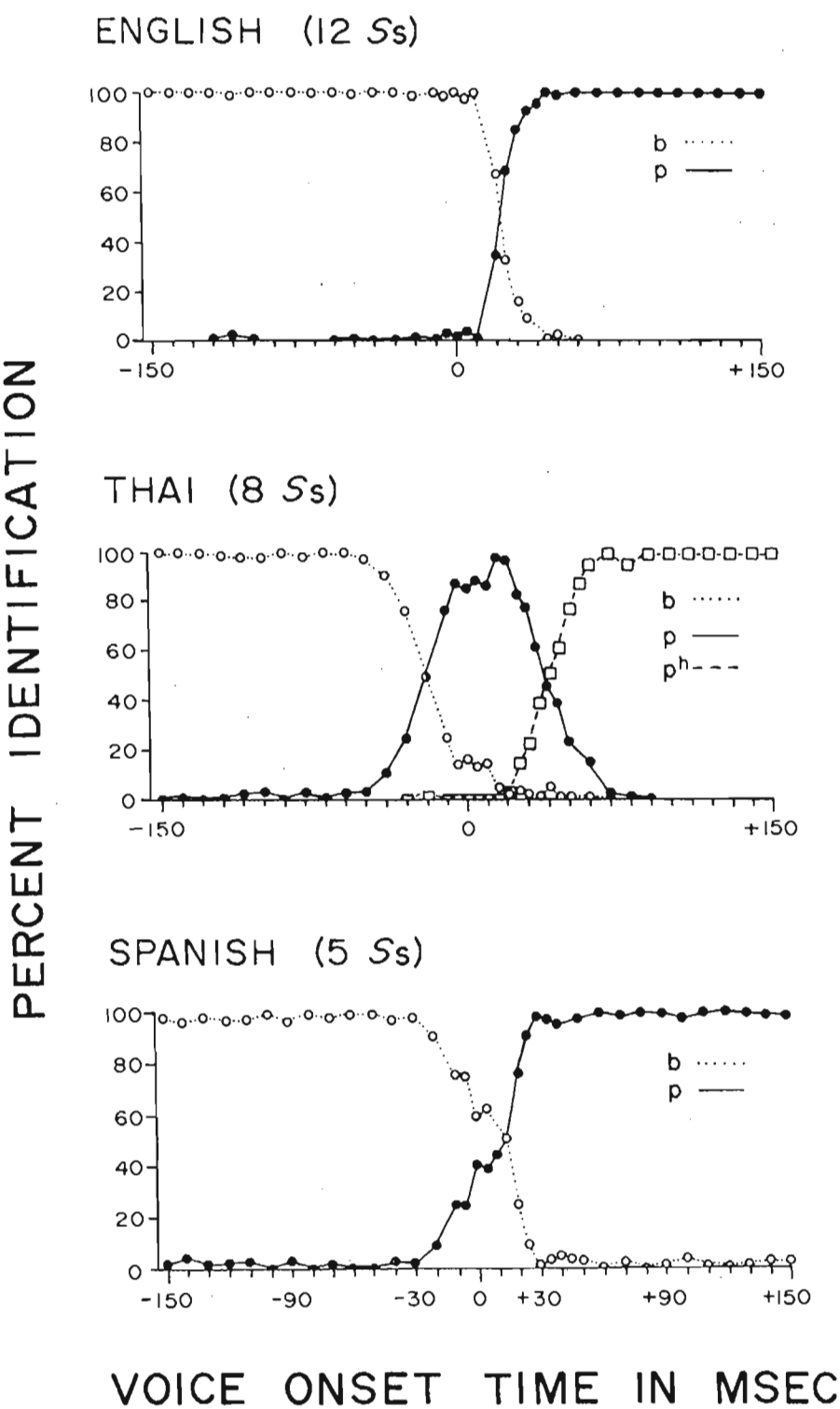


Fig. 2

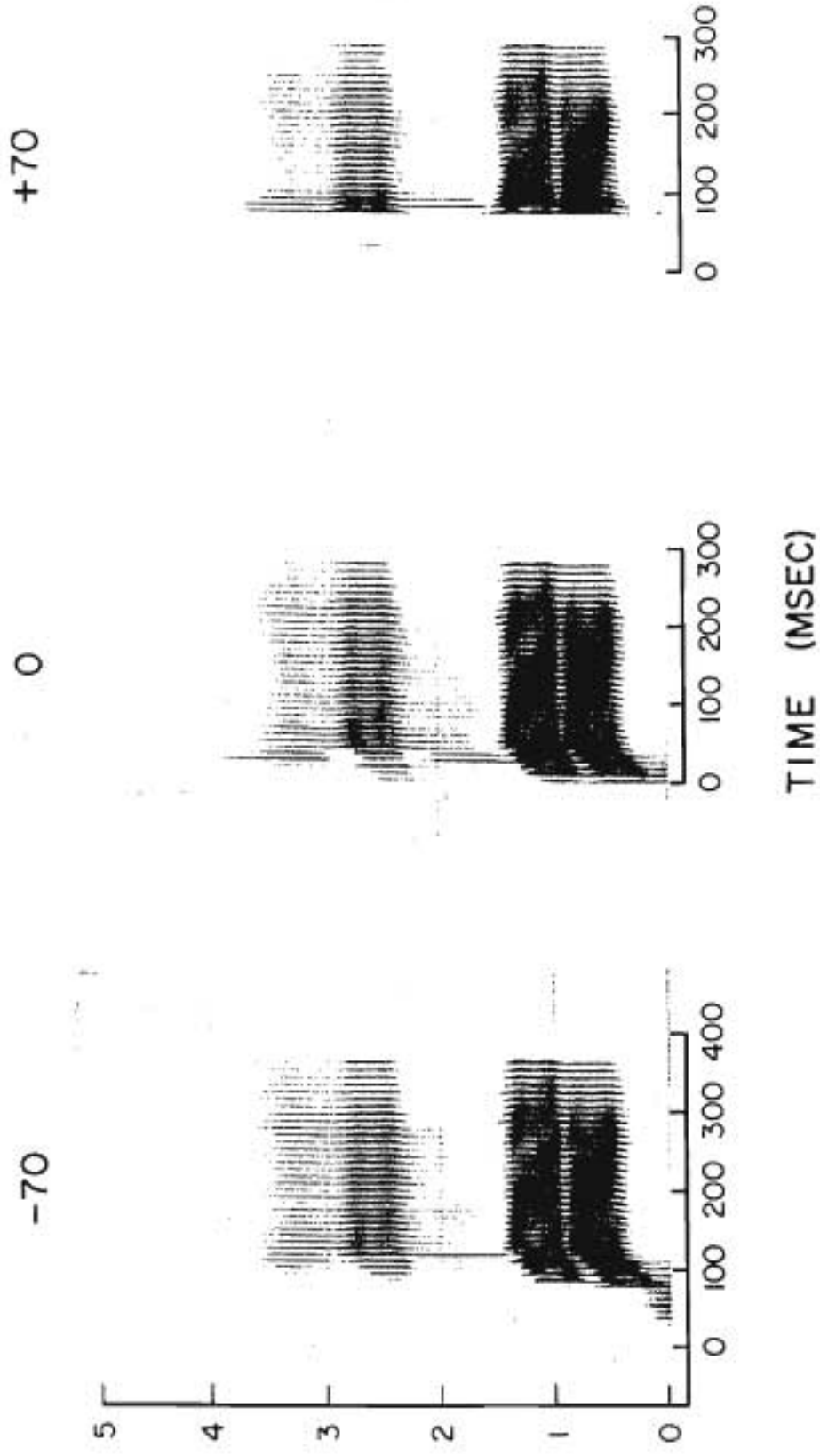

+70
0
-70
0 100 200 300
0 100 200 300
0 100 200 300 400
TIME (MSEC)
FREQUENCY (KHz)
0
1
2
3
4
5

Fig. 3

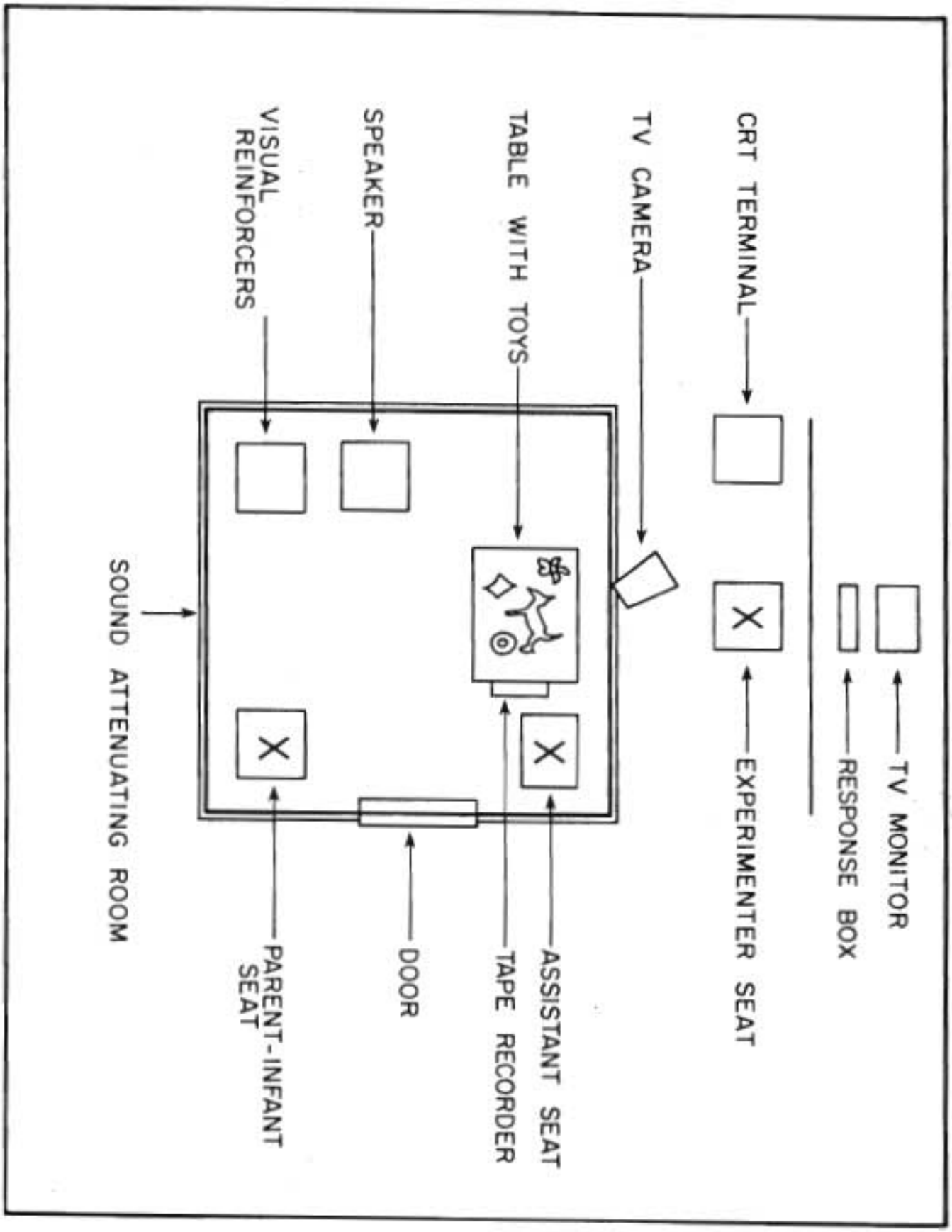

TV MONITOR
RESPONSE BOX
EXPERIMENTER SEAT
CRT TERMINAL
TV CAMERA
TABLE WITH TOYS
SPEAKER
VISUAL REINFORCERS
SOUND ATTENUATING ROOM
ASSISTANT SEAT
TAPE RECORDER
DOOR
PARENT-INFANT SEAT

Fig. 4

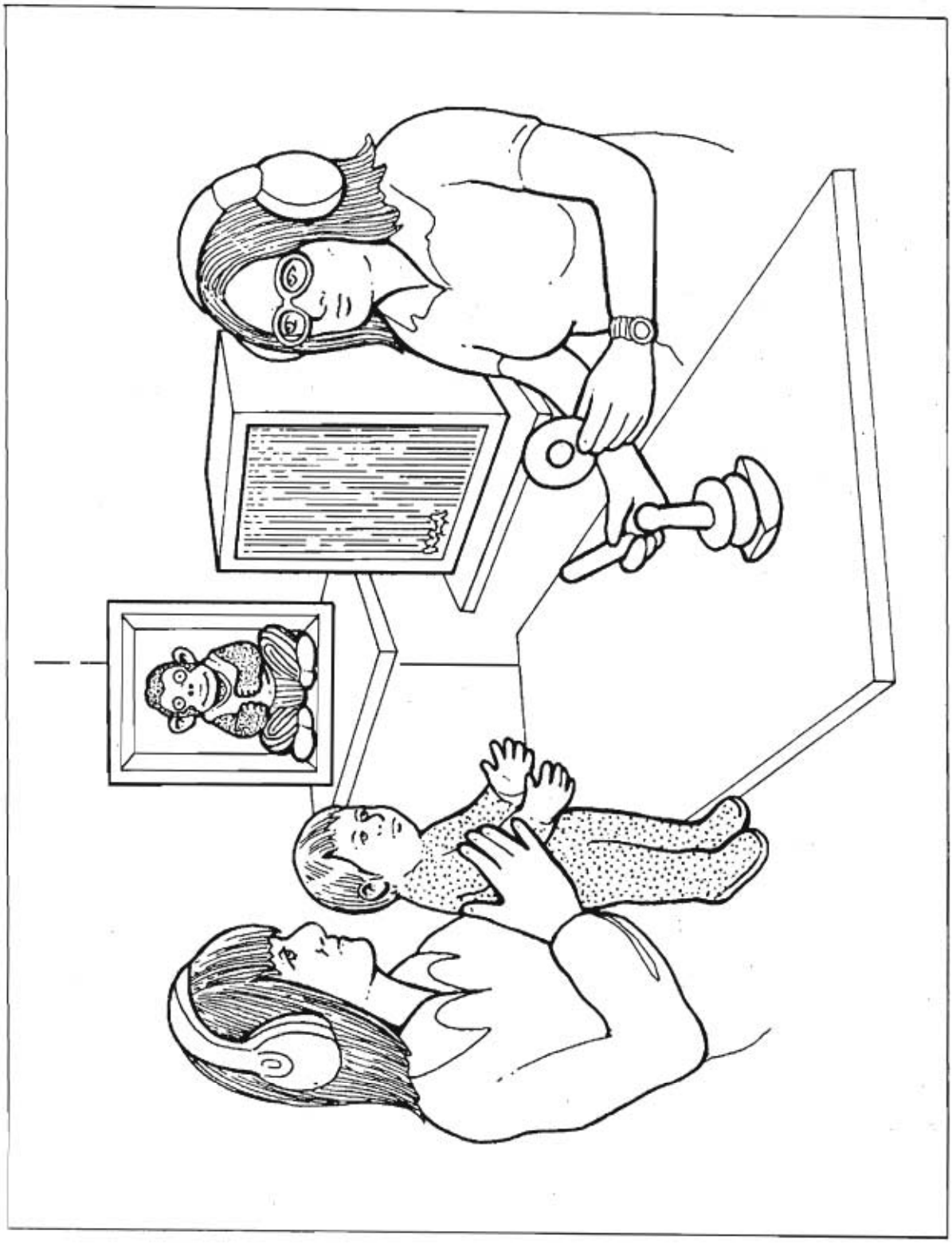
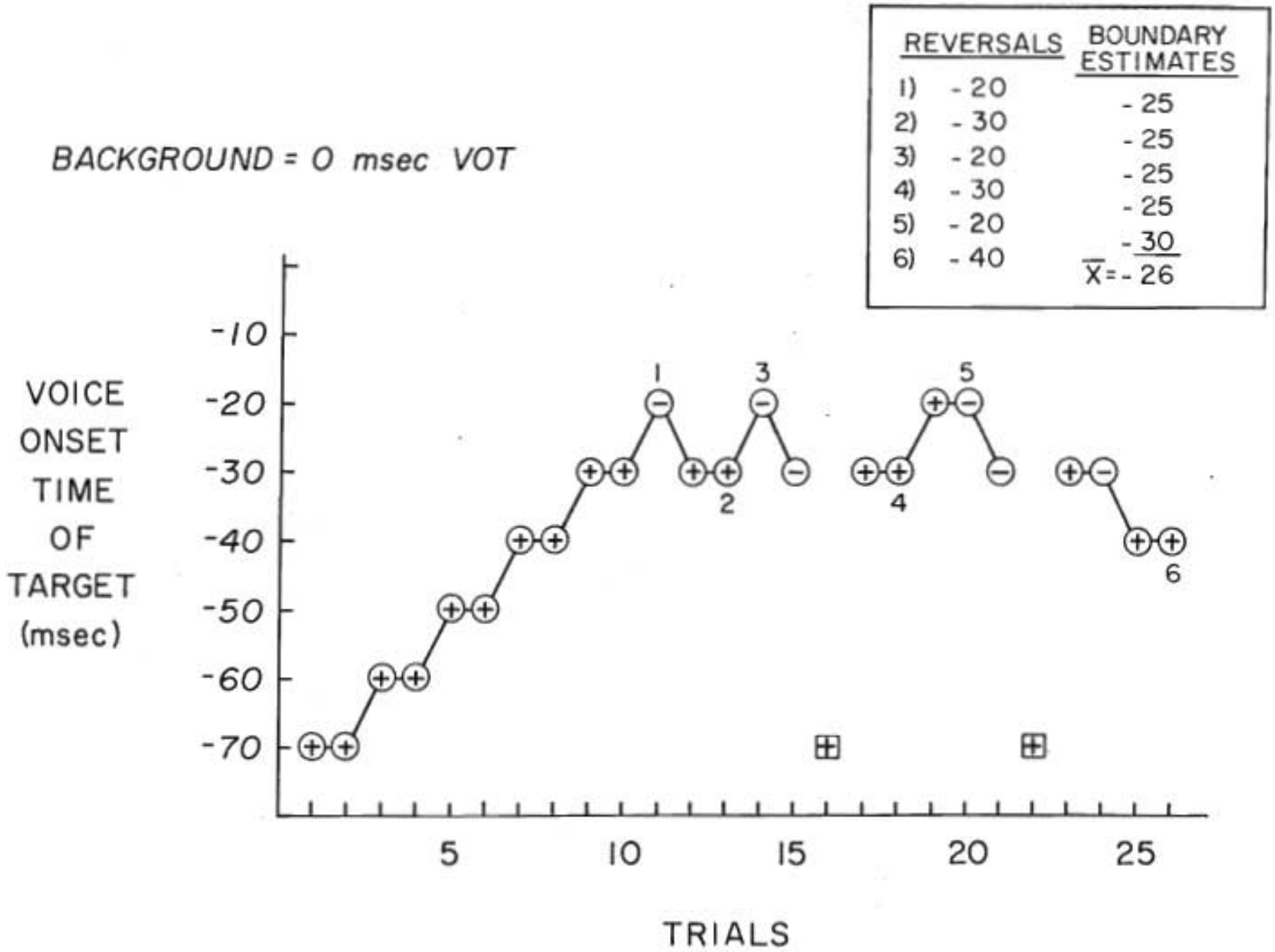
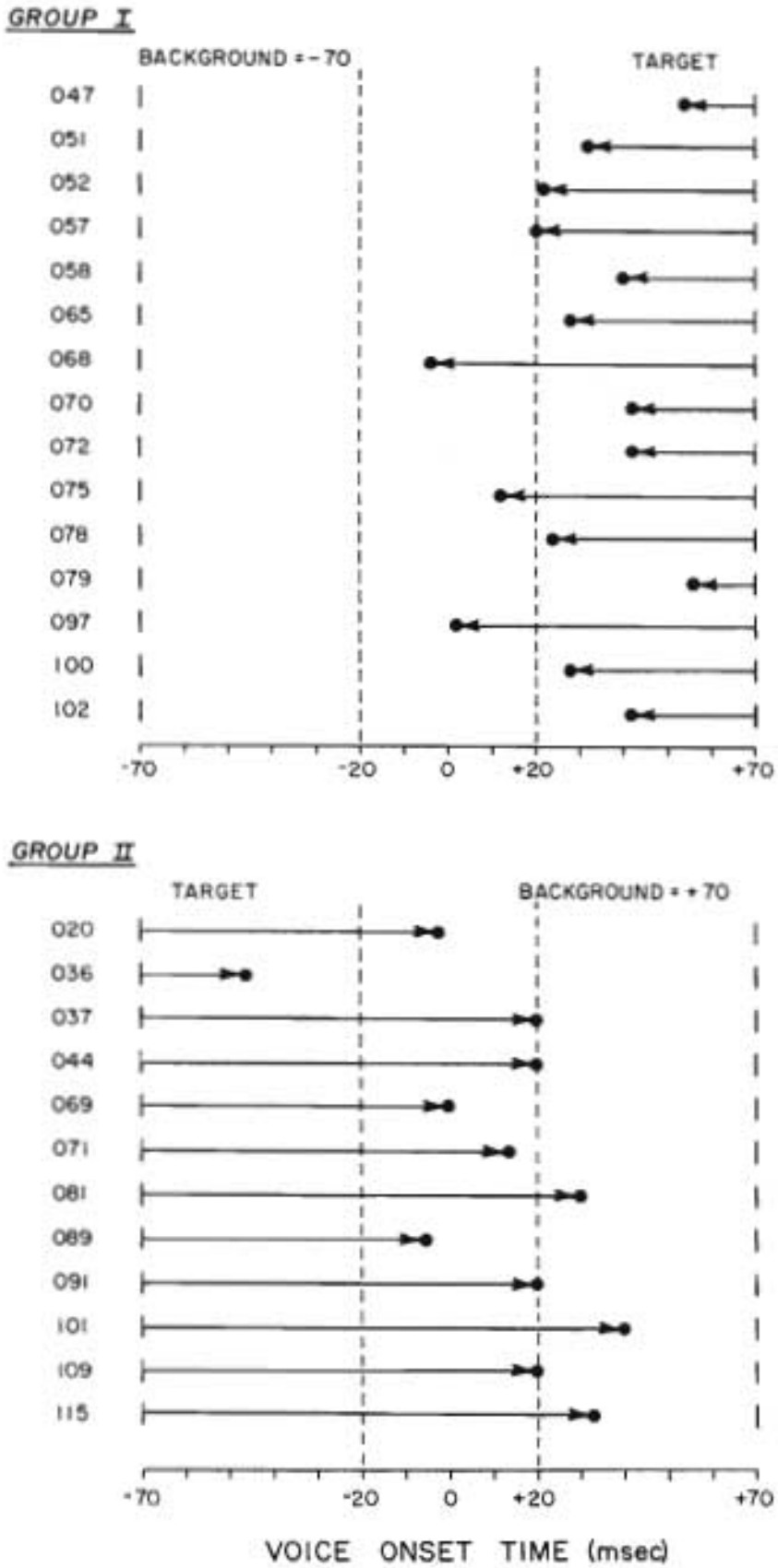


Fig. 5


BACKGROUND = 0 msec VOT
REVERSALS
BOUNDARY ESTIMATES
1) - 20
2) - 30
3) - 20
4) - 30
5) - 20
6) - 40
- 25
- 25
- 25
- 25
- 30
X̄ = - 26
VOICE ONSET TIME OF TARGET (msec)
-10
-20
-30
-40
-50
-60
-70
1
2
3
4
5
6
5
10
15
20
25
TRIALS


GROUP I
BACKGROUND = -70
TARGET
047
051
052
057
058
065
068
070
072
075
078
079
097
100
102
-70
-20
0
+20
+70
GROUP II
TARGET
BACKGROUND = +70
020
036
037
044
069
071
081
089
091
101
109
115
-70
-20
0
+20
+70
VOICE ONSET TIME (msec)

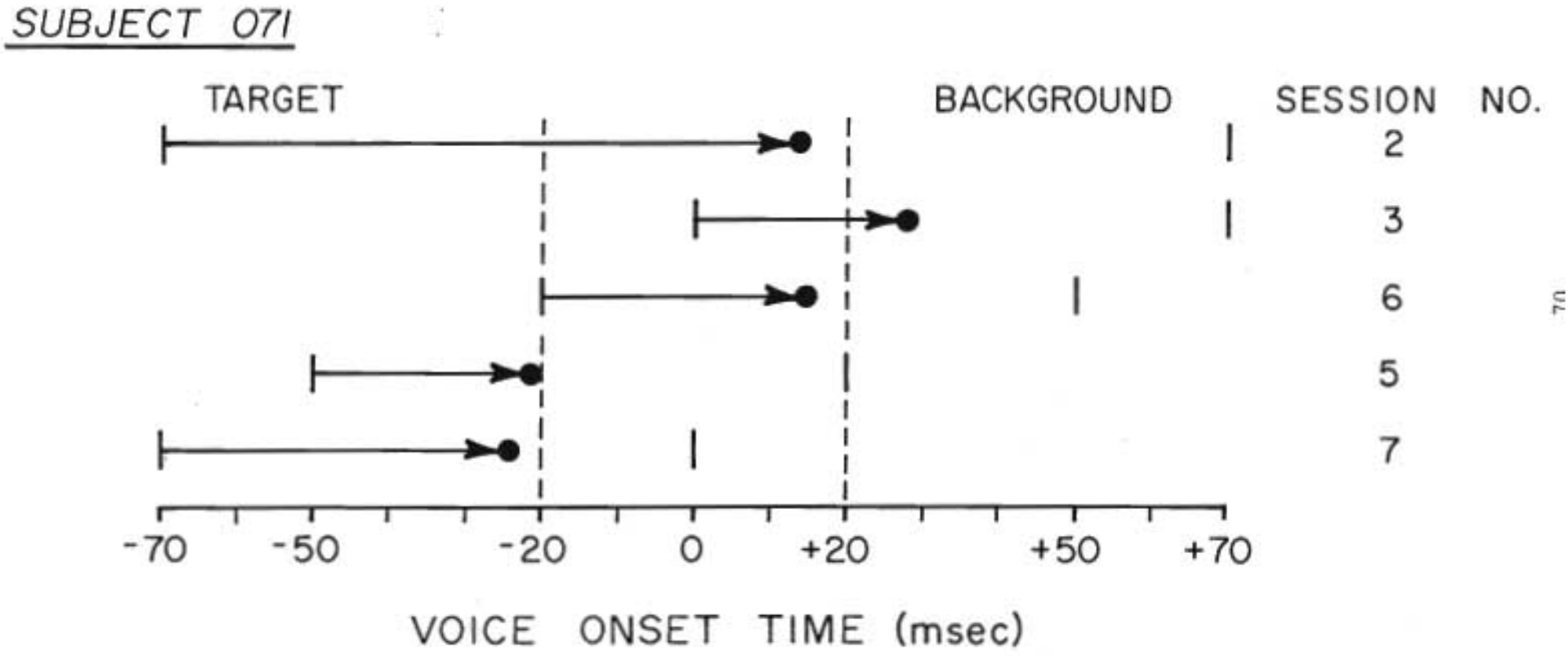

SUBJECT 071
TARGET
BACKGROUND
SESSION NO.
2
3
6
5
7
-70
-50
-20
0
+20
+50
+70
VOICE ONSET TIME (msec)

Fig. 7

Pure Tone Auditory Thresholds in Human Infants and Adults

Joan M. Sinnott

and

David B. Pisoni

Indiana University

Bloomington, Indiana 47405

This research was supported by grants from NICHHD (HD-11915-03) and NIMH (MH-24027-06) and by a postdoctoral traineeship awarded to JMS from NIH (PHS T32 NS7134-01). We thank Richard N. Aslin for valuable support and encouragement.

ABSTRACT

Pure tone auditory thresholds for frequencies from .250 to 8.0 kHz were obtained from human infants and adults for both 1.0 and .5 sec tone durations. An adaptive staircase (tracking) discrimination procedure was used in conjunction with a go - no-go version of an operant head-turning technique. Methods were devised for maintaining subjects under stimulus control during threshold testing. Overall, infant thresholds were 17-27 dB higher than adult thresholds. Adult audiograms were nearly flat between frequencies of .500 and 8.0 kHz with sensitivity ranging between 7 and 14 dB SPL. Infant thresholds were flat between frequencies of .500 and 4.0 kHz, with sensitivity ranging between 30 and 36 dB SPL. For infants, the most sensitive frequency was 8.0 kHz (25 dB SPL). Thresholds for 1.0 and .5 sec tones occurred within 5 dB of one another. Both sensory and attentional mechanisms may contribute to the observed diminished sensitivity of infants, relative to adults.

After a decade of infant speech research, important findings have emerged which are currently generating theoretical inquiry into the mechanisms governing phonological development (Aslin and Pisoni, 1980; Walley, Pisoni and Aslin, in press). For example, there is evidence that infants can discriminate some "foreign" speech contrasts which adults of the culture actually confuse (Trehub, 1976). Such a finding suggests that adults may lose a degree of sensitivity present in infancy. Yet other findings indicate that infants do not discriminate all the contrasts present in the native adult culture (Lasky, Syrdal-Lasky and Klein, 1975), arguing for some degree of either maturation or learning entering the developmental process.

However, before any definite theoretical conclusions can be offered concerning the development of infant speech discrimination, improved methods must be developed to directly compare discrimination capacities of infants and adults. Ideally, identical test procedures should be used and identical threshold criteria should be applied to both groups. It is difficult, if not impossible, to compare discrimination thresholds in infants and adults when infants are tested with sucking or heart rate techniques, and adults are tested on ABX or oddity discrimination procedures. Yet this has typically been the case in the infant speech perception literature. (See discussion of this problem in Jusczyk, Pisoni, Walley and Murray, 1980). Recently, the

development of an operant head-turning technique (Wilson, 1978) has provided a testing procedure which can be applied to infants and adults alike. In this procedure, a visual reinforcement (e.g. an animated toy) is paired with an auditory stimulus to be discriminated, and anticipatory head-turns by the subject to the stimulus alone are taken as an index of discrimination.

Using a go - no-go version of the head-turning technique, Aslin, Pisoni, Hennessey and Perey (Note 1) have made direct comparisons between infants and adults in speech discrimination capacities. Subjects listened to a repeating background stimulus representing one end of the voice-onset-time (VOT) continuum. They were then trained to make a head-turn response when the background changed to a stimulus from the opposite end of the continuum. When subjects had reached a criterion of >80% correct in a sequence of trials with randomly interleaved change and catch trials, VOT values between the end-points of the continuum were then presented for discrimination. An adaptive staircase (tracking) method was used based on earlier work of Levitt (1971). Staircase trials were presented according to a two-up, one-down algorithm, which specified that thresholds would converge upon the 70% detection point of the psychometric function. Checks on the infants' attentional state were made by objectively inserting "probe" trials, consisting of a maximum VOT difference, if two consecutive misses occurred in staircase

trials. Results showed that infants were sensitive to "phoneme boundaries" in both the prevoiced and voiceless regions of the continuum. Overall, however, infants were not as sensitive as adults to changes in VOT. Although delta-VOT thresholds varied somewhat for different regions of the continuum, adult thresholds averaged about 33 msec, while infant thresholds averaged about 53 msec.

A second methodology for making direct comparisons of infant and adult sensitivity has been recently developed by Trehub, Schneider and Endman (1980). These authors employed a go-right - go-left version of the head-turning technique to obtain absolute auditory thresholds for octave band noises. Subjects were reinforced for making appropriate head-turns to the right or left in response to stimuli randomly presented from either a right or left speaker. Such a forced-choice procedure allowed the auditory stimuli to remain on indefinitely, and did away with the need for catch trials. Stimuli were presented according to the method of constant stimuli, and fairly complete psychometric functions were obtained from infants and adults alike. However, since many infant functions never reached 100% correct detection for suprathreshold stimuli, the authors considered a lower value on the functions (65%) to represent infant thresholds, while the more conventional value of 75% was retained for adults. Using these analyses, infant thresholds were reported to be 20 - 30 dB higher than adults for frequencies below 4.0 kHz, but at higher frequencies

(10.0 kHz) infant sensitivity approached that of adults.

As this previous study verifies, experimenters are sometimes reluctant to apply identical threshold criteria to infants and adults, primarily because the human infant often has a tendency to resist all efforts to be brought under a reliable degree of stimulus control. Lapses of attention are common during testing. Yet, if valid sensory comparisons are to be made between infants and adults (as well as animals), behavioral techniques must be developed which will allow identical threshold criteria to be applied to all subjects. The present experiment represents another step in the evolution of the operant head turning technique, based primarily on the go - no-go procedure and staircase method of Aslin et al. (Note 1). Modifications were included relating to: a new method of assessing stimulus control during threshold testing; a more automated method of shaping; and a method of "punishing" false alarms. Using this procedure, infants and adults were directly compared in absolute auditory sensitivity to pure tone stimuli and identical threshold criteria were applied to both groups.

METHODS

Subjects

Subjects were 27 infants aged 7.2 - 11 months. Eleven were male and 16 were female. Three of these infants were naive, and the other 24 had had from 3 to 5 sessions of

previous testing experience in speech discrimination experiments. Seven infants had successfully completed testing in these experiments, and 17 had not. In assigning infants to the present experiment, no attempts were made to differentiate between those who had successfully completed the speech experiments and those who had not. All infants appeared in good health on test days. Infant's parents were paid $3 per visit to the laboratory. Adult subjects were 8 females and 1 male in their early twenties. All were students and/or assistants in the Infant Perception Laboratory at Indiana University, and volunteered their services. None reported any hearing difficulties at the time of testing.

Stimuli

Stimuli were sine wave tones of .250, .500, 1.0, 2.0, 4.0 and 8.0 kHz that were generated digitally on a PDP 11/34 computer by means of a specially designed software package (Kewley-Port, 1976). All tones had rise-fall times of 20 msec. Tones from .250 - 4.0 kHz were synthesized at a rate of 10.0 kHz, and during presentation were low-pass filtered at 4.8 kHz. The 8.0 kHz tone was synthesized at a rate of 20 kHz and presented through a 10.0 kHz low-pass filter. Two sets of tones were constructed, differing in duration. One set consisted of 1.0 sec tones from .250 - 4.0 kHz. The other set consisted of .5 sec tones from .500 - 8.0 kHz.

Apparatus

Digitized waveforms were permanently stored on disk and during the experiment they were transferred to the computer memory where they were accessable for on-line read out via a 12-bit D-to-A converter. The signals were then filtered, passed through a programmable digital attenuator (for on-line control), a manual attenuator, audio amplifier and loudspeaker (Radio Shack, No. 40-1980A), located inside the test booth.

A diagram of the testing area and arrangement of the apparatus is shown in Figure 1.

Insert Figure 1 about here

Sessions were conducted inside a single-walled (2m x 2m) sound attenuation chamber (IAC No. 402). In one corner was a chair for the parent and infant. Directly in front of this was a chair for the booth assistant and a table with toys. Underneath the table was a cassette tape recorder with two sets of headphones which provided masking music to the parent and assistant during all test sessions. Directly above the table was a closed-circuit TV camera which was focussed on the infant. To the left of the infant were the speaker and two visual reinforcers, one on top of the other, both enclosed in smoked plexiglass boxes so that they could not be seen until the lights within were illuminated. One

toy was a bear which beat a drum, and the other was a monkey which clapped its hands. Either reinforcer could be activiated via computer control. An inside view of the test booth in shown in Figure 2.

Insert Figure 2 about here

Outside the booth was a chair for the experimenter (E) and a table holding the TV monitor and the response box, which contained buttons which were sensed by the computer as input events to the threshold program. To E's left was a computer terminal which presented, after each completed trial, a record of the progress of the experiment, including trial number, intertrial interval duration, stimulus attenuation, response latency and occurrence of reversals during threshold testing. This information was also stored permanently on disk. The computer, which was located in an adjacent room, controlled all the experimental contingencies and recorded all the data.

Calibration

Calibration of the sound pressure level of the stimuli was accomplished with a General Radio meter (Type 1551-A, C scale), equipped with a circular rochelle salt crystal microphone (No. 9898). The meter was placed on the parent seat so that the microphone was pointed horizontally towards the speaker at the approximate position of the infant's

head. While reading the meter E sat in the assistant's seat. Six measurements were made approximately 2" apart in the general vicinity of the infant's head position, one measurement at right and left, front and back, and above and below center head position. The measurements are summarized in Table 1 for each of the six test frequencies.

Insert Table 1 about here

These measurements were made at 10 dB above the maximum level of the stimuli used during testing. The measurements were also verified on a daily basis with a portable Triplett Sound Level Meter (No. 370). The measurements from the Triplett never varied from those of the General Radio by more than 4 dB. For the final calibration, measurements from the General Radio meter alone were used.

Procedure

Each infant subject was seated on the parent's lap while the booth assistant played with toys to attract the infant's attention and maintain his/her gaze in a forward position. Adult subjects sat alone in the booth in the parent seat and assumed a slightly slouched position so that their head position approximated that of the infants, when viewed through the TV monitor. E viewed the subject (S) through the monitor and operated the response box which had three function buttons and two feedback lights. Button 1

was pressed by E to present trials to S. Button 2 was pressed whenever S made a head-turn. Button 3 was an abort button and was pressed only under unusual circumstances, for example to terminate a session if an infant began to cry. Light 1 was situated directly above button 1 and informed E when the program would allow trials to be initiated. Light 2 was located above button 2 and indicated when the program was delivering reinforcement to S.

The computer program which controlled the experimental contingencies consisted of a sequence of four states:

1) Intertrial interval: this state was normally of variable duration (10-15 sec). All lights on E's box were off. Button 1 was inoperative and E could not initiate trials. The purpose of this interval was to limit the rate at which E could present trials to S. Its variable duration ensured that E would not inadvertently provide timing cues to S during trial presentation. Button 2 was operative and each head-turn (false alarm) by S was recorded. Each head-turn response reset the intertrial interval to a maximum duration of 15 sec.

2) Observation interval: this state followed the intertrial interval and was of indeterminate length. During this state light 1 flashed at a rate of 5 Hz. Button 1 was operative and, if S was in a "ready" state, e.g. calm and maintaining a forward gaze, E pressed button 1 and the program presented a trial (see 3 below). During the

observation interval button 2 was operative and any head-turns by S prior to trial presentation (false alarms) returned S to the intertrial interval, which was reset to its maximum duration of 15 sec.

3) Trial interval: this state lasted for 3 sec and was cued by light 1 which assumed steady illumination for the duration of the trial. A 1.0 sec (or .5 sec) tone was followed by 2.0 sec (or 2.5 sec) of additional response time. Button 2 was operative and if a head-turn was recorded, the program proceeded to the reinforcement state (see 4). If no head-turn occurred, the program returned to the intertrial interval for another trial sequence.

4) Reinforcement state: this state lasted a total of 3 sec and was indicated by the illumination of light 2. The reinforcer inside the booth was activated and illuminated for S to view. Following this state, the program returned to the intertrial interval for another trial sequence.

These four basic states were used in the construction of two separate subroutines, which were distinguished on the basis of the types of trials which were presented.

Shaping routine: the purpose of the shaping routine was to train a naive infant S to turn his/her head to the tone stimulus, or, in the case of a trained S, to ensure that he/she had not forgotten previous training at the beginning of an experimental session. All experiments began

with the shaping routine, which presented two types of trials: 1) Probe trials - which presented suprathreshold tones, initially set to be 10 dB below the values presented in Table 1; and 2) Shaping trials - which were similar to probe trials except that the reinforcement was programmed to activate automatically at the termination of the 1.0 sec tone.

The shaping routine always began with the presentation of a probe trial. If the initial probe was not responded to, as was usually the case with a naive S, then the following trial sequence presented a shaping trial, in which the reinforcement occurred automatically. Normally, all Ss responded with a head-turn on a shaping trial. If S did not, shaping trials were presented repeatedly until a head turn response occurred. When this happened, the following trial sequence then presented a probe trial. The program recycled in this manner until two consecutive probe tones were responded to, at which point S entered the testing routine (see below). It should be noted that Ss who were already trained (as well as adults) received no shaping trials, since they would invariably respond to the first two probe trials and thus enter testing almost immediately. A state diagram of the shaping routine is shown in Figure 3 (top panel).

Insert Figure 3 about here

Testing routine: When S had responded to two successive probe trials in the shaping routine, he/she entered the testing routine. Here, three types of trials were presented: 1) test trials, 2) catch trials and 3) probe trials. For every sequence of eight trials, there were always four test trials, two probe trials and two catch trials. Their exact order of occurrence was randomly permuted after every sequence of eight, so that E was always blind as to what type of trial was being presented at any time.

Test trials presented tones according to an adaptive staircase (tracking) procedure. The first test tone was always presented at the same intensity level as the initial probe tones. Each test tone responded to caused the next scheduled test tone to be decremented by 10 dB. Each missed test tone caused the next one to be incremented by 10 dB. This one-up, one-down algorithm specified that the obtained threshold would converge on the 50% detection point of the psychometric function.

Catch trials were of equal duration (3.0 sec) as test trials, and the programmable attenuator was set to its maximum level (-127 dB) such that no tone was audible. If a head-turn occurred (false alarm), S was returned to the intertrial interval which was set to its maximum duration of 15 sec. If S did not respond to the catch trial (correct rejection), the immediately following intertrial interval

was set to a minimal value of 1.0 sec. This allowed another trial to be initiated almost immediately and thus ensured that long periods of time without audible signals would not occur during testing.

Probe trials presented suprathreshold stimuli and were similar to probe trials presented during shaping. However, during testing the intensity level of the probe tones was determined by S's prior response to test tones: probes were always set to be 20 dB above the last test tone correctly responded to. For example, if S had responded to a test tone of -40 dB, and then missed one at -50 dB, the next scheduled probe tone would be presented at a level of -20 dB. However, probe intensity levels could never exceed the initial intensity level used in shaping. A state diagram of the testing routine is shown in Figure 3 (bottom panel).

If S missed a probe trial during testing, further testing was discontinued and S was returned to the shaping routine. Normal shaping contingencies were in effect except for the following changes: the reinforcement was changed from bear to monkey or vice versa, and the probe tones were maintained at a level of 20 dB above the last test tone responded to. Furthermore, if S then missed two consecutive probes after re-entering shaping, the session was automatically terminated. However, if two consecutive probes were responded to after re-entering shaping, S would re-enter testing and continue where he/she left off. If a

second probe was missed during testing, the session was automatically terminated. In both of these cases, missed probes were taken as an indication that the infant was no longer operating under a reliable degree of stimulus control.

The experiment was also terminated automatically after six response reversals were obtained in the test trials. A threshold was calculated by averaging the six reversal points (Levitt, 1971). A threshold was considered valid if no more than one catch trial had been responded to, and no more than one probe trial had been missed during testing. Typically, four to six probe and catch trials occurred in a session. Therefore, a mistake on a single probe or catch trial would result in an error rate of 25% or less for these trials.

Infants and adults were tested in weekly sessions. No attempt was made to specify the order in which the various frequencies were tested. This was essentially random since a different frequency was tested each week and all subjects who were available for testing were presented with that frequency. Adults were tested for 5-10 sessions, and infants were tested for 3-12 sessions. Originally we planned to test all infants for at least four sessions, but three infants whose parents did not wish to continue were only tested for three sessions. After four sessions of testing, infants were not tested further if two consecutive

sessions occurred with terminations because of missed probe trials. Infants began a session with an alternate reinforcer each week. Of the 27 infant subjects, 16 were tested on 1.0 sec tones only. Eleven were tested with 1.0 sec tones and then transferred to .5 sec tones. Of the adults, 7 were tested with 1.0 sec tones, and three of these were then retested with .5 sec tones. Two adults were tested with .5 sec tones only.

RESULTS

Twenty-six out of 27 infants were successfully shaped and entered into testing in their first experimental session. One infant cried in his first session and the session was aborted prematurely. This same infant was successfully shaped in his second session. The mean number of trials spent in shaping before testing was initiated was 5.56, SD=4.25, Range = 2-21. It should be noted that the majority of infants were not naive to the experimental situation. For the three totally naive subjects (aged 6.7, 7.7, and 9.0 months), the number of shaping trials was 7, 21 and 4, respectively.

At least one session of usable threshold data was obtained from each infant tested. A record was set by one female infant (LH) from whom 10 consecutive sessions of usable data were obtained. Altogether, the 27 infants were tested for a total of 154 sessions. Of these, 91 were successful in yielding usable data, that is, the infant

missed no more than one probe trial and responded to no more than one catch trial during testing. Of these 91 usable sessions, 38 were "perfect" sessions, with 100% probe responses and 0% catch trial responses. Thirty-one were sessions with one catch trial response, and 12 sessions contained a single missed probe. Ten sessions contained both a missed probe and a catch trial response. Forty-five of the total 154 sessions did not yield usable data because more than one probe was missed (resulting in automatic termination), and 14 of the 154 sessions contained more than one catch trial response. Four sessions were aborted prematurely by E when the infant started to cry. Adult subjects never missed probe stimuli. Only one adult responded to a single catch trial during one test session.

Individual data

An example of a male infant's performance during the shaping and testing routines is shown in Figure 4.

Insert Figure 4 about here

This represents the subject's first session in the present experiment. The subject began the session in the shaping routine, and at trials 6 and 7 he responded to two consecutive probe trials and entered the testing routine at trial 8. During testing, his threshold was tracked with the one-up, one-down staircase algorithm until 6 response

reversals were obtained. Catch and probe trials were randomly interleaved with test trials. One catch trial was responded to at trial 20. No probe trials were missed. His reversal points were (in attenuation values): 60, 50, 70, 60, 70 and 60 dB. A threshold was calculated by averaging the 6 reversals for a mean of -61.6 dB, or 12.7 dB SPL. This subject was the most sensitive infant tested at 4.0 kHz.

Examples of individual pure-tone audiograms (thresholds as a function of frequency) are shown in Figure 5, for one sensitive adult (KD), one insensitive adult (SK), one sensitive infant (LH), one moderately sensitive infant (BH), and one insensitive infant (EG).

Insert Figure 5 about here

Also plotted for comparison purposes is the classic function of Sivian and White (1933) for binaural mimimum audible field data with single frequency tones (Group C), transformed into dB re .0002 dynes per cm 2. The sensitive adult overlapped the function of Sivian and White, but the insensitive adult approached the function of the sensitive infant. For both adults, no differences in sensitivity appeared with 1.0 and .5 sec tones. The infant LH was consisently 3-5 dB more sensitive with 1.0 sec than with .5 sec tones.

Average data

Mean thresholds and standard deviations for 1.0 sec tones from .250 to 4.0 kHz are shown in Table 2A, for both adults and infants.

Insert Table 2 about here

For infants, the mean probabilities of responding to probe trials (P) and catch trials (C) are also shown. For each frequency, infants responded to probe stimuli with a probability of greater than .90, while for catch trials, the probability was .10 or less. An additional analysis of the infant threshold data was performed using only "perfect" sessions (100% probe responses and 0% catch trial responses). These thresholds did not differ by more than 5 dB from the analyses derived from all sessions. Average infant and adult audiograms for 1.0 sec tones are displayed graphically in Figure 6 (top panel).

Insert Figure 6 about here

Both audiograms were flat from .500 to 4.0 kHz, with thresholds at .250 kHz slightly elevated. The infant audiogram was 21 to 27 dB higher than the adult.

Mean thresholds and standard deviations for .5 sec tones (.500 - 8.0 kHz) are shown in Table 2B. Infant

thresholds were 17-25 dB higher than those of adults. Average audiograms are shown in Figure 6 (bottom panel). Adult audiograms were flat from .500 to 8.0 kHz, while infants showed a slight increase in sensitivity at 8.0 kHz, relative to the other frequencies. 8.0 kHz was the only frequency tested in which the standard deviations of the infant and adult populations overlapped slightly, since one infant (LH) had a lower threshold (11 dB) than two of the adults tested (14 and 16 dB).

DISCUSSION

Methodology

Infants in the present experiment appeared to be operating under a high degree of stimulus control during threshold tracking, so that over 90% of the probe trials were responded to, and less than 10% of the catch trials were responded to. Such results compare favorably with stability criteria required from individual subjects in animal psychophysical experiments (e.g. Stebbins, 1970; Sinnott, Beecher, Moody and Stebbins, 1976). High false alarm rates, resulting in unusable data, occurred primarily in the initial sessions of testing. It is conceivable that "punishing" false alarms by returing the infant to the intertrial interval, and thus delaying further trial presentation, may have had some effect in extinguishing this behavior. This "timeout" strategy was borrowed directly from animal psychophysics (Stebbins, 1970).

After the initial sessions, as false alarms declined, it then became increasingly important to prevent the head turn response from extinguishing. Loss of stimulus control occurs frequently with animals on tracking procedures (Stebbins, 1970; Harrison and Turnock, 1975). The strategy of randomly interleaving suprathreshold probe stimuli with test stimuli served two purposes; first, these probe stimuli provided infants, as well as adults, with more salient discriminations during difficult periods of tracking stimuli near threshold. However, since probe stimuli were never set to be more than 20 dB above the test stimuli, these discriminations were not so salient as to distract subjects from making the more difficult ones. A second function of the probe stimuli was to catch moments of infant inattentiveness, after which an attempt was made to recapture the infant by returing to the shaping routine. If the infant had simply experienced a transitory lapse of attention, for example, had gotten too involved with the antics of the booth assistant (but was not permanently satiated), then upon returing to shaping, the infant would view the alternate reinforcer. The booth assistant would view this as well, and would proceed to reduce her activity. In 22 sessions, infants were successfully revived in this way. On the other hand, if the infant was becoming permanently satiated with both reinforcers, then "reshaping" was to no avail and the infant would continue to disregard probes, as well as shaping stimuli.

One observation from the present experiment indicated that infant inattentiveness to the stimuli was actually less of a problem than false alarms. For example, if sessions are examined where infants made only one "mistake", there were actually more sessions with a catch trial response (31 sessions) than with a missed probe (12 sessions). This suggests that, at least before the point in time when the infant became permanently satiated with the reinforcer (and had to be dropped from the experiment), temporary lapses of attention occurring on a moment-to-moment basis during testing did not appear to be a major probem in this experiment.

One final important aspect of the present methodology was that it provided a totally objective criteria for terminating an experimental session. Infant studies in the literature sometimes report that sessions were terminated when infants began to fuss or not attend to the assistant (e.g. Eilers, Gavin and Wilson, 1979). Allowing E to decide to terminate a session could potentially lead to experimenter bias, since some subjects might be terminated and data not analysed, while others might be left in the session, but would not appear to be optimally sensitive. In the present study, only four experimenter-induced terminations took place and all were initiated when infants began to cry. All other sessions were automatically terminated by the computer according to the criteria outlined earlier in the procedures.

Auditory sensitivity

The average audiogram obtained from adults in the present study did not provide a good match to the classic data of Sivian and White (1933). At 2.0 kHz, our average adult threshold was about 15 dB higher than that of Sivian and White. The adult audiograms of the present study appeared quite flat in the range from .500 - 8.0 kHz, while the Sivian and White audiogram shows an increase in sensitivity at 2.0 and 4.0 kHz. There are a number of obvious reasons for this mismatch. Sivian and White conducted their studies in an anechoic chamber which provided complete sound isolation as well as reduced acoustic refections. In contrast, our single-walled booth was far from ideal and although every attempt was made to reduce extraneous noise during testing, this was sometimes impossible due to activity in neighboring rooms. Therefore, we may have been measuring masked thresholds rather than absolute thresholds. Secondly, our calibrations may have been slightly inaccurate. When SPL measurements were being made, there was no adult sitting in the parent seat, since the SPL meter had to be placed there. Therefore, higher frequencies may have been refected differently in the calibration and test situations. Finally, our adult subjects were not repeatedly tested in an attempt to obtain their most sensitive thresholds at each frequency. Since this was not feasible with infants, we wished to maintain both groups under testing conditions as comparable as

possible.

The shapes of our audiograms are more similar to those of Trehub et al. (1980), since these latter also appear quite flat from .500 through 8.0 kHz. Overall, the subjects in Trehub's experiment appear slightly more sensitive than ours. One reason for this might be that Trehub used octave band noises which would be less susceptible to acoustic reflections in the test chamber than the single frequency tones of the present study. Also, the forced-choice aspect of the Trehub experiment allowed the stimuli to remain on indefinitely until the subject responded. In the present study, stimuli were only 1.0 or .5 sec long and subjects were allowed only three seconds to respond.

No indications of developmental changes in low frequency sensitivity were found in the present study, as Trehub et al. reported for their infant subjects. It is possible that the age range covered by our subjects was not large enough for such a trend to clearly emerge. Trehub found the greatest difference in sensitivity between the ages of 6 and 12 months. We did not test many infants less than 7 months, or greater than 10 months of age. However, the present study provides support for another finding of Trehub that relates to the convergence of infant and adult sensitivity as frequency was increased (see also Schneider, Trehub and Bull, 1980). Whereas our adult audiogram was flat from .500 - 8.0 kHz, our infant audiogram showed a

slight increase in sensitivity at 8.0 kHz, relative to the other frequencies. It is conceivable that infants, with this relative gain in high frequency sensitivity, may be exhibiting a more "primitive" characteristic of primate hearing, according to the phylogenetic trends observed by Masterton, Heffner and Ravizza (1969). These authors reported that, in the primate line leading to man, e.g. from prosimians to monkeys to apes to man, there is a progressive loss in high frequency sensitivity.

We did not observe any reliable differences in auditory sensitivity between infants who had successfully completed the previous speech discrimination experiments and those who had not. Although the infant shown in Figure 4 had successfully completed speech testing, another infant (not shown) who had also completed went on to exhibit the highest auditory thresholds obtained from any infant in the present study (54 and 55 dB SPL at .250 and .500 kHz, respectively). None of the infants shown in Figure 5 had completed the speech experiments, including LH, the most sensitive infant. In general, it appeared that many infants who did not complete speech testing were subsequently able to complete at least one or two sessions of pure tone threshold testing. Thus, the present pure tone detection task may have involved an overall easier discrimination than the speech tasks. Also, the methodological modifications in the pure tone detection task may have aided in obtaining additional data from infants who were unable to complete speech testing.

Clearly, more research in this area is necessary. For future infant studies, we are considering a strategy of "screening" naive infants with a pure tone detection task before assigning them to a speech discrimination test. This would serve to adapt the infants to the general auditory test situation as well as determine if their hearing was within normal limits. Such a strategy might aid in reducing the high attrition rates sometimes observed in infant speech studies.

Infant vs. adult differences

It is of interest to consider possible hypotheses for the diminished auditory sensitivity of infants, relative to adults. First of all, it should be emphasized that reliable infant "go" and "no-go" responses to probe and catch trials (respectively) during threshold testing indicated that their diminished sensitivity was not due to lapses of attention, or loss of stimulus control in general. Therefore, one possibility is that there are real sensory differences between infants and adults. Supporting this hypothesis is a study by Suzuki and Taguchi (1968) in which cerebral evoked potentials to auditory stimuli were recorded in a number of infants, children and adults. Subjects were put to sleep by sedation and presented with free-field stimuli in a soundproof room. Threshold data indicated that, with a stimulus consisting of 1.2 kHz tone pips presented at a level of 30 dB SPL, 80% of the adults exhibited an evoked

response. However, only 67% of the children (1-5 years of age) responded, and only 50% of the infants (.5 - 3 months of age) responded. Furthermore, adult evoked responses were of the order of 32 microvolts, while the infant and child responses were only 22-26 microvolts. On the basis of these data, the authors concluded that infants required a stimulus roughly 20 dB more intense than adults for eliciting a response of equivalent amplitude.

A second possibility is that threshold differences between infants and adults in the present study were related to differing attentional mechanisms elicited by the test situation. One important difference between infants and adults was that the adults were presumably selectively attending to the auditory stimuli. For them, the situation was akin to a vigilance task where they had nothing much to do except wait for tone presentations. In contrast, the infants were being visually entertained during threshold testing. Infants therefore may have been in a situation demanding divided attention between the visual and auditory stimulation. One classic theory of human selective attention maintained that unattended stimuli were somehow "attenuated" relative to attended stimuli (Treisman, 1964). (For a recent review considering evidence for human performance decrements during tasks requiring divided attention, see Navon and Gopher, 1979).

Implications and conclusions

Differences in sensory and/or attentional mechanisms between infants and adults affecting basic auditory acuity may contribute to differences observed between these groups in speech discrimination (Aslin et al., Note 1). Therefore, one implication of the present study is that hearing sensation levels for infants in psychophysical, as well as speech experiments should be set approximately 20 dB above those of adults. In general, it is also clear that much data is needed concerning the more basic auditory-sensory capacities of infants. This is particularly important now that studies of adult speech perception are implicating psychophysical mechanisms (Pisoni, 1978; Stevens and Blumstein, 1978).

So far, speculations as to mechanisms of infant speech perception have included both "phonetic feature detectors" (Eimas, 1978) and "general mammalian mechanisms" (Kuhl, 1979). The former view maintains that infants process phonetic units in a manner similar to adults, while the latter view argues that infant processing is similar to that of a chinchilla or monkey. Both of these theoretical views emerged in the complete absence of any precise, quantitative data based on direct auditory-sensory-threshold comparisons between infants and adults or animals. We are optimistic that the use of the operant head-turning technique will lead to further methodological improvements in infant auditory

testing. Ultimately, these should yield a hard-core data base of infant discrimination capacities for speech, as well as non-speech stimuli. Such a data base will provide a much needed ingredient for future theories of infant speech perception.

REFERENCE NOTE

1. Aslin, R. N., Pisoni, D. B., Hennessy, B. L, and Perey, A. J. 1980. Discrimination of voice-onset-time by human infants: New findings concerning phonetic development. Submitted to Child Development.

REFERENCES

Aslin, R. N. and Pisoni, D. B. 1980. Some developmental processes in speech perception. In Child Phonology: Perception and Production, edited by G. Yeni-Komshian, J. F. Cutting and C. A. Ferguson, Academic Press, New York.

Eilers, R. E., Gavin, W. J. and Wilson, W.R. 1979. Linguistic experience and phonemic perception in infancy: A cross-linguistic study. Child Development, 50, 14-18.

Eimas, P. D. 1978. Developmental Aspects of Speech Perception. In Handbook of Sensory Physiology: Perception, Vol. 8, edited by R. Held, H. Leibowitz and H. L. Teuber. Springer-Verlag, New York.

Harrison, J. M. and Turnock, M. J. 1975. Animal psychophysics: Improvements in the tracking method. Journal of the Experimental Analysis of Behavior, 23, 141-147.

Jusczyk, P. W., Pisoni, D. B., Walley, A. and Murray, J. 1980. Discrimination of relative onset time of two-component tones by infants. Journal of the Acoustical

Society of America, 67, 262-270.

Kewley-Port, D. 1976. A complex tone generating program. In Research on Speech Perception: Progress Report No. 3. Indiana University, Bloomington, Indiana.

Kuhl, P. K. 1979. Models and mechanisms in speech perception. Brain, Behavior and Evolution 16, 374-408.

Lasky, R. E., Syrdal-Lasky, A. and 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, 215-225.

Levitt, H. 1971. Transformed up-down methods in psychoacoustics. Journal of the Acoustical Society of America 49, 467-477.

Masterton, B., Heffner, H. and Ravizza, R. 1969. The evolution of human hearing. Journal of the Acoustical Society of America 45, 966-985.

Navon, D. and Gopher, D. 1979. On the economy of the human processing system. Psychological Review 86, 214-255.

Pisoni, D. B. 1978. Speech perception. In Handbook of Learning and Cognitive Processes, Vol. 6, edited by W. K. Estes. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Schneider, B., Trehub, S. E. and Bull, D. 1980. High frequency sensitivity in infants. Science 207, 1003-1004.

Sinnott, J. M., Beecher, M. D., Moody, D. B. and Stebbins, W. C. 1976. Speech sound discrimination by monkeys and humans. Journal of the Acoustical Society of America 60, 687-695.

Sivian, L. J. and White, S. D. 1933. On minimum aubible sound fields. Journal of the Acoustical Society of America 4, 288-321.

Stebbins, W. C. 1970. Studies of hearing and hearing loss in the monkey. In Animal Psychophysics: the Design and Conduct of Sensory Experiments, edited by W. C. Stebbins, Appleton Century Crofts, New York.

Stevens, K. N. and Blumstein, S. K. 1979. Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America 64, 1358-1368.

Suzuki, T. and Taguchi, K. 1968. Cerebral evoked responses to auditory stimuli in young children during sleep. Annals of Otology, Rhinology and Laryngology 77, 102-110.

Trehub, S. E. 1976. The discrimination of foreign speech contrasts by infants and adults. Child Development 47, 466-472.

Trehub, S. E., Schneider, B. A. and Endman, M. 1980. Developmental changes in infants sensitivity to octave band noises. Journal of Experimental Child Psychology 29, 282-293.

Treisman, A. M. 1964. Selective attention in man. British Medical Bulletin 20, 12-16.

Wilson, W. R. 1978. Behavioral assessment of auditory function in infants. In Communicative and Cognitive Abilities - Early Behavioral Assessment, edited by F. D. Minifie and L. L. Lloyd, University Park Press, Baltimore.

Walley, A. C., Pisoni, D. B. and Aslin, R. N. The role of early experience in the development of speech perception. In Sensory and Perceptual Development, edited by R. N. Aslin, J. Alberts and M. R. Petersen, Academic Press, New York (in press).

Table 1

Calibrations in dB SPL of the test area

	Frequency (kHz)					
	.250	.500	1.0	2.0	4.0	8.0
Mean	79.0	70.1	69.3	72.3	74.3	69.8
SD	2.37	2.71	3.08	2.66	2.25	2.40

Table 2A

Mean thresholds and standard deviations for infant and adult subjects for 1.0 sec tones.

	Frequency (kHz)				
	.250	.500	1.0	2.0	4.0
Infants					
N	13	13	11	14	16
Mean	37.7	30.3	31.2	36.1	34.3
SD	8.53	11.8	10.2	9.01	12.3
P	1.00	.96	.94	.92	.92
C	.04	.07	.06	.09	.10
Perfect Sessions					
N	9	5	5	5	4
Mean	40.1	29.2	33.6	31.8	31.7
SD	8.91	12.4	13.6	4.92	15.2
Adults					
N	7	7	7	7	7
Mean	15.5	8.50	8.00	12.4	7.29
SD	9.61	5.80	7.42	8.73	5.53
Diff	22.2	21.8	23.2	23.7	27.0

Table 2B

Mean thresholds and standard deviations for infants and adults for .5 sec tones.

	Frequency (kHz)				
	.500	1.0	2.0	4.0	8.0
Infants					
N	5	7	2	5	5
Mean	33.6	35.0	36.5	31.2	25.0
SD	4.98	7.51	10.6	7.79	9.97
P	1.00	.95	1.00	.93	.94
C	.06	.09	.20	.07	.09
Adults					
N	5	5	5	5	5
Mean	11.0	9.20	14.2	7.60	7.2
SD	5.48	10.4	9.23	9.04	8.07
Diff	22.6	25.8	22.3	23.6	17.8

FIGURE LEGENDS

1. Diagram of the testing area.

2. Inside view of the test booth.

3. State diagrams of the shaping and testing routines. ITI = intertrial interval; OBS = observation interval; RF = reinforcement; PRB = probe trial; TEST = test trial; CATCH = catch trial.

4. Trial-by-trial record of an experimental session from one infant, showing performance in the shaping routine and threshold tracking during the testing routine. Test trials are linked by a solid line. Numbers (1-6) with test trials indicate occurrence of response reversals.

5. Individual audiograms obtained from three infants and two adults.

6. Average audiograms for infants and adults for 1.0 sec and .5 sec tones.

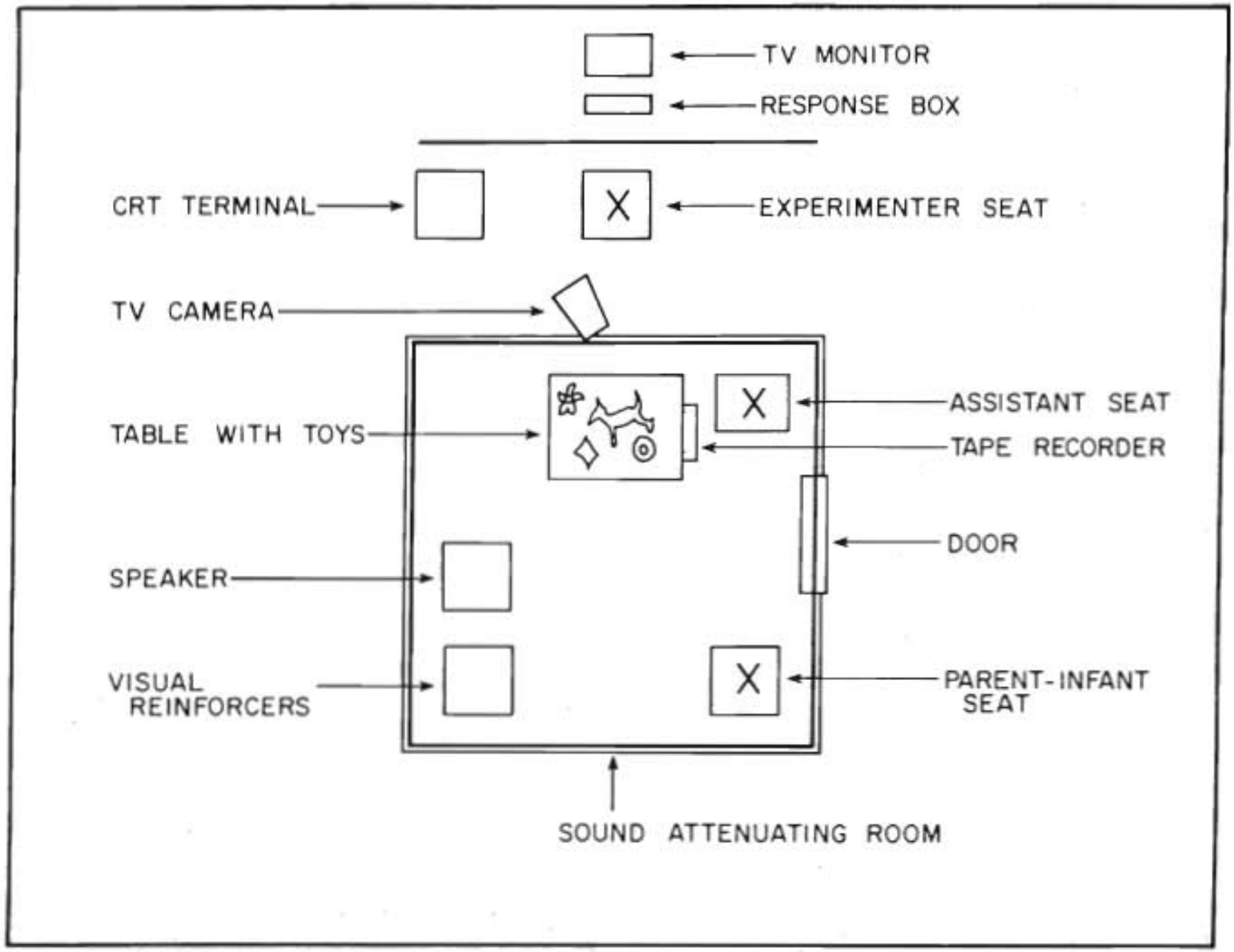

TV MONITOR
RESPONSE BOX
CRT TERMINAL
EXPERIMENTER SEAT
TV CAMERA
ASSISTANT SEAT
TABLE WITH TOYS
TAPE RECORDER
DOOR
SPEAKER
VISUAL REINFORCERS
PARENT-INFANT SEAT
SOUND ATTENUATING ROOM

Figure 1.

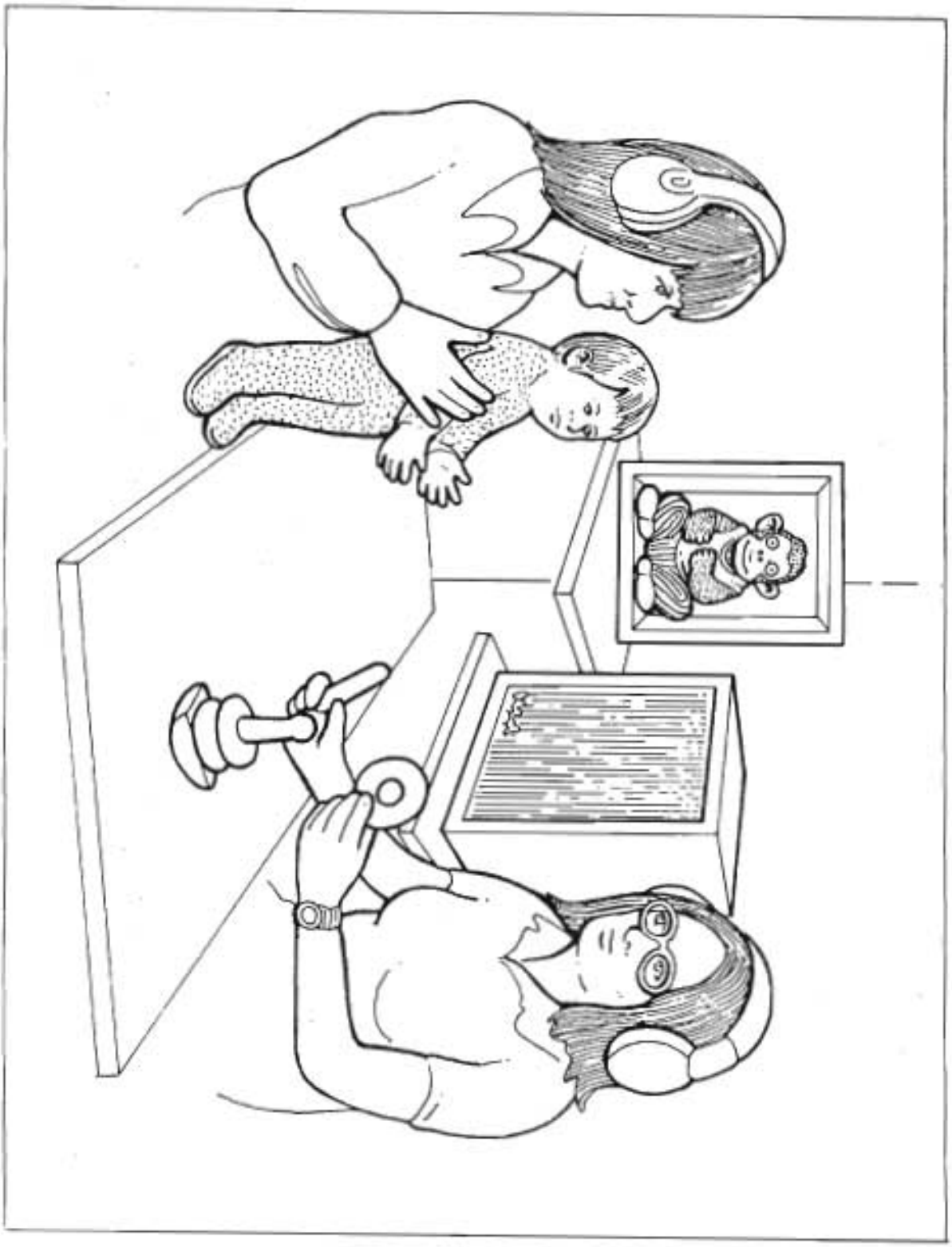


Figure 2.

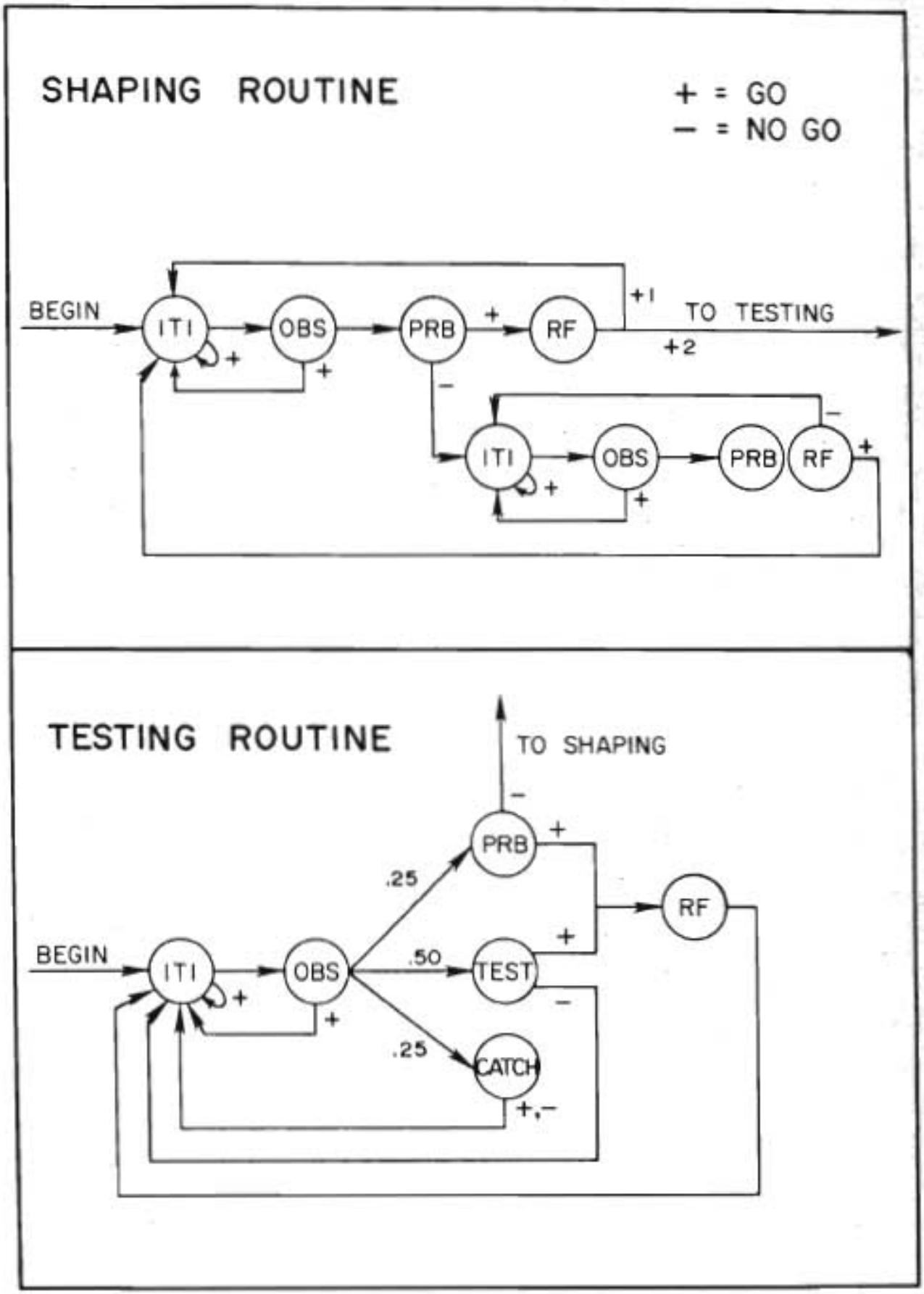

SHAPING ROUTINE
+ = GO
— = NO GO
BEGIN
ITI
OBS
PRB
RF
+1
TO TESTING
+2
TESTING ROUTINE
TO SHAPING
.25
.50
TEST
CATCH
+,—

Figure 3.

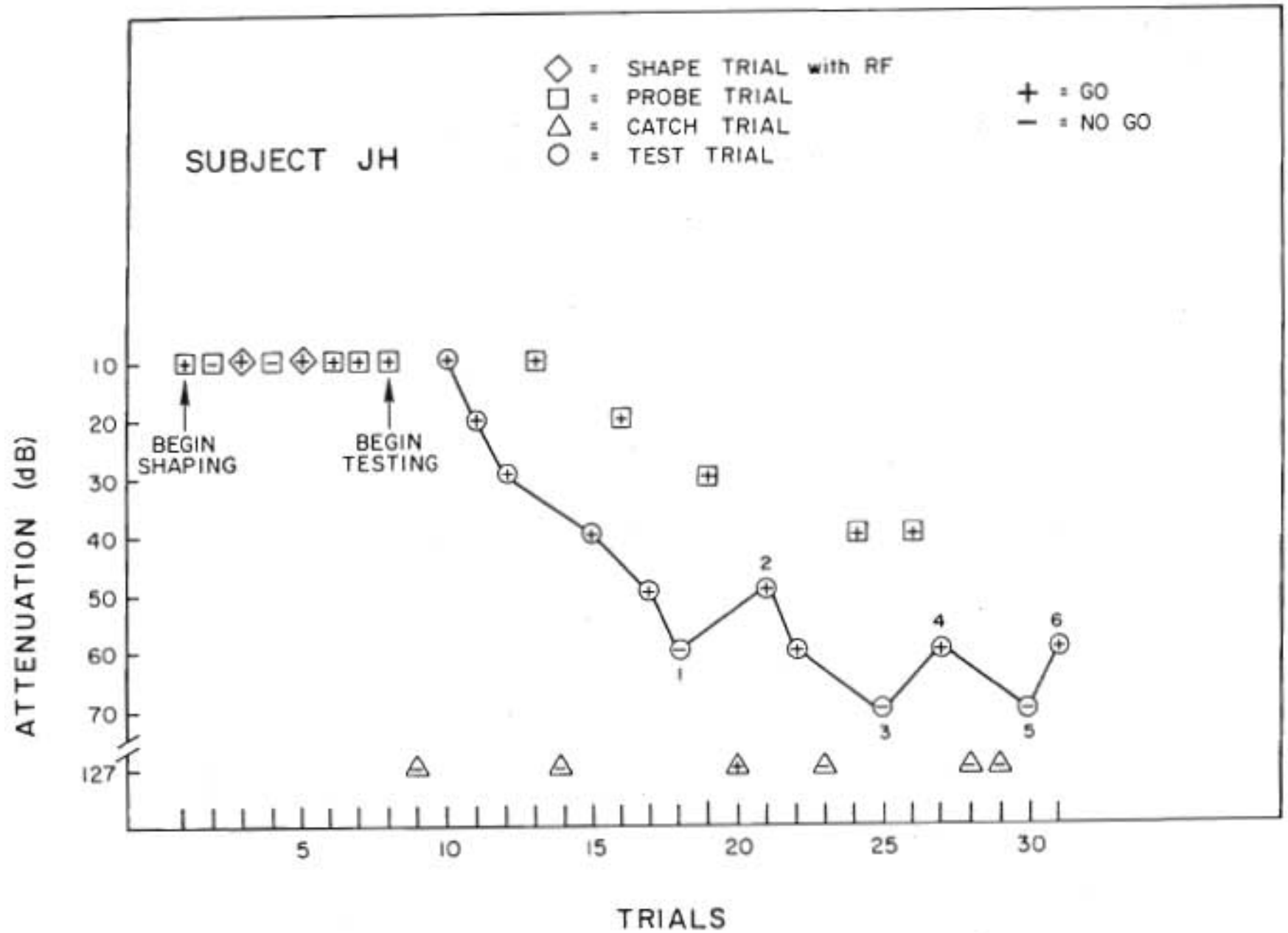

◇ = SHAPE TRIAL with RF
□ = PROBE TRIAL
△ = CATCH TRIAL
○ = TEST TRIAL
+ = GO
— = NO GO
SUBJECT JH
BEGIN SHAPING
BEGIN TESTING
ATTENUATION (dB)
10
20
30
40
50
60
70
127
1
2
3
4
5
6
5
10
15
20
25
30
TRIALS

Figure 4.

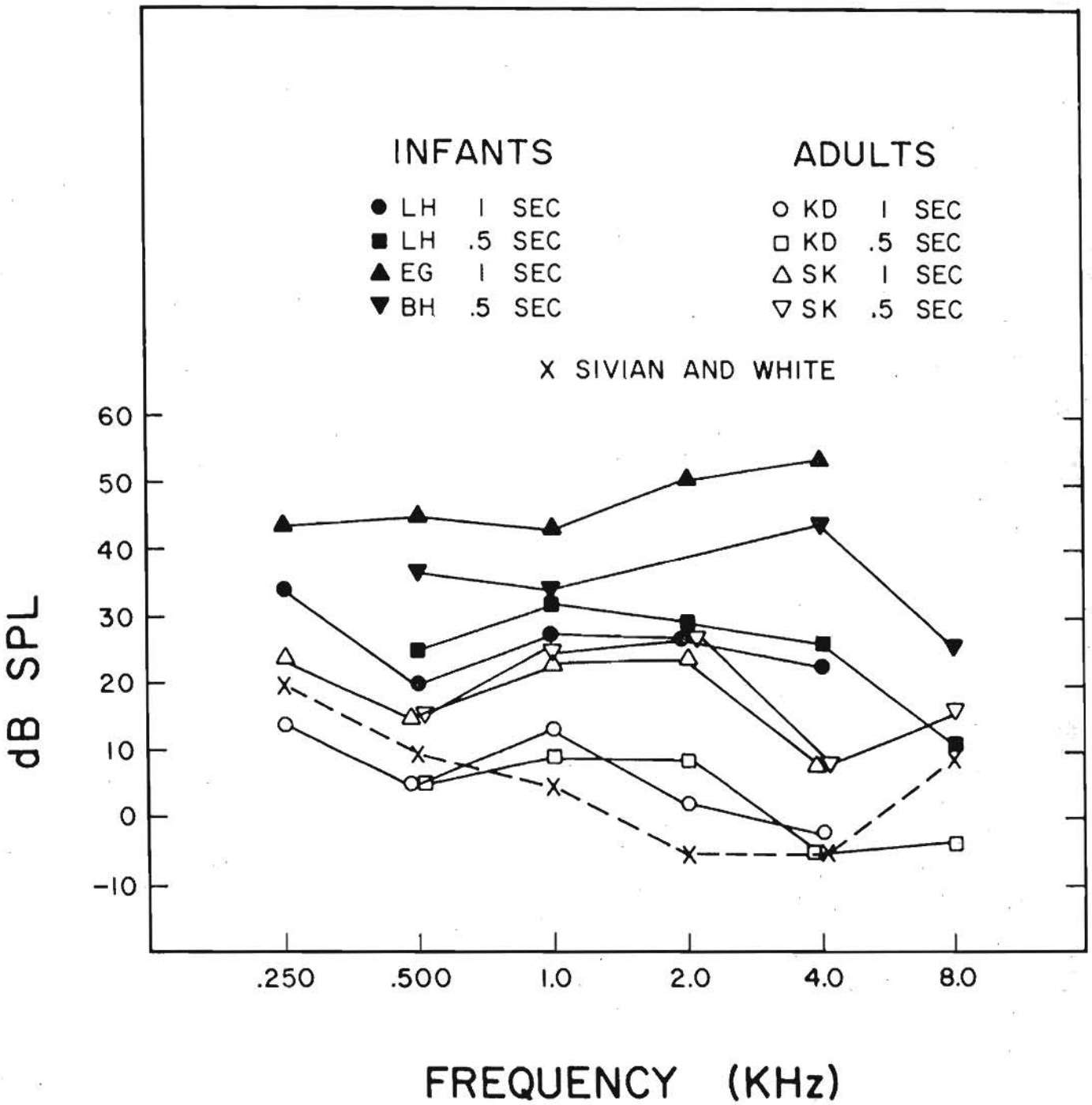

INFANTS
● LH 1 SEC
■ LH .5 SEC
▲ EG 1 SEC
▼ BH .5 SEC
ADULTS
○ KD 1 SEC
□ KD .5 SEC
△ SK 1 SEC
▽ SK .5 SEC
X SIVIAN AND WHITE
dB SPL
60
50
40
30
20
10
0
-10
.250
.500
1.0
2.0
4.0
8.0
FREQUENCY (KHz)

Figure 5.

FREQUENCY (KHz)
.250
.500
1.0
2.0
4.0
8.0
dB SPL
-10
0
10
20
30
40
50
60
.5 SEC TONES
ADULTS
INFANTS
1.0 SEC TONES
ADULTS
PERFECT SESSIONS
INFANTS

Figure 6

The Role of Early Experience in the Development of Speech Perception

Amanda C. Walley, David B. Pisoni and Richard N. Aslin

Indiana University

Bloomington, Indiana 47405

This is a draft of a chapter to appear in R. N. Aslin, J. Alberts and M. R. Petersen (Eds.), Sensory and Perceptual Development. New York: Academic Press, 1981.

Preparation of this chapter was partially supported by grants from NICHHD (HD-11915-03) and NIMH (MH-24027-06) and by a doctoral fellowship awarded to ACW from the Social Sciences and Humanities Research Council of Canada.

The Role of Early Experience in the Development of Speech Perception

Amanda C. Walley, David B. Pisoni and Richard N. Aslin
Indiana University
Bloomington, Indiana 47405

I. INTRODUCTION

A substantial body of information has accumulated over the past ten years concerning the speech processing capacities of prelinguistic infants. Some of the major theoretical issues surrounding the precise nature of these capacities, particularly the extent of their innate specification and the role of early experience in their development, will be discussed in this chapter. In addition, several theories that attempt to explain these capacities will be summarized. A conceptual framework will be presented for evaluating these theories and, within this context, some of the processes and mechanisms underlying the perception of segmental contrasts will be examined. Particular emphasis will be placed in this chapter on the perception of voicing and place of articulation in stop consonants - two phonetic distinctions that have received considerable attention in both adult and infant speech perception research over the last few years. However, the general approach to problems of perceptual development to be advocated here can be extended to other classes of speech sounds and other aspects of the phonology of natural language that eventually become part of the linguistic knowledge of all mature speaker-hearers.

II. BACKGROUND

The pioneering work of Eimas and his colleagues (Eimas, Siqueland, Jusczyk & Vigorito, 1971) demonstrated that prelinguistic infants could discriminate speech sounds differing in voice-onset time (VOT) - a major cue to voicing. VOT has been defined for word-initial stop consonants as the interval between the onset of the release burst and the onset of laryngeal pulsing (Lisker & Abramson, 1964). Stop consonants are those speech sounds which are produced by achieving complete closure of the relevant articulators and thus complete and brief obstruction of the airstream. The Eimas et al. study was motivated, in part, by two previous empirical findings concerning VOT. The first of these was that adults perceive variations in VOT categorically (e.g., Liberman, Harris, Kinney & Lane, 1961); in contrast to the perception of other auditory stimuli, the discriminability of variations in these speech sounds appears to be limited by the listener's ability to differentially identify them. The second was that speakers' productions, sampled from eleven diverse languages, cluster around the same three modal values of VOT (Lisker and Abramson, 1964). These two findings suggested the existence of innate constraints on perception and production of the voicing contrast. Because infants represent a linguistically naive population, they seemed to provide an excellent opportunity for studying the roles of genetic and experiential factors in perceptual development. The finding that infants discriminated VOT differences categorically led Eimas et al. to conclude that

infants were perceiving these sounds in a manner approximating categorical perception in adults. This in turn was interpreted as support for the inference that the perceptual categories of the young infant closely resemble those of the adult and that these perceptual categories are genetically specified, since infants have had little experience in the language learning environment. Moreover, since the predominant view at the time was that categorical perception was unique to the perception of speech signals, Eimas et al. also argued that the infants perceived the sounds in a linguistic mode and that the mechanisms underlying speech perception were therefore part of the human's biological makeup - i.e., phonetic categories were thought to be innately specified. This view acknowledged, therefore, little, if any, influence of nongenetic, experiential factors in the development of speech perception. In a more recent discussion of infant speech perception, Eimas (1980) has essentially reasserted these claims and ignored theoretical arguments for integrating the role of early experience in the development of speech perception (Aslin & Pisoni, 1980b).

A. LEVEL OF ANALYSIS

The results from several lines of research in speech perception necessitate a substantial tempering and modification of the conclusions originally drawn from the Eimas et al. study. For example, it is now well established that categorical perception does not necessarily imply the operation of a linguistic mode of processing. First, categorical perception is not, as was once

thought, unique to the perception of speech signals. Several studies (Cutting & Rosner, 1974; Miller, Wier, Pastore, Kelley & Dooling, 1976; Pisoni, 1977) have demonstrated quite conclusively that there are classes of complex nonspeech signals which can be perceived categorically by adults. Moreover, it has been found that infants discriminate these same nonspeech signals categorically (Jusczyk, Rosner, Cutting, Foard & Smith, 1977; Jusczyk, Pisoni, Walley & Murray, 1980). In fact, this line of research has led to the proposal that psychophysical constraints on the resolution of temporal order underlie VOT perception (Hirsh, 1959; Pisoni, 1977). Recently, Jusczyk (1980, and chapter in this volume) has offered a psychophysical explanation of the stop/glide (e.g., /ba/ vs. /wa/) distinction. Thus, the original claim that the infant's categorical indicates processing at a linguistic level is certainly not supported by evidence of uniqueness, since nonspeech signals are also discriminated categorically.

A second argument against the claim that categorical perception is mediated by linguistic analysis follows from the demonstration that categorical perception is not limited to human perception. Kuhl and Miller (1975, 1978) have shown that chinchillas exhibit categorization of stop consonants that is similar to humans even though such perception in chinchillas is obviously not mediated by the phonological system of any natural language. These two findings concerning categorical perception clearly render untenable the strong contention (Eimas et al., 1971) that categorical-like perception implies perception in a linguistically relevant manner. The demonstration of categorical

discrimination in infants is not, therefore, sufficient to warrant the claim that speech signals are processed by specialized perceptual mechanisms or that the infant's discrimination is directly constrained by the phonological structure of any particular natural language. Rather, it appears that there are general constraints on the mammalian auditory system and that infants (and chinchillas, for that matter) may simply be responding to the psychophysical or sensory properties of speech signals without any subsequent linguistic interpretation of these signals. It seems quite plausible, however, that certain languages have, in the manner that Stevens (1972) has suggested, exploited various properties of the auditory system in selecting the inventory of speech sounds to be used as phonological distinctions.

B. INNATE SPECIFICATION OF PERCEPTUAL CATEGORIES

These findings concerning categorical perception raise serious doubts about the inference that specific adult-like phonetic categories exist as such for the infant. The second of Eimas' claims - that the perceptual categories underlying phonetic categories, at least those for stop consonants, are innately specified is also subject to criticism. The original basis for this nativistic claim was twofold. First, although the precise locations of the boundaries between phonological categories for stops differed somewhat across languages, Lisker and Abramson (1964, 1967) found that production and perceptual categories tended to fall around at least one of three modal values along the

VOT continuum corresponding to long lead, short lag or long lag distinctions in voicing. These investigators proposed therefore that VOT is a universal dimension for realizing voicing distinctions and is closely tied to the way in which laryngeal and supralaryngeal events are controlled in speech production. Second, because of the close correspondence between the infant's discrimination performance for the synthetic VOT stimuli and the Lisker and Abramson adult English perceptual data, Eimas (1975, 1978) argued that prelinguistic infants are predisposed to process VOT information and that this processing is achieved via the operation of an innately specified linguistic feature detector system which is independent of the infant's linguistic environment.

Despite the apparently sound reasoning behind this strong nativistic position, Eimas overlooked several important empirical findings. First, the precise location of the voicing boundaries described by Lisker and Abramson do differ somewhat from language to language, indicating that some fine tuning or realignment of perceptual categories must occur in development - an implication which Eimas (1975) later realized and discussed briefly. Indeed, this sort of perceptual modification might explain why Eimas failed to find any evidence for the discrimination of the prevoiced/voiced distinction among English infants; i.e., even the limited exposure which two-month-olds have to their native language might have exerted a change in sensitivity to this particular contrast. On the other hand, several more recent studies have suggested that such a period of exposure is unlikely to produce any pronounced change in sensitivity along the VOT

continuum and that infants are, in fact, sensitive to this contrast (but see Eilers, Gavin & Wilson, 1979, for an exception to this position). Lasky, Syrdal-Lasky & Klein (1975) found that infants raised in a Spanish-speaking environment do discriminate voicing contrasts that are not discriminated by Spanish speaking adults, but that these infants fail to discriminate the contrast that straddles the Spanish adult voicing boundary. Additional evidence for the discrimination of voicing contrasts not found in the linguistic environment has been obtained by Streeter (1976) for Kikuyu infants and by Aslin, Hennessy, Pisoni and Perey (1979) for English infants. Although Eimas' innate feature detector model can account for the ability of all infants to discriminate all three of the VOT contrasts, there certainly must be some influence exerted by experience during development. Moreover, the features used in discrimination are most likely auditory, and not phonetic, ones.

A second problem with Eimas' argument is his use of the correspondence between the English adult and infant discrimination data to support the claim that voicing categories are perceived by infants in a linguistically relevant manner and are, therefore, genetically specified. If such a correspondence between the data from English infants and adults supports a linguistic level of analysis in these infants, at what level do Spanish infants, whose data do not correspond to that of Spanish adults, analyze these same stimuli? Apparently, Eimas did not see any conflict in the lack of correspondence between the Spanish adult and infant data. Yet, this lack of correspondence would seem to invalidate the claim that English infants process speech sounds (at least those differing in VOT) at a linguistic level.

Although the nativistic account of speech perception proposed by Eimas might be thought of as representing a significant advance over previous views of perceptual development which have assumed that speech production precedes or parallels perception (e.g., Fry, 1966), it seems clear that experience must, at some point in development, exert an influence on perception to produce the adult perceptual categories appropriate for speech. Indeed, the necessity of such environmental influence language acquisition was obvious to Hockett (1955), who, identified several essential 'design features' that characterize all spoken languages and set them apart from communication systems of other organisms. Specifically, Hockett observed that inherent in all natural languages is a 'duality of patterning' - i.e., there is a syntactic level of representation which consists of the arrangement of meaningful elements (morphemes or words), differences in which are realized by variations in the arrangement of meaningless elements (phonemes) at the phonological level of representation. The arbitrary relationship between sound and meaning in human language, and the resulting variability that exists across linguistic communities, necessitate the cultural transmission of any particular language. Cultural transmission of a language in turn requires a substantial amount of plasticity in learning and susceptibility to the influence of experiential factors in the language learner. However, it remains unclear what mechanisms underlie such responsiveness to environmental input, what precise effects linguistic experience exerts on perception, and when during development these experiential influences are most significant.

C. MECHANISMS

Clearly, there is some selective modification during the course of phonological development. This is evidenced by the fact that different languages have different phonological systems and by the apparent difficulty adults have in recognizing phonetic contrasts which are phonologically irrelevant in their native language. The cross- language research of Lisker and Abramson (1967) supports the contention that only phonologically distinctive perceptual categories are perceived by adults and that linguistic experience plays a significant role in the categorization of speech sounds. As can be seen in Figure 1,

---

Insert Figure 1 about here

English, Thai and Spanish adults are only able to differentially label those speech sounds which are used contrastively in their language. More recent cross-language research by Miyawaki, Strange, Verbrugge, Liberman, Jenkins and Fujimura (1975) also provides support for the view that the phonologically distinctive /r/-/l/ contrast in English is perceived by English adults, but not by Japanese adults who do not use this contrast in their language. Training studies such as those summarized recently by Strange and Jenkins (1978) suggest further that, although the ability to perceive a phonologically irrelevant contrast may have been present at birth, adults who have lost or failed to develop the ability to discriminate that contrast are probably incapable

of (re)acquiring it. Findings such as these have been interpreted as evidence that a neural substrate for the perception of phonologically irrelevant contrasts either failed to form during a critical or sensitive period or atrophied in the absence of experience with that contrast (Eimas, 1975). This neural theory of phonological development is passive, in the sense that it assumes that little, if any involvement, either attentional or productive, is required to maintain or create a particular perceptual ability. It cannot, therefore, explain how a child eventually realizes that different acoustic cues signal differences at the morphological level of language.

An alternative to this passive or strictly receptive account of the role of early experience in speech perception is the view that the failure to actively engage an attentional or productive system in the use of a particular phonetic contrast only depresses or attenuates subsequent discrimination performance on that contrast. The apparent difficulty adults have in discrimination may not be due to the degradation of any sensory or neural process per se, but may simply be a consequence of an attentional deficit similar to the process of acquired equivalence - a perceptual mode that involves learning to ignore distinctive differences among stimuli (Riley, 1968). This view assumes that perception of the relevant distinctive contrasts is so automatic as a result of processing strategies or mechanisms previously acquired by the listener that reacquisition of a phonologically-irrelevant contrast is difficult, if not impossible, to obtain reliably in untrained adults who have had little experience in the laboratory (see Shiffrin and Schneider, 1977).

Recently, Pisoni, Aslin, Perey and Hennessy (1978) have conducted several experiments which show that monolingual adult speakers of English are able to acquire a new voicing contrast. In one condition, the subjects had three response categories corresponding to /ba/, /pa/ and prevoiced /ba/. All subjects were very consistent in labeling the synthetic VOT stimuli into three perceptual categories, despite the absence of highly prevoiced stops in syllable-initial position in English and despite the very limited exposure and training experience that preceded the labeling task. Prior to testing in this condition, subjects listened to several repetitions of the -70, 0 and +70 msec VOT stimuli (prevoiced /ba/, /ba/ and /pa/, respectively) to familiarize themselves with the stimulus contrasts and the appropriate responses. However, no overt response was required at this time and no feedback was provided, nor was any attempt made to train subjects in any explicit way. The identification results from this three-category labeling task are particularly striking when compared to the identification results from the more traditional two-alternative forced-choice task in which subjects simply categorized the stimuli into two groups corresponding to English /ba/ and /pa/ (see Figure 2).

Insert Figure 2 about here

Note the classic two-category identification functions obtained in this task for /ba/ and /pa/ responses. The ABX discrimination functions for subjects in both conditions reveal two peaks in discrimination (see Figure 2); i.e., even subjects in

the two-category labeling condition discriminate stimuli in the voicing lead region of the stimulus continuum despite the fact that these stimuli were all identified as belonging to the same phonological category (/ba/). This result was also obtained for subjects tested without prior labeling experience and without feedback. Apparently, phonologically irrelevant categories (such as prevoiced /ba/, for speakers of English) can be consistently categorized by adults even without very extensive training. Thus, the lack of exposure to specific phonetic contrasts during infancy and childhood does not appear to result in a complete neural loss or atrophy of the feature detectors which have been assumed to underlie phonetic categorization (Eimas, 1975). These new findings call into question the recent conclusions of Strange and Jenkins (1978) concerning the negligible effects of laboratory training studies in speech perception. Moreover, given that subjects could use three response categories consistently and without extensive training in the Pisoni et al. study, it is difficult to argue that there was any appreciable "selective" loss in perceptual sensitivity by these subjects in processing voicing information. The performance decrements observed in earlier studies on voicing discrimination may simply have been the result of criterion shifts and response constraints that were a part of the different subject strategies used in these tasks (see Pisoni and Lazarus, 1974; Carney, Widin & Viemeister, 1977).

III. ROLE OF EARLY EXPERIENCE IN PERCEPTUAL DEVELOPMENT

There still exists a strong tendency toward theoretical simplification in describing the ontogeny of various infant speech processing capacities - i.e., toward explaining them either in vague terms of learning or by recourse to strong nativistic accounts. In contrast, several researchers working in the area of visual system development have begun to appreciate the many diverse and interactive roles that genetic and experiential factors can play in the development of sensory and perceptual systems. For example, some of the neural mechanisms underlying visual functioning are not present at birth, nor do they emerge during development as a simple consequence of a genetically controlled plan or schedule. Instead, early visual experience does have some influence on the course of visual system development (see the chapters by Mitchell, Aslin, Norton & Blake, this volume). This experience does not, however, totally control the outcome of visual system development since some genetically specified limits are clearly placed on how much and at what point in development such early experience can influence visual system development(for general reviews, see Blakemore, 1976; and Grobstein and Chow, 1976).

It has become clear from the study of visual system function and its development that a simple dichotomy between nativistic and empiricist accounts of the process of development is simply inadequate to capture the multiple and seemingly complex genetic and environmental interactions that underlie normal perceptual

development (see the chapter by Aslin, this volume). Similarly, the following discussion is motivated primarily by the concern for providing a more explicit and coherent framework from which to view the course of perceptual development - particularly the development of speech perception. The need for such a framework in understanding the processes underlying the development of speech perception is particularly pressing in light of the many seemingly diverse and conflicting empirical findings that have appeared in the infant speech perception literature in recent years, some of which will be reviewed below (see also Jusczyk, 1980, and chapter this volume).

Recently Gottlieb (1976a,b, and chapter this volume) has provided an account of some of the possible roles that early experience can play in behavioral development. His conceptualization of these experiential processes seems particularly relevant and amenable to discussions of the development of speech perception (Aslin & Pisoni, 1980b). According to our application of Gottlieb's framework, there are four basic ways in which early experience could influence the development of speech processing abilities. These alternatives are illustrated in Figure 3.

Insert Figure 3 about here

First, a perceptual ability may be present at birth but require certain specific types of early experience to maintain the integrity of that ability. The absence or degradation of the requisite early experience can result in either a partial or a

complete loss of the perceptual ability, a loss which may be irreversible despite subsequent experience. For example, the work of Hubel and Wiesel (1965, 1970) on the visual system of the kitten showed, among other things, that the full complement of neural cells responsible for binocular vision was present at birth, although they lost their function if the kittens were deprived of binocular vision during a sensitive period. Early experience in this case served then to maintain the functional integrity of the mechanisms underlying binocular vision (see also Blakemore, 1978).

Second, an ability may be only partially developed at birth, requiring specific types of early experience to facilitate or attune the further development of that perceptual ability. The lack of early experience with these stimuli which may serve a facilitating function could result either in the absence of any further development or a loss of that ability when compared to its level at birth. As an example of a facilitating effect of experience, Gottlieb himself has shown that ducklings modify their subsequent preference and recognition of species-specific calls by their own vocalizations prior to and shortly after hatching (Gottlieb, 1976a). If these self- produced vocalizations are prevented from occurring (through devocalization techniques) in the early stages of development, the developmental rate of preference for species-specific calls declines and the ability to discriminate and recognize particular calls is substantially reduced (Gottlieb, 1975).

Third, a perceptual ability may be absent at birth, and its development may depend upon a process of induction based on

specific early experiences of the organism. The presence of a particular ability, then, would depend to a large extent upon the presence of a particular type of early experience. For example, it is well known that specific early experience presented to young ducklings leads to imprinting to a particular stimulus object and can be taken as an instance of inducing a behavioral preference (Hess, 1972). Thus, in this case, the presence of a particular early experience is necessary for the subsequent development of a particular preference or tendency.

Finally, early experience may, of course, exert no role at all in the development of a particular perceptual ability. That is, the ability may be either present or absent at birth and it may remain, decline or improve in the absence of any particular type of early experience. Absence of experiential effects is difficult to identify and often leads to unwarranted conclusions, especially those that assume that an induction process might be operative. For example, it is quite common for investigators to argue that if an ability is absent at birth, but then observed to be present sometime after birth, the ability must have been learned (see Eilers et al., 1979). In terms of the conceptual framework outlined above, this could be an instance of induction. Yet it is quite possible that the ability simply unfolded developmentally according to a genetically specified maturational schedule - a schedule that required no particular type of early experience in the environment. (Fantz, Fagen and Miranda, 1975, provide an example of this by demonstrating a preference for patterned stimuli in both full-term and premature infants.) This unfolding of an ability may be thought of as adhering to the

general class of maturational theories of development. As an example, although general motor activity is necessary to prevent the atrophy of various muscle systems, many of the classic studies by Gesell in the 1930s demonstrated that no specific training experience was necessary for infants to acquire the ability to walk (Gesell and Ames, 1940). Thus, the complexity of these numerous alternatives - maintenance, facilitation, induction and maturation - and their possible interactions should serve to caution researchers against drawing any rash or premature conclusions about the developmental course of specific perceptual abilities.

In order to make clear the relevance of Gottlieb's scheme of the roles of early experience to the development of speech perception, four general classes of theories of perceptual development that seem appropriate to a discussion of phonological development and that parallel the effects of early experience as described by Gottlieb will be outlined. After the assumptions of these perceptual theories are described, several examples from the infant speech perception literature will be selected to illustrate the usefulness of this conceptualization. The classes of theories of perceptual development to be considered below are Universal Theory, Attunement theory, Perceptual Learning Theory and Maturational Theory.

Universal Theory assumes that, at birth, infants are capable of discriminating all the possible phonetic contrasts that may be used phonologically in any natural language. According to this view, early experience functions to maintain the ability to discriminate phonologically relevant distinctions - those actually

presented to the infant in the environment. However, the absence of exposure to phonologically-irrelevant contrasts results in a selective loss of the ability to discriminate those specific contrasts. The perceptual mechanisms responsible for this loss of sensitivity may be either neural or attentional or both. These two alternatives also make several specific predictions concerning the possible reacquisition of the lost discriminative abilities in adults, an important topic in its own right, as mentioned previously.

Attunement Theory assumes that at birth all infants are capable of discriminating at least some of the possible phonetic contrasts contained in the world's languages, but that the infant's discriminative capacitites are incompletely developed and/or possibly quite broadly tuned. Early experience therefore functions to align and/or sharpen these partially developed discriminative abilities. Phonologically-relevant contrasts in the language-learning environment would then become more finely tuned with experience and phonologically-irrelevant contrasts would either remain broadly tuned or become attentuated in the absence of specific environmental stimulation.

In contrast with the other two views, Perceptual Learning Theory assumes that the ability to discriminate any particular phonetic contrast is dependent upon specific early experience with that contrast in the language-learning environment. The rate of development could be very fast or very slow depending on the relative importance of the phonetic contrasts during early life, the relative psychophysical discriminability of the acoustic cues underlying the phonetic contrast compared with other phonetic

contrasts, and the attentional state of the infant. According to this view, however, phonologically-irrelevant contrasts would never be discriminated better than the phonologically-relevant ones present in the language-learning environment.

Finally, Maturational Theory assumes that the ability to discriminate a particular phonetic contrast is independent of any specific early experience and simply unfolds according to a predetermined developmental schedule. All possible phonetic contrasts would be discriminated equally well irrespective of the language environment, although the age at which specific phonetic contrasts could be discriminated would be dependent on the developmental level of the underlying sensory mechanism. For example, if infants did not show sensitivity to high frequencies until later in development, one would not expect them to discriminate phonetic contrasts that are differentiated by high frequency information.

These classes of theories of perceptual development make rather specific predictions concerning the developmental course of speech perception in infants and young children. It is important to note here that probably no single class of theory will uniquely account for the development of all speech contrasts. Rather, it may be the case that some hybrid of the theories provides the best description of the development of specific classes of speech sound discrimination. In fact, this view of parallel developmental processes appears to be supported by current empirical findings.

A. Voicing in Stop Consonants.

The Eimas et al. study (1971) generated a great deal of interest in the infant's speech processing capacities (see Jusczyk, 1980, and chapter this volume). Since then, infants' discrimination of over two dozen VOT contrasts has been studied and positive evidence of discrimination has been obtained for all contrasts that crossed the English voiced-voiceless boundary. However, for contrasts that crossed a prevoiced-voiced boundary, the only positive evidence of discrimination was obtained with infants whose native language environment contained this phonological contrast for stop consonants (see Eilers et al., 1979; see also Aslin & Pisoni, 1980a, for an important critique of this study).

While these results on the discrimination of voicing contrasts by infants might appear to provide strong support for Perceptual Learning theory, there are findings that clearly conflict with the theory's predictions. For example, several contrasts have been tested with infants whose native language environment was not English (e.g., Lasky et al., 1975; Streeter, 1976). Discrimination performance on the majority of these contrasts was observed despite the fact that these specific contrasts were not phonologically distinctive and therefore unlikely to occur in the infants' language learning environment. However, other VOT studies have failed to provide evidence that infants discriminate contrasts that are present in their language learning environment (e.g., Lasky et al., 1975). Within the

conceptual framework being proposed here, these seemingly contradictory results can be reconciled and a systematic account of them offered in terms of what is currently known about the psychophysical properties of these speech signals and the developmental processes responsible for realizing the discrimination.

There is now sufficient evidence to suggest that the basis for VOT discrimination by infants is probably not directly related to phonetic categorization or a linguistic mode of analysis (see also Stevens and Klatt, 1974). Recently, Pisoni (1977) has demonstrated that when the relative onset times of two-component tones are varied, adults perceive such variations in these nonspeech stimuli categorically. The identification data indicate that the tone-onset-time (TOT) continuum is parsed into three discrete categories; stimuli with onset differences greater than 20 msec are perceived as having either leading or lagging onsets, those with onset differences less than 20 msec as having simultaneous onsets. Pisoni (1977) has also shown that the peaks in the discrimination functions for these stimuli coincide quite closely with these values.

Similar discrimination performance with these TOT stimuli has been observed in infants, although the precise location of the infants' category boundaries (as inferred from the discrimination data) differs somewhat from the adults' (Jusczyk et al., 1980). The adult category boundary values observed by Pisoni for the TOT continuum also correspond very closely to the loci along the VOT continuum of the three voicing categories found earlier by Lisker and Abramson (1964, 1967) across a wide variety of languages. This

correspondence, together with the infant and chinchilla data, suggests that the categorical perception of VOT information may simply reflect an inherent limitation of the mammalian auditory system to resolve temporal differences between two acoustic events - specifically, in the case of voicing, between laryngeal and supralaryngeal ones. The resolution of the temporal relation between these two events is greater at certain regions (+/-20msec) along the VOT stimulus continuum which correspond roughly to the the psychophysical threshold for resolving these differences (Hirsh, 1959).

This psychophysical account of VOT perception is attractive in that it can account for both the infant and chinchilla data without recourse to the assumption of innate linguistic (i.e., phonetic) knowledge. Moreover, it can account for the cross-language similarities that have been observed for infant perception. However, two questions are immediately apparent from this analysis. First, why is there so little evidence for the discrimination of VOT in the -20 msec region of voicing lead in the infant literature? Second, what role does the environment play in tuning the perceptual mechanism responsible for processing temporal order information? In other words, what accounts for cross-language differences in the adult perceptual data?

The first question can be addressed by observing that even with nonspeech signals differing in relative onset time, discrimination of onset differences is better in the positive region of the stimulus continuum than in the negative one. The same relation can be found in the original Lisker and Abramson (1967) discrimination data obtained with Thai subjects (see top panel of Figure 4). The smaller incidence of discrimination of VOT

Insert Figure 4 about here

---

differences in the region of voicing lead is probably then due to the generally poorer ability of the auditory system to resolve temporal differences in which a lower frequency component precedes a higher one (Danaher, Osberger & Pickett, 1973).

Lower discriminability of stimuli in the minus region of the VOT continuum cannot completely account for the overall performance of infants since all three positive instances of discrimination of prevoiced and voiced stop contrasts reported in the literature involved infants from linguistic environments where this contrast is used. Thus, it can be further argued that early linguistic experience must play some role in modifying the discriminability of speech stimuli depending on the relative predominance of certain VOT values in the productions of adults.

Differences in the relative discriminability of VOT contrasts are yet another indication that early environmental experience plays an important role in perceptual development. Although there are two regions of high discriminability even in the functions obtained from speakers of English, a language which does not have the prevoiced/voiced contrast, the peak in the minus region is substantially lower than the Thai discrimination data (see top panel of Figure 4). A very similar finding is apparent in the discrimination data of Williams (1974) for Spanish and English subjects and in the more recent data of Pisoni et al. (1978) with naive adult English subjects (see Figure 2). The Spanish subjects in the Williams' study displayed a much broader region of

heightened discriminability extending well into the area encompassing the location of the English voicing boundary.

The available data on the development of voicing perception, therefore, provide good support for the Attunement Theory outlined earlier, since there appears to be a partially specified ability to process temporal order information present at birth. Perceptual sensitivity to temporal order differences such as those present in synthetic stimuli is, however, susceptible to the influence of early experience, which apparently selectively modifies the distinctiveness and location of the regions of sensitivity along a stimulus continuum such as VOT.

Jusczyk (1980; see also chapter in this book) has suggested how such modification might be achieved. Experience in a particular linguistic environment could direct the infant or child to make use of other prominent acoustic cues to voicing that occur regularly in certain phonetic environments as a consequence of phonological constraints specific to that language. This experience would result in a differential perceptual weighting of the various cues to voicing and produce a change in the perceptual salience and location of regions of sensitivity along the VOT continuum. Although Jusczyk has proposed that the infant may only begin to attend to relevant acoustic cues for phonetic contrasts when speech begins to assume a communicative (i.e., meaning-related) purpose, it is also quite possible that a communicative context (i.e., dyadic interchange) acts as a vehicle for directing the child's attention to the subtle acoustic features of the speech signal. If the latter explanation is correct, then the social role in the acquisition of speech

processing skills is to focus the child's perceptual capacities and attention, rather than to trigger a new level of analysis related to referential skills. Clearly, these issues require extensive study of infants during the early months of speech production (12-18 months of age) and, unfortunately, such studies are virtually absent from the speech perception literature.

B. Place of Articulation for Stop Consonants.

Because the acoustic cues to place of articulation have in the past so successfully eluded any simple characterization, this phonetic contrast has been the subject of extensive research and the resulting information about place perception has been of major importance in the formulation of speech perception theories. Although several acoustic properties, such as formant transitions, burst spectra, and direction of rapid spectrum change following consonantal release, have been implicated in the perception of place of articulation (e.g., Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967; Cooper, Delattre, Liberman, Borst & Gerstman, 1952; Stevens, 1975; Stevens and Blumstein, 1975), attempts to state unequivocally which acoustic properties or correlates constitute the primary perceptual cue for this feature have been complicated by the fact that all of these acoustic features may vary for a given place of articulation in different phonetic contexts. The failure to find an absolute, invariant set of acoustic properties which correspond to place of articulation in all environments has led some theorists to argue more generally that invariant cues do not exist for phonetic features - that the

relation between the acoustic signal and the phonetic percept is not a direct one (e.g., Liberman et al., 1967). Rather, it is claimed that the invariance of the phonetic percept results from the interpretation of acoustic cues in a manner that is context dependent. In order to explain how such contextually determined interpretation could be achieved, one class of speech perception theories has therefore found it necessary to view the speech perception process as an active one in which the identification of phonetic segments depends on some sort of computational or look-up procedure that involves higher level linguistic knowledge that imposes structure on the incoming speech waveform (e.g., Chomsky & Miller, 1963; Chomsky & Halle, 1968; Liberman et al., 1967; Stevens & Halle, 1967; Stevens & House, 1972).

In support of context dependent views of speech perception is the finding that vastly different acoustic cues may give rise to the same phonetic percept, and the converse finding that the same acoustic segment in different contexts may give rise to the perception of different phonetic segments (Liberman et al., 1967; Liberman, Delattre & Cooper, 1952; Shatz, 1954). Moreover, the finding that potential cues for place of articulation are context dependent has been an important justification for the claim that speech perception involves specialized processing mechanisms (e.g., Liberman & Studdert-Kennedy, 1978). Several studies have shown that infants discriminate place contrasts categorically in speech contexts, but not in nonspeech contexts (for a summary, see Jusczyk, 1980), providing additional support for this type of theory.

There are, however, two potential sources of evidence against such context dependent views of speech perception. First, if linguistic knowledge is actually a prerequisite for the perception of speech, this would seem to require that experience in speech perception and/or production is necessary for the perception of specific contextual dependencies. However, this implication is challenged to some extent by the results of infant speech perception studies, which show that infants, with only very limited exposure to the numerous phonetic distinctions employed by various languages and with virtually no experience in the consistent articulation of these distinctions, are capable of discriminating certain acoustic variations across adult phonemic categories while ignoring within-category variations. Although active theories of speech perception might still account for these abilities by positing innate knowledge of phonological rules, Eimas (1975) has proposed instead that the infant data is accommodated better by simply assuming the existence of a linguistic feature detector system. He has argued that the human auditory system might be endowed with feature or property detectors which are sensitive to the restricted ranges of acoustic information that signal phonetic features (see also Stevens, 1975). Thus, infants may be predisposed to perceive certain speech stimuli in a linguistically relevant way. As discussed previously, it is more parsimonious to attribute categorical perception, at least of the VOT continuum, to basic psychophysical constraints on the auditory system, rather than linguistically-oriented feature detectors.

A second source of evidence against context-dependent theories of speech perception derives from research that attempts to specify in detail what these psychoacoustic constraints on the mammalian auditory system are. The importance of these constraints is that they provide the basis for a mode of perception which accounts, in part, for the invariance of the phonetic (voicing) percept and thus the child's ability to acquire voicing as a phonemic contrast. Although it seems reasonable to assume that a psychophysical basis for the categorical perception of variations in the cues to place of articulation may also exist, such a basis is less obvious in view of the greater contextual variability of the hypothesized cues for this feature. Indeed, Bailey, Summerfield and Dorman (1977) have provided evidence that the psychophysical boundaries obtained for a set of nonspeech frequency- and amplitude-modulated sinewaves, which were modelled after the formant structure of stop CV syllables, do not correspond to the phonetic boundaries obtained for the speech stimuli. Thus, they interprete this finding as support for the existence of some specialized speech processing mechanism. However, Pisoni (1979), using comparable nonspeech stimuli, has shown that the location and extent of perceptual categories are not necessarily rigidly controlled by any simple physically defined invariant, such as the direction of frequency change, but rather that these categories are influenced by contextual information as well. This finding may account for the boundary differences obtained by Bailey et al. In any event, a number of more recent investigations (Kewley-Port, 1979; Searle, Jacobson & Rayment, 1979; Stevens & Blumstein, 1978, 1980), which employ new

methods of speech analysis and which attempt to incorporate known psychophysical and psychophysiological properties of the human peripheral auditory system into models of the initial stages of speech processing, have been more successful in finding invariant acoustic cues for place of articulation. Such findings would seem to argue strongly against active, context-dependent theories of speech perception.

While the results of studies of infant perception of place of articulation might be expected to address the viability of active, context-dependent theories of speech perception, only a few studies have actually addressed the question of whether or not the perceptual equivalence of different and/or contextually varying acoustic features exists for infants (Eilers, 1977; Fodor, Garrett and Brill, 1975; Kuhl, 1979). Investigations of infant speech perception have, for the most part, only examined discrimination of stimuli varying along a single acoustic dimension. Thus, while it has been shown that infants are capable of categorically discriminating place distinctions (for a review, see Jusczyk, 1980), it is not yet clear from these investigations that the perceptual equivalence of different and/or contextually varying acoustic segments exists for infants, nor that these infant data can be used to refute active, context-dependent theories.

The recent work of Stevens and Blumstein (1978,1980) represents the most substantial theoretical account of the infant's perception of place of articulation. Because several studies have shown that infants can discriminate place of articulation differences (e.g., Bush and Williams, 1978; Eimas, 1974; Leavitt, Brown, Morse and Graham, 1976; Miller and Morse,

1975; Moffitt, 1971; Morse, 1972), Stevens and Blumstein object to the view, entailed in active speech perception theories, that only after learning to organize contextually diverse and variable acoustic features into their appropriate adult phonemic categories does the child come to perceive place of articulation distinctions. Instead, they propose that some innate mechanism must mediate the invariance which they assume is entailed in the discrimination of such distinctions by infants. It should be emphasized, however, that studies to date have merely shown that infants are capable of discriminating place of articulation differences in stimuli varying along a particular acoustic dimension - i.e., formant starting frequency and direction. They have not demonstrated that infants perceive syllables such as /da/ and /di/ as being similar with respect to the initial phonetic segment and evidence for perceptual constancy cannot, therefore, be inferred from any of these simple discrimination studies. Until it is shown that infants are able to sort different and/or contextually variant acoustic features into their appropriate adult phonemic categories, theories which require experience in the perception and/or production of these features are not, as Eimas (1975) contends, and Stevens and Blumstein implicitly assume, invalidated on the basis of the current data from infant speech perception research (but see Pisoni, 1978, for other criticisms of these theories).

While it has not yet been demonstrated that infants have any initial basis for recognizing that contextual variations in acoustic features belong to certain phonetic categories, Stevens and Blumstein are reluctant to abandon the notion that some

invariant property exists in the acoustic correlates of each particular place of articulation category. They have argued that even though various context-dependent features, such as starting frequency and direction of formant transitions and release bursts, are separately observable in a spectrogram, the auditory system does not necessarily process these features independently of one another. Instead, Stevens and Blumstein claim that the auditory system integrates these features in such a way that the gross spectral properties associated with each place of articulation category provide the acoustic invariance which underlies the constancy of the phonetic percept and which must, in their opinion, mediate infant perception. The search for invariant acoustic correlates of phonetic features and thus for a means of automatic, passive recognition of phonetic distinctions represents a notable digression from earlier proposals that speech perception proceeds primarily via the active operations entailed, for example, in analysis-by-synthesis(Stevens and Halle, 1967; Stevens and House, 1972) or by reference to motor-articulatory patterns (e.g., Liberman et al., 1967).

Stevens and Blumstein's (1978, 1980) assertion that there are distinctive and context-independent acoustic properties associated with different places of stop consonant articulation derives from both theoretically based expectations about the gross shape of the short-term spectrum sampled at consonantal release and spectral analyses they have carried out on natural speech. According to these criteria, labials (/b/, /p/) are characterized by a diffuse-falling spectrum, alveolars (/d/, /t/) by a diffuse-rising spectrum and velars (/g/, /k/) by a prominent mid-frequency spectral peak. Examples of these are shown in Figure 5. These

---

Insert Figure 5 about here

---

putatively invariant acoustic cues for place of articulation - location and diffuseness of spectral energy at stimulus onset - are, of course, very similar to the compact vs. diffuse and grave vs. acute features originally proposed by Jakobson, Fant and Halle (1952). Stevens and Blumstein have also maintained that these spectral properties may be used to characterize nasals of different places of articulation and that the spectrum sampled at vowel offset of a CV syllable should exhibit the same properties as the onset spectrum for a given place of articulation. The characteristics of onset spectra are determined by the burst spectrum and the initial portions of the formant transitions at voicing onset. The same spectral shapes can be obtained for synthetic stimuli containing only formant transitions and no burst, but these shapes are enhanced by the presence of the burst. Stimuli with only the release burst present do not yield these distinctive spectral shapes.

Because Stevens and Blumstein (1978) found that only those synthetic stimuli with distinctive spectral characteristics were identified consistently by adults according to place of articulation, they proposed that the auditory system also performs a short-term spectral analysis at stimulus onset for a stop consonant. According to their account, formant transitions are not the primary cue to place of articulation in CV syllables. Rather, identification of this phonetic feature is achieved through the operation of property detectors which, at the peripheral stage of

auditory processing, are tuned to the invariant properties of the onset spectrum. They argue that it is the operation of these detectors which accounts for the infant's ability to discriminate stimuli with different places of articulation - particularly when these stimuli contain both formant transitions and release bursts. This assertion concerning the relative discriminability of stimuli with and without release bursts is based on the earlier work of Bush and Williams (1978) and may not be valid since these investigators failed to actually examine discrimination of pairs of stimuli with and without bursts.

The claim that the context-independent properties described above provide the basis for perception of a given place of articulation might be challenged by the finding that adults are able to differentially identify two-formant synthetic stimuli with respect to place of articulation (Cooper et al., 1952; Delattre, Liberman & Cooper, 1955; Liberman et al., 1952). These stimuli do not, so Stevens and Blumstein report (1978), yield spectra with the distinctive, contextually invariant shapes which purportedly underlie the perception of this phonetic feature. Stevens and Blumstein agree that, in two-formant stimuli, only the second-formant transition can signal differences in place of articulation. They attempt to explain the adult's ability to use this context-dependent cue in terms of the co-occurrence of the primary, invariant and secondary, context-dependent features in the full formant stimuli. Because adults have learned the co-occurrences between primary and secondary cues through repeated exposure to them in the linguistic environment, they can, in the absence or distortion of the primary attributes of the stimulus,

use the secondary cue of starting frequency and direction of formant transitions to identify place of articulation. In terms of the theories previously outlined, Stevens and Blumstein's account incorporates certain aspects of both Universal and Perceptual Learning Theory. Perception of place of articulation is claimed to be mediated by innately specified mechanisms sensitive to the onset spectra of stimuli. By this account, experience presumably functions to maintain the integrity of these perceptual categories. However, perception of place of articulation is eventually also mediated by the detection of formant transition information, once sensitivity to this information is induced by linguistic experience - specifically, by virtue of the co-occurrence of these secondary cues with the primary stimulus attributes that determine the overall shape of the spectrum at stimulus onset.

By proposing that formant transitions constitute a secondary, learned cue to place of articulation, Stevens and Blumstein's theory makes several predictions about infants' perception of place of articulation (see Walley, 1979). Foremost of these is the prediction that infants should not be able to discriminate place of articulation differences in two-formant stimuli, since these stimuli do not, according to Stevens and Blumstein, yield the distinctive, contextually invariant spectra of their full-formant counterparts and only contain formant transition information. According to Stevens and Blumstein's theory, formant transitions provide only a secondary, learned cue to place of articulation and infants should not, therefore, be able to use this cue to discriminate two-formant stimuli differing in place of

articulation in the same way that adults do. If, on the other hand, formant transitions do constitute the major cue for this phonetic feature, infants should be able to discriminate differences in two formant stimuli even without specific experience with them.

With regard to previous demonstrations that infants discriminate place of articulation differences (Miller & Morse, 1976; Moffitt, 1971; Morse, 1972; Leavitt et al., 1976), Stevens and Blumstein would, of course, assert that it is the distinctive shape of the onset spectra of the three-formant stimuli employed in these studies which underlies the infant's discrimination. However, Eimas (1974) found that infants presented with two-formant stimuli could discriminate labial vs. alveolar stops which were differentiated solely by the second-formant transition. It cannot be asserted that discrimination here is mediated by a divergence in spectral shape if stimuli containing only the first two-formants do not possess the distinctive and invariant spectra that three-formant stimuli do. Rather, this suggests that it is the second-formant transition which provides the basis for the infant's discrimination, although, according to Stevens and Blumstein, infants should not be able to use this cue exclusively in discrimination of place of articulation.

Walley (1979) recently attempted to establish which of these two theories offers the best account of place perception by conducting a more extensive examination of infants' discrimination of place differences in synthetic two-formant stimuli. After the two-formant stimuli for this test were constructed and their onset spectra analyzed, it was observed, however, that the onset spectra of the labial and velar stimuli (see Figure 6) were, contrary to

---

Insert Figure 6 about here

Stevens and Blumstein's report, very similar in overall shape to those of their full-formant counterparts - a discrepancy which is perhaps due to the superior quality of more recent digital synthesizers, the extensive spectrographic analyses and the adult subject feedback used in construction of Walley's stimuli. Of course, the two-formant alveolar stimulus differed from the full-formant one (an obvious consequence of removing the upper formants) and was actually very similar to the two-formant velar stimulus.

These initial findings concerning the onset spectra of the two-formant stimuli clearly render Stevens and Blumstein's proposal that formant transitions constitute secondary, learned cues to place of articulation of little predictive value in considering the perception of place of articulation. It may well be that spectral attributes mediate place perception, but since the putatively primary, invariant spectral cues typically are present even in so-called degraded (i.e., two-formant) stimuli, there would seem to be no necessity for an infant to learn to use contextually diverse formant transition starting frequency and direction as an additional cue to place of articulation (at least in the case of the labial vs. alveolar and labial vs. velar two-formant contrasts). Thus, an account in terms of perceptual learning seems to be ruled out. Moreover, if, as Stevens and Blumstein maintain, it is differential sensitivity to spectral shape that mediates perception of place of articulation

differences, then infants should indeed be able to discriminate the labial vs. alveolar and labial vs. velar contrasts in the two-formant stimuli. Although Stevens and Blumstein's theory actually makes the same predictions about infants' discrimination of two-formant stimuli as does the the notion that formant transitions provide the primary cue to this feature, Walley still hoped to differentiate the two accounts on the basis of their predictions concerning infants' discrimination of the alveolar vs. velar two-formant contrast. Because the spectra for these two stimuli are very similar, infants should not, according to Stevens and Blumstein's theory, be able to make this discrimination. If formant transition starting frequency and direction are used as cues to place of articulation, infants might be expected to discriminate this contrast as well.

Using a variation of the operant head-turning (OHT) procedure (Eilers, Wilson and Moore, 1977), in which the infant learns to make a directional head-turn in anticipation of the presentation of a visual stimulus, which serves as a reinforcer, whenever a change in a speech stimulus occurs (see Figure 7 for an illustra-

---

Insert Figure 7 about here

---

tion of the experimental set-up), Walley (1979) assessed infants' discrimination of place of articulation differences. She was able to verify previous reports that prelinguistic infants can discriminate full cue exemplars of CV syllables differing in place of articulation (/ba/, /da/ and /ga/). In addition, infants were also found to be capable of discriminating these contrasts when

they are specified by only two-formant, partial cue stimuli. Most important with respect to Stevens and Blumstein's theory, a number of infants were able to discriminate the partial cue alveolar vs. velar contrast. This result is not predicted by Stevens and Blumstein since the onset spectra for these stimuli are very similar. Moreover, this result would seem to suggest that for the other two contrasts (labial vs. alveolar and labial vs. velar) where onset spectra and formant transition cues are confounded, it may be the latter cue that mediates discrimination.

The infant data on discrimination of place of articulation differences in two-formant stimuli argue rather strongly against one of the main assertions of Stevens and Blumstein's theory - namely, that formant transitions constitute only secondary, learned cues to place of articulation perception in stop consonants and that onset spectra differences are the primary mediator of place discrimination. If formant transition starting frequency and direction can cue place of articulation, then obviously these cues are not learned in any traditional sense, since infants can, indeed, discriminate two-formant stimuli at an early age. Because formant transition information is contextually variable, presumably the child would have to learn how contextual variations in these cues can continue to signal the same place of articulation - a requirement which is probably most in accord with either Universal or Attunement Theory. (However, note that it has not been determined that infants can discriminate contextual variations in formant transitions for a given place of articulation. For example, it is not known whether they can discriminate /da/ vs. /di/ on the basis of information contained

only in the formant transitions.) Even if onset spectra do mediate place perception and experience functions simply to maintain sensitivity to this cue, it appears that contextually invariant and distinctive spectral cues generally characterize even such degraded stimuli as two-formant ones and logically there would, therefore, be little reason for learning to use the secondary cues. It would, of course, be important to examine the spectra of other two-formant stimuli containing different vowels to ensure that this observation can be generalized to other contexts. Regardless of which cue infants use to discriminate place differences then, Stevens and Blumstein's distinction between primary, innate vs. secondary, learned cues does not seem to be a valid or useful one for understanding the perception of place of articulation or its development. (For further criticisms of Steven's and Blumstein's static theory, see Kewley-Port, 1979).

A strong test of Stevens and Blumstein's theory would entail manipulating spectral and transitional cues independently of one another in labial, alveolar and velar stimuli and ascertaining which cue controls perception. Although Walley's (1979) work revealed that this is not possible with two-formant stimuli, Walley and Carrell (1980) have gathered identification data from adults for synthetic stimuli which contained conflicting spectral onset and transition cues to place of articulation. In general, their results indicate that listeners use transition information more often than they do spectral information in identifying place of articulation. The notion that listeners use the dynamic information contained in formant transitions - i.e., <u>changes</u> in spectral energy over time, as opposed to the static spectral

sample at stimulus onset proposed by Stevens and Blumstein, has served as the basis for some of Kewley-Port's (1979) recent work. Figure 8 includes examples of the running spectral displays

---

Insert Figure 8 about here

---

used by Kewley-Port to characterize place of articulation in stop consonants.

IV. Summary and Conclusions

From this consideration of voicing and place perception, it should be apparent that no one major role of experience discussed previously can be uniformly invoked to account for the development of the abilities needed to discriminate the various speech contrasts found in spoken language. Perception of voicing appears to be best described by Attunement Theory and is subject to retuning and alignment effects as a result of linguistic experience. Perception of place of articulation, on the other hand, may best be accounted for by Universal Theory, where the effect of experience is to maintain or collapse already specified perceptual categories. Unfortunately, there is, at the present time, no crosslinguistic adult or infant data on the perception of place of articulation to further clarify the role of experience in the development of this distinction. However, this view of the diverse effects of experience on the development of speech perception is buttressed by an examination of the course of perceptual development for other classes of speech sounds, such as

fricatives, liquids and vowels, within the framework being proposed (for a discussion, see Aslin and Pisoni, 1980b). For example, the auditory system of humans may well be specialized for processing certain very specific types of acoustic attributes at an early age. If some phonetic contrasts in language happen to have these distinctive acoustic properties in common, the infant should be able to discriminate these speech signals with practically no experience in the language learning environment short of sensory deprivation. If, on the other hand, a certain amount of neural maturation or specific early experience is required for discrimination, then a delay or developmental lag in discrimination of these contrasts might be anticipated.

This parallel interactive view of the role of early experience in the development of speech perception has as a precedent empirical support from recent work on visual system development (see Aslin & Dumais, 1980). This work emphasizes the fact that parallel developmental mechanisms (analogous to three of the general classes of speech theories discussed) can operate upon different aspects of the same sensory input. However, only a very detailed description of the development of discriminative abilities will permit a distinction between the various types of complex interactions between genetic and experiential factors in determining speech processing capacities. Such a description requires a better understanding of the infant's basic auditory capacities and more sophisticated experimental procedures for measuring discrimination and other aspects of perceptual analysis. For example, more detailed information about the infant's frequency resolving capabilities is necessary before a really

complete understanding of the infant's perception of more complex speech stimuli, for example, those involving place contrasts, can be obtained. Fortunately, these problems are beginning to be addressed by some researchers (Sinnott, personal communication; Trehub, Schneider & Endman, 1980). In addition, if questions about the discriminative abilities of infants to resolve small differences between speech signals at a purely sensory level are to be answered, many more data points from individual subjects need to be obtained to produce more detailed and reliable discrimination functions (see Aslin et al., 1979; Aslin & Pisoni, 1980b). Furthermore, to study questions surrounding the development of perceptual constancy and categorization of speech signals - abilities that are probably more relevant to the larger task of language acquisition, measures of generalization and perceptual similarity analogous to adult identification tasks are required. These measures may prove to be particularly important to the understanding of language acquisition if they involve the active engagement of the infant's perceptual, attentional and cognitive capacities. All of the current measures of infant speech discrimination capitalize on reflexive (orienting) response systems to indicate perceptual sensitivity. More volitional response systems, such as manual responding, may provide a better indication of when the infant is attentive to the task of responding to various auditory stimuli in a two-choice paradigm. Such measures, however, will of necessity involve older infants, and this work has only recently been initiated (e.g., Sinnott, personal communication).

Finally, the issue of species-specificity in human speech perception demands careful analysis and a cautious approach to the interpretation of data gathered from infants. The analogy between song learning in various species of birds and speech perception by human infants (e.g., Marler, 1970) is quite seductive. Songbirds are apparently predisposed to encode, at a preproductive age, a prototypic "song" that is characteristic of their species. Human infants also appear capable of perceiving the "appropriate" categories of human speech long before they can actually produce these sounds. Yet, there is not a single piece of evidence that any sub-group of the human species, when raised in the natural language environment of another sub-group, fails to acquire the speech perception skills required for that "new" language. Moreover, in the case of VOT, it is quite clear that the basic species-specific perceptual abilities are essentially invariant across a wide variety of linguistic environments. The cross-language data suggest only that these basic perceptual sensitivities are modified slightly and occasionally suppressed if they are not required by a particular language system. The simple fact that members of a particular language environment can, even at later ages, acquire new phonological systems and communicate effectively (albeit with some productive deficiencies) argues strongly against the "template" analogy used in the birdsong literature.

The issue of species-specificity in human speech perception is also related to the debate over what level of analysis is operative in early speech perception. If an aspect of a human adult's performance, when presented with speech signals, appears

to be unique to speech, it is absolutely essential that appropriate controls are used before concluding that speech signals are processed in a special way and that this specialized processing is related to the phonological rule system of a natural language. For example, categorical perception was initially thought to be unique to speech. Later, however, three types of control studies completely eliminated this claim. First, nonspeech signals, perceived by adults as meaningless auditory signals, yielded categorical identification and discrimination functions. Second, nonhumans, that do not use human speech signals in a communicative manner, produced similar categorical data. Third, studies of adults from different language environments, when presented with phonologically-irrelevant speech signals, also yielded categorical data. To claim, therefore, that the categorical nature of the infant speech perception data supports a specialized, phonetic mode of processing that has a biological basis (i.e., is species-specific), is clearly unjustified at the present time. Similarly, more recent claims (Eimas, 1980; Eimas & Miller, 1980) concerning the "unique" context dependent nature of the perception of certain speech sounds must again be evaluated with caution until the appropriate control studies have been conducted (e.g., see Pisoni, Carrell & Gans, 1980). If the context dependency is present in nonspeech signals, in nonhumans, and in adults from other language communities, then this "specialized" speech mechanism will also have to join the ranks of those general characteristics of the mammalian auditory system that are used by the phonological systems of natural languages to support the perception (and perhaps production) of speech. To claim that human

infants perceive speech phonetically because their discrimination data indicate contextual dependency effects makes the same premature and logical error that characterized the first infant studies on voicing perception. A remarkable amount of empirical data on infant speech perception has been gathered since the first studies in the early 1970's, but research in this field must, however, strive for the methodological and interpretive rigor that is found in other areas of experimental psychology.

References

Aslin, R. N. & Dumais, S. T. Binocular vision in infants: A review and a theoretical framework. In L. Lipsitt and H. Reese (Eds.), Advances in Child Development and Behavior, Volume 15, 1980 (In press).

Aslin, R. N., Hennessy, B. L., Pisoni, D. B. & Perey, A. J. Individual infants' discrimination of VOT: evidence for three modes of voicing. Research on Speech Perception Progress Report No. 5, Indiana University, 1979, 347-369.

Aslin, R. N. & Pisoni, D. B. Effects of early linguistic experience on speech discrimination by infants: A critique of Eilers, Gavin, and Wilson (1979). Child Development, 1980, 51, 107- 112. (a)

Aslin, R. N. & Pisoni, D. B. Some developmental processes in speech perception. In G. Yeni-Komshian, J. F. Kavanagh & C. A. Ferguson (Eds.), Child Phonology: Perception and Production. New York: Academic Press, 1980 (In press). (b)

Bailey, P. J., Summerfield, Q. & Dorman, M. On the identification of sine-wave analogues of certain speech sounds. Haskins Laboratories: Status Report on Speech Research, SR-51/52, 1977.

Blakemore, C. The conditions required for the maintenance of binocularity in kittens' visual cortex. Journal of Physiology (London), 1976, 261, 432-444.

Bush, L. & Williams, M. Discrimination by young infants of voiced stop consonants with and without release bursts. Journal of the Acoustical Society of America, 1978, 63 (4), 1223-1226.

Carney, A. E., Widin, G. P. & Viemeister, N. F. Noncategorical perception of stop consonants differing in VOT. Journal of the Acoustical Society of America, 1977, 62, 961-970.

Chomsky, N. & Halle, M. The Sound Pattern of English. New York: Harper and Row, 1968.

Chomsky, N. & Miller, G. A. Introduction to the formal analysis of natural languages. In R. D. Luce, R. Bush & E. Galanter (Eds.), Handbook of Mathematical Psychology, Vol. 2. New York: John Wiley and Sons, 1963, 269-231.

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 (6), 597-606.

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

Danaher, E. M., Osberger, M. J. & Pickett, J. M. Discrimination of formant frequency transitions in synthetic vowels. Journal of Speech and Hearing Research, 1973, 16, 439-451.

Delattre, P. C., Liberman, A. M. & Cooper, F. S. Acoustic loci and transitional cues for consonants. Journal of the Acoustical Society of America, 1955, 27 (4), 769-773.

Eilers, R. E. Context sensitive perception of naturally produced stop and fricative consonants by infants. Journal of the Acoustical Society of America, 1977, 61 (5), 1321-1336.

Eilers, R. E., Gavin, W. & Wilson, W. R. Linguistic experience and phonemic perception in infancy: a cross linguistic study. Child Development, 1979, 50, 14-18.

Eilers, R. E., Wilson, W. R. & Moore, J. M. Developmental changes in speech discrimination in infants. Journal of Speech and Hearing Research, 1977, 20, 766-780.

Eimas, P. D. Auditory and linguistic processing of cues for place of articulation by infants. Perception and Psychophysics, 1974, 16 (3), 513-521.

Eimas, P. D. Developmental aspects of speech perception. In R. Held, H. W. Leibowitz & H.-L. Teuber (Eds.), Handbook of Sensory Physiology Vol VIII: Perception. Berlin: Springer Verlag, 1978.

Eimas, P. D. Infant speech perception: Issues and models. Paper presented at the C. N. R. S. Conference in Paris, France, June 15-18, 1980.

Eimas, P. D. Speech perception in early infancy. In L. B. Cohen & P. Salapatek (Eds.), Infant Perception, Vol. 2. New York: Academic Press, 1975.

Eimas, P. D. & Miller, J. Contextual effects in infants speech perception. Science, 1980 (In press).

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

Fantz, R. L., Fagan, J. F., III, & Miranda, S. B. Early visual sensitivity. In L. B. Cohen & P. Salapatek (Eds.), Infant Perception: From Sensation to Cognition, Volume I, Basic Visual Processes. New York: Academic Press, 1975.

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

Fry, D. B. The development of the phonological system in the normal and deaf child. In F. Smith & G. A. Miller (Eds.), The Genesis of Language. Cambridge, Mass.: M.I.T. Press, 1966.

Gesell, A. L. & Ames, L. B. The ontogenetic organization of prone behavior in human infancy. Journal of Genetic Psychology, 1940, 56, 247-263.

Gottlieb, G. Development of species identification in ducklings: I. Nature of perceptual deficit caused by embryonic auditory deprivation. Journal of Comparative and Physiological Psychology, 1975, 89, 387-399.

Gottlieb, G. Conceptions of prenatal development: Behavioral embryology. Psychological Review, 1976(a), 83, 215-234.

Gottlieb, G. The roles of experience in the development of behavior and the nervous system. In G. Gottlieb (Ed.), Neural and Behavioral Specificity. New York: Academic Press, 1976(b).

Grobstein, P. & Chow, K. Receptive field organization in the mammalian visual cortex: The role of individual experience in development. In G. Gottlieb (Ed.), Neural and Behavioral Specificity. New York: Academic Press, 1976.

Hess, E. H. "Imprinting" in a natural laboratory. Scientific American, 1972, 227, 24-31.

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

Hockett, C. F. A Course in Modern Linguistics. New York: MacMillan, 1958.

Hubel, D. H. & Wiesel, T. N. Binocular interaction in striate cortex of kittens reared with visual squint. Journal of Neurophysiology, 1965, 28, 1041-1059.

Hubel, D. H. & Wiesel, T. N. The period of susceptibility to the physiological effects of unilateral eye closure in kittens. Journal of Physiology (London), 1970, 206, 419-436.

Jakobson, R., Fant, C. G. M. & Halle, M. Preliminaries to Speech Analysis. Technical Report No. 13 Acoustics Laboratory, Massachusetts Institute of Technology, May, 1952.

Jusczyk, P. W. Infant speech perception: A critical appraisal. In P. D. Eimas & J. L. Miller (Eds)., Perspectives on the Study of Speech. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

Jusczyk, P. W., Pisoni, D. B., Walley, A. & Murray, J. Discrimination of relative onset time of two-component tones by infants. Journal of the Acoustical Society of America, 1980, 67, 262-270.

Jusczyk, P. W., Rosner, B. S., Cutting, J. E., Foard, C. F. & Smith, L. B. Categorical perception of non-speech sounds by two-month-old infants. Perception and Psychophysics, 1977, 21, 50-54.

Kewley-Port, D. Continuous spectral change as acoustic cues to place of articulation. Research on Speech Perception Progress Report No. 5, Indiana University, 1979, 327-346.

Kuhl, P. K. Speech perception in early infancy: The acquisition of speech-sound categories. In S. K. Hirsh, D. H. Eldredge, I. J. Hirsh & S. R. Silverman (Eds.), Hearing & Davis; Essays Honoring Hallowell Davis. St. Louis, Mo.: Washington University Press, 1976.

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

Kuhl, P. K. and Miller, J. D. Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. Journal of the Acoustical Society of America, 1978, 63, 905-917.

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

Leavitt, L. A., Brown, J. A., Morse, P. A. & Graham, F. K. Cardiac orienting and auditory discrimination in 6-week infants. Developmental Psychology, 1976, 12, 514-523.

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

Liberman, A. M. & Studdert-Kennedy, M. Phonetic perception. In R. Held, H. Leibowitz & H. L. Teuber (Eds.), Handbook of Sensory Physiology: Perception. New York: Springer-Verlag, 1978.

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

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.

Marler, P. A comparative approach to vocal learning; song development in white crowned sparrows. Journal of Comparative and Physiological Psychology, Monograph, 1970, 71 (2), Part 2, 1-25.

Miller, C. L. & Morse, P. A. The "heart" of categorical speech discrimination in young infants. (Research Status Report No. 1) Madison: University of Wisconsin, Infant Development Laboratory, August, 1975.

Miller, J. D., Wier, C. C., Pastore, R., Kelley, 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.

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 and Psychophysics, 1975, 18, 331-340.

Moffitt, A. R. Consonant cue perception by twenty-four-week-old infants. Child Development, 1971, 42, 717-731.

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

Pisoni, D. B. Identification and discrimination of the relative onset of two-component tones: Implications for the perception of voicing in stops. Journal of the Acoustical Society of America, 1977, 61, 1352-1361.

Pisoni, D. B. Some remarks on the perception of speech and nonspeech signals. Research on Speech Perception Progress Report No. 5, Indiana University, 1979, 305-325.

Pisoni, D. B. Speech perception. In W. K. Estes (Ed.), Handbook of of Learning and Cognitive Processes. Volume 6. Hillsdale, N. J.: Lawrence Erlbaum Associates, 1978.

Pisoni, D. B., Aslin, R. N., Perey, A. J. & Hennessy, B. L. Identification and discrimination of a new linguistic contrast: Some effects of laboratory training on speech perception. Research on Speech Perception Progress Report No. 4, Indiana University, 1978, 49-112.

Pisoni, D. B., Carrell, T. D. & Gans, S. J. Perception of the duration of rapid spectrum changes in speech and nonspeech signals. Research on Speech Perception Progress Report No. 6, Indiana University, 1980 (In press).

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

Riley, D. A. Discrimination Learning. Boston: Allyn and Bacon, 1968.

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

Searle, C. L., Jacobson, J. E. & Rayment, S. G. Phoneme recognition based on human audition. Journal of the Acoustical Society of America, 1979, 65, 799-809.

Shiffrin, R. M. & Schneider, W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. Psychological Review, 1977, 84, 127-190.

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

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.

Stevens, K. N. & Blumstein, S. E. Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America, 1978, 64 (5), 1358-1368.

Stevens, K. N. & Blumstein, S. E. Perceptual invariance and onset spectra for stop consonants in different vowel environments. Journal of the Acoustical Society of America, 1980, 67 (2), 648-662.

Stevens, K. N. & Blumstein, S. E. Quantal aspects of consonant production and perception: A study of retroflex consonants. Journal of Phonetics, 1975, 3, 215-234.

Stevens, K. N. & Halle, M. Remarks on analysis by synthesis and distinctive features. In W. Wathen-Dunn (Ed.), Models for the Perception of Speech Visual Form. Cambridge, Ma.: Academic Press, 1967.

Stevens, K. N. & House, A. S. Speech perception. In J. Tobias (Ed.), Foundations of modern auditory theory: Volume II. New York: Academic Press, 1972.

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

Strange, W. & Jenkins, J. J. The role of linguistic experience in the perception of speech. In H. L. Pick, Jr. & R. D. Walk (Eds.), Perception and Experience. New York: Plenum Publishing Corp., 1978.

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

Trehub, S. E., Schneider, B. A. & Endman, M. Developmental changes in infants' sensitivity to octave-band noises. Journal of Experimental Child Psychology, 1980, 29, 282-293.

Walley, A. Infants' discrimination of full and partial cues to place of articulation in stop consonants. Research on Speech Perception Progress Report No. 5, Indiana University, 1979, 85-145.

Walley, A. C. & Carrell, T. D. Onset spectra vs. formant transitions as cues to place of articulation. Unpublished manuscript, 1980.

Williams, C. L. Speech perception and production as a function of exposure to a second language. Doctoral Dissertation, Harvard University, 1974.

Figure Captions

Figure 1. Adult labeling functions for synthetic labial, apical and velar stop consonants differing in VOT obtained from native speakers of English, Thai and Spanish (redrawn from Lisker & Abramson, 1967).

Figure 2. Average identification and ABX discrimination functions for two category (Group I) and three category (Group II) labeling of synthetic VOT (from Pisoni et al., 1978).

Figure 3. Illustration of the major roles that early postnatal experience can play in modifying the relative discriminability of speech sounds. Three general classes of theories are shown here to account for the development of speech sound discrimination: Universal theory, Attunement theory and Perceptual Learning theory (from Aslin & Pisoni, 1980b).

Figure 4. Oddity discrimination data obtained from adult speakers of Thai and English for synthetic bilabial stop consonants differing in VOT (redrawn from Lisker & Abramson, 1967).

Figure 5. Representation of the context-independent spectra associated with the labial, alveolar and velar places of articulation with release bursts present (after Stevens and Blumstein, 1978).

Figure 6. Onset spectra for the full and partial cue labial, alveolar and velar stimuli used by Walley (1979).

Figure 7. Testing set-up in the sound-attenuated booth employed for testing infants' discrimination in the operant head-turning paradigm (from Aslin et al., 1979).

Figure 8. Running spectral displays of the three voiced stops, /b/, /d/ and /g/ in different vowel environments (from Kewley-Port, 1979).

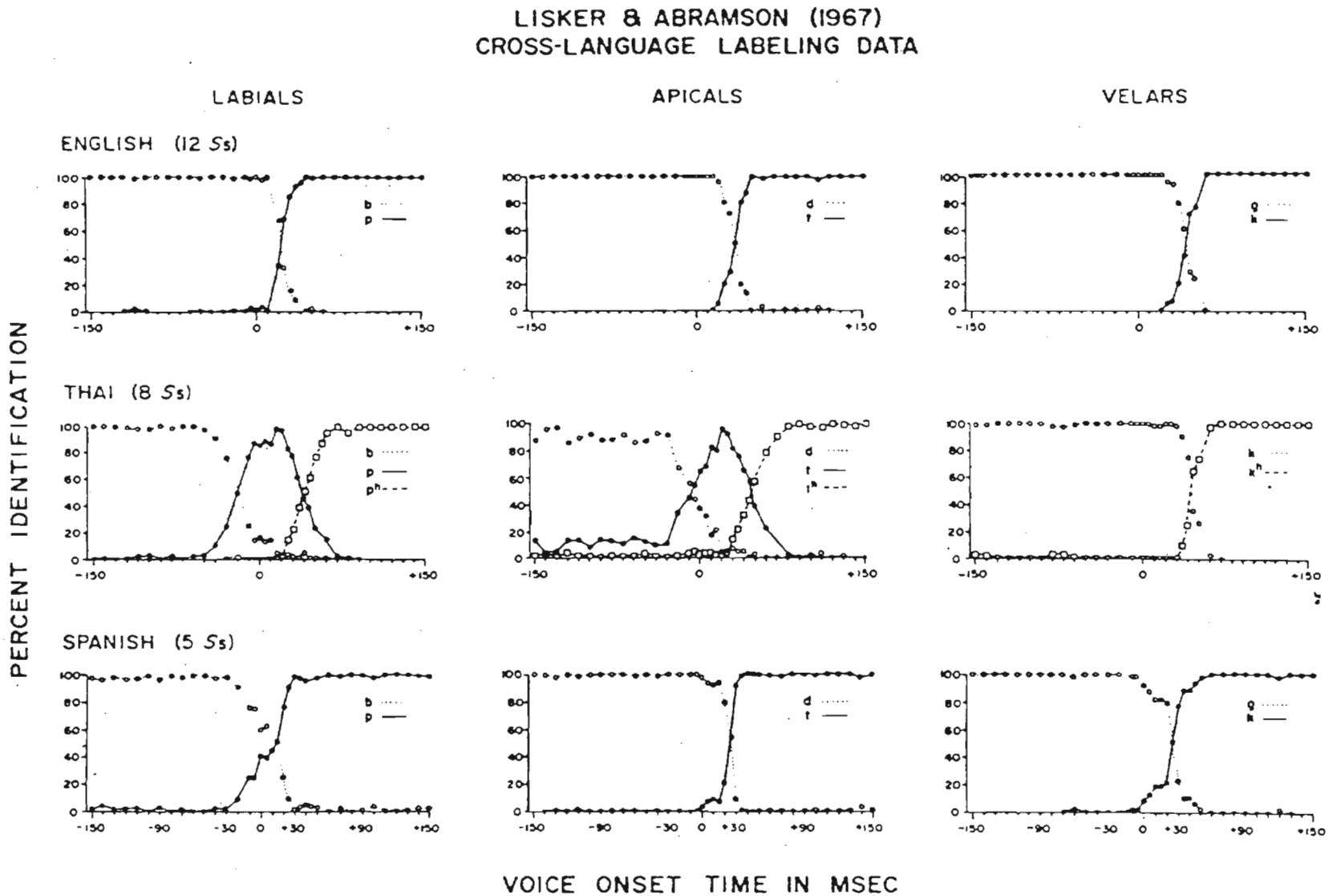

LISKER & ABRAMSON (1967)
CROSS-LANGUAGE LABELING DATA
LABIALS
APICALS
VELARS
ENGLISH (12 Ss)
THAI (8 Ss)
SPANISH (5 Ss)
PERCENT IDENTIFICATION
VOICE ONSET TIME IN MSEC
b
p
d
t
g
k
ph
th
kh
-150
-90
-30
0
+30
+90
+150

Figure 1.

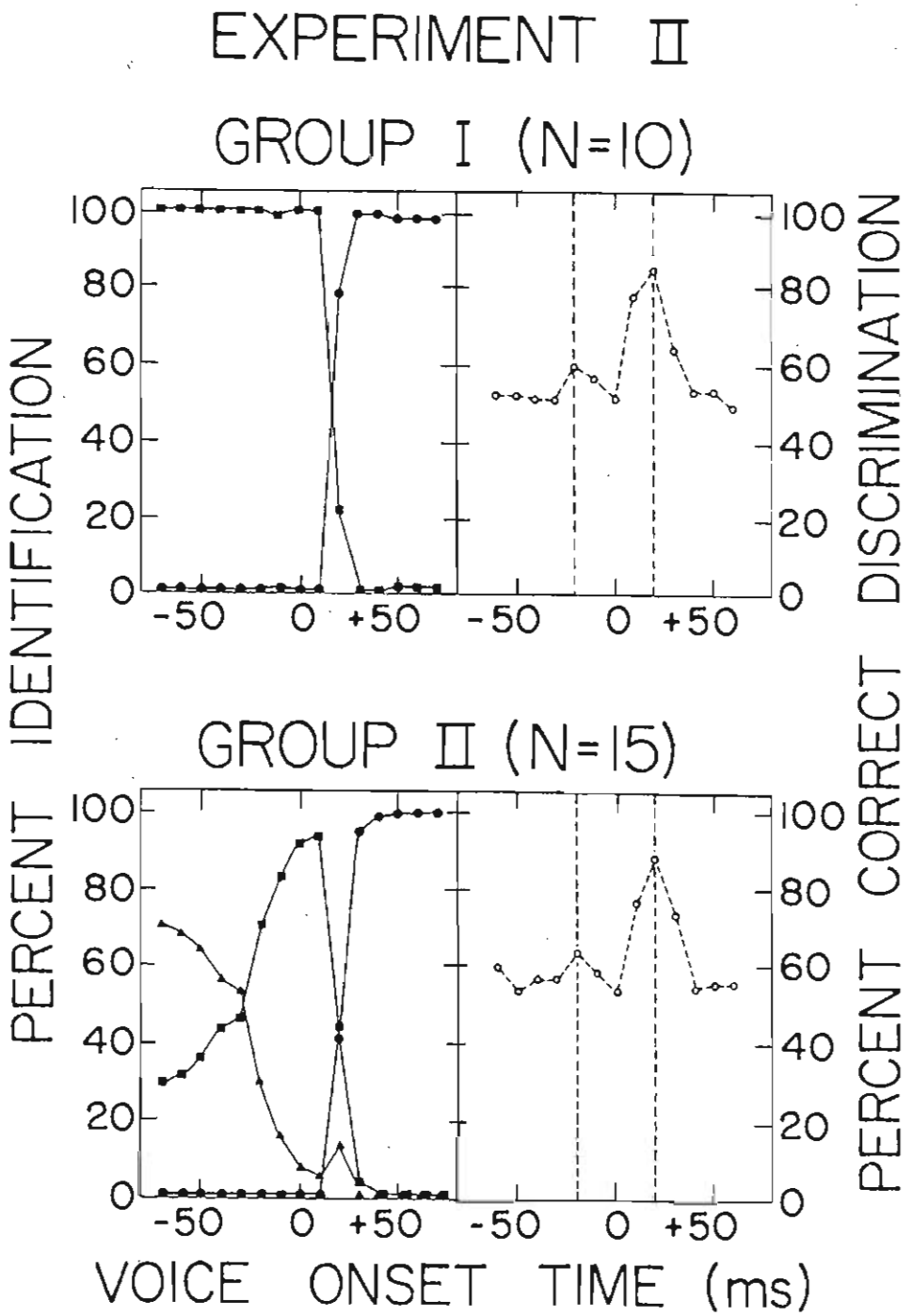

EXPERIMENT II
GROUP I (N=10)
GROUP II (N=15)
PERCENT IDENTIFICATION
PERCENT CORRECT DISCRIMINATION
VOICE ONSET TIME (ms)
100
80
60
40
20
0
-50
0
+50

Figure 2

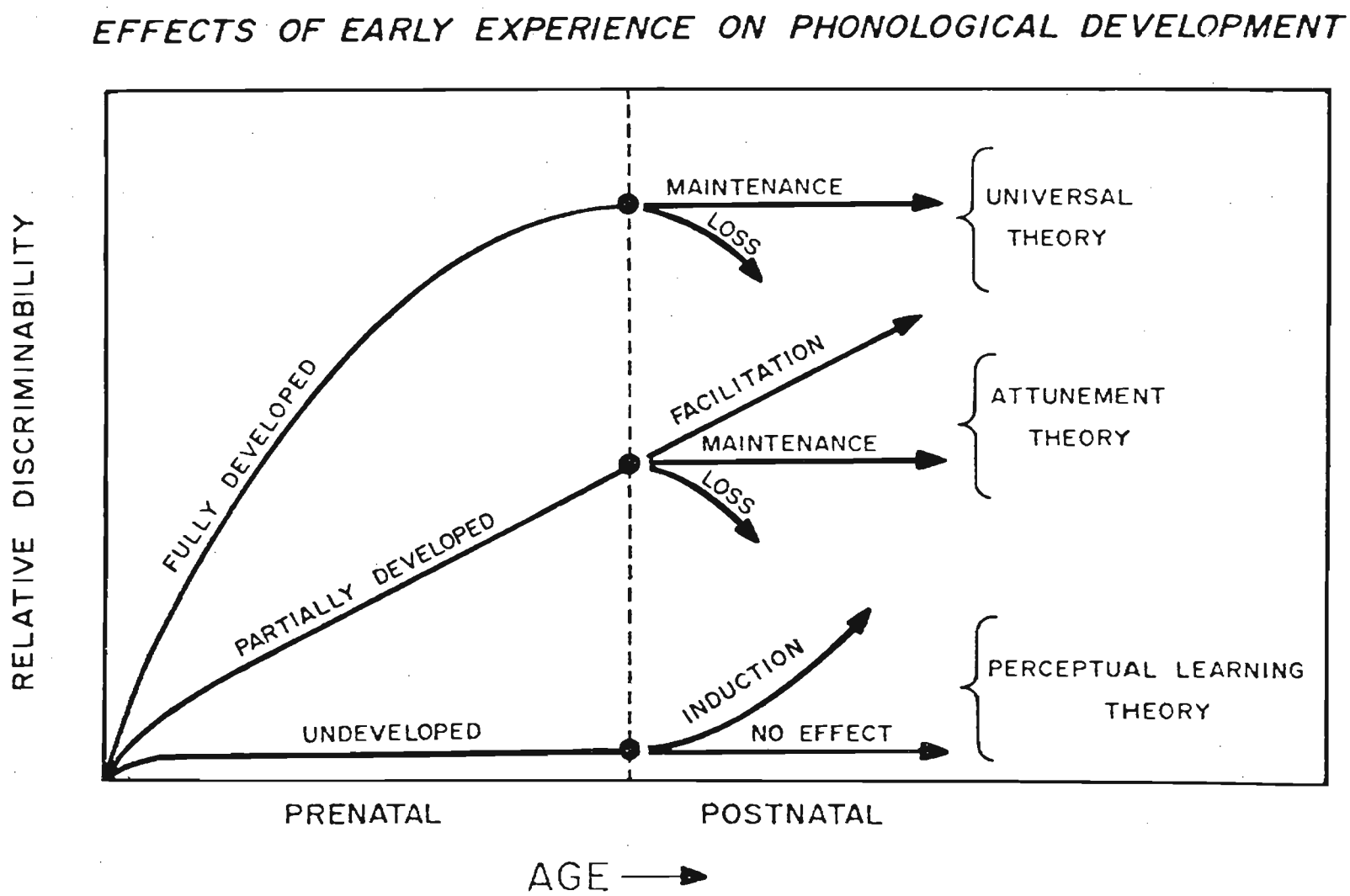

EFFECTS OF EARLY EXPERIENCE ON PHONOLOGICAL DEVELOPMENT
RELATIVE DISCRIMINABILITY
FULLY DEVELOPED
PARTIALLY DEVELOPED
UNDEVELOPED
MAINTENANCE
LOSS
FACILITATION
MAINTENANCE
LOSS
INDUCTION
NO EFFECT
UNIVERSAL THEORY
ATTUNEMENT THEORY
PERCEPTUAL LEARNING THEORY
PRENATAL
POSTNATAL
AGE

Figure 3.

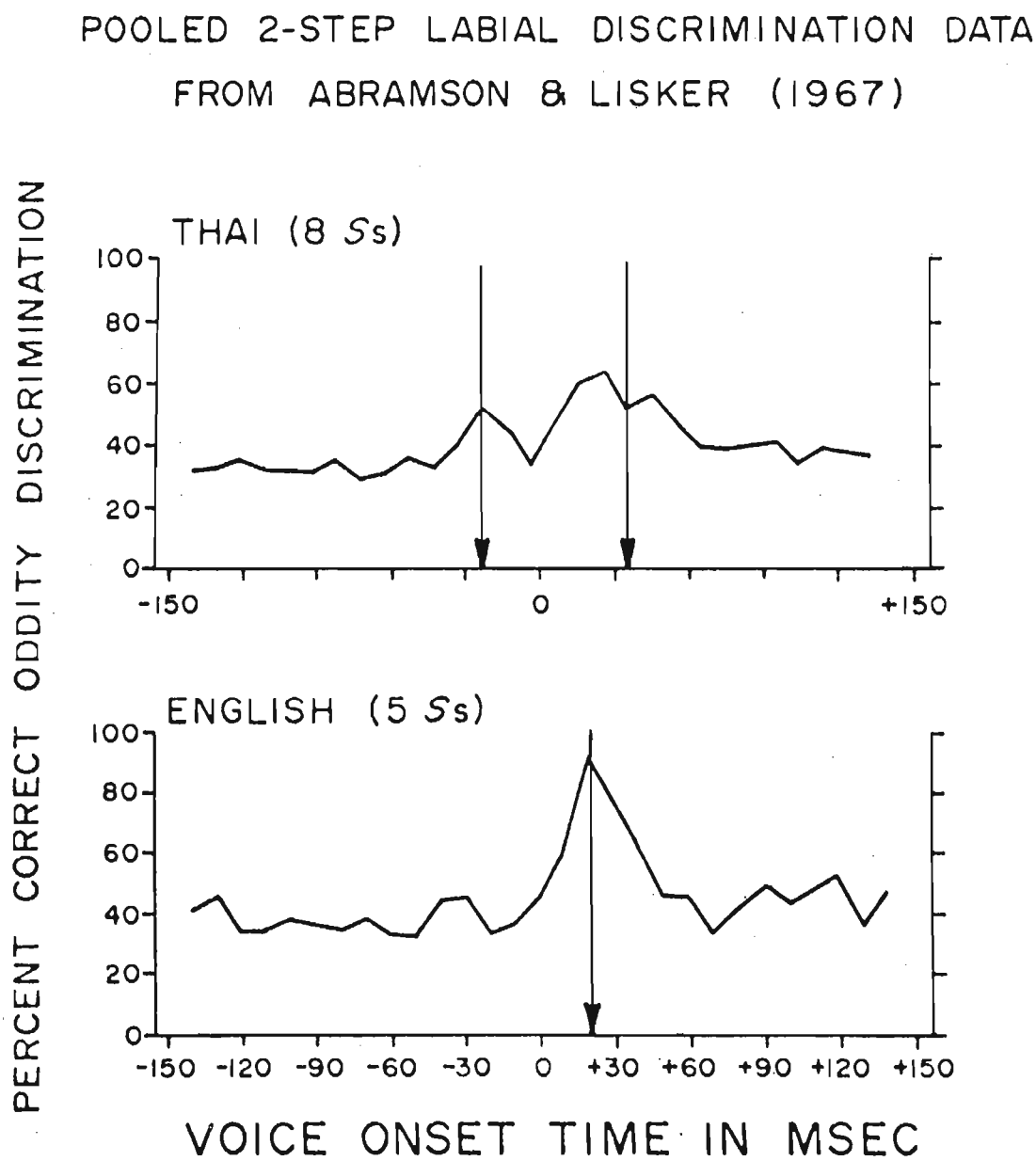

POOLED 2-STEP LABIAL DISCRIMINATION DATA
FROM ABRAMSON & LISKER (1967)
PERCENT CORRECT ODDITY DISCRIMINATION
THAI (8 Ss)
ENGLISH (5 Ss)
100
80
60
40
20
0
-150
0
+150
-150 -120 -90 -60 -30 0 +30 +60 +90 +120 +150
VOICE ONSET TIME IN MSEC

Figure 4.

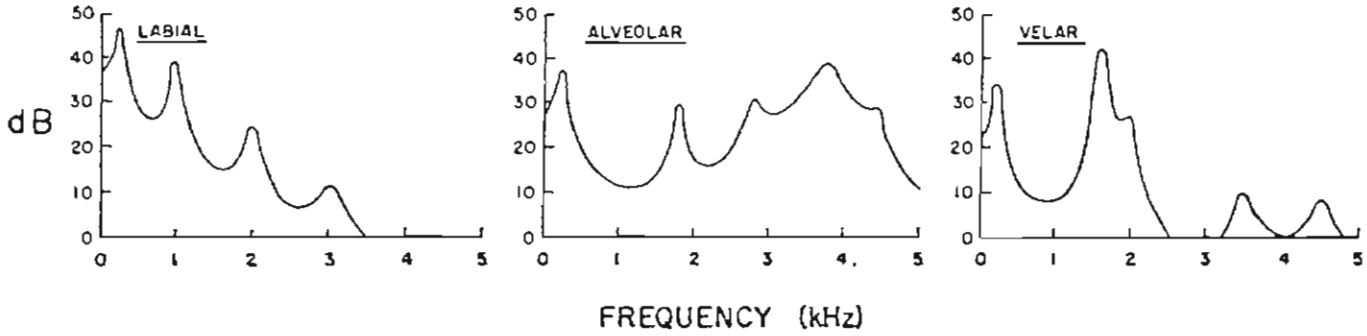

LABIAL
ALVEOLAR
VELAR
dB
50
40
30
20
10
0
0
1
2
3
4
5
FREQUENCY (kHz)

Figure 5.

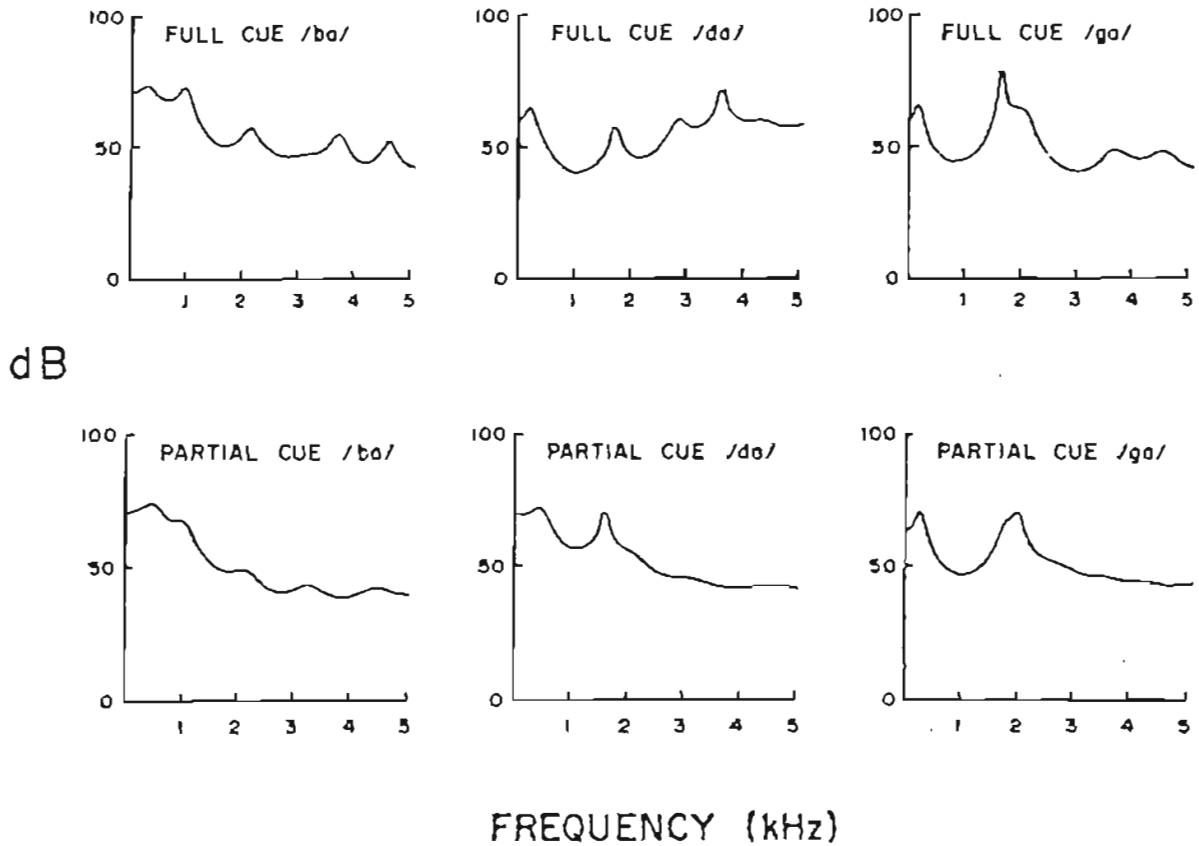

FULL CUE /ba/
FULL CUE /da/
FULL CUE /ga/
PARTIAL CUE /ba/
PARTIAL CUE /da/
PARTIAL CUE /ga/
100
50
0
1 2 3 4 5
dB
FREQUENCY (kHz)

Figure 6.

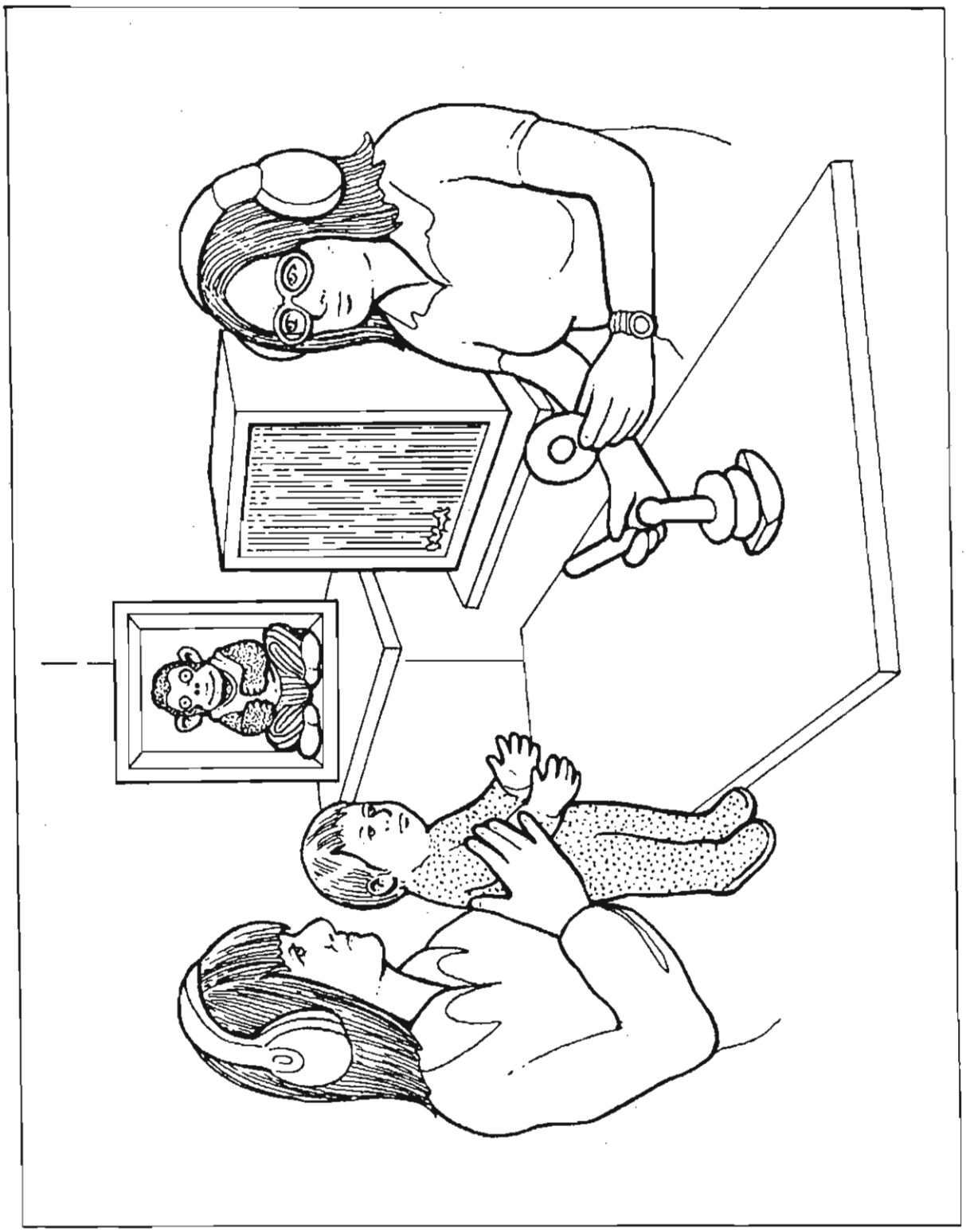


Figure 7.

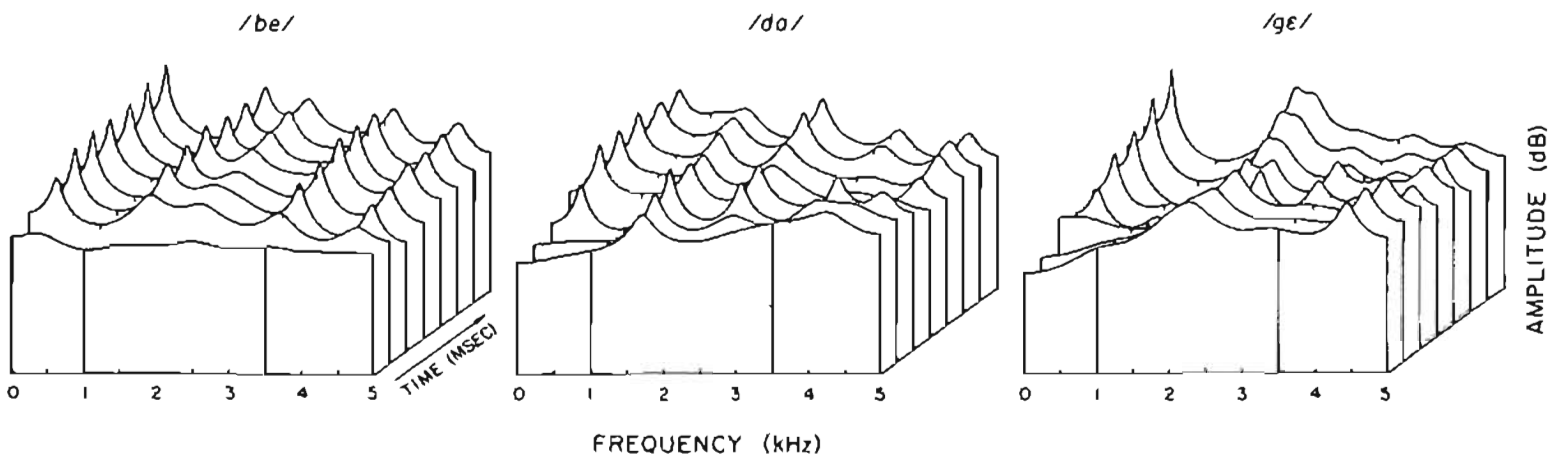

RUNNING SPECTRA FOR STOPS b, d, g
/be/
/do/
/gɛ/
TIME (MSEC)
0 1 2 3 4 5
0 1 2 3 4 5
0 1 2 3 4 5
FREQUENCY (kHz)
AMPLITUDE (dB)

Figure 8.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 6 (1979-80) Indiana University]

Effects of Target Monitoring on Understanding Fluent Speech

Michelle A. Blank, David B. Pisoni and Cynthia L. McClaskey

Department of Psychology
Indiana University
Bloomington, Indiana 47405

Short Title: Effects of Target Monitoring

Abstract

Phoneme-monitoring and word-monitoring are two experimental tasks that have frequently been used to assess the processing of fluent speech. Each task is purported to provide an "on-line" measure of the comprehension process, and each requires listeners to pay conscious attention to some aspect or property of the sound structure of the speech signal. The present study is primarily a methodological one directed at the following question: Does the allocation of processing resources for conscious analysis of the sound structure of the speech signal affect on-going comprehension processes, or the ultimate level of understanding achieved for the content of the linguistic message? Our subjects listened to spoken stories. Then, to measure their comprehension, they answered multiple-choice questions about each story. During some stories they were required to detect a specific phoneme; during other stories they were required to detect a specific word; during other stories they were not required to monitor the utterance for any target. The monitoring results replicated earlier findings showing longer detection latencies for phoneme-monitoring than for word-monitoring. Somewhat surprisingly, the ancillary phoneme- and word-monitoring tasks did not adversely affect overall comprehension performance. This result undermines the specific criticism that on-line monitoring paradigms of this kind should not be used to study spoken language understanding because these tasks interfere with normal comprehension.

Effects of Target Monitoring on Understanding Fluent Speech*

Michelle A. Blank, David B. Pisoni and Cynthia L. McClaskey

When faced with the task of understanding spoken language, listeners are rarely consçious of the sound structure of an utterance. The primary focus of the listeners conscious awareness of the speech signal is directed towards understanding the content of the message and not in analyzing its constituent elements (viz. the individual phonemes, syllables, or words). Despite this observation, subjects are nevertheless able to make reliable judgments about the detailed properties of the sound structure of an utterance while at the same time also devoting efforts toward comprehending the message. This has been demonstrated numerous times in experiments using paradigms such as phoneme-monitoring, word-monitoring, and mispronunciation detection (see, e.g., Cole & Jakimik, 1980; Foss & Blank, 1980; Marslen-Wilson & Tyler, 1980).

Phoneme-monitoring, word-monitoring, and mispronunciation detection are representative of a class of experimental techniques that have been used quite often to assess various components of fluent speech comprehension. Each of these tasks has been assumed to provide a measure of on-going comprehension processes. All involve latency measures and all are assumed to index "momentary processing load" during fluent speech perception. Each task explicitly entails directing the subjects' attention to some property of the sound structure of the speech signal while at the same time requiring listeners to comprehend

the utterance. (See Levelt, 1978, for an extensive review of studies usng tasks of this kind.)

Even though these tasks have been used extensively in the past, only recently has an interest developed in specifing the perceptual and cognitive processes involved in the tasks themselves (see, e.g. Blank, 1980a; Foss & Blank, 1980; Rudnicky, 1980). Other than some speculation among theorists, there is still relatively little known about the effects of the task demands of target monitoring on comprehension performance. This lack of knowlege on our part is by no means trivial. After all, task demands in psychological experiments, particularly experiments involving linguistic materials, have been shown to affect the perceptual organization and encoding of the stimuli (e.g., Ammon, Ostrowski and Alward, 1971; Aaronson, 1976; Cary, 1971). Thus, the validity of phoneme-monitoring, word-monitoring and mispronunciation detection tasks as measures of on-going language comprehension would appear to be limited without much more detailed knowledge about how these tasks affect the normal processes of spoken language understanding. We would not want to base our theoretical accounts of spoken language comprehension on experimental paradigms which may disrupt the integrity of the very processes we wish to study.

In this paper, we are interested in the following specific issue: Does the conscious allocation of processing resources to different levels of the sound structure of the speech signal affect the way listeners process the utterance and the ultimate level of understanding achieved for the content of the message? The experiment reported here is a preliminary investigation that

examines the potential interferring effects of phoneme-monitoring and word-monitoring on the normal comprehension process. Suppose we find that comprehension is impaired by performing the ancillary task of monitoring the speech signal for a phoneme or word throughout a passage of connected discourse. This result would then have to be taken into account when drawing inferences about language comprehension from monitoring data of this type. If we find deleterious effects of phoneme- and word- monitoring on comprehension in the present experiment, then we will have empirical justification to extend our criticisms to other popular on-line measures of fluent speech decoding, such as mispronunciation detection and speech shadowing.

By examining listeners' performance on various kinds of comprehension questions as a function of different monitoring conditions, we hope to learn something about the specific task demands of phoneme- and word-monitoring and how they may interact with ordinary comprehension processes. Of primary interest in the present study is the comparison of comprehension performance in the two monitoring conditions, on the one hand, with a non-monitoring control condition, on the other hand.

The other important question which this study addresses is whether monitoring at the word level will have the same effects on comprehension as monitoring at the phoneme level. The answer to this question bears directly on the roles of lexical and phonemic representations in speech processing. Several theorists have argued that the computation of phonemic information is not a basic, or even necessary, process in the perception and comprehension of fluent speech (Klatt, 1979; Marslen-Wilson &

Tyler, 1980; Morton & Long, 1976; Warren, 1976). Instead, these investigators claim that lexical and not phonemic representations play a primary role in understanding spoken language. If this view of the primacy of lexical interpretation in speech understanding is legitimate, then we might expect phoneme-monitoring to interfere with on-going comprehension processes more than word-monitoring. On the other hand, if phonemic as well as lexical representations are normally computed during fluent speech processing (see, Foss and Blank ,1980; Blank ,1980b), then both word- and phoneme-monitoring tasks might be expected to have more-or-less comparable effects on comprehension processes. Both targeting conditions would, nevertheless, be expected to produce selective decrements in comprehension performance when compared to the non-monitoring control condition. Both targeting tasks explicitly require a listener to make an overt response about a specific property or attribute of the sound structure of the speech signal that is not typically brought to conscious awareness during the usual course of sentence processing and language comprehension activities. A monitoring task that requires explicit attention to sound attributes may promote and perhaps even require the use of special perceptual and cognitive strategies. This, in turn, may adversely affect comprehension in ways that are currently unknown.

Method

Design: Twelve narrative stories were choosen from several published adult reading or listening comprehension tests. (See Table 1 for the exact details of the passages.) In order for each of the stories to occur in the three experimental conditions (viz. non- monitoring, word-monitoring, and phoneme-monitoring), three sets of tapes were constructed. Each set contained all 12 stories; four of the stories in the three sets came from each of the three conditions. The experiment was therefore a three (monitoring: none/word/phoneme) x three (tape sets) factorial design with the former variable within-subjects and the latter between-subjects. The comprehension test questions for each story were identical across the three tape sets.

Insert Table 1 about here.

Materials: A female speaker (MAB) recorded all 12 stories on one track of an audio tape with a professional quality microphone and tape recorder in a sound attenuated IAC booth. Each story on this master tape was preceded by the word "Ready" and three targets specifications (viz. "Do not listen for any target"; "Listen for the target word _____"; "Listen for the target sound ___"). Initially, four (of the total 12) stories were randomly assigned to each of the monitoring conditions. Then, using a roll-over design, the three tape sets were made by cross-recording the master tape and editing the target specifications so that each story occurred in each monitoring condition across the tape sets.

Presentation of the stories for each tape set was blocked by monitoring condition. Order of presentation of the monitoring condition was counter-balanced for each tape set. Thus, a tape set consisted of three tapes which differed only in the order of condition presentation. A total of nine tapes were cross-recorded.

The phoneme targets in the phoneme-monitoring condition consisted of all and only the word-initial phonemes of the word targets in the word-monitoring condition. A marking tone, inaudible to subjects, was placed on the second track of the audio tapes at the beginning of each word-initial target phoneme (or target word). The tone started a timer which stopped when subjects pressed a reponse button. Timing and data collection were controlled by a PDP 11/05 computer.

Response booklets for measuring comprehension of the stories were prepared for each tape. The booklets contained a varying number of multiple-choice questions keyed to each story. The order of pages was determined by the presentation schedule of the stories on a given audio tape. Performance on these post-test questions was used to provide an objective measure of listening comprehension. There were a total of 48 questions. Twenty questions were factual in nature, requiring nothing more than recall of some explicitly stated information contained in the story; the remaining 28 questions required listeners to understand the implications of ideas and propositions expressed in the passages and to integrate these ideas with general knowledge.

Subjects: Forty-two naive students at Indiana University in Bloomington served as paid subjects in this study. They were recruited by means of an advertisement, and each reported no history of a hearing or speech disorder at the time of testing. The subjects were all right-handed, native speakers of English. Fourteen subjects were assigned to each tape set.

Procedure: Subjects were tested in groups of one to five. Each subject was seated in a booth out of direct sight of the others in a small testing room used for speech perception experiments.

Prerecorded instructions were presented at the beginning of each tape. The instructions and stories were presented binaurally over TDH-39 headphones. A typed copy of the instructions was placed at the front of each booklet to allow subjects to read along as the instructions were read aloud. Typed copies of the stories were not available to subjects.

The instructions emphasized that the primary concern of the experiment was to study how listeners understand and remember spoken stories. Subjects were told that they would hear several short stories about a wide variety of topics and that their task was to answer the multiple-choice questions keyed to each story. They were told to do as well as they could based on the information contained in the story they heard.

Subjects were also told that for some stories they would also be asked to listen for a particular target; either a word target or a word-initial sound target would be specified before the start of each story. In these cases each subject was required to press a response button in front of them whenever

they detected the presence of a particular target. The instructions emphasized that it was important to listen for the target throughout the entire story because it would occur several times. Speed and accuracy of responding were also stressed. Subjects were explicitly told, however, not to let the task of listening for a target interfere with their attempts to understand the story because they would still have to answer comprehension questions about stories with targets in them.

Subjects were presented with the test stories in a self-paced format. The experimenter was present in the testing room and operated the tape recorder via remote control. Each story was presented only once for listening. After each story, subjects immediately turned their booklets to the appropriate set of test questions and answered them by circling one of several response choices in pencil.

Subjects heard three practice stories, one from each monitoring condition before actual testing began. They answered two comprehension questions for each practice story. After the experimenter answered questions clarifying the procedures and instructions, the test stories were presented. The entire experiment lasted about 45 minutes.

Results

Mean latencies for the phoneme- and word-monitoring conditions were computed for each subject. These means were based on all reaction times that were longer than 100 msecs and shorter than 1600 msecs. Reaction times outside this range were presumed to reflect anticipation, momentary inattention, or some other type of unusual processing strategy on the part of the

listener. The overall means for the phoneme- and word-monitoring conditions are shown in Table 2. The total number of missed targets for these conditions was also computed for each subject. Table 2 presents the mean number of misses for both monitoring conditions.

Insert Table 2 about here.

As shown in Table 2, monitoring latencies are shorter for detecting words than phonemes. The observed pattern of reaction times is consistent with other reported findings of shorter latencies to word targets than to phoneme targets (see, e.g., Foss & Swinney, 1973; Savin & Bever, 1970). The results of a t-test for matched samples showed that the 93 msec difference between the two conditions was highly significant [$t(41) = 4.68$, $p<.001$].[1] The difference between the mean number of misses for word and phoneme targets, though small, was also reliable by a t-test for independent samples [$t(82) = 2.91$, $p<.01$].

The multiple-choice comprehension questions for each of the twelve stories were scored separately for each subject. A composite error score was then obtained by accumulating across subjects, the individual error scores for all the stories within each monitoring condition. This value was then expressed as a percentage of the total possible errors. The overall error scores for the three monitoring conditions were word-monitoring: 27.3%; phoneme-monitoring: 30.5%; control: 26.5%. These data are shown in panel A of Figure 1. None of the planned comparisons using independent t-tests resulted in significant

differences among the three conditions.

Insert Figure 1 about here.

Panel B of Figure 1 presents comprehension performance for each monitoring condition broken down in terms of errors on inferential and factual questions. For inferential questions error scores were: word-monitoring: 30.8%; phoneme-monitoring: 31.1%; control: 28.8. For factual questions error scores were: word-monitoring: 22.5%; phoneme-monitoring: 25.7%%; control: 21.1%. None of the error scores on the inferential questions were significantly different from one another. This was also true for performance on the factual questions.

Discussion

Overall performance on the comprehension questions suggests that conscious focusing of a listener's attention on properties of the sound structure of the speech signal does not adversely affect spoken language understanding (at least under the conditions examined in the present experiment). Comprehension questions were responded to at similar levels in the word-, phoneme- and non-monitoring conditions. It is noteworthy that the level of comprehension performance observed in this experiment, about 70% correct, is similar to that obtained in a recent listening comprehension study reported by Pisoni (1979) using synthetic speech produced by rule. The approximate 30% error rate indicates that the level of difficulty of these stories was relatively high. Such performance levels reduce the possibility that the comprehension task was so easy for listeners

that the expected monitoring by comprehension interaction would not be observed due to the presence of ceiling effects.[2]

The absence of a significant monitoring by comprehension interaction in this study is important because critics of on-line monitoring paradigms as measures of listening comprehension have assumed, without empirical support, that conscious attention to the sound stucture of an utterance at any level interferes with normal comprehension. Since the existence of a monitoring x comprehension interaction was an intuitively sound prediction, it is quite appropriate to examine any, and all, reasonable alternative interpretations of the present findings.

Perhaps subjects chose to sacrifice performance on the monitoring task to insure the availability of processing resources for comprehension. This possibility seems ruled out on several accounts. First, as already pointed out, the pattern of monitoring latencies reported here replicates earlier findings of shorter RTs to word targets than to phoneme targets. Second, absolute RTs for word and phoneme targets are also quite similar to earlier studies (about 500 msecs). Third, the number of detection misses obtained for each target type are in reasonable accord with other monitoring studies (about four percent). Fourth, as in other monitoring studies, there is no evidence of a speed-accuracy trade-off.

The reasonable conclusion to draw from the present data, then, is that listening comprehension does not appear to be impaired by simultaneous monitoring of single words or phonemes. This is an unexpected result for it undermines the specific criticism that on-line monitoring paradigms of this kind

interfere with normal speech processing. More specifically, phoneme- and word-monitoring data may not be neglected on the grounds that the tasks interfere with normal comprehension.

In the introduction we asked whether monitoring at the word level would have the same effects on comprehension as monitoring at the phonemic level. We hoped that the answer to this question would provide some insights into the respective roles of lexical and phonemic representations in speech processing. The present findings do not support the view that only lexical representation play a critical role in understanding spoken language (as Marslen-Wilson & Tyler, 1980, have claimed). This view would predict that word- and phoneme-monitoring should have differential effects on comprehension performance; our results show they did not. The finding of equivalent effects is consistent with the view that both phonemic and lexical representations are normally computed during fluent speech processing, and that both are important for spoken language understanding (see, e.g., Foss & Blank, 1980). We find little evidence to support the contention that one level (lexical) is more primary or central than another (phonemic) to understanding fluent speech.

The present results indicate that the ultimate level of understanding achieved for the content of a spoken message is unaffected by an ancillary monitoring task. The present results do not, however, shed light on whether on-going comprehension processes are affected by the task demands of word- or phoneme-monitoring. It may be, for example, that monitoring slows down the speed at which listeners can execute the processes

involved in comprehension, while leaving the end-product (namely, understanding) intact. An empirical test of this requires collecting latencies to comprehension questions as well as error rates. Experiments of this kind currently underway in our laboratory.

In summary, the results of the present study suggest that spoken language understanding is not adversely affected by simultaneous, conscious monitoring of phonemic or lexical properties of the speech signal. Subject's understanding of spoken stories, as measured by performance on multiple-choice comprehension questions, did not differ across phoneme-, word-, and no-monitoring conditions. This result calls into question the specific claim that target monitoring tasks selectively interfere with understanding fluent speech.

Footnotes

* This paper is based on research conducted at Indiana University in Bloomington. The work reported here was supported by NIH training grant NS-07134-01 and NIMH research grant MH-24027-06. We thank Jerry C. Forshee and David Link for their technical help, and Robert E. Remez for comments on an earlier draft of the paper. Reprints may be obtained from either Michelle Blank at Bell Laboratories, Whippany, New Jersey, 07981, or David Pisoni at Indiana University.

1 The interested reader is referred to Foss, Harwood and Blank (1980) for a recent discussion of previous interpretations ascribed to this reaction time difference. In particular, Foss et al discuss why it is a mistake to assume the following: (1) that the order of reaction times obtained in monitoring experiments reflects the order in which perceptual entities (like phoneme and words) are derived by listeners, and (2) that the perceptual entities derived the earliest are the primary units for speech processing. These two theoretically naive assumptions have led some theorists to conclude, apparently prematurely, that monitoring latencies are shorter for words than for phonemes because the former, not the latter, are the basic units of speech perception.

2 Were we to have increased the difficulty of the subsidary monitoring task by requiring detection of several different target items, we might have succeeded in affecting comprehension performance (see, e.g., Logan, 1979). Such a finding, while interesting in and of itself, would not weaken any of the conclusions drawn from the present study. After all, the

standard methodology used in monitoring experiments investigating fluent speech processing does not involve listening for several different target items.

References

Aaronson, D. Performance theories of sentence coding: Some qualitative observations. Journal of Experimental Psychology: Human Perception and Performance, 1976, 2, 42-55.

Ammon, P.R., Ostrowski, B., & Alward, K. Effects of task on the perceptual organization of sentences. Perception and Psychophysics, 1971,10, 361-363.

Blank, M. A. Measuring lexical access during sentence processing. Perception and Psychophysics, 1980a,

Blank, M. A. Decoding fluent speech: The role of sensory- and knowledge-based proceses. Manuscript submitted for publication, 1980b.

Carey, P.W. Verbal retention after shadowing and after listening. Perception and Psychophysics, 1971, 9, 79-83.

Cole, R.A., & Jakimik, J. A model of speech perception. In R.A. Cole (Ed.), Perception and production of fluent speech. New Jersey: Lawrence Erlbaum Ass., 1980, 133-164.

Foss, D.J., & Blank, M.A. Identifying the speech codes. Cognitive Psychology, 1980, 1, 1-31.

Foss, D.J., Harwood, D.A., & Blank, M.A. Deciphering decoding decisions: Data and devices. In R.A. Cole (Ed.), Perception and production of fluent speech. New Jersey: Lawrence Erlbaum Ass., 1980, 165-200.

Klatt, K.S. Speech perception: A model of acoustic-phonetic analysis and lexical access. Journal of Phonetics, 1979, 7, 279-312.

Levelt, W.J.M. A survey of studies in sentence perception: 1970-1976. In W.J.M. Levelt and G.B. Flores d'Arcais (Eds.), Studies in the perception of language. New York: John Wiley & Sons, 1978, 1-73.

Logan, G.D. On the use of a concurrent memory load to measure attention and automaticity. Journal of Experimental Psychology: Human Perception and Performance, 1979, 5, 189-207.

Marslen-Wilson, W.D., & Tyler, L.K. The temporal structure of spoken language understanding. Cognition, 1980, 8, 1-71.

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

Pisoni, D.B. Some measures of intelligibility and comprehension. Research on Speech Perception Progress Report No. 5, Indiana University, 1979, 3-48.

Rudnicky, A. Structure and familiarity in the organization of speech perception. Unpublished dissertation thesis, Carnegie- Mellon University, Pittsburgh, Pennsylvania, 1980.

Warren, R. Auditory illusions and perceptual processes. In N.J. Lass (Ed.) Contempory issues in experimental phonetics, New York: Academic Press, 1976, 389-417.

Table 1

Description of Stories and Questions

Used to Measure Comprehension

Story Number	General Topic	Number of Words	Target Phoneme	Target Word	Target Frequency	No. of Comprehension Questions		
						Factual	Inferential	Total
1	Measuring Star Distances	161	/d/	distance	4	1	2	3
2	Verbal Communication	315	/g/	group	5	1	5	6
3	An Outdoor Scene	278	/g/	green	3	2	1	3
4	Insufficient Oxygen	363	/b/	behavior	8	1	4	5
5	Aluminum	273	/m/	metal	5	3	3	6
6	Carbon Dating	374	/c/	carbon	14	2	3	5
7	The American Party System	490	/b/	both	5	2	3	5
8	Effective Communication	343	/t/	talk	12	3	1	4
9	Basketball	270	/m/	men	4	1	2	3
10	Science Fiction	271	/b/	books	5	2	0	2
11	Self-Actualization	228	/g/	goal	2	1	2	3
12	Hair Today; Gone Tomorrow	313	/b/	bald	10	1	2	3

Table 2

Mean latencies (msecs) and mean number of misses

for word- and phoneme-monitoring conditions

	Word Monitoring	Phoneme Monitoring
Mean Reaction Times	574	667
Mean misses*	4.5	7.5

*26 was the total possible mean number of misses.

Figure Caption

Error rates for comprehension questions as a function of monitoring condition. Overall performance is shown in Panel A; preformance broken down in terms of factual and inferential questions is shown in Panel B.

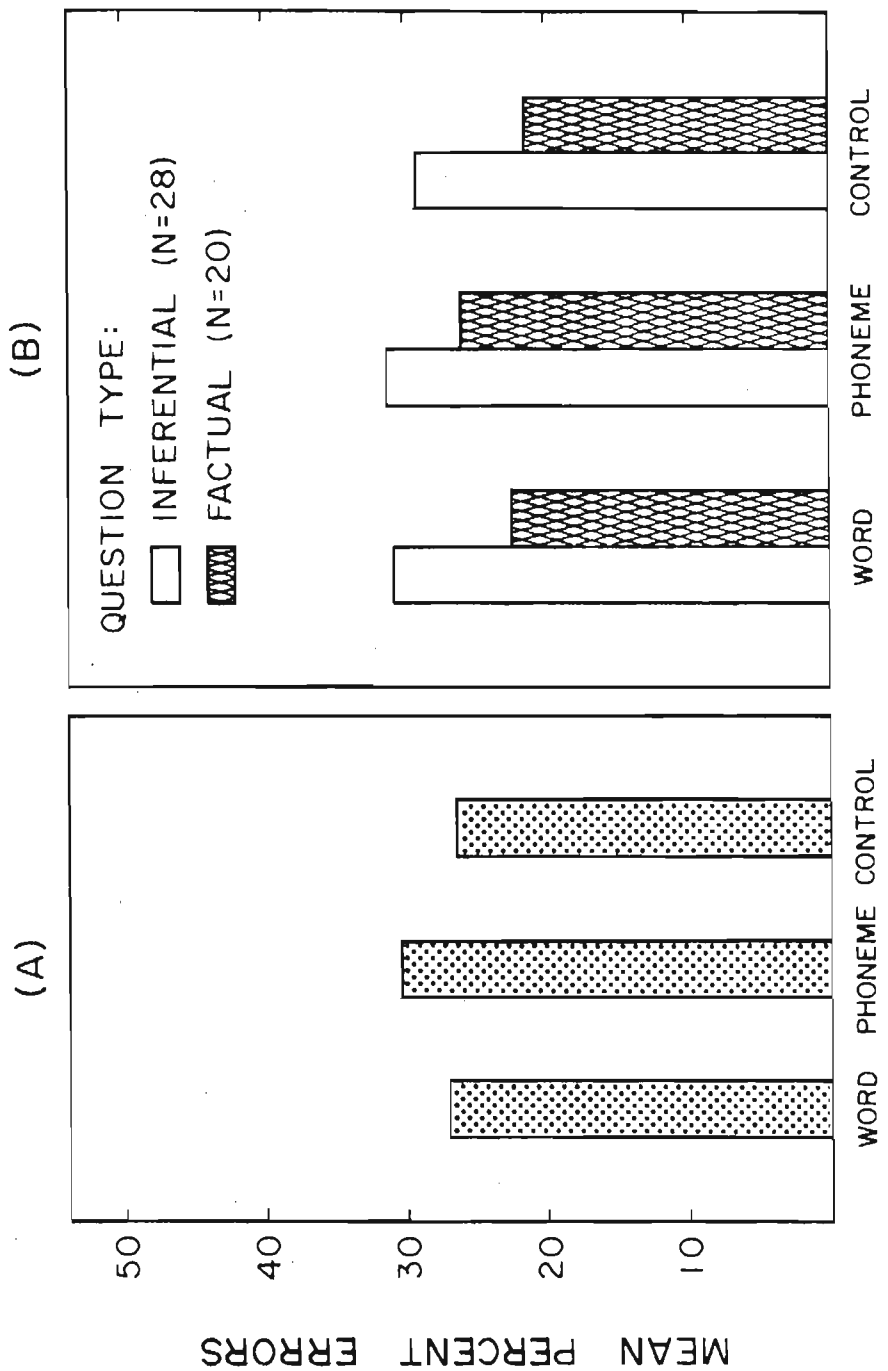

(A)
(B)
QUESTION TYPE:
INFERENTIAL (N=28)
FACTUAL (N=20)
MEAN PERCENT ERRORS
50
40
30
20
10
WORD
PHONEME
CONTROL
TARGET MONITORING CONDITION

Figure 1.

Effects of Transfer of Training on Identification of a New Linguistic Contrast in Voicing

Cynthia L. McClasky, David B. Pisoni and Thomas D. Carrell
Indiana University
Bloomington, Indiana 47405

Running Head: Transfer of Training of a New Linguistic Contrast

Abstract

The present study examined the plasticity of the human auditory system by means of laboratory training procedures that were designed to modify the perception of the voicing dimension in synthetic speech stimuli. Although the results of earlier laboratory training studies have been ambiguous, Pisoni, Aslin, Perey, and Hennessy (1978) have succeeded recently in altering the perception of labial stop consonants from a two-way contrast in voicing to a three-way contrast. The present study extended these initial results by demonstrating that knowledge of VOT gained from discrimination training on one place of articulation (e.g., labial) can be transferred to another place of articulation (e.g., alveolar) without additional training on the specific test stimuli. Quantitative analyses of these transfer effects in identification showed that the perceptual categories which emerged were stable and displayed well-defined labelling boundaries between categories. Taken together with the earlier findings, these results imply a greater degree of plasticity in the adult human perceptual system than has generally been recognized in past studies. Although the linguistic environment exerts an important influence on the perception of speech sounds, laboratory training methods designed to selectively focus the listener's attention to the relevant acoustic cues were highly successful in modifying the perception of at least one synthetic continuum in a relatively short amount of time. These results have implications for previous conceptions of the role of linguistic experience in perceptual development and the usefulness of specific laboratory training procedures in modifying the perception of speech sounds.

Effects of Transfer of Training on Identification of a New Linguistic Contrast in Voicing[1]

This report is concerned with the perceptual categorization of human speech sounds, particularly the categorization of the voicing dimension in stop consonants. The voicing dimension has been studied extensively in recent years in an attempt to characterize the interaction between genetic and environmental influences in speech perception. In the word initial English stop consonants /b, p, d, t, g, k/, voice onset time (VOT) is defined as the interval between the onset of vocal cord vibration and the release of articulatory closure. The English phonemes, however, represent only a subset of the modes of voicing that are possible in human language systems. By measuring the production of stop consonants in eleven diverse languages, Lisker and Abramson (1964) found three universal modes of voicing. The first is long lead, or prevoiced stops in which vocal cord vibration precedes the release. Lead values generally occur at -25 msec VOT or greater. The minus sign indicates that the release burst occurs after the onset of laryngeal pulsing. The category of short lag stops is observed when laryngeal and supralaryngeal events are nearly simultaneous. It falls between approximately -20 and +25 msec VOT. Long lag or voiceless aspirated stops in English are characterized by substantial voicing delay until after the release. These stops have VOT values greater than +25 msec. The English stops /p, t, and k/ are exemplars of the long lag mode of voicing, while /b, d, and g/ are referred to as voiced. Long lead and short lag stops

are allophonic in English so that, for example, a /b/ produced with -30 msec VOT and a /b/ at 0 msec VOT are both still perceived as the same phoneme. Lisker and Abramson found that each language they studied used only a subset of the three voicing modes that are possible at each place of articulation.

In subsequent cross-language experiments using synthetic speech stimuli differing in VOT, Lisker and Abramson (1967) found that subjects perceived the stimuli categorically. The identification and discrimination functions both corresponded closely to the phonemic categories observed in each language. The identification or labeling functions were consistent and displayed steep slopes at the boundaries separating perceptual categories. In addition, subjects showed poor discrimination of stimuli within the same phonemic category and improved discrimination of stimuli at the category boundary. This combination of results suggests that the ability to discriminate between speech stimuli is closely related to their identification (Liberman, Harris, Hoffman, and Griffith, 1957). Thus, although the occurence of three primary modes of voicing implies that certain regions of the voicing continuum may be naturally discriminable (Streeter, 1976), the perceptual categories used by adults appear to be determined largely through linguistic experience.

The effects of linguistic experience on speech perception have been studied to a lesser extent with other phonemic contrasts. Goto (1971) and Miyawaki, Strange, Verbrugge, Liberman, Jenkins, and Fujimura (1975) have examined the abilities of Japanese and English subjects to identify and discriminate the liquids /r/ and /l/. These consonants are phonemically distinctive

in English but not in Japanese. In both studies, the Japanese subjects showed poorer discrimination of /r/ and /l/ than the English subjects. In Goto's study, the Japanese even had difficulty understanding recorded tokens of their own speech. He attributed this result to the lack of experience with the target phonemes and stated that "auditory discrimination of (non-native) syllables must be acquired by daily practice from childhood" (Goto, 1971, p. 322).

In general, these cross-language studies have demonstrated that linguistic experience plays an important role in facilitating the discrimination of linguistic contrasts. However, the permanence of this environmental influence and the ease with which the perceptual system of mature adults can be modified is a more controversial issue. Strange and Jenkins (1978) recently reviewed a number of laboratory training experiments which tried to alter the perception of the voicing continuum. Although the ability of some adults to discriminate non-native contrasts in voicing showed improvement, Strange and Jenkins concluded that "significant modification of phonetic perception is not easily obtained by simple laboratory training techniques" (Strange and Jenkins, 1978, p. 153). Moreover, they observed that the modifications that did occur appeared to be limited to the same test stimuli and the same perceptual tasks that were used in training. The possibility exists, of course, that even those subjects whose perceptions were altered were not learning to modify the voicing continuum as such. Hence the generalizability of these results is severely restricted. Strange and Jenkins also concluded, as have most researchers over the past two decades, that the adult perceptual

system is extremely resistant to change and that selective retuning or modification through laboratory procedures is an arduous if not impossible process.

In an attempt to add a voiced/voiceless distinction to the repertoire of native speakers of Russian, Lisker (1970) presented exemplars of the two categories (+10 and +60 msec VOT) alternately for comparison. The Russian subjects were not required to respond during this training phase and they were never provided with feedback about their responses. Following training, all subjects were presented with the +10, +20, +30, +40, +50 and +60 msec VOT in random order and they were required to judge whether the stimuli were more similar to the +10 or the +60 alternative. From the results of this experiment, Lisker concluded that the Russians were not able to successfully learn contrasts that were not distinctive in their native language.

In another study, Strange (1972) measured the ability of a small number of subjects to identify and discriminate stop consonants that varied in VOT. Subjects were first presented with exemplars from each of two categories, prevoiced and voiced. Then they were trained using an oddity discrimination task with verbal feedback after each trial. Subjects were then tested in the same task but without feedback. Although all four subjects in this study showed some general improvement in discrimination, their performance was poorest at the prevoiced/voiced boundary which was the focus of training. Strange concluded that her subjects were unable to learn this three-way contrast in VOT. Other studies carried out by Strange (1972) included identification and transfer of training tasks and these also showed no consistent improvement in learning a third category along the VOT continuum.

It is possible that the failure of these previous attempts to produce new linguistic contrasts reflects a genuine absence of the required perceptual abilities. However, other studies indicate that the problem may be more methodological. For example, Carney, Widin, and Viemeister (1977) have demonstrated that discrimination within phonemic categories is possible when the relevant acoustic cues are emphasized during training. In addition, the subjects were able to change their category boundaries to arbitrary values chosen by the experimenters. However, the subjects in the Carney et al. study were highly experienced veterans of psychophysical experiments and they had been participants of many testing sessions before the data were collected.

In contrast, the subjects used in a recent study by Pisoni, Aslin, Perey, and Hennessy (1978) were naive undergraduates who were able to modify their perception of VOT in less than one hour of exposure to the new contrast. By providing immediate feedback during training in order to emphasize the relevant acoustic cues, Pisoni et al. were able to show that naive listeners could learn a new voicing category. However, one weakness of the experimental design employed by Pisoni et al. was first pointed out by Strange (1972) in reference to her own work. Because the subjects were both trained and tested on the same synthetic labial stop consonants, it is not clear whether the subjects were actually learning about the specific place of articulation or about the distinctive feature (in this case VOT) in general.

In a recent study dealing with the discrimination of VOT, Edman, Soli, and Widin (1978) examined the ability of subjects to transfer knowledge gained on one place of articulation to another.

Subjects were given initial and final discrimination and identification tests using the /b-p/ and /g-k/ continua of Lisker and Abramson, but they were trained on only one of the series. For a subject who was trained on the labial continuum, the discrimination training effects were substantial for the labial stimuli as might be expected. However, the subject also demonstrated transfer of training in discrimination suggesting that subjects can learn "VOT per se and not some unique properties of the training series" (Edman, 1978, p. 5).

The present investigation was designed to extend these findings on discrimination to absolute identification of a new linguistic category in voicing. Although Pisoni et al. demonstrated that naive subjects could learn to identify a new linguistic category they did not establish that the discrimination training and experience with the new linguistic category transferred to other places of articulation. Although Edman et al. (1978) showed transfer of training in discrimination, they did not assess whether the discrimination training transferred to identification of a new voicing category.

Method

Subjects. Fifteen subjects were drawn from a paid subject pool of undergraduate students attending Indiana University. The subjects were originally recruited through advertisements and were paid $3.00 per hour for each testing session. All subjects were monolingual speakers of English with no history of a speech or hearing problem as determined by a pretest questionnaire. All subjects were naive to the experimental procedures used in the experiment.

Stimuli. The stimuli consisted of two sets of fifteen synthetic stop consonant-vowel syllables corresponding to labial and alveolar places of articulation. Each stimulus set was generated on the Klatt (1980) cascade-parallel software synthesizer as implemented in the Speech Perception Laboratory at Indiana University (Kewley-Port, 1978). The stimuli differed in ten millisecond steps of VOT from -70 to +70 msec. The steady-state portion of the stimuli consisted of the vowel /a/ which was 255 msec in duration. The formant values chosen were: F1 = 700 Hz, BW1 = 90 Hz; F2 = 1200 Hz, BW2 = 90 Hz; F3 = 2600, BW3 = 130 Hz; F4 = 3300 Hz, BW4 = 400/Hz; F5 = 3700 Hz, BW5 = 500Hz. The labial formant transitions were 40 msec in duration and had starting frequencies of F1 = 438 Hz, F2 = 1025 Hz, F3 = 2425 Hz which were chosen to simulate the production of natural speech as measured by Klatt (1978) and from measurements of the spectrograms of a male speaker in our laboratory. The alveolar formant transitions were 50 msec. The starting frequencies were: F1 = 400 Hz, F2 = 1550, F3 = 2600 Hz. To simulate the burst that occurs at release of stop closure, a turbulent noise source (AF) was passed through the bypass channel of the parallel branch of the synthesizer. The burst was ten msec. in duration with the amplitude carefully chosen on the basis of pilot listening tests. The spectrum of the labial release burst was distributed fairly uniformly across all frequencies, while the energy of the alveolar burst was centered around 3300-3850 Hz. Voicing lead was simulated by setting F1 at 180 Hz with a bandwidth of 150 Hz. The sinusoidal voicing source (ASV) was then passed through F1 to simulate the low frequency energy. Voiceless

stimuli have an aspirated component produced by turbulence at the opening of the vocal cords. Aspiration was simulated by passing a noise source through the cascade branch of the synthesizer. Aspiration noise was also added to the final 35 msec of the syllable and F1 was widened somewhat to make the offsets sound more natural. At the onset of release of stop closure, the pitch contour rose briefly from 120 to 125 Hz and then fell linearly to 100 Hz and remained there for the duration of the steady-state portion of the vowel.

Procedure. The experiment was conducted with groups of two to six subjects seated in separate testing carrels in a small experimental room. The stimuli were presented binaurally over matched and calibrated TDH-39 headphones. Voltage levels were maintained with a VTVM at a constant level of 80 dB SPL during presentation of stimuli. The collection of all responses and presentation of feedback to subjects was controlled on-line in real-time by a PDP-11/05 computer. All experimental trails were paced to the slowest subject in each group.

On Day 1, the subjects were presented with four different phases of the experiment, each phase was described with a typed page of instructions which were also read aloud by an experimenter just prior to the presentation of stimuli. In the first phase, the subjects were asked to identify the fifteen stimuli (each presented ten times in random order) into two categories corresponding to the English categories /ba/ and /pa/. The onset of a trial was signalled by the illumination of a cue light on top of the response box.

In the next phase, familiarization, subjects merely listened to several ordered presentations of the -70, 0 and +70 msec VOT stimuli. Subjects did not respond to the stimuli but were instructed to concentrate on listening to the beginning of the consonant-vowel syllable. The third phase consisted of 40 presentations of the three exemplary stimuli (i.e., -70, 0 and +70 msec VOT) in a random order. Subjects were required to use three response categories labeled /mba/, /ba/, and /pa/. Immediately following each identification response, a light above the correct response button was illuminated to provide feedback to the subject indicating the correct response. Those subjects who met a predetermined criteria of 85% correct in this phase of the training were asked to return for a second day of testing. Finally, in the last phase of Day 1, all fifteen stimuli were again presented ten times each in random order. However, now subjects were required to identify the stimuli using all three response categories without feedback.

On Day 2, the last two phases of the previous day were repeated again using the labial CV series. When these trials were completed, the subjects were presented with the alveolar stimuli (i.e., the transfer series) using -70, 0 and +70 msec VOT exemplars arranged in order. Subjects did not respond overtly to these stimuli but merely listened to several presentations of these test signals. The last phase of Day 2, the transfer phase, consisted of ten randomly ordered presentations of each of the fifteen alveolar CV syllables. Subjects were now required to identify these stimuli into the three new categories corresponding to /nda/, /da/, and /ta/. One group of subjects received training

on the labial stimuli first and then were tested for transfer of training on the alveolar series. Another group received the stimuli in reverse order; that is, training on alveolars first followed by transfer to labials.

Results

Figure 1 shows the average labelling functions for the group of subjects trained first on the labial stimuli and then transferred to alveolar stimuli; Figure 2 shows the functions of the second group which was trained on alveolar and then tranferred to labial stimuli. In each figure there are four panels corresponding to the four conditions of the experiment.

Insert Figures 1 & 2 about here

The two category identification functions are shown in the upper left hand panel of each figure. The data shown in this panel indicate that the labelling of two sounds that are phonemically distinct in English is done consistently and with sharp category boundaries. The second panel, in the upper right of each figure, corresponds to the last condition of Day 1. Inspection of these functions reveals that after familiarization and training with one set of stimuli that varied in VOT, the subjects demonstrated consistent identification of a third category which is not functionally distinctive in English. This result was accomplished with less than one hour of training. Panel 3, in the lower left of each figure, shows a replication of the three category identification obtained on Day 2 of the experiment. A visual

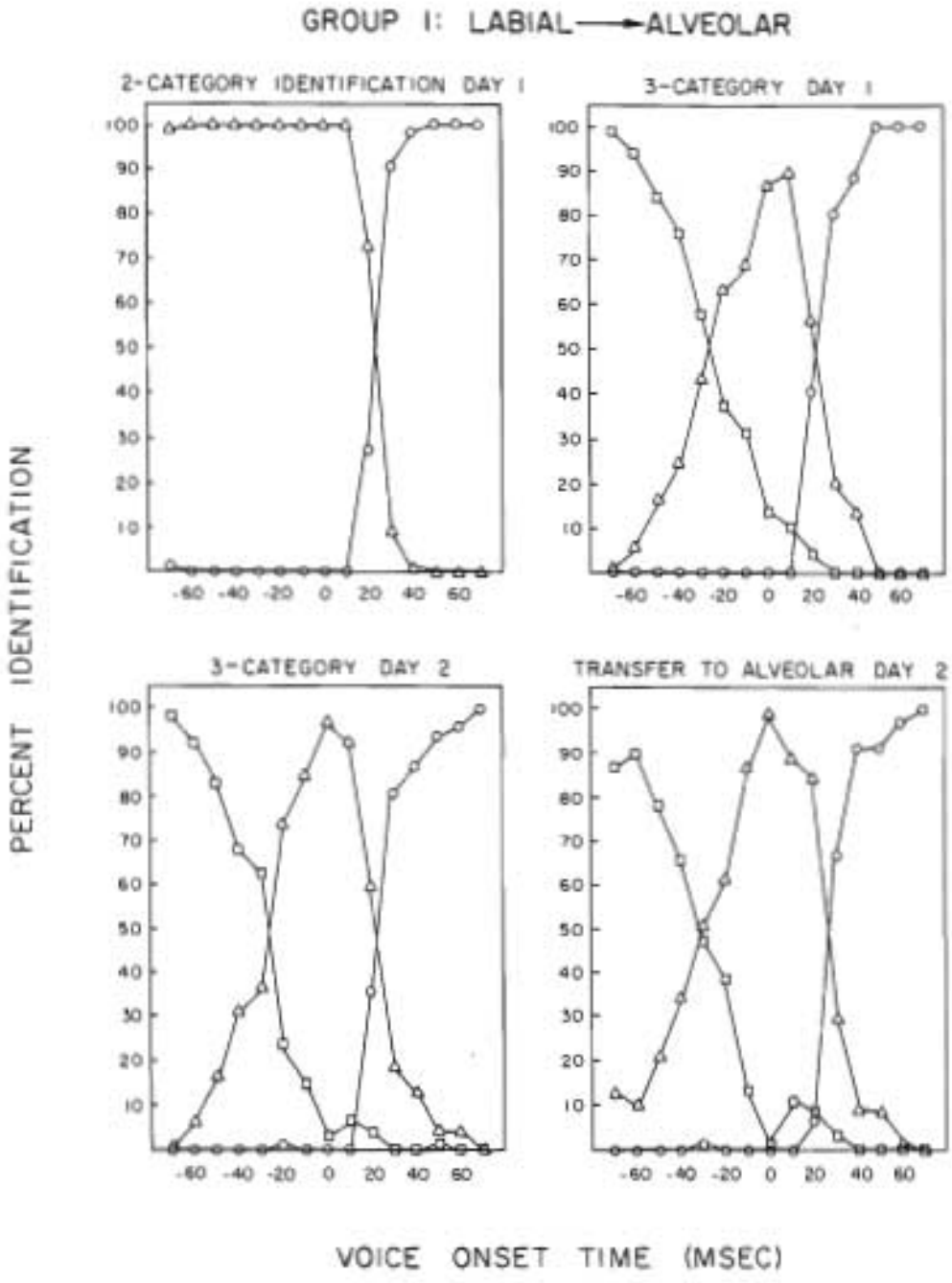


Figure 1. Average identification functions for two and three category labelling and three category transfer labelling for subjects in Group 1.

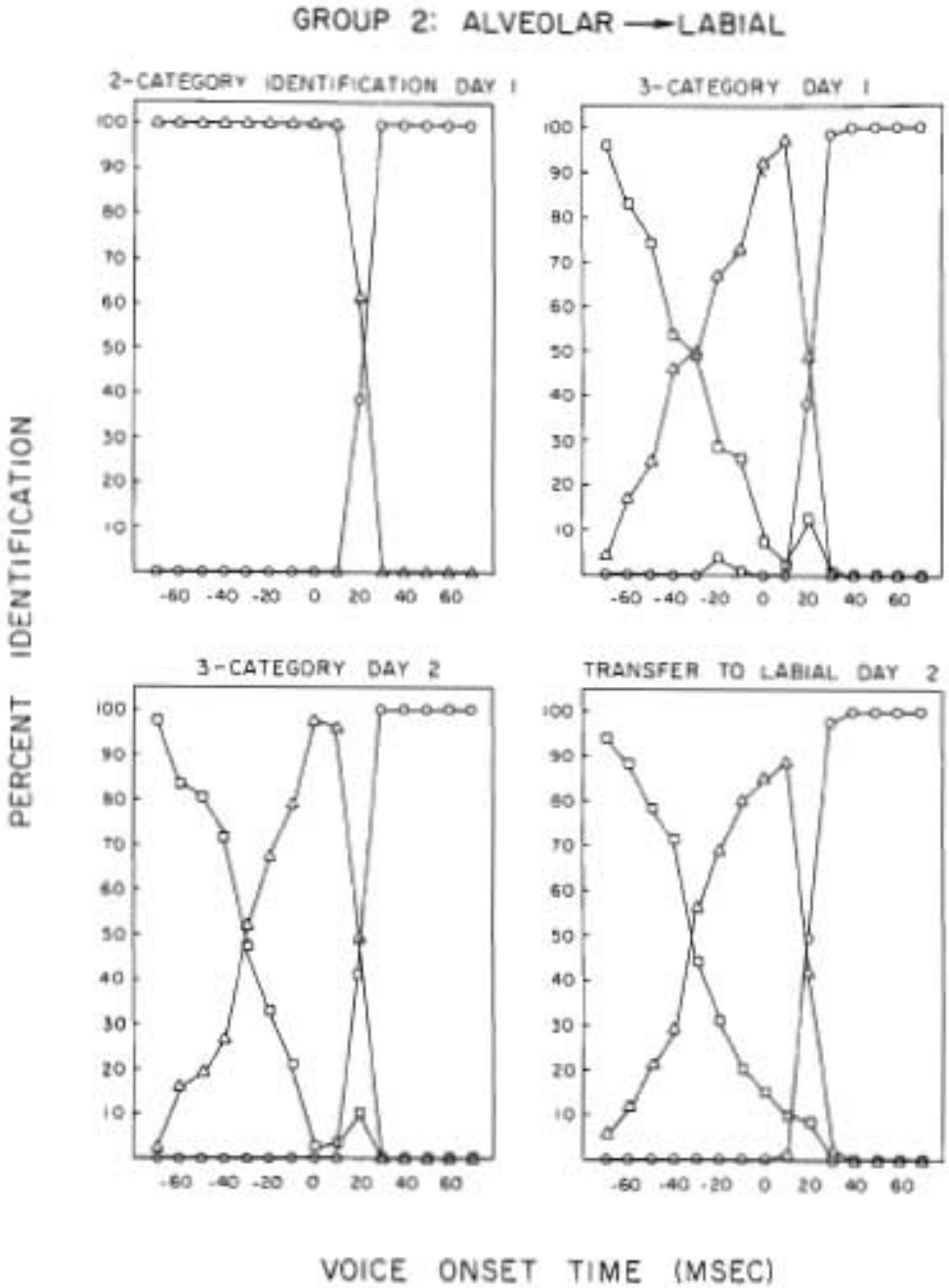


Figure 2. Average identification functions for two and three category labelling and three category transfer labelling for subjects in Group 2.

comparison of subjects across these two conditions indicates the consistency of the identification responses although there is some variability across subjects. This consistency is due, in part, to the criterion used for inclusion in the study. Subjects were required to meet an 85% correct level of performance on the three category exemplar stimuli in order to participate in the second phase of the experiment.

The transfer of training data, shown in Panel 4 in the lower right of each figure, closely resembles the data shown in Panels 2 and 3. It is obvious from inspection these data that subjects were able to use three labelling categories in a consistent manner even though one of the categories is not phonemically distinctive in their native language. Moreover, these subjects were able to reliably perceive differences in voicing lead even though they were tested with stimuli on which they received no specific training experience.

Examination of the individual subjects data shows that subjects were able to identify three distinct categories along the voicing continuum. Inspection of these figures also shows the occurrence of a substantial transfer effect in perception of VOT from one place of articulation to another. In order to quantify these observations more precisely, several analyses of the identification data were carried out to obtain an estimate of the strength of the transfer of training effects. First, a measure of the consistency of responses at each of the fifteen stimulus values was computed and then averaged across the stimulus values to obtain an overall consistency score for each subject (Attneave, 1959). These values are shown in Table 1 for each subject in Conditions 1-4 along with the mean values for each condition.

Insert Table 1 about here

The values of this index, the index of response uncertainty, range from .999 (very consistent) to .593 (fairly consistent). In general, the response consistency values of each subject are higher for Condition 1, the two category identification responses, than for any other condition. This result reflects the fact that two categories of voicing are typically identified by monolingual speakers of English while the ability to identify three categories is a product of less than two hours of laboratory training.

The next analysis that we carried out was designed to measure the sharpness or steepness of the labelling functions. Individual slopes and crossover points were calculated for each labelling function by fitting a normal ogive to the identification data via the methods outlined in Woodworth (1938). Thus, for the two-category identification phase there is only one slope for each subject because only the voiced/voiceless distinction is phonemic in English. For the three-category conditions two slopes were computed, one for prevoiced/voiced boundary and the other for the voiced/voiceless boundary.

Insert Tables 2 & 3 about here

Tables 2 and 3 present the slope values in milliseconds. A low value represents a steeper slope than a high value in this analysis. The slope values are consistently smaller for the voiced/voiceless boundary (Wilcoxon test. $p < .01$). Thus, the

Table 1

Response Consistency Values for Identification

Group 1: Labial to Alveolar Series

	Conditions:			
Subject	1	2	3	4
1	.999	.791	.821	.842
2	.919	.712	.773	.737
3	.940	.813	.813	.870
4	.999	.747	.757	.879
5	.813	.800	.865	.593
6	.934	.728	.794	.739
7	.951	.695	.845	.706
Means:	.936	.755	.815	.767

Group 2: Alveolar to Labial Series

	Conditions:			
Subject	1	2	3	4
1	.934	.807	.846	.849
2	.999	.726	.726	.678
3	.940	.838	.850	.806
4	.967	.714	.768	.736
5	.951	.786	.742	.847
6	.940	.752	.882	.721
7	.951	.834	.979	.861
8	.999	.719	.706	.712
Means:	.960	.772	.821	.776

Table 2

Slope Values for Voiced/Voiceless Boundary in Identification

Group 1: Labial to Alveolar Series

	Conditions:			
Subject	1	2	3	4
1	20.82	21.87	20.82	20.82
2	21.25	26.24	35.60	21.52
3	19.87	19.87	19.94	20.39
4	20.82	21.04	21.23	21.49
5	24.13	20.64	20.90	24.30
6	19.94	19.87	20.05	20.39
7	20.26	19.78	20.91	32.36
Means:	21.01	21.33	22.78	23.04

Group 2: Alveolar to Labial Series

	Conditions:			
Subject	1	2	3	4
1	19.94	19.78	20.91	20.17
2	20.82	20.39	20.82	19.98
3	19.87	19.28	19.28	19.07
4	20.39	20.82	20.81	21.71
5	20.26	20.39	20.02	19.66
6	30.17	21.28	20.39	20.26
7	20.26	20.39	20.82	20.39
8	19.28	20.26	19.28	19.66
Means:	21.37	20.32	20.29	20.11

Table 3

Slope Values for Prevoiced/Voiced Boundary in Identification

Group 1: Labial to Alveolar Series

	Conditions:		
Subject	2	3	4
1	-22.07	-22.80	-22.74
2	-25.92	-20.92	-24.13
3	-24.27	-22.56	-22.70
4	-27.33	-30.23	-22.42
5	-24.19	-21.30	-35.78
6	-27.21	-26.79	-26.18
7	-26.95	-36.35	-73.07
Means:	-25.42	-25.84	-32.43

Group 2: Alveolar to Labial Series

	Conditions:		
Subject	2	3	4
1	-28.54	-23.89	-23.49
2	-35.39	-27.64	-27.76
3	-32.80	-40.06	-42.89
4	-32.54	-29.63	-30.00
5	-24.85	-27.23	-28.37
6	-24.99	-22.77	-24.80
7	-22.18	-18.76	-20.28
8	-32.30	-29.33	-30.13
Means:	-29.20	-27.43	-28.47

voiced/voiceless labelling functions have steeper slopes than the prevoiced/voiceless boundary. Other analyses, across conditions and between groups, revealed no other reliable differences in the slope of the labelling functions.

A third analysis was also carried out to quantify the precise location of the category boundaries. A cross-over point was obtained from the fitted ogives used to calculate the slopes for each boundary. In addition, a midpoint for each boundary was obtained directly by inspection from the graphs by selecting the point at which each function first straddled the 50% identification value. The two values in milliseconds for the voiced/voiceless boundary are given in Table 4 while the prevoiced/voiced boundary values are given in Table 5.

Insert Tables 4 & 5 about here

Scores obtained in these three analyses (response consistency, slope, and crossover) were compared across the four conditions for each group by means of the Wilcoxon matched pairs test (Siegel, 1956). Comparisons of response consistency across the four conditions indicated that only two conditions differed significantly from each other ($p < .01$, two-tailed). For both groups, the consistency of responses during identification of two categories on Day 1 showed marked differences from the consistency of responses in the transfer test on Day 2. These conditions correspond to the data shown in Panels 1 and 4 in each figure. Although the values of response consistency for the data shown in panel 4 are fairly high with a mean value of .767 for Group 1 and

Table 4

Midpoint and Crossover Values for Voiced/Voiceless Boundary in Identification

Group 1: Labial to Alveolar Series

	Conditions:							
	1		2		3		4	
Subject No.	Midpt	Mean	Midpt	Mean	Midpt	Mean	Midpt	Mean
1	56	18.67	58	22.77	56	18.67	55	18.67
2	26	29.68	24	37.43	47	64.39	30	20.50
3	17	13.56	17	13.56	19	31.97	25	16.37
4	25	18.67	25	18.86	26	19.57	26	21.29
5	27	22.07	20	16.74	25	18.13	37	31.01
6	18	13.97	18	13.56	16	10.79	24	16.37
7	24	15.67	16	13.08	18	16.23	53	46.59
Means:	27.57	20.86	25.42	19.42	29.57	23.11	35.72	24.40

Group 2: Alveolar to Labial Series

	Conditions:							
	1		2		3		4	
Subject No.	Midpt	Mean	Midpt	Mean	Midpt	Mean	Midpt	Mean
1	18	13.97	15	13.08	17	16.23	23	15.17
2	25	18.67	24	16.37	24	18.67	19	14.14
3	17	13.56	15	10.37	16	10.37	14	8.30
4	25	16.37	25	18.67	26	18.67	26	22.14
5	23	15.67	23	16.37	20	14.36	16	12.42
6	22	15.17	24	12.01	25	16.37	24	15.67
7	24	15.67	26	16.37	25	18.67	26	16.37
8	16	10.37	16	14.70	15	10.37	16	12.42
Means:	21.55	14.93	21	14.74	21	15.46	20.50	14.58

Table 5

Midpoint and Crossover Values for Prevoiced/Voiced Boundary in Identification

Group 1: Labial to Alveolar Series

	Conditions:					
	2		3		4	
Subject No.	Midpt	Mean	Midpt	Mean	Midpt	Mean
1	-17	-12.74	-22	-15.75	-15	-17.65
2	- 6	-19.61	-21	- 5.36	-29	-22.55
3	-30	-26.62	-26	-23.41	-30	-23.94
4	-35	-28.26	-53	-41.03	-30	-23.08
5	-25	-23.76	-23	-11.40	-34	-31.24
6	-40	-30.23	-20	-33.25	-15	-21.36
7	-20	-17.15	- 4	-66.41	-60	-158.58
Means:	-24.71	-23.63	-21.14	-28.09	-34.63	-42.63

Group 2: Alveolar to Labial Series

	Conditions:					
	2		3		4	
Subject No.	Midpt	Mean	Midpt	Mean	Midpt	Mean
1	-36	-40.47	-26	-27.22	-33	-21.68
2	-47	-47.24	-35	-29.90	-26	-22.87
3	-46	-51.92	-46	-73.56	-57	-78.05
4	-40	-34.05	-40	-32.29	-32	-28.71
5	-20	-22.45	-26	-29.72	-30	-41.01
6	-30	-26.55	-32	-23.74	-35	-18.80
7	- 3	- 7.67	- 5	- 5.13	- 9	- 8.15
8	-44	-31.59	-60	-29.60	-34	-34.06
Means:	-33.25	-32.74	-33.75	-31.40	-32	-31.67

a mean value of .776 for Group 2, this result is expected due to the relative unfamiliarity of the transfer stimuli. Another cross-condition consistency comparison which was significant occured only in Group 1. A significant change ($p < .01$, two-tailed) in consistency of responses was found from Day 1 to Day 2 on the labial three category identification stimuli shown in Panels 2 and 3. It appears that the brief amount of practice provided on Day 2 improved the consistency of the responses to labial stimuli (Group 1) but had no comparable effect on the alveolar stimuli (Group 2). The reason for this asymmetry in performance is not clear at the present time.

Analysis of cross-series differences in terms of the slope of the labelling functions showed that, in general, the slopes did not change significantly across conditions. However, there were two exceptions to this observation. For Group 1, the voiced/voiceless boundary differed significantly ($p < .01$) for the two category identification condition (Panel 1) compared to the transfer condition (Panel 4). Although this result indicates a steeper slope for the two category condition, a comparison of the slope values in Table 2 indicates that, with the exception of Subject 7, most values are fairly consistent across the two conditions. Also for Group 1, a prevoiced/voiced VOT boundary comparison was significant. When Day 2 of three category identification was compared to the transfer function (Panel 3 vs 4), the slopes of the identification function for the transfer series were less steep than the slopes of the functions for the training stimuli.

Turning to comparisons of the crossover values, no significant differences were observed for comparisons of Group 2 indicating that the location of the category boundaries did not change across the various conditions of the experiment. These results were further confirmed by analysis of the midpoint comparisons of Group 2 which also showed no change across condition. Group 1, however, showed a significant difference in the location of the voiced/voiceless boundary which shifted slightly from the two category identification to the transfer condition (panels 1 vs 4). The precise reason why a shift was observed in one series and not the other is not clear at this time.

In order to determine if there were any between-group differences, the data were compared by means of Mann-Whitney U tests for each condition (Siegel, 1956). A comparison of the U values showed no significant differences between Group 1 and Group 2 for three of the four conditions: two category identification (Condition 1), three category identification Day 1 (condition 2), and three category identification Day 2 (condition 3). This was true for both the slope analysis and the response consistency analysis. However, the fourth condition, the transfer phase, showed one statistically significant difference between Group 1 and Group 2 ($U = 6$). This result was found in the voiced/voiceless region and reflects the finding that Group 1 (transferred to alveolar stimuli) had a steeper slope than Group 2 (transferred to labial stimuli).

In summary, the results of all three analyses--slope, response consistency and cross-over point--show that three

perceptual categories were reliably identified by our naive subjects in less than two hours of laboratory training and, further, that when presented with stimuli on which they have not received training, our subjects generalized knowledge of VOT gained during training to identification of the new stimuli. A reliable and consistent perceptual category emerged clearly for these subjects with relatively simple laboratory procedures in a short period of time.

Discussion

The present experiment demonstrated that a new perceptual category can be acquired in a relatively short amount of time through relatively simple laboratory training procedures. The subjects consistently identified the VOT stimuli into three perceptual categories and showed strong evidence of transfer of training to another place of articulation from the one they were originally trained on. Such findings demonstrate clearly that the perceptual capacities required to learn these contrasts are still present in experimentally naive monolingual speakers of English. English speakers typically divide stop consonants differing in VOT into only two categories at each place of articulation corresponding to /b-p/, /d-t/ and /g-k/. In the present study, after training on stimuli from one place of articulation, our subjects were able to generalize their knowledge of the three-way labelling contrast to another place of articulation without additional training. The present findings on transfer of identification are complementary to the results reported by Edman et al. (1978) who demonstrated improved intra-phonemic

discrimination of stop consonants after training. Edman et al. also showed that discrimination training generalized from one place of articulation during training (e.g. /b-p/) to a second place of articulation at the time of final testing (e.g. /g-k/). However these investigators did not obtain identification data from their subjects.

The results of the present study also extend the earlier findings of Pisoni et al. (1980) by demonstrating a robust transfer of training effect from one synthetic stimulus continuum to another. In the Pisoni et al. study, subjects were trained and tested with only labial consonant-vowel syllables. By demonstrating three category identification for stimuli which were not used in training, the present study has established that naive subjects are able to acquire specific knowledge about the complex temporal-spectral dimension of voicing as cued by VOT and not just about the distinctive correlates of the specific stimuli used in training.

The findings obtained in this study have several implications for the previous conclusions concerning the effects of laboratory training in speech perception, namely, that laboratory training procedures appear to have little effect on modifying speech perception in adult listeners. In this regard, several methodological aspects of the present study are worth emphasizing. First, we were able to control the subjects' attentive processes to selectively focus on those distinctive aspects of the stimuli that were relevant to successful identification. This was accomplished by familiarization and training with immediate feedback on exemplars of the three perceptual categories. Although

our verbal instructions also directed the subjects attention to the initial part of each stimulus, the use of on-line computer techniques for discrimination training and feedback was also an important aspect of the success of the present study.

Second, we imposed a strict training criterion on subjects which effectively reduced intrasubject variability. Although the subjects who were eliminated from testing on Day 2 did not identify the stimuli into three categories at the required performance levels, they did respond well above chance. While not a specific concern of the present study, it would be of some interest to measure the amount of additional training that these non-criterion subjects would require to reach the criterion level of performance.

In their review of laboratory training studies, Strange and Jenkins (1978) concluded that modifications of the VOT dimension is not easily accomplished by the use of laboratory training methods in a short period of time. Their conclusions have broad implications for the plasticity of the adult perceptual system and the acquisition of new linguistic contrasts at the phonological level. The data obtained in this study suggest that previous failures to acquire a new linguistic contrast do not appear to be related in any way to permanent changes in the perceptual mechanisms used by mature adult listeners. Our listeners did not appear to have "lost" any of their speech processing abilities due to environmental exposure or realignment of their perceptual mechanism. In light of the positive results obtained in this study there is strong impetus to extend the use of these discrimination training procedures to other linguistic contrasts involving

English /r/ vs /l/ in Japanese subjects and various tonal distinctions in certain dialects of Chinese. Research of this kind is currently underway in our laboratory and reports of this work will be forthcoming.

Footnotes

[1]This research was supported by NIMH research grant MH-24027 to Indiana University. Special thanks go to Amanda C. Walley for her help and assistance in preparing the synthetic stimuli used in this study.

References

Attneave, F. Applications of information theory to psychology. New York: Holt, Rinehart and Winston, 1959.

Carney, A. E., Widin, G. P., & Viemeister, N. F. Noncategorical perception of stop consonants differing in VOT. Journal of the Acoustical Society of America, 1977, 62, 961-970.

Edman, T. R., Soli, S. D., & Widin, G. P. Learning and generalization of intraphonemic VOT discrimination. Paper presented at the 95th Meeting of the Acoustical Society of America, Providence, RI, May, 1978.

Eimas, P. D. Developmental aspects of speech perception. In R. Held, H. Liebowitz and H. L. Teuber (Eds.), Handbook of sensory physiology (Vol. VIII: Perception). New York: Springer-Verlag, 1978.

Goto, H. Auditory perception by normal adults of the sounds "l" and "r". Neuropsychologia, 1971, 9, 317-323.

Klatt, D. H. Analysis and synthesis of CV syllables in English. Unpublished manuscript, 1978.

Klatt, D. H. Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America, 1980, 67, 971-995.

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

Lisker, L. On learning a new contrast. Status Report on Speech Research SR-24. New Haven: Haskins Laboratories, 1970, 1-15.

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

Lisker L. and Abramson, A. S. The voicing dimension: Some experiments in comparative phonetics. Proceedings of the 6th International Congress of Phonetic Sciences, Prague, 1967.

Miyawaki, K., Strange, W., Verbrugge, R. 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.

Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennessy, B. L. Identification and discrimination of a new linguistic contrast: Some effects of laboratory training on speech perception. Research on Speech Perception Progress Report No. 4. Indiana University, 1978, 49-112.

Strange, W. The effects of training on the perception of synthetic speech sounds: Voice onset time. Unpublished doctoral dissertation, University of Minnesota, 1972.

Strange, W. & Jenkins, J. J. Role of linguistic experience in the perception of speech. In R. D. Walk & H. L. Pick (Eds.), Perception and Experience. New York: Plenum Press, 1978.

Woodworth, R. S. Experimental Psychology. New York: Holt, 1938.

Identification and Discrimination of Durations of Silence in Nonspeech Signals

A. J. Perey and D. B. Pisoni
Indiana University
Bloomington, IN 47405

Short Title: Perception of Duration of Silence

Abstract

Early research on the perception of speech and nonspeech signals appeared to reveal a form of perception that was unique to speech stimuli. The differences in perception between speech and nonspeech signals were interpreted as evidence for the operation of a specialized speech mode of processing. However, recent research demonstrating categorical perception for nonspeech stimuli and other studies on the perception of synthetic speech stimuli by the chinchilla have raised the possibility that certain phenomena in speech perception may be based in the psychophysical attributes of the stimuli themselves rather than on some additional specialized coding. In intervocalic position the voiced-voiceless distinction can be cued solely by differences in closure duration. This cue was simulated in a set of nonspeech tonal stimuli by varying the duration of a silent interval between two component tones. The results of identification and discrimination tests indicated sharp and consistent labeling functions but flat and near chance ABX discrimination functions for these stimuli. The addition of simulated formant transitions into and out of the silent interval did not alter these functions in any systematic way. The results of these preliminary tests could not be interpreted as unambiguous support for either a speech mode of processing or an account based on the psychophysical attributes of complex acoustic stimuli. Additional experiments manipulating closure duration in nonspeech contexts are proposed to provide additional evidence to differentiate between these two alternatives.

Identification and Discrimination of Durations of Silence in Nonspeech Signals

A. J. Perey and D. B. Pisoni

A major theoretical issue in speech perception has focused on the question of whether a highly specialized mode of processing, a speech mode, is needed to account for the perceptual results observed with speech signals. This perceptual mode is assumed to be one in which speech is perceived in unique fashion, presumably because at some point during perceptual processing, the acoustic signal is routed through some sort of specialized speech processor which operates in a linguistically relevant manner (Liberman, 1970a). The speech mode of processing was originally proposed to account for the perception of speech signals under certain conditions. In particular, it was noted that the shortest of the meaningful segments in language, the phones or consonants and vowels, are not distinctly separable as acoustic segments in the speech signal itself. In fact, because of coarticulation in speech production, features from successive phonetic segments overlap and are often produced simultaneously. Thus, multiple features of a single phoneme are transmitted over several segments resulting in parallel transmission of feature information often spanning several acoustic segments (Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967; Liberman, 1970a, 1970b; Liberman, Mattingly and Turvey, 1972; Mattingly, 1972; Liberman and Pisoni, 1977). Finally, different values of the same phonetic feature may be differentiated by a number of different acoustic attributes.

The speech mode of processing is assumed to extract and decode this contextually variable acoustic information in a linguistically relevant manner (Liberman, 1970a).

Support for the notion of a speech mode of processing has been gathered from numerous studies using synthetically produced speech signals. One distinction that has been studied extensively is voicing, a phonetic feature which serves to distinguish the voiced consonants /b/, /d/ and /g/ from the voiceless consonants /p/, /t/ and /k/. The voicing distinction has been investigated in various syllable positions, in speech and nonspeech environments, in adults and human infants, in chinchillas and in monkeys (Lisker, 1957; Liberman, Harris, Kinney and Lane, 1961; Lisker, 1970; Eimas, Sigueland, Jusczyk and Vigorito, 1971; Kuhl and Miller, 1975; Miller, Weir, Pastore, Kelly and Dooling, 1976; Waters and Wilson, 1976; Pisoni, 1977).

In initial syllable position the voicing distinction is primarily based on the acoustic dimension of voice-onset-time (VOT), the temporal interval between the release of stop closure in the vocal tract and the onset of laryngeal pulsing (Lisker and Abramson, 1964). In a synthetic speech context VOT is generally perceived categorically; discrimination can be made only in so far as the stimuli have been labeled as different phonemes (Lisker and Abramson, 1970; Abramson and Lisker,1970). That is, discrimination functions are at a maximum across a phoneme boundary and at a minimum, presumably close to chance, within a phoneme category. The labeling functions associated with the phonological categories show a sharp crossover point from one category to another. In addition, the peaks of the discrimination functions coincide with

the crossover point in the labeling functions. Since nonspeech signals are typically perceived continuously and subjects can discriminate many more differences than they can absolutely identify, categorical perception was generally assumed to be a result of labeling processes associated with phonetic categorization accomplished in the speech mode.

Initial comparisons of the discrimination functions of speech and nonspeech stimuli were primarily directed at determining whether the nonmonotonic discrimination functions obtained with the speech stimuli were innate or a function of learning and experience (Liberman, Harris, Kinney and Lane, 1961; Liberman, Harris, Eimas, Lisker and Bastian, 1961). These investigators explicitly accepted the learning assumption and attempted to determine whether the learning was due to acquired distinctiveness (AD), whereby the ability to discriminate differences at the phoneme boundary is sharpened, to acquired similarity (AS), whereby the ability to discriminate differences within a phoneme category is attenuated, or to a combination of both of these two processes (see Liberman et al., 1961). The nonspeech control stimuli used in these early studies were designed to preserve the relevant acoustic features of the speech stimuli. Studies were then carried out to establish baseline discrimination functions against which the extent and direction of learning could be determined.

Liberman, Harris, Kinney and Lane (1961) investigated the F1 cutback, the delay in the onset of the first formant relative to the second and third formants, in synthetic speech stimuli and in nonspeech control environments. The synthetic speech stimuli

varied from /do/ to /to/ by manipulating the duration of the F1 transition. A set of nonspeech control stimuli were also constructed which preserved this cue by inverting the spectrograms of the synthetic speech stimuli before they were converted into sound on the Pattern Playback. However, since inverting the speech stimuli resulted in a rising nonspeech first formant which sounded speechlike, it was also necessary to replace this rising transition with a falling transition. Identification and discrimination tests were then carried out with both sets of stimuli. Labeling functions were sharp and consistent for the speech stimuli. However, none were obtained for the nonspeech stimuli. The ABX discrimination functions for the speech stimuli showed the characteristic peaks and troughs of categorical perception while those for the nonspeech controls were relatively flat and rarely above chance level of performance. These findings provided strong support for the notion that speech and nonspeech signals are processed differentially and led these investigators to conclude that the discrimination functions obtained with speech are due to acquired distinctiveness, whereby the ability to discriminate differences at a phoneme boundary is enhanced through experience in dealing with the distinction in the environment.

Early infant research using the High Amplitude Sucking (HAS) technique also provided some additional support for the notion of a speech mode of processing. Basically the HAS procedure permits an investigator to examine differences in sucking rate between a stimulus to which a subject is habituated and a new stimulus. An increase in sucking after a change in stimulation is taken as evidence for discrimination while no change is considered evidence

of a lack of discrimination. Using this technique, Eimas and his associates (1971) discovered that 1-3 month old infants from English speaking environments exhibit categorical-like discrimination of VOT differences in synthetic speech stimuli. These data were considered to be quite similar to the adult data on VOT found earlier by Lisker and Abramson, (1970) and Abramson and Lisker (1970). The infant results led Eimas et al. to conclude that either categorical perception is innate or it develops very rapidly after birth. The important result of this study, however, was that the infants appeared to be responding to the stimuli in a linguistically relevant manner, much as adults do, with only very limited experience in the language-learning environment.

Earlier we noted that the acoustic cues for a single phonetic feature often vary with context. Such is the case for the voicing distinction. Lisker (1957) studied spectrograms of sentences containing trochees in a primary stress position and trochees in isolation. These trochees contained either a /p/ or a /b/ in intervocalic position. From these spectrograms he observed a number of acoustic differences between words containing /p/ or /b/, one of which was closure duration, the time interval between the termination of the vowel formant preceding the stop and the onset of the subsequent vowel. Lisker also carried out a series of tape splicing experiments with natural speech stimuli in which he demonstrated that the perception of a /p/ in medial position could be altered to a /b/ by simply reducing the closure duration. And Lisker showed that a /b/ could be changed to a /p/ by simply lengthening the closure duration. The labeling functions obtained from these two procedures showed crossover points at approximately

75 msec and 105 msec respectively. Both experiments established that the acoustic cue of closure duration is sufficient for perception of the voiced-voiceless distinction in intervocalic position.

In one of the earliest perceptual experiments using nonspeech controls, Liberman, Harris, Eimas, Lisker and Bastian (1961) examined the perception of closure duration in synthetic speech and nonspeech contexts and found substantial differences in the discrimination functions for each stimulus series. Closure duration was simulated in the synthetic speech context by manipulating the duration of the silent interval between the two syllables of a synthesized word, i.e., "rabid" vs. "rapid." In the nonspeech control condition, the duration of the silent interval separating two bursts of band-limited noise was manipulated to match the onset, offset and duration characteristics of the original speech stimuli. The purpose of this study was to determine whether the discrimination of speech was a function of learning, and if so, whether this learning was due to acquired distinctiveness or acquired similarity. Liberman et al. determined that the synthetic speech stimuli were perceived categorically while the nonspeech ABX discrimination functions were noncategorical and basically flat across the entire stimulus continuum although slightly above chance. These findings led Liberman et al. to conclude that the discrimination functions obtained with the nonspeech signals could be thought of as representing "true" discriminability without the benefit of any linguistic experience. In this regard, the discrimination peaks typically obtained with speech could be accounted for as a result

of learning, specifically learning that involved the process of acquired distinctiveness.

Evidence from investigations of the perception of different acoustic cues thought to be important for other phonetic distinctions has also supported the notion of a specialized speech mode of processing. Mattingly, Liberman, Syrdal and Halwes (1971) reported that the synthetic syllables /bae/, /dae/ and /gae/, which differ in place of articulation cued by the direction and extent of the second formant transition, are perceived categorically. This study also showed that the same cue, the formant transition, was perceived noncategorically when presented in isolation. The ABX discrimination functions were irregular and close to chance, when the F2 transition cue was presented as a "chirp" or in conjunction with an F2 steady-state as a "bleat". The latter two stimulus contexts were designed to be nonspeech control conditions. In addition, a reversal of the speech and nonspeech patterns in time differentially affected the speech and nonspeech discrimination functions. In the backward speech condition, the discrimination function lost its peaks and troughs but remained at about the same overall level. In the "backward" nonspeech condition, the function remained flat but performance was at a higher level than in the forward nonspeech conditions. The differences in perception between speech and nonspeech as well as the differences within the various speech and nonspeech conditions led these authors to conclude that their results, as well as the others discussed above, cannot resolve the questions about acquired distinctiveness or acquired similarity. Instead, Mattingly et al. argued that their results should be taken as

evidence for two basically different modes of perception, a speech mode and an auditory mode.

Using the same synthetic stimuli, Eimas (1974) also investigated perception of place of articulation in infants as cued by the direction and extent of the second formant transitions. His results showed that when the second formant transitions are embedded in the context of synthetic speech stimuli they are discriminated categorically by infants but when they are presented in isolation as "chirps" they are discriminated noncategorically by infants. These results are based solely on discrimination data obtained with the HAS technique. Eimas concluded that his results with infants could also be used as further evidence for the existence of a specialized speech processor which operated in a linguistic mode.

In another study, Miyawaki, Strange, Verbrugge, Liberman, Jenkins and Fujimura (1975) studied the perception of the /r/ - /l/ contrast in Japanese and English adults. In English, /r/ and /l/ can be differentiated by the acoustic cue of a rising or falling third formant transition, while in Japanese the phonetic distinction is not used phonologically. Miyawaki et al. reported that English speakers perceive synthetic stimuli varying from /r/ to /l/ categorically while Japanese speakers do not. When the third formant transitions were presented in isolation as nonspeech controls which sounded like "chirps," the discrimination functions were relatively flat and just above chance. Furthermore, the Japanese adults did not differ substantially from the English adults on the nonspeech control stimuli. Thus Miyawaki et al. concluded that linguistic experience selectively affects only the speech mode of processing and not a more general auditory mode.

The two series of speech and nonspeech stimuli from the Miyawaki et al. (1975) study were also used by Eimas (1975) to examine discrimination of the /r/ - /l/ distinction in two and three month-old infants with the HAS procedure. The discrimination data obtained for the speech stimuli indicated that discrimination of stimuli from within an adult category was less than discrimination of stimuli from different categories; discrimination of the isolated chirps was approximately the same along the entire stimulus continuum. Thus, these results also indicate that for infants speech is perceived categorically while comparable acoustic cues in nonspeech contexts are not. Eimas concluded that the categorization of speech is due to speech specific processing mechanisms and therefore is evidence against alternative hypotheses which suggest that phonetic boundaries are simply located at regions of increased discriminability along either simple or complex acoustic continua.

The investigators discussed above have all interpreted their results as support for either the effects of perceptual learning or, even more recently, for the operation of a specialized speech mode of processing. Recently, however, new experimental findings from research with animals, young infants, and further studies using nonspeech stimuli suggest that categorical perception per se may not be appropriate evidence to support the distinction between the processing of speech and nonspeech signals. These recent findings have also called into question the postulation of a specialized speech mode of perception and have suggested a more general framework for understanding the perception of complex acoustic signals in terms of processing limitations on the auditory system.

In an important study, Kuhl and Miller (1975) demonstrated that chinchillas can be trained to respond differentially to natural and synthetic tokens of the consonants /d/ and /t/. They obtained labeling functions whose shape and boundaries coincided quite closely to those obtained for humans. Since chinchillas do not have spoken language, it is unlikely that they use a phonological coding system or a speech mode of processing. Given these results, the claims surrounding categorical perception have to be reevaluated in terms of the acoustic or psychophysical attributes of the specific stimuli in question not by recourse to the postulation of a specialized mode of perceptual processing.

Additional investigations of infant speech perception of VOT have also cast doubt on the notion that categorical perception implies the operation of a speech mode of perception. Cross-language studies conducted by Lasky, Syrdal-Lasky and Klein (1975) and Streeter (1976) suggest that phonetic categories resulting from differences in the perception of VOT may be the result of naturally defined regions of high discriminability along the VOT continuum rather than the consequence of specialized speech processing or phonetic categorization.

Lasky et al. (1975) investigated the perception of the bilabial stop consonants /ba/ and /pa/ which can be distinguished by differences in VOT. Their subjects were four to six and a half month old infants from monolingual Spanish environments. They used a heart rate habituation/dishabituation technique in which discrimination is inferred from changes in heart rate correlated with stimulus changes after the infant habituates to an initial stimulus. The infants were presented with stimulus pairs having

VOT values of -20 and +20 msec, -20 and -60 msec and +20 and +60 msec. Since only the latter two pairs were reliably discriminated, the results suggest category boundaries were located between each of those pairs, therefore providing evidence for at least three phonetic categories. Most interesting, however, was the finding that neither of these boundaries corresponded to the adult boundary found for Spanish speakers.

Streeter (1976) examined the discrimination of VOT differences in the bilabial stops /b/ and /p/ in Kikuyu speaking environments using the HAS technique. The infants were presented with stimulus pairs having VOT values of -30 and 0 msec, +10 and +40 msec and +50 and +80 msec. Only the former two were reliably discriminated by her infants. Thus, Kikuyu infants also show evidence for three phonetic categories which do not correspond precisely to the boundaries observed in adults who do not have a voiced- voiceless distinction for bilabial stops.

Finally, data collected in two separate investigations of VOT simulated with nonspeech signals has shown evidence of categorical perception. Miller et al. (1976) simulated VOT in nonspeech signals by varying the duration of a noise burst preceding the onset of a buzz. The labeling task consisted of a choice between noise and no noise present before the buzz. The discrimination task used an oddity paradigm. Labeling results showed consistent identification with a sharp crossover between the two categories. Discrimination was best for stimuli selected from different categories but close to chance for stimuli within a category. Miller et al. interpreted these results as evidence that categorical perception should be considered in terms of

psychophysical boundaries for perceptual effects obtained when components of a complex stimulus change continuously relative to the remainder of the stimulus. At a threshold, the changing component results in increased discriminability. Across a threshold, discrimination of various changes in the magnitude of the component follow Webers law until another threshold is encountered.

Pisoni (1977) reasoned that such an account of the Miller et al. data suggests the presence of naturally defined category boundaries along the VOT continuum which occur when new perceptual attributes emerge as a function of continuous variations in a parameter of a complex signal. In this study, Pisoni (1977) simulated VOT in nonspeech tonal stimuli by varying the relative onset of two component tones. The stimulus complex consisted of a 500 Hz and a 1500 Hz tone. The onset of the 500 Hz tone was varied between +50 and -50 msec relative to the 1500 Hz tone. He found that the identification functions for these stimuli were very consistent with sharp crossover points separating categories. Moreover, the ABX discrimination functions showed peaks at the crossover points and regions of low discriminability within a single perceptual category. Thus, these results provided additional evidence that nonspeech stimuli could be perceived categorically. Furthermore, Pisoni found evidence for three perceptual categories which corresponded to the judgments of three temporal events; simultaneity, leading and lagging. The leading and lagging events had boundaries at roughly -20 and +20 msec. Pisoni suggested that the results obtained for VOT perception in animals, human infants and adults could be explained in a

consistent way by assuming a natural limitation on the temporal resolving power of the auditory system. The continuous variations in one or more dimensions of a complex stimulus are judged in relation to the temporal attributes of other events. Thus, the listener appears to be sensitive to the presence or absence of a particular attribute resulting from a change in some configuration rather than the magnitude of the difference. As a consequence, the categorical perception of VOT in initial position could be interpreted in terms of the relative discriminability of the temporal order of two or more events.

Other nonspeech stimuli have also been shown to be perceived in a categorical manner. Cutting and Rosner (1974) demonstrated that musical tones differing in rise time could be consistently labeled as either a "pluck" or a "bow" by adults. The ABX discrimination functions showed the characteristic peak at the crossover point in the labeling function and low discrimination within the two categories. Jusczyk, Cutting, Rosner, Foard and Smith (1977) presented "pluck" and "bow" stimuli to two-month-old infants using the HAS technique and obtained a change in sucking only between stimuli from different categories. These results were interpreted as evidence that human adults and infants can perceive at least some nonspeech stimuli in a categorical fashion. Jusczyk et al. suggested that the perceptual properties of speech such as categorical perception may well have developed around the existing perceptual properties of the auditory system and not as the result of an entirely new speech processing system.

These recent findings with complex nonspeech signals such as "plucks" and "bows" and the simulations of the timing events in

VOT stimuli suggest that categorical perception may be based on the acoustic or psychophysical attributes of the stimuli rather than phonetic categorization that is accomplished by mechanisms specific to a specialized speech mode of processing. However, support for such a hypothesis requires the investigation of acoustic cues other than VOT that are known to be important in making phonetic distinctions in various contexts. The experiments reported below were conducted to examine the perception of the acoustic cue of closure duration. This cue was simulated in a nonspeech environment by varying the duration of a silent interval between two component tones. Since closure duration is a temporal cue known to be sufficient to distinguish voicing in intervocalic position in a categorical manner, it was of interest to determine whether the same cue is perceived categorically in a nonspeech environment while still preserving the stimulus structure that occurs in speech environments. Previous experiments that examined discrimination of closure duration in nonspeech environments showed near chance performance. As a consequence, it was difficult to draw any firm conclusions about inherent discriminability of the nonspeech continuum. Evidence that closure duration is also perceived categorically in a nonspeech environment would be an important finding to support the development of a more comprehensive theory of the perceptual processing of speech based on the psychophysical properties of the stimuli themselves rather than one based primarily on the operation of a specialized speech mode of processing.

Experiment 1

As noted above, Lisker (1957) demonstrated that a series of stimuli differing from "rabid" to "rapid" could be generated by manipulating the duration of the silent interval between the first and second syllables. In this first experiment closure duration was simulated by manipulating the duration of silence between two complex tonal stimuli. Subjects were trained to identify stimuli along this continuum in order to determine whether closure duration exhibits categorical labeling functions when presented in a nonspeech environment.

Method

Subjects. Sixteen paid volunteers were obtained through an advertisement in a student newspaper. All were right-handed native speakers of English and were paid a base rate of $2.00 an hour plus whatever they could earn during the experiment.

Stimuli. Three complex tones each composed of a triad of 250 msec sine waves were generated digitally on a PDP-11/05 computer using a program called Tone (Kewley-Port, 1976). The first tone A1, was composed of three sine waves having frequencies of 470 Hz, 1500 Hz and 250 Hz. The respective amplitudes of these three components were 66dB, 57dB and 51dB SPL. These values were chosen to simulate those for the neutral schwa vowel. The second tone, A2, was composed of the same three sine waves as A1 except the final 50 msec of each component simulated a formant transition or rapid spectrum change into the closure period. This was done by

decreasing the 470 Hz component to 350 Hz, the 1500 Hz component to 1250 Hz and the 2500 Hz component to 2100 Hz linearly during the last 50 msec while still maintaining the same relative amplitudes. The third tone, A3, was composed of the same parameters as A2 but the simulated formant transitions out of the closure period consisted of a frequency rise during the initial 50 msec rather than a terminal decline as in A2.

Two sets of thirteen stimuli were generated by pairing A1 with A1 (the A1-A1 series) and A2 with A3 (the A2-A3 series). The stimuli varied in the duration of the silent interval between the offset of the first tone and the onset of the second tone from 40 msec to 160 msec in 10 msec steps. This experimental variable will hereafter be referred to as the duration of the silent interval. These stimuli are represented schematically in Figure 1.

Insert Figure 1 about here

The A1-A1 series was designed to investigate the acoustic cue of closure duration in two complex tones simulating one aspect of the structure of speech, namely, the presence of at least three formants. The A2-A3 series was designed to examine the perceptual effects that occur when other features contained in speech signals are added such as formant transitions. If subjects are sensitive to changes in the overall stimulus complex leading into and out of closure, we would expect concurrent changes in the labeling and discrimination functions.

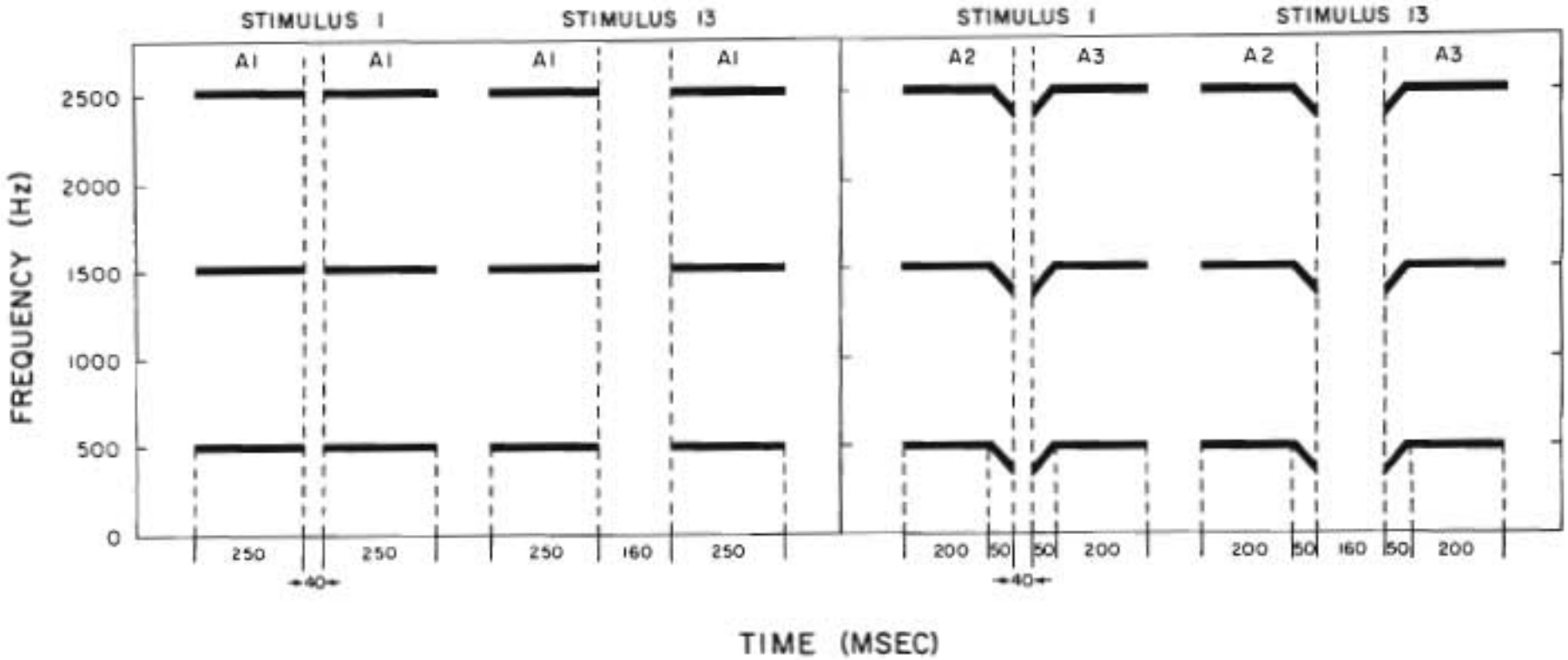


Figure 1. Schematic representations of two stimuli differing in the duration of the silent interval; Stimulus 1 (40 msec), Stimulus 2 (160 msec). The panel on the right represents stimuli without simulated formant transitions; the panel on the left shows the same stimuli with the added transitions into and out of the closure period.

Procedure. The presentation of stimuli, data collection and feedback to subjects was controlled on-line by PDP-11/05 computer. The digitized waveforms were reconverted to analog form via a 12-bit D/A converter running at a sampling of 10 kHz. The signals were presented to subjects binaurally through Telephonics (TDH-39) matched and calibrated headphones at a comfortable listening level of about 80 dB (re 0.002 dyn/cm2) throughout these experiments.

The present experiment consisted of two 1 hour sessions conducted on separate days. Each session began with an initial shaping phase during which the stimuli with silent intervals of 50 msec and 150 msec were presented in short alternating blocks with feedback. The size of these blocks decreased from five stimuli to one stimulus. The subjects were instructed that their task was to learn which of two buttons was associated with each stimulus pattern. Immediate feedback for the correct response was provided by illuminating a small light above the correct response button. No explicit coding or labeling instructions were given to subjects. This procedure left the subjects free to adopt their own strategies. After 60 trials these two stimuli were presented in random order with feedback. All subjects were required to meet a criterion of 90 percent correct for identification of the training stimuli to be included in the subsequent identification phase of the experiment.

In the identification or labeling phase of each session, the subjects were presented with all 13 stimuli from a given stimulus series in random order. They were instructed to respond exactly as they did in the preceding shaping phase but feedback was no longer provided after each trial. Six subjects were presented with the

A1-A1 series on Day 1 and the A2-A3 series on Day 2; ten subjects received the A2-A3 series followed by the A1-A1 series.

Results and Discussion

All sixteen subjects learned to respond to the endpoint stimuli with a probability greater than 0.90 during the initial shaping phase of the experiment. The group labeling function for the A1-A1 and A2-A3 stimulus series are shown in Figure 2. The A1-A1 series is shown by the open circles and dashed lines; the A2-A3 series is indicated by the filled circles and solid lines.

Insert Figure 2 about here

In this figure it can be seen that the functions for the A1-A1 series and the A2-A3 series are essentially the same. The mean crossover point of the A1-A1 series is 106 msec with a standard deviation of 18 msec. The mean crossover point of the A2-A3 series is 105 msec with a standard deviation of 16 msec. The difference between the means was not significant ($t_{14} = .776$; $p > .05$). These results demonstrate that the addition of simulated formant transitions into and out of the silent interval in the A2-A3 series has no systematic effect on the crossover point in these labeling functions for the nonspeech stimuli.

As mentioned above, Liberman et al. (1961) studied perception of closure duration in synthetic speech and nonspeech stimuli. However, they did not collect any labeling data for their nonspeech stimuli although they inferred the presence of a boundary between /b/ and /p/ at approximately 70 msec and,

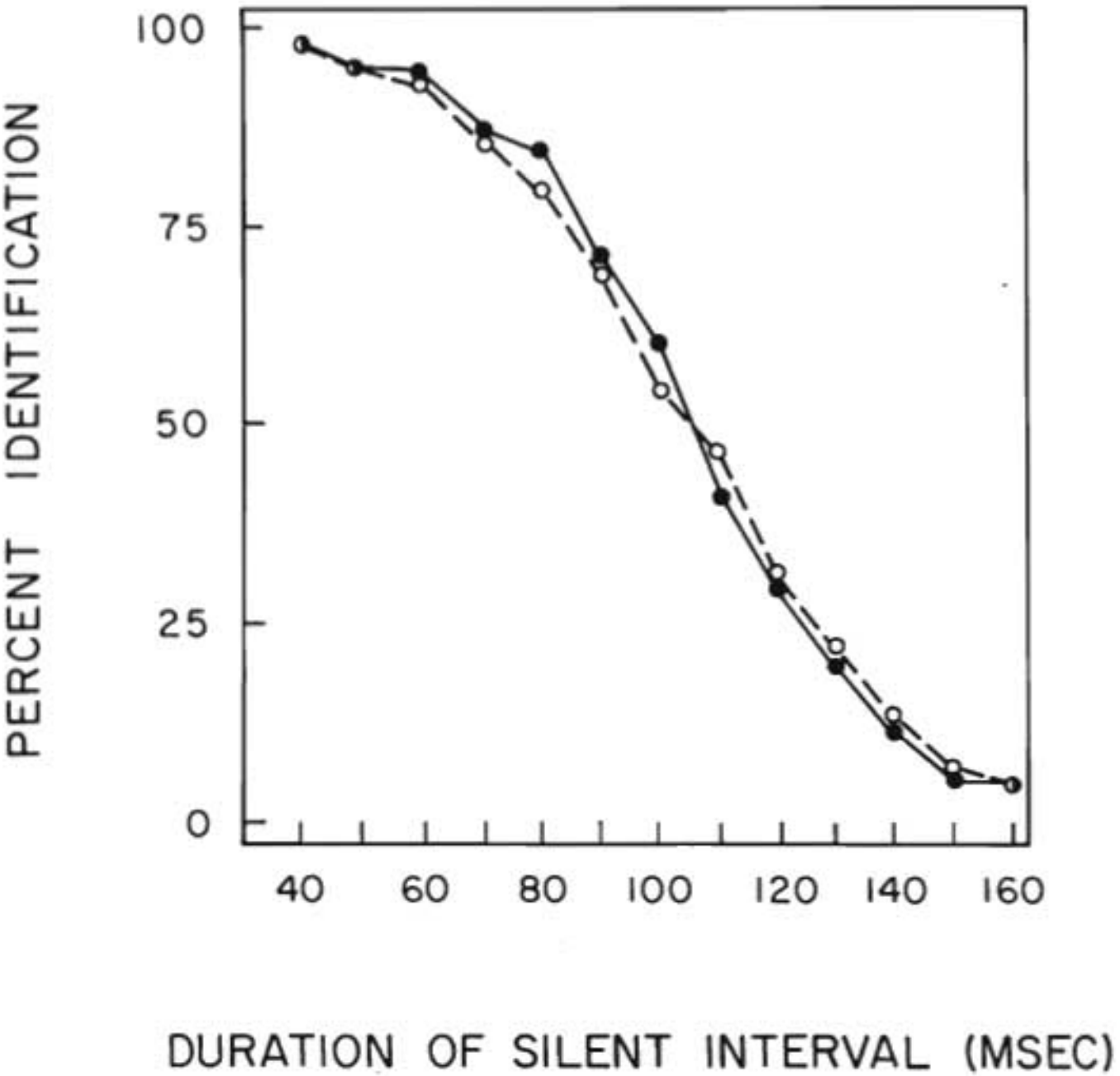


Figure 2. Group labeling functions for the 16 subjects in Experiment 1. The A1-A1 series is indicated by open circles and dashed lines, the A2-A3 series is indicated by filled circles and solid lines.

surprisingly, a second boundary between 100 and 110 msec. This second boundary at 100-110 msec was not as well defined as the one between /b/ and /p/ but it was definitely present in discrimination data. Interestingly, the crossover points obtained in the present experiment appear to correspond to the less well defined boundary rather than the one observed at the shorter duration. Perhaps if a third response category was available to the subjects in the present tests, they would also show evidence of three categories. However, it is not clear why the less well-defined boundary appeared in these labeling functions and not the more well-defined boundary associated with /b/ and /p/ at 70 msec.

These labeling data on nonspeech stimuli demonstrate that subjects can assign nonspeech stimuli differing in the duration of a silent interval into two discrete perceptual categories. However, it is also necessary to obtain ABX discrimination functions in order to compare the observed performance with that expected from identification. Experiment 2 was designed to measure discrimination of the duration of the silent intervals in the same stimuli.

Experiment 2

In this experiment subjects were required to discriminate between pairs of stimuli differing in the duration of the silent interval. Our goal was to determine whether the observed discrimination functions are nonmonotonic and show the characteristic peaks and troughs of stimuli that are typically perceived categorically.

Method

Subjects. Fifteen paid volunteers were obtained in the same manner as Experiment 1.

Stimuli. The two sets of 13 tonal stimuli that were used in Experiment 1 were also used in this experiment.

Procedure. The presentation of stimuli, data collection and feedback was conducted as in Experiment 1. The ten three-step pairs, resulting in 30 msec differences between test stimuli were arranged in the four ABX permutations and presented to each subject with feedback for the correct response. Subjects were instructed to determine whether the third sound in each triad was most like the first sound or most like the second sound. The timing and sequencing in the experiment was self-paced to the slowest subject in each group.

The experiment was conducted in two 1 hour testing sessions occuring on separate days. On each day subjects received eight replications of each stimulus comparison for a total of 320 ABX trials. Seven subjects were presented with stimuli from the A1-A1 series on Day 1 followed by the A2-A3 series on Day 2. The

remaining eight subjects received the two stimulus series in the reverse order.

Results and Discussion

The group ABX functions for the A1-A1 series and the A2-A3 series are displayed in Figure 3. As in the previous experiment, the A1-A1 series is shown by open circles and dashed lines while the A2-A3 series is shown by solid circles and solid lines. Both ABX functions are relatively flat and very close to chance. The fact that these discrimination functions are so close to chance makes any conclusions tenuous at this time. Apparently, our subjects had great difficulty discriminating these differences.

Insert Figure 3 about here

An examination of the individual discrimination functions for both the A1-A1 and A2-A3 series revealed substantial variability in the location of the peaks and troughs. The functions were erratic with the peaks rarely rising above 60 percent correct and the troughs often dropping below the 50 percent chance level. Such findings suggest that subjects were unable to focus their attention on the distinctive acoustic cues in these stimuli. The overall performance for each individual was very close to the 50 percent chance level.

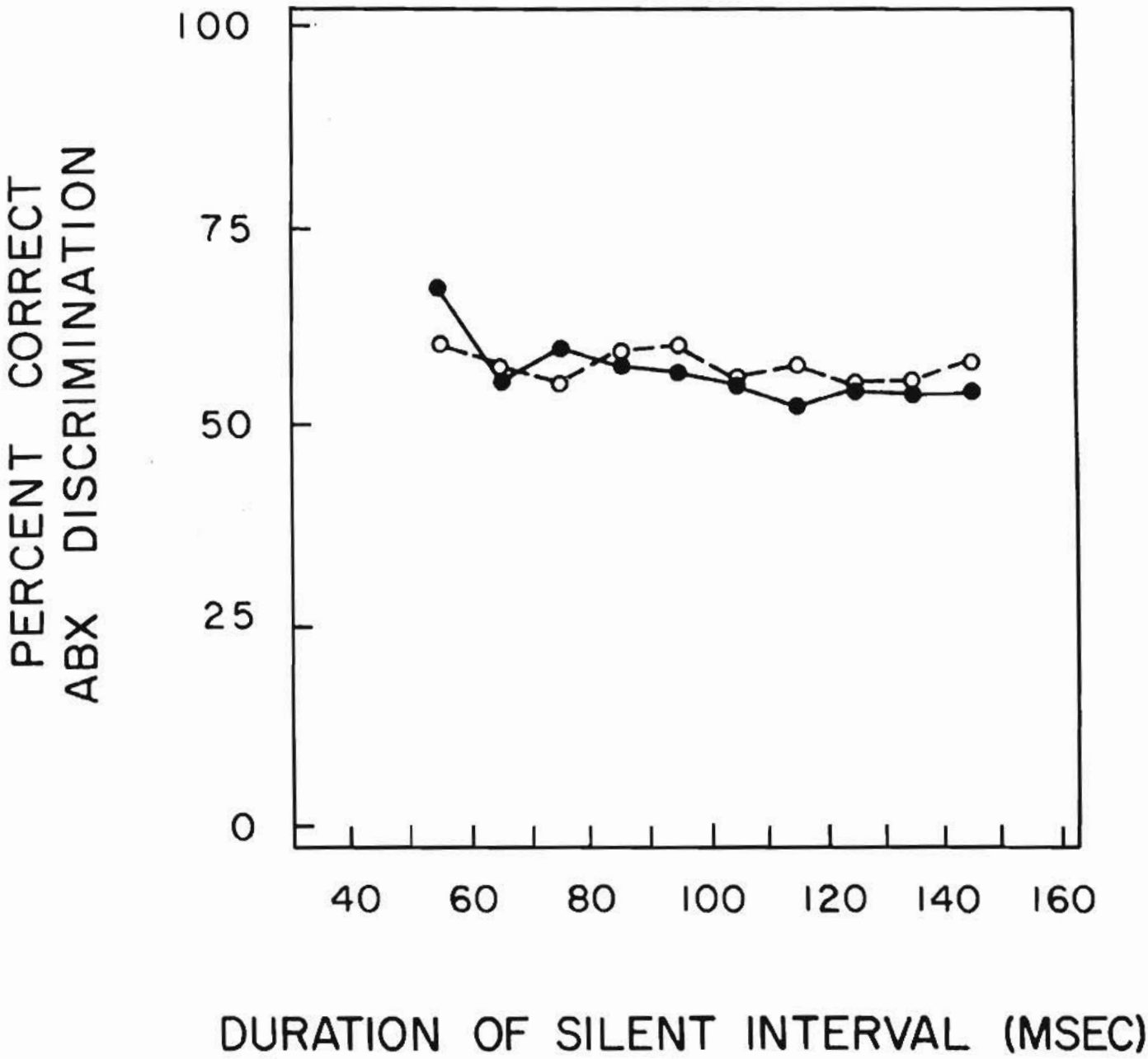


Figure 3. Group ABX discrimination functions for the 15 subjects in Experiment 2. The A1-A1 series is shown by open circles and dashed lines, the A2-A3 series is shown by filled cirlces and solid lines.

General Discussion

The first experiment described above demonstrated clearly that subjects can classify nonspeech stimuli differing in the duration of a silent interval into two discrete categories with a relatively sharp crossover point between the two categories. The addition of simulated formant transitions does not appear to alter the shape or the boundary of these labeling functions. However, these results taken alone are not sufficient to argue for the categorical perception of this acoustic cue. The additional condition that these stimuli demonstrate nonmonotonic discrimination functions with a peak at the crossover point in the labeling functions and low discriminability within categories was not observed in the second experiment. The ABX discrimination performance was very poor and quite close to chance at almost all stimulus values along the continuum suggesting that these subjects could not discriminate the differences in the silent interval.

In the earlier study of Liberman et al., (1961) labeling functions were not obtained for the nonspeech stimuli which were two noise bursts separated by various durations of silence. The nonspeech ABX discrimination functions, however, were also flat and very close to chance. Liberman et al. considered the nonspeech discrimination functions as representing the "true" discriminability of the acoustic cue of closure duration without the benefit of any linguistic experience. The discrimination peaks observed in their synthetic speech stimuli varying in closure duration were therefore ascribed to learning through a process of acquired distinctiveness.

The discrimination results of Experiment 2 would appear, at first glance, to confirm the earlier results of Liberman et al. (1961). Furthermore, it might be argued that such results are additional evidence that certain acoustic cues to speech are perceived differently in speech contexts than when they appear in nonspeech contexts. The problem with this interpretation is that it is based upon data representing chance level performance obtained with the nonspeech stimuli. Such low overall discrimination performance makes a direct comparison with speech stimuli that exhibit very high overall discrimination performance quite difficult. It may well be that some other factor is responsible for the low performance and the absence of peaks in the nonspeech discrimination functions.

One possibility that we considered was that the subjects were not attending to the relevant attributes in these complex stimuli necessary to distinguish these sounds. Furthermore, two different groups of subjects received exposure to the labeling and discrimination experiments reported here. It is possible that higher performance in the ABX discrimination tasks might have been obtained by prior exposure to the labeling task and/or by more explicit instructions that would direct the subjects attention to the silent interval in the stimuli. Such instructions might include explicit labels such as "gap" vs. "no gap" or "short" vs. "long gap" to draw attention to the differentiating attributes of the stimuli. Another possibility would be to instruct subjects to listen to the center of the stimuli when they determine whether the third stimulus was most like the first or the second in the ABX task. We found in listening to these stimuli that it was

difficult to determine where the gap was. The two components of each stimulus were not perceived as a unitary event at all but rather they appeared as separate events in time.

If an increase in the ABX discrimination performance could be obtained through these manipulations at least two different outcomes are possible. First, the nonspeech stimuli differing in gap intervals might then exhibit the nonmonotonic functions that contain the characteristic peaks and troughs of categorical-like discrimination functions. Such a result would then allow for a direct comparison with the speech stimuli and lend further support for the notion that the perceptual processing of speech may be based on the psychophysical attributes of the stimuli themselves, at least in the case of certain temporal cues.

A second possibility is that the discrimination performance may be raised substantially but still retain a basically flat monotonic shape across the entire stimulus continuum. This outcome would imply that the acoustic cue of closure duration is, in fact, perceived differently when it appears in nonspeech contexts than when it appears in speech contexts. To test this notion other experiments should be carried out in which the same durations of silence are enclosed by vowels which match, at the very least, the durational characteristics of the nonspeech signals. Comparable performance combined with the appearance of categorical labeling and discrimination functions would provide additional support for the notion of differential processing of speech and nonspeech signals and for the operation of a speech mode of processing.

Several findings have suggested that closure duration alone may not be the only cue for the voiced-voiceless distinction in

stop consonants in medial or intervocalic position. Port (1979) has suggested that the relevant cue for this distinction is closure duration relative to the duration of the preceding syllable. Thus, in case of "rabid" versus "rapid", a subject will hear the former when the ratio of the silent interval to the duration of the first syllable is less than one and will hear the latter when the ration is approximately equal to one. To test this possibility we plan to generate synthetic tokens of the words "rabid" and "rapid" so that duration of the final syllable is held constant while the duration of the first syllable and the closure duration are varied systematically over a range of values. If the critical perceptual cue is the ratio of closure duration to the duration of the first syllable, any stimulus ratio less than one should yield judgements of "rabid" while those approximately equal to one should yield judgements of "rapid", regardless of the absolute value of the closure interval.

A nonspeech analogue of the "rapid-rabid" experiment should also be constructed using stimuli similar to those of the present experiments. The stimuli will be exactly the same in all respects except that the duration of the initial component in each series will be either 100 or 150 msec. The gap interval will then be varied. If results show labeling and discrimination performance like those observed for the speech version, this outcome would lend further support to the notion that speech perception is based upon the psychophysical attributes of complex acoustic stimuli. If these stimuli show comparable but flat monotonic discrimination functions relative to the speech stimuli differing in the same cue, a case could be made for the existence and operation of a specialized speech mode of processing for speech stimuli.

Almost all the research conducted on closure duration as a cue to voicing has removed the low frequency voicing energy during the closure interval. Since Lisker (1957) demonstrated that closure duration alone is sufficient to cue the voiced-voiceless distinction in minimal pairs, the presence of voicing pulses has been largely neglected. It is possible that the presence of voicing during closure may have a substantial effect on distinguishing "rabid" from "rapid" in natural speech. The voicing response might then be based on a temporally continuous versus a temporally discontinuous event. This possibility could be examined in a nonspeech environment by manipulating the duration of the silent interval with and without low frequency energy occurring during this interval.

In summary, the two experiments reported here were concerned with the question of whether the acoustic cue of closure duration, utilized in distinguishing voicing distinctions between stop consonants in intervocalic position, is perceived in a categorical manner when it appears in a nonspeech context. Such a demonstration would be useful in detailing the extent to which the perceptual categories of speech are based on the psychophysical attributes of the stimuli themselves. The results of this study indicated that while individuals can reliably identify the nonspeech stimuli differentially based in differences in the duration of a silent interval, discrimination performance was flat and very close to chance. Further experiments are currently underway to determine the basis of the poor discrimination performance observed with these nonspeech stimuli and to draw appropriate comparisons with speech stimuli differing in the duration of the silent interval.

References

Abramson, A. S. & Lisker, L. Discriminability along the voicing continuum: Cross language tests. Proceedings of the Sixth International Congress of Phonetic Sciences, Prague, 1967. Prague: Academic, 1970.

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

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.

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

Jusczyk, P. W., Rosner, B. S., Cutting, J. E., Foard, C. F. & Smith, L. B. Categorical perception of nonspeech sounds by 2-month-old infants. Perception & Psychophysics, 1977, 21, 50-54.

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

Lasky, R. E., Syrdal-Lasky, A. & Klein, R. E. VOT discrimination by four and six and a half month old infants from Spanish environments. Journal of Experimental Child Psychology, 1975, 20, 213-255.

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 Williams Co., 1970. (a)

Liberman, A. M. The grammars of speech and language. Cognitive Psychology, 1970, 1, 301-323. (b)

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

Liberman, A. M., Harris, K. S., Eimas, P., Lisker, L. & Bastian, J. An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. Language and Speech, 1961, 4, 1975-195.

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

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

Liberman, A. M. & Pisoni, D. B. Evidence for a special speech-perceiving subsystem in the human. In T. H. Bullock (Ed.), Recognition of complex acoustic signals. Berlin: Dahlem Konferenzen, 1977, Pp. 59-76.

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

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

Lisker, L. & Abramson, A. S. The voicing dimension: Some experiments in comparative phonetics. Proceedings of the Sixth International Congress of Phonetic Sciences, Prague, 1967. Prague: Academic, 1970.

Mattingly, I. G. Speech cues and sign stimuli. American Scientist, 1972, 60, 327-337.

Mattingly, I. G., Liberman, A. M., Sydral, A. K. & Halwes, T. 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, 410-417.

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

Pisoni, D. B. Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. Journal of the Acoustical Society of America, 1977, 61, 1352-1361.

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

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

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

Waters, R. A. & Wilson, W. A., Jr. Speech perception by rhesus monkeys: The voicing distinction in synthesized labial and velar stop consonants. Perception & Psychophysics, 1976, 19, 285-289.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 6 (1979-80) Indiana University]

The Perceptual Classification of Speech and Nonspeech Sounds

Peter W. Jusczyk
Department of Psychology, University of Oregon

Linda B. Smith and Christopher Murphy
Department of Psychology, Indiana University

Abstract

By employing new methods of analysis to the physical signal, a number of researchers have provided evidence which suggests that there may be invariant acoustic cues which serve to identify the presence of particular phonetic segments (e.g. Kewley-Port, 1980; Searle, Jacobson & Rayment, 1979; Stevens & Blumstein, 1978). Whereas previous studies have focused upon the existence of invariant properties present in the physical stimulus, the present study examines the existence of any invariant information available in the psychological stimulus. For this purpose subjects were asked to classify either a series of full-CV syllables ([bi], [bɛ], [bo], [bɔ], [di], [dɛ], [do], [dɔ]) or one of two series of chirp stimuli consisting of information available in the first 30 msec of each syllable. The full-formant chirp stimuli consisted of the first 30 msec of each syllable, whereas the two-formant chirps were composed of the first 30 msec of only the second and third formants. The object of the present study was to determine whether or not there was sufficient information available in either the full- or two-formant chirp series to allow subjects to group the stimuli into two classes corresponding to the identity of the initial consonant of the syllables (i.e. [b] or [d]. A series of classification tasks were used, ranging from a completely free sorting task to a perceptual learning task with experimenter-imposed classifications. The results suggest that there is information available in the full-formant chirps, but not the two-formant chirps, which allows subjects to group the sounds into classes corresponding to the identity of the initial consonant sounds. Thus, the present results show that within the first 30 msec of CV-syllables invariant cues to the identity of initial consonants are available in a form usable by the perceiver.

A central problem in speech perception is how a listener is able to recognize the presence of a particular phone in all of its possible utterance contexts. The most straightforward solution to this problem would be to isolate a particular phone in the speech stream and to identify the property or properties which signal its presence in all contexts. However, as is well-known, there are several difficulties with this approach. First of all, it is not always possible to isolate and segment phones from the speech stream. Because of coarticulation, information about more than one phone is frequently present over any given interval in time. Thus, attempts to divide a syllable like [di] into two segments corresponding to [d] and [i] have been unsuccessful (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). As larger and larger sections of the vocalic portion of the syllable are removed, the listener's perception of the remaining portion changes from hearing a shortened [di] to hearing a nonspeech sound. At no point does the listener report perceiving only [d] (but see Kewley-Port, 1980).

Despite the fact that it is not always possible to isolate individual phonetic segments, one could still ask whether there are invariant properties in the utterance which indicate the presence of a particular phone. Early attempts to isolate such properties were based upon measurements obtained via spectrographic analyses of speech. These analyses indicated that while some phones appear to exhibit relatively stable properties in different phonetic contexts, such as the characteristics of the noise spectra accompanying [s] and [š] (Harris, 1958), other phones appear to undergo considerable context-conditioned variation. For example, studies with two-formant synthetic speech sounds indicate that [d] can be signaled by a rising second formant transition in some contexts like [di], and a falling one in others like [du] (Liberman et al., 1967). Not only are invariant properties lacking in the spectrographic representations of [d] across different

phonetic contexts, but the same acoustic property may even cue entirely different phones depending upon the folloiwng vowel context. For example, the same burst of sound is heard as [p] when placed before an [i], but as [k] when before an [a] (Liberman, DeLattre, & Cooper, 1952; Schatz, 1954).[1] Findings such as these led some researchers to suppose, for at least some phones, there are no invariant physical properties present in the acoustic signal (Liberman et al., 1967). Instead, the source of any invariants for speech sounds was deemed to occur in the way in which the acoustic signal was decoded by the perceiver. In particular, it was hypothesized that the invariance between the different manifestations of a particular phone in various contexts could be traced to neural correlates of the articulatory gestures used to produce it.

The belief that some phones lack invariant properties in the acoustic signal was propagated by the failure to find such properties in spectrographic analyses of speech (Liberman et al., 1967). However, as useful as such analyses have been in investigating speech perception, they constitute only one of a number of ways in which the speech signal can be analyzed. Recently, some investigators have begun to employ other methods of analysis to the speech signal. Three efforts which have achieved some measure of success in this area are those by Kewley-Port (1980), Searle, Jacobson, and Rayment (1979) and Stevens and Blumstein (1978; in press). The methods used by Kewley-Port (1980) and by Searle et al. (1979) are based upon current understanding of the physiology and psychophysics of the auditory system. Searle et al. compute a series of running spectra for speech sounds from which certain features are extracted to determine what phone was uttered. Although encouraging, the recognition accuracy of their model (77% for detection and classification of stop consonants in initial position) still falls short of that obtained in human speech perception. Using slightly different filtering parameters and a

more fine-grained analysis of running spectra, Kewley-Port (1980)'s model approached the accuracy of human speech perception. Under similar test conditions to those employed by Searle et al., Kewley-Port's model obtained an overall recognition accuracy score of 88%.

The approach taken by Stevens and Blumstein (1978, in press; Blumstein & Stevens, 1979) is based upon considerations derived from the acoustic theory of speech production (Fant, 1960). The basic notion behind this approach is that owing to the acoustic properties of the cavities in the vocal tract, the spectrum sampled over the 10-20 msec immediately following consonantal release will have distinctive overall characteristics for different places of articulation. The time window that Stevens and Blumstein use in computing the onset spectrum is long enough to include both burst and some formant transition information. However, unlike earlier analyses which tended to examine each of these cues independently, the cues occur together in an integrated form in the onset spectra. Hence, invariants for speech sounds are presumed to lie in the overall gross shape of the spectrum at onset. As a test of the accuracy of their model, Blumstein and Stevens (1979) examined a large number of CV and VC tokens containing both voiced and voiceless stops produced by several speakers. Despite some variability across vowel contexts, the model attained an overall correct identification for CV syllables. However, recognition accuracy for VC syllables was only 76%. Moreover, a later test of the generality of the model to nasal stops was considerably less successful as the false alarm rate for the model was as high as 67% in some instances. Nevertheless, even the limited success which Stevens and Blumstein have achieved has spawned renewed interest in the search for invariant acoustic cues in the speech signal.

The recent gains towards identifying potential acoustic invariants have come about because investigators have employed new ways of describing the physical signal. The goal of these investigators is to discover a description of the physical signal under which physical properties associated with a particular phone remain constant across various contexts. The hope is that if such a description were found, it might correspond to the one used by the perceiver in recognizing speech. However, the best physical description of a set of stimuli need not correspond to the best psychological description (Boring, 1926, 1933; Gibson, 1966; Smith & Kemler, 1978). A description of the speech signal in terms of some invariant physical properties will provide a successful account of speech perception, only insofar as it isolates those properties that the perceiver operates on. What is required is that any proposed set of invariant properties for phones derived from a description of the physical stimulus also remain invariant under a description of the psychological stimulus. Therefore, an alternative and hopefully converging way of approaching the invariants issue would be to begin with a description of the psychological stimulus. At what point, if any, in the analysis of the speech signal does the listener behave as though there is invariant information common to all instances of a particular phone?

Recent research has begun to address this issue of the psychological reality of onset-spectra cues (Blumstein & Stevens, 1980; Kewley-Port, 1980). In a series of studies Blumstein and Stevens (1980) collected identification data for brief segments excerpted from the initial portions of synthetic consonant-vowel stimuli. Subjects labeled these stimuli as either [b], [d], or [g], with a high level of accuracy even when the segments were as short as 10-20 msec. Moreover, in most instances, subjects were also able to correctly label the vowel of the syllable from which the stimulus had been excerpted.

Blumstein and Stevens' (1980) results indicate that there is information available in the onset spectra of CV syllables which the perceiver can use to determine the identity of a particular phone. These results taken together with their earlier work (Stevens & Blumstein, 1978) would seem to suggest that the invariant properties derived from their analyses of the physical stimulus are present as well in some form in the psychological stimulus. However, before this conclusion can be accepted, several factors must be considered. First of all, since Blumstein and Stevens (1980) always conducted their identification tests separately for each vowel context, it is possible that subjects employed different cues to establish the identity of a particular phone in each context. Certainly, the fact that subjects were quite successful in recognizing the following vowel is consistent with the use of context-dependent cues to establish the identity of the consonant. Second, by instructing their subjects to label the stimuli as [b], [d], or [g], Blumstein and Stevens imposed a constraint on the kinds of perceptual categories which subjects could form. The imposition of such constraints makes it difficult to know whether a categorization according to [b], [d], and [g] represents the most natural psychological grouping of the sounds.

The main objective of the present study was to determine whether brief segments of speech at the onsets of syllables contain invariant information for phones in a form which is usable to the perceiver. While the basic objectives of the two studies are similar, the approach taken in the present study differs in several notable ways from that of Blumstein and Stevens (1980). First we employed free classification tasks to determine subjects' natural groupings for brief segments of speech. In such tasks, subjects are essentially free to group the stimuli in any way they wish. Second, in instances in which subjects were

directed to classify according to phonetic categories, their performance was assessed by comparing it to that obtained when they were instructed to sort according to an arbitrary classification scheme. Inclusion of such a comparison made it possible to test whether a classification according to a phonetic invariant was any more psychologically real for subjects than an arbitrarily chosen grouping. Third, a variety of vowel contexts were present in all classification tasks. Fourth, in addition to utilizing brief segments of synthetic speech with a full complement of formants, the present study also employed both complete syllables and brief two-formant patterns from which the first formant had been deleted. Classification of the latter stimuli was examined because it has been claimed that the deleted first formant information is redundant across different places of articulation (e.g., Mattingly, Liberman, Syrdal, & Halwes, 1971). The effect of deleting such "redundant" material on the perceptual categories of brief segments is of interest--especially since these "chirplike" patterns have been used as nonspeech controls in many experiments (e.g., Mattingly, et al., 1971; Morse, 1972). The specific questions addressed in the present study include the following: (1) Are brief segments of speech corresponding to one consonant (e.g., [b]) perceived to be more similar to each other than ones corresponding to different consonants (e.g. [b] and [d])? (2) Might such segments be perceived as falling into categories isomorphic to ones organized by initial phonetic segments? (3) More specifically, if subjects are presented with brief speech segments corresponding to [be, bɛ, bo, bə^] and [di, dɛ, do, də^], will they perceive these segments as belonging to two distinct perceptual categories--"b" and "d"?

Experiment I

The hypothesis that brief segments of speech contain invariant information leading to the recognition of particular phones implies that all segments containing

the same invariants are perceived as being similar. This hypothesis does not imply that the listener is necessarily able to label all the segments corresponding to a particular phone (say [d]) as sounding like "d". The critical question is whether listeners perceive brief segments corresponding to [d] as distinct from those corresponding to [b]. Thus, Experiment I consisted of a classification task in which subjects were free to classify the stimuli in any way which they chose. Our question was whether subjects easily and naturally group brief segments of speech into categories corresponding to phonetic ones.

Three types of stimuli were employed to address this issue: complete CV syllables, truncated versions of these syllables containing only the first 30 msec of each (hereafter referred to as "full-formant chirps"), and truncated versions of the syllables containing only the first 30 msec of the second and third formants (hereafter called "two-formant chirps"). Full CV syllables were included since it was expected that subjects might easily sort these stimuli into categories based on the identity of their initial consonants. The full-formant chirps were employed because these segments preserved the onset spectra information which Stevens and Blumstein (in press) have claimed provides the invariant phonetic cues. Finally, two-formant chirps were examined because such patterns have often been used to present listeners with the "same acoustic information" in a nonspeech context that is present in a speech context, (e.g., Eimas, 1975; Mattingly et al., 1971; Miyawaki, Strange, Verbrugge, Liberman, Jenkins, & Fujimura, 1975; Morse, 1972). If the two-formant chirps really do preserve the same acoustic information as the syllables, then one might expect to observe similar sorting patterns for these two types of stimuli.

Method

Stimuli

The stimuli consisted of eight synthetic speech syllables [bi, bɛ, bo, bə^ , di, dɛ, do, də^] plus the two truncated versions of each. All stimuli were prepared on a PDP 11/03 computer in the Speech Perception Laboratory at Indiana University and were generated with the cascade-parallel synthesizer designed by Klatt (1980) and modified by Kewley Port (Note 1). The eight natural speech tokens spoken by P. W. J. that served as models for constructing the synthetic tokens are displayed in figure 1. The specific range of vowel contexts included in the syllables was chosen to maximize differences between the

---
Insert Figure 1 about here
---

relationship of the first, second, and third formants. This measure was adopted so as to provide the strongest possible test of potential invariant cues to the initial consonants.

The syllable stimuli were all generated without release bursts and were equated for overall duration (295 msec) and pitch contour. The latter had an initial value of 121 Hz, rose to a peak of 125 Hz after 45 msec, and then fell linearly to a terminal value of 100 Hz.[2] Syllables sharing a common vowel (e.g., [bo] and [do]) were equated in all respects except for their second and third formant transition values. Table 1 presents the values of the first, second, and third formant values sampled at four points in the duration of each test syllable.

Insert Table 1 about here

Full-formant chirps for each syllable were produced by truncating the syllable after 30 msec., at which point the transitions of the first, second and

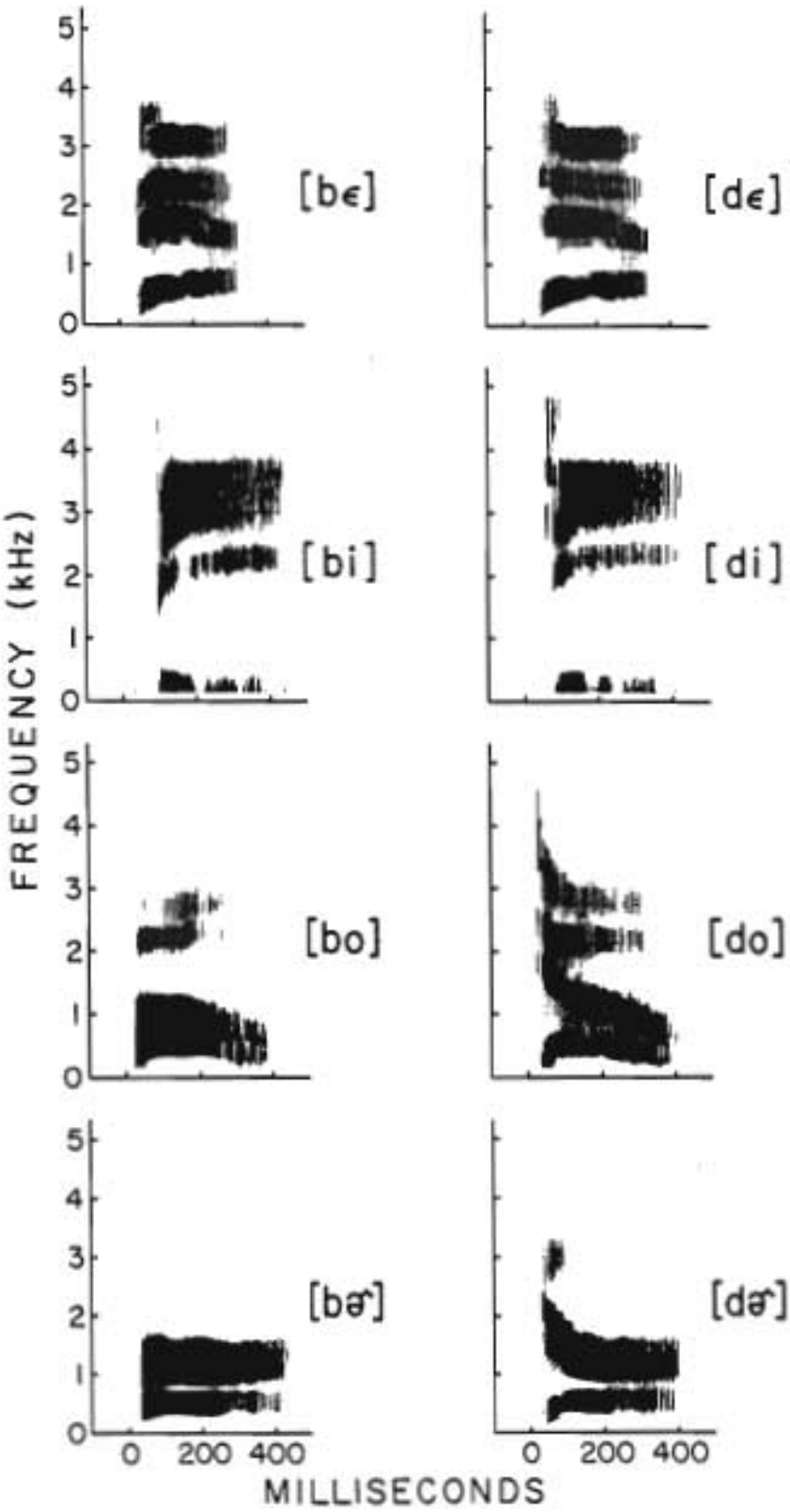


Figure 1 - A spectrographic display of the natural CV syllables which served as models for the synthetic speech tokens used in the present

Table 1

First, second, and third formants transitions (0-30 msec) and steady state frequencies (corresponding to the following vowel: 31-295 msec) in Hz for the 8 syllables sampled at 4 points. The relationship between the frequencies of all adjacent samples is linear.

		0 msec	30 msec	200 msec	295 msec
	F1	200	300	249	220
/bi/	2	1500	2000	2160	2250
	3	2200	2800	3057	3200
	F1	200	300	249	220
/di/	2	1900	2000	2160	2250
	3	3300	2800	3057	3200
	F1	200	550	678	750
/bɛ/	2	1400	1850	1850	1650
	3	2100	2500	2500	2500
	F1	200	550	678	750
/dɛ/	2	2250	1850	1850	1650
	3	3200	2500	2500	2500
	F1	250	500	500	350
/bo/	2	650	1050	1050	800
	3	1700	2240	2240	2240
	F1	250	500	500	350
/do/	2	1800	1050	1050	800
	3	3000	2240	2240	2240
	F1	200	600	600	600
/bɝ/	2	800	1200	1200	1200
	3	1100	1600	1600	1600
	F1	200	600	600	600
/dɝ/	2	1600	1200	1200	1200
	3	2600	1600	1600	1600

third formants were complete. Thus, the relevant formant trajectories are identical to those of the full syllables and are displayed in the first two columns of Table 1. Moreover, since the full-formant chirps are merely truncated versions of the complete syllables, the onset spectrum for a given full-formant chirp is identical to that of the syllable from which it is derived. The onset spectra for the syllables were analyzed in the same way as Stevens and Blumstein (1978) using the linear prediction method and by pre-emphasizing the higher frequencies and employing a 26 msec time window (Kewley-Port, Note 2).[3] The resulting spectra for the stimuli are displayed in Figure 2.

Insert Figure 2 about here

The two-formant chirps were generated by essentially removing the first, fourth, and fifth formant information from the full-formant chirps. However, since the removal of this information can result in a drastic change in the amplitude relations between the second and third formants, measurements of the amplitudes of the transition portions of these formants were made from each syllable using the KLAMP program devised by Kewley-Port. The two formant chirp patterns were then generated on the parallel branch of the Klatt synthesizer taking care to maintain the appropriate amplitude relations of the formants throughout the duration of the chirps. Owing to the lack of acoustic energy in the regions of the first, fourth, and fifth formants, the onset spectra for the two-formant chirps differ considerably from those of the syllables and full-formants. Figure 3 presents the onset spectra for the two-formant chirps.

Insert Figure 3 about here

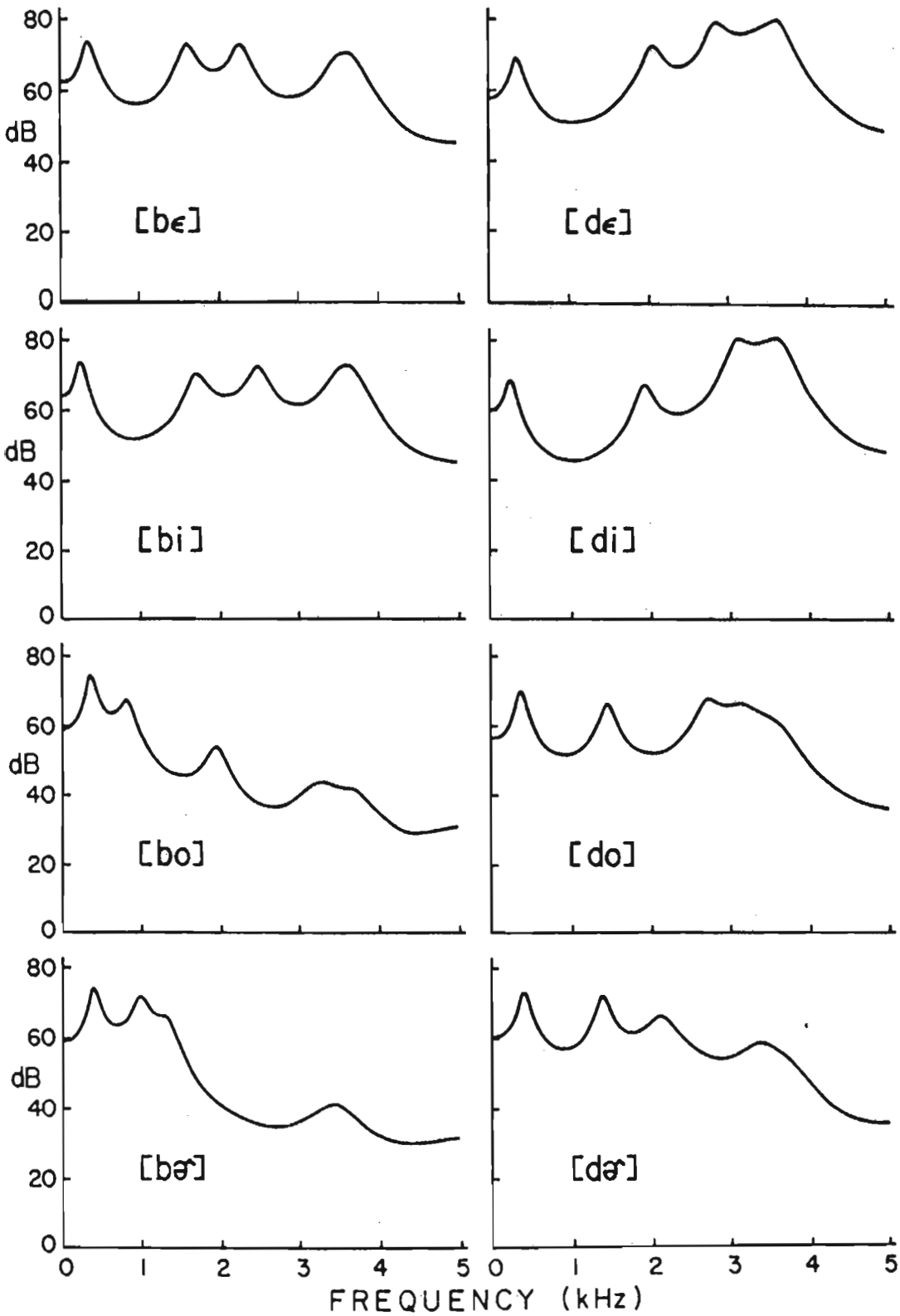


Figure 2 - A display of the onset-spectra for both the full-syllable and full-formant chirp stimuli.

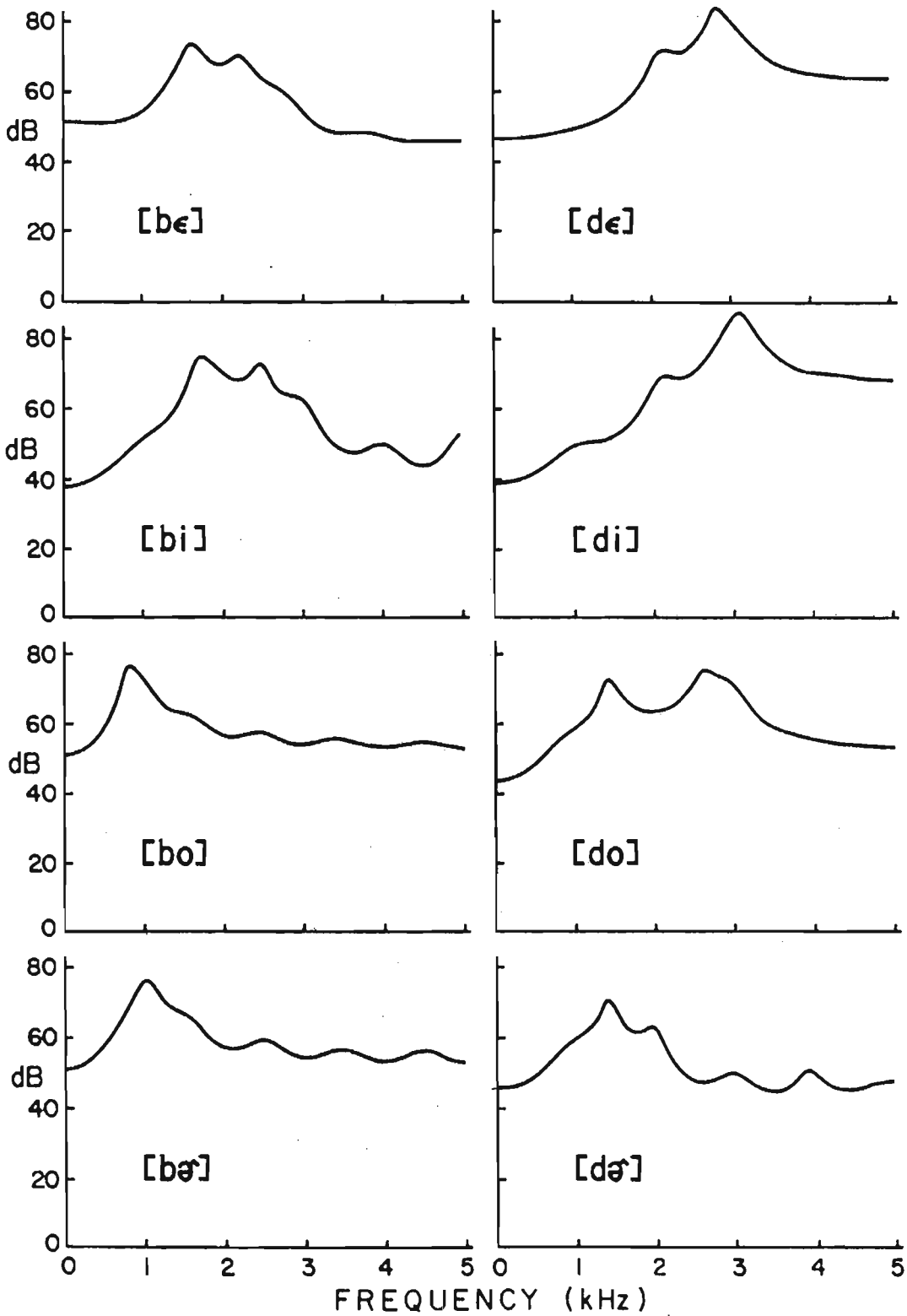


Figure 3 - A display of the onset-spectra of the two-formant chirp stimuli.

Nine test tapes were generated using an audio tape making program at Indiana University. The stimuli were converted to analogue form in real-time via a 12-bit digital-to-analogue converter, low pass filtered at 4.8 KHz and recorded on a Crown model 822 tape recorder at 7-1/2 ips. There were three tapes for each of the three stimulus conditions (i.e., full-CV syllables, full-formant chirps, and two-formant chirps). For the syllable condition, each tape included randomized sequences of the eight syllables [bi, bɛ, bo, bɔ̂, di, dɛ, do, dɔ̂]. There were ten occurrences of each syllable for a total of 80 trials per tape. A four-second response interval separated successive syllables on each tape. The tapes for the full-formant and two-formant chirp conditions were prepared in an identical fashion with the appropriate chirps being substituted for syllables. In addition to the test tapes, four tapes were prepared for each stimulus condition for the purpose of familiarizing the subjects with the stimulus sets.

Subjects

The subjects were fifteen undergraduates at Dalhousie University who received course credit for participating in the experiment. All were native speakers of English and reported no history of either speech or hearing disorder.

Procedure

Each subject was tested individually in a small quiet room for one half-hour session. Practice and test tapes were played on a Tandberg TB-3500 X tape recorder equipped with Koss PRO 4AAA headphones. The volume was adjusted with reference to a sound level meter (General Radio model 1565-A) so that the stimuli were played at a level of approximately 72 dB (A) SPL.

An equal number of subjects was assigned randomly to each of the three stimulus conditions. Subjects were instructed that they would be hearing a series

of eight sounds and that they would have to sort the stimuli into groups. No explicit rule was provided for the subjects, rather they were told to "put together the stimuli which sounded the most alike." Subjects were informed that they were free to form as many groups as they desired. Following these instructions, subjects were played one of the demonstration tapes containing one occurrence of all eight stimuli so that they might familiarize themselves with the entire set of sounds. Two more demonstration tapes with the eight test items in different random orders were then played and subjects were asked to practice assigning the stimuli to groups. Subjects were told to use numbers to designate the different groups and to indicate the assignment of a particular stimulus to a group by writing down the group number in an appropriately marked space on their answer sheet. After the practice trials, subjects were administered one 80-item test tape. A short five-minute break ensued during which time subjects were told that they were free to change the way in which they grouped the stimuli if they wished. Upon completion of the break, the demonstration and practice sequences were repeated prior to the administration of a test tape (with a different random ordering of the stimuli). Before departing, all subjects completed a short questionnaire regarding the criteria and strategies they employed in grouping the stimuli.

Results and Discussion

Owing to the fact that subjects were often still familiarizing themselves with the stimulus materials during the initial test series (especially with the chirp stimuli), only the data from the second 80-item test series were submitted to analysis. The first consideration was whether subjects were consistent in putting a given stimulus in a particular group or whether they changed the group

to which the stimulus was assigned from trial to trial during the test. Inconsistent classifications of particular sounds might result from difficulties in identifying the stimuli and/or remembering one's groups. To assess the overall consistency of subjects' classification of the stimuli, a consistency measure was devised using the Relative H statistic which provides a measure of the amount of uncertainty present in subjects' categories (Attneave, 1959; Garner, 1962). The consistency with which a subject classified a given stimulus is equivalent to 1-RelH, where RelH = $\frac{\text{Iplogp}}{\text{\# of bits}}$. Consistency scores for each of the eight stimuli were computed separately for each subject. These scores were then submitted to an ANOVA of a 3 (condition) X 8 (stimulus) mixed design. Neither of the main effects (condition $F_{(2,12)} = 1.55$; stimulus $F_{(7,84)} = 1.04$) nor the interaction ($F_{(14,18)} = 1.22$) was significant indicating there was no difference in the consistency with which subjects sorted the syllables or either type of chirp patterns. More importantly, the overall consistency score of .78 (where minimum and maximum possible scores are 0.0 and 1.0 respectively) demonstrates that subjects were able to reliably classify the stimuli into categories.

The kinds of groups that subjects formed depended somewhat on the stimulus condition. Each of the five subjects in the syllable condition sorted the eight sounds into four groups based on vowel identity: (bi, di), (bɛ, dɛ), (bo, do), (bɑ, dɑ). There was no indication of any other pattern prevalent in subjects' sorting of the syllables. Even the subject whose sorting fit the vowel identity rule least well was accurate by this rule on 82% of the trials.

By contrast, no particular classification pattern predominated in the full-formant condition. However individual subjects were all internally consistent in following the idiosyncratic classification schemes they had chosen. One subject

actually followed the same pattern as those who sorted the syllables, but the remaining subjects grouped them in very different ways. Similarly, for the two-formant chirps, there was no particular pattern extant in the sorting behavior of all five subjects. Instead, a variety of patterns occurred. Interestingly enough, none of the subjects followed the vowel identity rule with these stimuli, although there was some indication, present strongly in the data from one of the subjects, of division according to what might be termed "vowel similarity." On 100% of the trials, this subject sorted the chirps into two groups-- (bi, bɛ, di, dɛ) and (bɚ , bo, dɚ , do)--which conforms to a split based upon a front-back vowel distinction.

Thus, in no condition did subjects group the sounds by categories corresponding to the initial consonant--though information about the nature of the initial consonant sounds was almost certainly available to subjects in the syllable condition. Thus, the present results do not rule out the possibility that there is invariant information concerning the identity of consonants available to the perceiver in the chirp stimuli. Rather, it can only be said that if such information is available, it does not form the most salient basis for classifying the stimuli into perceptual categories.

The salience of the vowels in the syllable condition is perhaps not so surprising. After all, the vocalic portion of the syllables was much longer in duration than the consonantal portion. Interestingly, there is some indication that the full-formant chirps also preserve vowel information. One of the subjects in the full-formant condition employed a vowel identity rule in grouping the sounds. This finding is consistent with those of Blumstein and Stevens (1980) who employed comparable stimuli and found that subjects were quite successful in

labeling vowels. The picture for the two-formant patterns is less clear. No subject in this condition divided the stimuli up into groups based upon vowel identity, although several subjects appeared to be following some sort of strategy based on vowel similarity. Thus it is possible that both the full syllables and the brief beginning segments of those syllables contain both consonant and vowel information. This possibility was pursued in the second experiment by requiring subjects to sort the sounds into only two categories.

Experiment II

In the previous study, no subject sorted the stimuli into groups based on the identity of the initial consonant sounds, even in the syllable condition. Instead, the most prevalent grouping chosen was a division according to vowel identity. One factor which may have contributed to the choice of this classification scheme was the fact that subjects were not restricted as to the number of categories they could use in sorting the stimuli. The lack of such restriction plus the apparent salience of the vocalic portions of the stimuli, may have encouraged subjects to group the stimuli into four equal-sized categories based on vowel identity.

As noted earlier, a given set of stimuli can be divided up in a number of different ways. Just which way a subject chooses depends on a number of things including the number of categories to be employed (Garner, 1974). The most obvious common factor for an unconstrained classification may not be the most obvious one for a division into <u>two</u> categories. Given the present set of stimuli, one might hypothesize that a classification according to the identity of the initial consonant would be enhanced if subjects were constrained to divide the stimuli up into only two classes. Hence, in the present experiment, subjects

were once again required to sort syllables, full-formant chirps or two-formant chirps into groups, but this time, they were instructed to use only two groups.

Methods

Stimuli

The stimulus materials were identical to those used in Experiment I.

Subjects

The subjects were thirty undergraduates at Dalhousie University who received course credit for participating in the experiment. All were native speakers of English and reported no history of either speech or hearing disorder.

Procedure

The details of the procedure were virtually identical to that of Experiment I. The only change which was introduced was that subjects were instructed that they should divide the stimuli into two classes. Otherwise, there were no other restrictions on the kinds of classes which subjects could form (e.g., the classes did not necessarily have to contain an equal number of members).

Results

Once again, only the data for the second 80-item test block were analyzed. Consistency scores for each subject were computed using the Rell measure described in Experiment I. The average consistency score across all conditions was .79. The scores were then submitted to an ANOVA of a 3 (condition) X 8 (stimulus) mixed design. In contrast to the previous experiment, there was a reliable main effect of stimulus ($F_{(7,189)} = 4.08$, $p < .05$) and a marginal main effect of condition ($F_{(2,27)} = 3.21$). However, the interaction between these two factors ($F_{(14, 189)} < 1.00$) did not attain significance. Post hoc analysis conducted with Newman-Keuls tests ($p < .05$ or better) indicated that subjects were more consistent in classifying the syllables than the full-formant chirps, suggesting that the syllables

were easier to remember. In addition, the consistency scores attained for [bi] stimuli (.91 overall) were reliably better than those for either [bɛ] or [dɚ] (.68).[4]

For purposes of describing the classifications used by individual subjects, each subject's data were scored according to the proportion of trials which followed a particular sorting rule. In order to be classified as having used a particular sorting rule, a subject had to correctly sort the stimuli according to the rule on at least 80% of the trials. By this criterion, there were two prevalent patterns observed for the full-syllables. Three of the subjects divided the syllables into two groups based on the identity of the initial consonant sounds--i.e., ([bi], [bɛ], [bɚ], [bo], vs. ([di], [dɛ], [dɚ], [do]), while five chose groups based on vowel similarity--i.e., ([bi], [bɛ], [dɪ], [dɛ] vs. ([do], [bɚ], [do], [dɚ]). The grouping patterns followed by the remaining two subjects appeared to be largely idiosyncratic. With respect to the full-formant chirps, one subject sorted according to consonant identity and two subjects by vowel similarity. The sorting pattern of two of the remaining seven subjects was notable in that they divided the stimuli as follows: ([bi], [di]) vs. (all others). This classification scheme pits the two stimuli having the highest second and third formant offset values against all the other stimuli. The remaining five subjects in the full-formant chirp condition employed idiosyncratic sorting patterns. In contrast to the other two conditions, the data for the two formant chirp condition are easily described since all ten subjects sorted according to the vowel similarity rule (i.e. [bi], [bɛ], [di], [dɛ] vs. [bo], [bɚ], [do], [dɚ]).

The predominant sorting patterns--vowel similarity and consonant identity--which emerged in the examination of the individual subject data were confirmed

in an analysis of the group results. Figure 4 displays the mean proportion of trials correctly sorted by the vowel and consonant rules for each of the three conditions. It is readily apparent that across all three conditions, the preferred grouping was a division according to vowel similarity. In each condition sorting according to a vowel similarity rule was reliably better than chance ($p < .05$ or better by a t-test). However, the degree to which the vowel

---

Insert Figure 4 about here

---

similarity rule was applied differed across conditions as the results of a one-way ANOVA revealed. There was a significant main effect of Condition ($F_{(2,27)} = 6.02$, $p < .05$) which a Newman-Keuls test indicated could be traced to greater use of this rule in the two-formant condition than in either of the other two conditions.

The consonant identity rule was used much less widely by subjects. In fact, only in the syllable condition did its usage reliably exceed that expected by chance.

Discussion

There are several points to be made about the results of this experiment which required subjects to sort the stimuli into two categories. First, there is some evidence that the restriction to sort the stimuli into two categories increased the likelihood that subjects would form perceptual categories based on the identity of the initial consonants. More importantly, one of the four subjects who employed the consonant identity rule did so with the full-formant chirp stimuli. This result alone suggests that there is usable information concerning the identity of the initial consonant present in the full-formant chirps. By contrast, there was no evidence that any of the subjects in the two-formant chirp condition utilized the consonant identity rule.

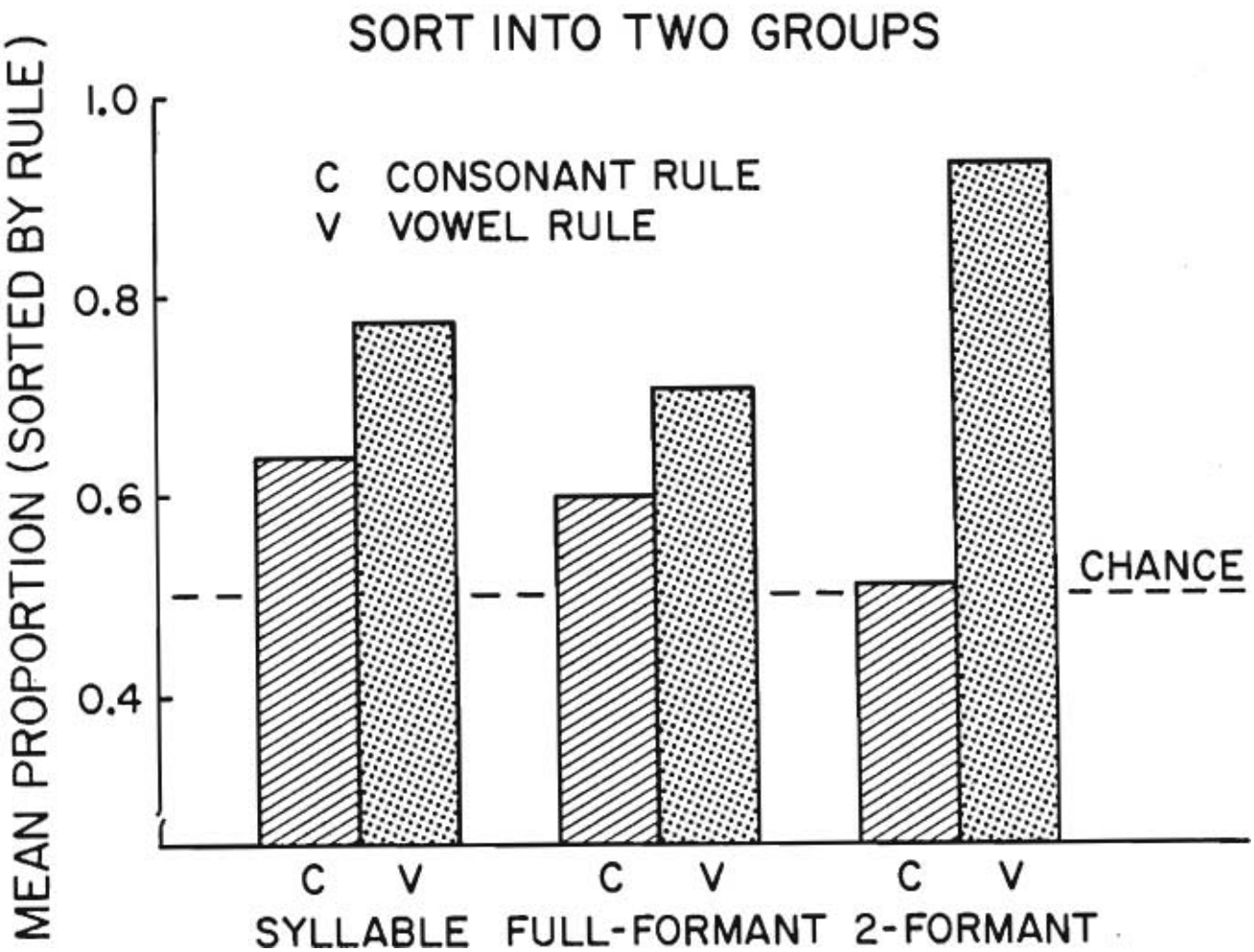


Figure 4 - The mean proportion of trials correctly sorted by the consonant-identity and vowel-similarity rules for each of the three types of stimuli when subjects were instructed to sort the stimuli into two groups.

Despite the fact that some subjects grouped the stimuli on the basis of consonant identity, the majority of subjects appeared to follow a different sorting rule--based on vowel similarity. This grouping scheme which was manifest in the behavior of a few subjects in Experiment I was the only rule employed by subjects in the two-formant chirp condition. Moreover, half of the subjects in the syllable condition also appeared to follow this strategy.

What underlies this sorting by vowel similarity in the chirp conditions when only the formant transitions of the syllables were present? One clue to the basis for these groupings comes from previous investigations in which similarity judgments have been collected for brief, rapidly changing acoustic segments. Brady, House, and Stevens (1961) required subjects to match the frequency of a steady tone to that of a short stimulus which underwent a rapid shift in frequency. They observed that there was a strong tendency for subjects to choose a frequency value which nearly matched the offset value of the rapidly changing stimuli. Similarly, in an experiment involving formant transition patterns not unlike the two-formant chirps in the present study, Shattuck and Klatt (1976) found that subjects favored matches between stimuli which shared offset values in the region of the second formant. Consequently, the results of these two studies suggest that the most likely grouping strategy for subjects in the present study would be to put together those stimuli most similar in their offset values. A classification according to vowel similarity would be expected in the present experiment if the categories were based upon the offset values of the formants, especially those of the second formant.

By manipulating the instruction set the present experiment succeeded in getting at least one subject in the full-formant chirp condition to detect a commonality among stimuli which share the same initial consonant. Still, the detection of this commonality was apparently overshadowed by a more powerful and salient property

which linked the stimuli in terms of similar offset frequency values. Moreover, even subjects who were asked to classify full syllables often followed a classification rule (vowel similarity) which linked the stimuli in terms of their offset frequency values. In this second experiment, then, classifications of the syllables and the chirps do correspond well. However, the correspondence does not appear to be based on invariant consonant information.

Experiment III

The previous experiment suggests that there may be invariant cues for the identity of initial consonants available to the perceiver in the full-formant chirps. Yet, the use of these cues was not widespread even among subjects who classified syllables. The chief factor here seems to be the greater salience of the offset values of the formants as a standard for classifying the stimuli into natural categories. One possible way of countering the predominance of the offset value cue would be to direct subjects to sort the stimuli according to their "initial sounds." In this way, subjects might be more apt to employ information concerning the identity of the initial consonants in their groupings to the extent that that information is available in the syllables and chirps.

Method

Stimuli

The stimulus materials were identical in all respects to those used in the previous two experiments.

Subjects

The subjects were fifteen undergraduates at Dalhousie University who received course credit for participating in the experiment. All subjects were native speakers of English and reported no history of either speech or hearing disorder.

Procedure

The procedure was virtually identical to that of Experiment II with the exception that subjects were instructed to sort the stimuli into two groups on the basis of their initial sounds.

Results

Only the data from the second 80-item test block were analyzed. Again, consistency scores were computed for each subject using the RelH measure. An ANOVA of a 3 (condition) X 8 (stimuli) mixed design was used to analyze the consistency data. Neither of the main effects (condition: $F_{(2,12)} = 2.08$; Stimulus: $F_{(7,84)} = 1.59$) was reliable, nor was the interaction between these factors ($F_{(14,18)} < 1.00$). The average consistency score across all conditions was .82.

An examination of the classifications imposed by individual subjects was conducted using the criterion of 80% correct sorts by a particular rule, as in Experiment II. In the syllable condition, all five subjects grouped the stimuli by the consonant identity rule. In the full-formant condition, one subject grouped the stimuli by consonant identity, two by vowel similarity, and two by a grouping which pitted the [bi] and [di] stimuli against all others. Thus, although there was further confirmation that the full-formant chirps could be partitioned according to consonant identity, the change in instruction set did not appear to greatly lessen the tendency of most subjects to group according to the offset values of the stimuli. Finally, in the two-formant condition four of the five subjects grouped according to the vowel similarity rule.[5] Once again, no subject in this condition gave evidence of using the consonant identity rule.

Figure 5 displays the mean proportion of trials correctly sorted by the vowel and consonant rules in each condition. The most notable change between the patterns

displayed here and those of the previous experiment occurs in the syllable condition. Usage of the vowel similarity rule dropped to chance level, whereas usage

---

Insert Figure 5 about here

---

of the consonant identity rule increased to the point where 99% of the sorts accorded with this rule. Yet, although the instructions set obviously altered the predominant mode of classifying in the syllable condition, it did not appear to affect either of the two chirp conditions. In both of these conditions sorting according to a vowel-similarity rule was reliably better than chance ($p < .05$ or better by a t-test), whereas sorting by the consonant-identity rule was not.

Further tests of the degree to which each sorting rule was used by subjects in each condition were conducted using one-way ANOVAs. For the consonant identity rule there was a reliable main effect of Condition ($F_{(2,12)} = 45.06$, $p < .001$) which Newman-Keuls tests indicated could be traced to decreased usage of the rule from syllables to full-formant chirps to two-formant chirps. A similar analysis conducted on the vowel-similarity rule also found a main effect of Condition ($F_{(2,12)} = 8.52$) attributable solely to the infrequent usage of this rule in the syllable condition.

Discussion

The change in instruction set appears to have had its greatest impact on the performance of subjects in the syllable condition. All five subjects in this condition formed classes consistent with a division by initial consonant sound. By contrast, there was no observable change in the sorting patterns chosen by subjects in either of the two chirp conditions. Once again, similarities in the offset values of the stimuli were the dominant factor in the organization of stimulus groups, although as in the previous experiment, one subject in the full-formant

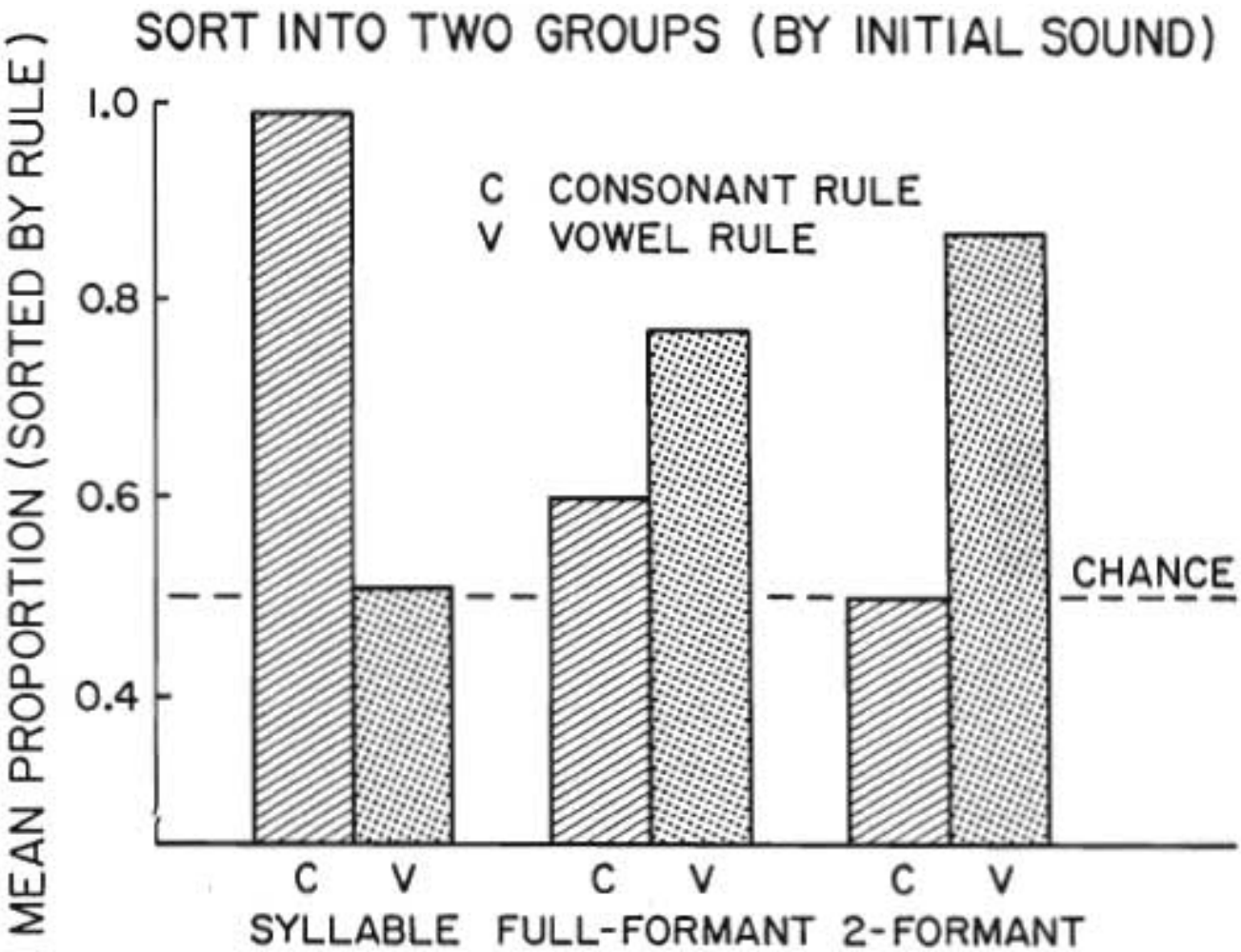


Figure 5 - The mean proportion of trials correctly sorted by the consonant-identity and vowel-similarity rules for each of the three types of stimuli when subjects were instructed to sort the stimuli into two groups based on the identity of the initial sounds.

condition did follow a consonant identity pattern. Thus, the change in instruction set was not sufficient to increase the use of the consonant identity strategy in either of the chirp conditions.

To this point, then, there is little evidence that information is available in full-formant chirps that allows subjects to sort these stimuli into groups corresponding to the identity of initial consonants and no indication that any such information is available in the two-formant chirps.

Experiment IV

Although there is not a strong tendency for subjects to spontaneously group the chirps according to consonant identity, they still might be able to abstract the consonant information from these stimuli if required to do so. One way of requiring subjects to use information about consonant identity is to employ a perceptual learning task in which subjects are instructed to categorize sounds according to this specific rule. Grunke and Pisoni (1979) have shown that such a task may provide a more sensitive measure of the perceptual relatedness of auditory patterns than a similarity judgment task. To the extent that there is information available to the perceiver for a partitioning of the stimuli along the lines of consonant identity, then subjects should be able to learn to sort by this rule. More importantly, if the information in the chirps concerning consonant identity has any psychologically real status for the perceiver, it should be easier to learn to sort by this rule than by one which imposes an arbitrary organization on the stimuli.

The present experiment was undertaken in order to determine the degree to which subjects might be able to learn to classify the chirps according to a consonant identity rule. There were two components to this experiment. The first of these

was a learning task in which each subject was required to classify a given set of sounds according to three different rules: consonant-identity, vowel-similarity, and arbitrary.[6] The second phase of the present experiment was a speeded classification task in which subjects had to assign the stimuli to groups according to one of the three rules as rapidly as possible. Differences in the speed with which subjects were able to apply either the consonant-identity or the vowel-similarity rules relative to the arbitrary rule would provide a further index of the psychological status of these rules.

Methods

Stimuli

The stimuli themselves were identical to those used in the previous experiments. However, new types of demonstration and training tapes were employed. The demonstration tapes were constructed so as to correspond to one of the three rules. For example, the demonstration tape for the consonant-identity grouping of the syllables consisted of all the members of one category (e.g., [bi], [bɛ], [bo], [bɚ] presented consecutively at one-second intervals. After a short pause, the members of the other category were presented in the same manner. Following another pause, the category one items recurred but in a different order, followed again by the category two items also in a different order. Demonstration tapes for each stimulus type (syllable, full-formant chirp, two-formant chirp) by grouping rule (consonant-identity, vowel-similarity, arbitrary) combination were prepared in a similar fashion.

The training tapes employed in this experiment consisted of randomized sequences of the eight stimuli in a given stimulus condition. There were a total of three occurrences of each syllable for a total of 24 trials per tape spaced at 4-second

intervals. Three separate training tapes with different random orders were prepared for each stimulus condition.

Subjects

The subjects were eighteen undergraduates at Dalhousie University who received course credit for their participation in the experiment. All were native speakers of English and reported no history of either speech or hearing disorders.

Procedure

Each subject was assigned randomly to one of the three stimulus conditions (i.e., full-CV syllables, full-formant chirps, or two-formant chirps). Within a given stimulus condition, subjects were trained to sort the stimuli according to each of three sorting rules. These rules are diplayed in Table 2. Thus, the consonant identity rule required subjects to group together stimuli which shared the same initial consonant (i.e. [b] vs. [d]). The vowel similarity rule required subjects to group stimuli from syllables containing the [i] and [ɛ] vowels into one category and the [o] and [ɔ] vowels into another. Finally, the arbitrary rule grouped together stimuli which shared neither a consistent consonant identity nor vowel similarity. Training took place separately for each rule and proceeded in the following manner. A subject was informed that he or she would be hearing

---

Insert Table 2 about here

---

eight different sounds and that four of these were to be assigned to category 1 and four to category 2. The experimenter then played the four stimuli corresponding to category 1, stopped the recorder, and informed the subject that the next four items were members of category 2. After the category 2 members had been presented, the tape was stopped again, at which point the experimenter repeated the instructions

TABLE 2

THREE SORTING RULES FOR LEARNING AND SPEEDED CLASSIFICATION TASKS

CONSONANT		VOWEL SIMILARITY		ARBITRARY	
[bi]	[di]	[bi]	[bo]	[bi]	[di]
[bɛ]	[dɛ]	[di]	[do]	[bo]	[do]
VS		VS		VS	
[bo]	[do]	[bɛ]	[bɚ]	[dɛ]	[bɛ]
[bɚ]	[dɚ]	[dɛ]	[dɚ]	[dɚ]	[bɚ]

and demonstrated the categories for the subject a second time. The subject was then informed that he or she would receive a set of training trials during which each stimuli was to be assigned to either category 1 or 2, and that feedback as to the correctness of a response would be provided on each trial. The assignment of a given stimulus to a particular category was to be made by pressing buttons labeled 1 and 2 located on a box situated next to the subject's right hand. The subject was instructed to place the right index finger on one button and the right middle finger on the other. For a given subject, the location of the category 1 and category 2 buttons remained constant throughout the experiment. However, the ordering of the buttons was counterbalanced across subjects.

A subject was assumed to have successfully learned a given rule if he or she answered correctly on 20 of the 24 training trials in a block. If at the end of the first block of training trials the subject had not met the learning criterion, the demonstration tapes were played once again for the subject and a second block of training trials was run. Testing continued in this manner until either a subject successfully learned the rule or a total of four training blocks were completed without the subject having learned the rule.[7] In the latter circumstance, testing on the rule was terminated and the subject began training on any other rule which remained to be learned.

If a subject had successfully learned a particular rule, then he or she was immediately tested on the corresponding speeded classification task prior to learning another classification rule. The instructions for the speeded classification task were essentially the same as for the learning task with the exception that subjects were told to respond as rapidly as possible without making errors and that no feedback would be provided regarding the correctness of responses. An 80-item test block (10 occurrences of each of the eight stimuli) consisting of a

random arrangement of the stimuli spaced at four-second intervals was then presented to the subject. On each trial, a digital timer was activated by the first voicing pulse from the stimulus. The timer was halted as soon as the subject pressed one of the two response keys or if the subject failed to respond within three seconds (at which point a reaction time of 3000 msec was recorded). On each trial, the experimenter recorded both the subject's response and the reaction time. Two small lights indicated to the experimenter which response key had been pressed on a given trial. Both the response lights and timer, though visible to the experimenter, were out of view for the subject. Upon completion of the 80th trial, a five-minute break period ensued, after which training began on a new classification rule. Training and testing proceeded as before until a subject had completed training on all three rules. The order of learning the three rules was counterbalanced within each stimulus condition. The entire experimental session took approximately one hour to complete.

Results

The learning data were examined in two ways. First, the number of subjects who successfully completed training on a given rule in a particular condition was calculated. Whereas in the syllable condition, all six subjects passed the training phase for each rule, the situation was considerably different for the two-chirp conditions. In the full-formant condition, all subjects passed the training phase for both the vowel-similarity and consonant-identity rules, but only one subject successfully completed training with the arbitrary rule. In the two-formant condition, all subjects successfully mastered the vowel similarity rule, but not a single subject passed the training criterion with either the consonant-identity or arbitrary rules. These results are quite informative because they

indicate that though subjects can learn to classify the full-formant chirps on the basis of consonant-identity, they are unable to classify two-formant chirps on this basis at all.

A more extensive analysis of the learning data is possible when the number of errors made during the training tasks are considered. The mean number of errors made in learning each rule is displayed for the three stimulus conditions in Figure 6. The error data for individual subjects were submitted to an ANOVA of a 3 (Condition) X 3 (Classification Rule) mixed design. Both main effects (Condition ($F_{(2,15)} = 112.98$, $p < .01$); Classification ($F_{(2,30)} = 73.99$, $p < .01$); and the interaction ($F_{(4,30)} = 21.40$, $p < .01$) were significant. Newman-Keuls

---

Insert Figure 6 about here

---

tests with a criterion level of $p < .05$ were used to conduct a post hoc analysis of the results. Recall that for each stimulus condition the critical comparison in each case is how performance with the consonant-identity or vowel-similarity rule differs from that with the arbitrary rule. In both the syllable and full-formant conditions, significantly more errors were made in learning the arbitrary rule than in learning either the consonant-identity or vowel-similarity rules which did not differ from each other. Hence, both rules appear to have a psychologically real status for subjects in the syllable and full-formant conditions. This was not the case for the two-formant condition. Only errors in learning the vowel-similarity rule were reliably less than errors in learning the arbitrary rule. Thus, it appears that if there are any invariant cues to consonant identity in the two-formant chirps, they may be inaccessible to the perceiver. That is, the psychological status of any potential invariants in the two-formant chirps is questionable.

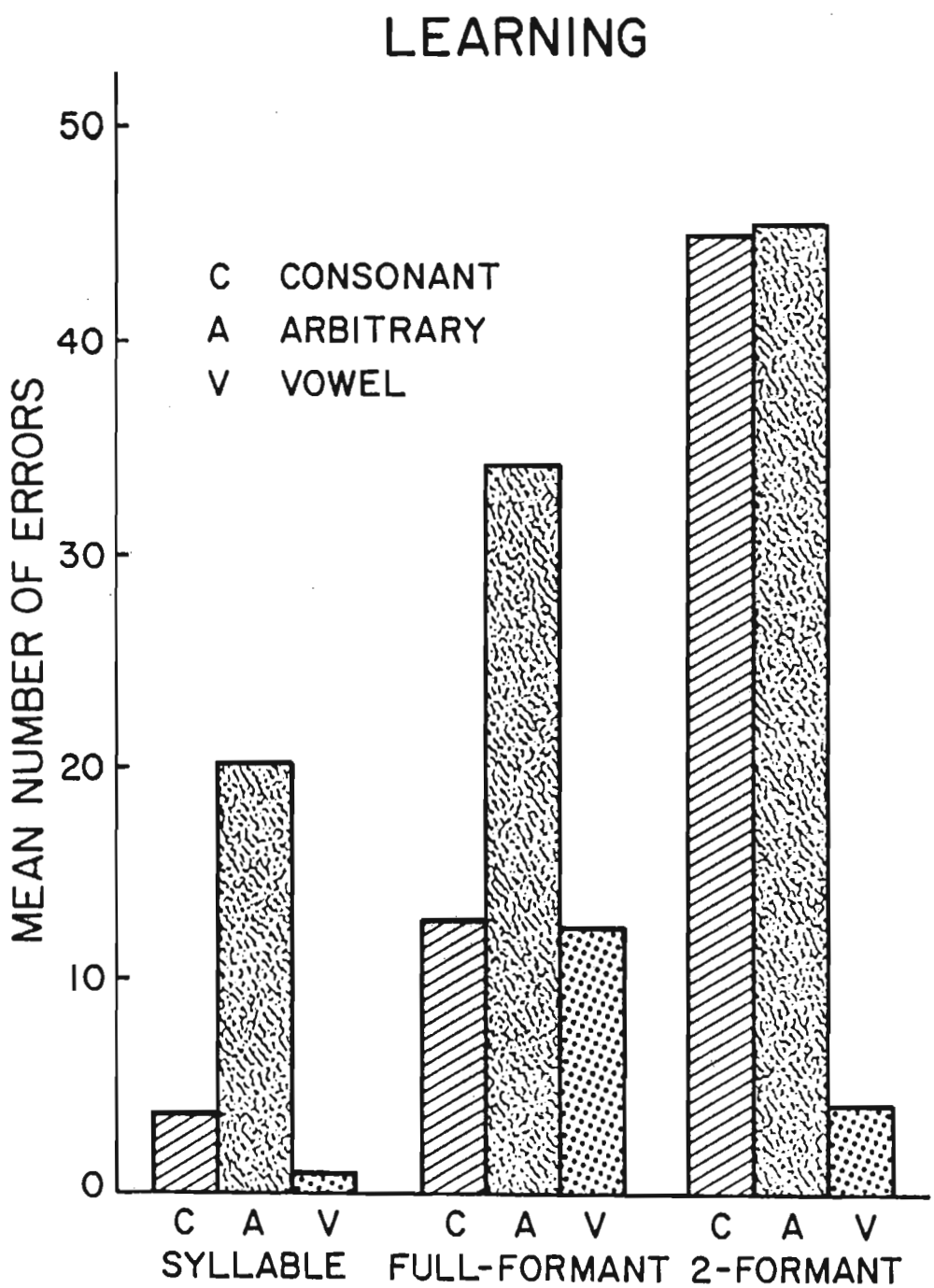


Figure 6 - Mean number of errors for each type of stimulus when subjects were forced to sort the stimuli according to the consonant-identity, arbitrary and vowel-similarity rules.

An indication of the ease with which a given rule was learned across the three stimulus conditions was obtained from additional post hoc tests on the error data. With the consonant-identity rule there were no reliable differences between the syllable and full-formant conditions, though both were obviously superior to the two-formant condition. With the vowel-similarity rule, there were no reliable differences between any of the conditions. Finally, with the arbitrary rule, there were considerably fewer errors in the syllable condition than in either of the other two.

The speeded classification data, displayed in Table 3, exhibit much the same pattern as the learning data. The reaction time data reinforce the conclusion that there is psychologically usable information regarding consonant identity available in both the syllables and full-formant chirps. A one-way ANOVA conducted on the syllable condition indicated a reliable main effect ($F_{(2,12)} = 9.82$, $p < .01$) traceable to the fact that reaction times were significantly slower with the arbitrary rule. The pattern of no reliable differences in reaction time for the consonant-identity and vowel-similarity rules present in the syllable condition, also held for the full-formant condition ($t_{(5)} < 1.00$). Finally, a comparison of the reaction time data across stimulus conditions was not conducted because the different lengths of the syllables and chirps rendered such a comparison uninterpretable.

Insert Table 3 about here

Discussion

The present results provide very clear evidence that the full-formant chirps contain sufficient information, in a form accessible to the perceiver, to allow

TABLE 3

SPEEDED CLASSIFICATION EXPERIMENT

(MEDIAN REACTION TIMES IN MILLISECONDS)

	RULE USED TO SORT STIMULI		
	VOWEL (n=6)	CONSONANT (n=6)	ARBITRARY (n=5)
SYLLABLES	787	857	1057
FULL-FORMANT PATTERNS	905	843	–
2-FORMANT PATTERNS	681	–	–

them to be grouped according to consonant-identity. That this information has a psychologically real status for the perceiver is shown by the fact that a grouping by consonant-identity was much more readily learned than one by an arbitrary classification of the stimuli. This performance difference between the consonant-identity and arbitrary rules is crucial since it serves to establish that the use of the consonant identity rule involved more than mere rote memorization of the stimulus sets. At the same time, there is every indication that the two-formant chirps do not contain sufficient acoustic information to allow for sorting, on the basis of consonant-identity.

But the information specifying consonant-identity is not the only psychologically usable information in the stimuli for the perceiver as the results for the vowel-similarity rule demonstrate. Across all stimulus conditions, subjects' mastery of this rule was clearly superior to the arbitrary one. Not only was this rule easier for subjects to learn, but in the one condition in which subjects did learn the arbitrary rule (i.e., syllable condition), speeded sorting was reliably faster with the vowel-similarity rule than with the arbitrary rule. Thus, the present results confirm the observation made on the basis of the earlier experiments that there is information in the stimuli which leads to a psychologically real division along the lines of vowel similarity.

There is one additional point that needs to be made about the present experiment. The perceptual learning task proved to be a more sensitive measure of whether or not subjects could access information which would allow them to sort by a consonant-identity rule. It seems likely that a task of this type is apt to be most useful in those situations in which there is potentially a very powerful perceptual classification scheme which may overwhelm alternative ways of organizing the stimulus set. In such circumstances, a perceptual learning task, wherein performance

under a given classification scheme is compared to that under an arbitrary scheme, may prove to be the most appropriate means of establishing the psychological status of the classification scheme and the stimulus dimensions.

General Discussion

The main objective of the present investigation was to determine whether brief portions of the speech signal present at the onset of CV syllables contain some invariant information which listeners can use to categorize utterances of the same consonant occurring in different vowel contexts. The results, especially those of Experiment IV, demonstrate convincingly that such information is available to the listener in the onsets of brief segments of speech containing five formants. Hence, the present investigation nicely complements those which have employed new analyses of the acoustic signal (e.g., Kewley-Port, 1980; Searle et al., 1979; Stevens & Blumstein, 1978), in that it shows that there is invariant information present in the psychological stimulus as well as the physical stimulus.

Although the present results and those of Blumstein and Stevens (1980) imply that there is information present in the first 30 msec of the syllables that allows them to be grouped according to initial consonant, they do not indicate the physical basis on which subjects so categorize the sounds. As such, the present results cannot be used to distinguish which, if any, of the alternative descriptions of invariant properties in the physical signal best corresponds to the one used by the perceiver. Thus, accounts which derive physical invariants from computations over running spectra (e.g., Kewley-Port, Note 2; Searle et al., 1979) are just as consistent with the present results as those based on averaged onset spectra (e.g., Stevens & Blumstein, 1978; in press) provided that they locate the source of the invariance within the first 30 msec of the signal.

Yet, despite the fact that there is information available in the full-CV syllables and full-formant chirps that allows them to be grouped by consonant-identity, it is clear that such a grouping is not the one most preferred by subjects. Throughout this series of experiments, and with all stimulus sets, subjects appeared to favor classifications which grouped together stimuli having similar offset values. This grouping was designated "vowel-similarity" since it divided the stimuli in a way which corresponded to a front-back vowel distinction.

As noted earlier, the finding that subjects grouped the stimuli in terms of offset values was not without precedent in the literature. The work of Brady et al., (1961) and Shattuck and Klatt (1976) indicated that subjects tend to match brief, rapidly changing acoustic patterns in terms of their offset values. Nor is this tendency limited solely to brief patterns as Grunke and Pisoni (1979) discovered. Using patterns of similar duration to those in the present study, Grunke and Pisoni found that subjects most readily learned to match patterns having similar offset values. Moreover, they also discovered that adopting a procedure in which the differences in the offset values of the stimuli were reduced, enabled subjects to match rather dissimilar rise-fall patterns as readily as mirror-image acoustic patterns. Hence, it is clear that offset frequency information is particularly salient for the listener. In this regard, it is worth noting that in their answers on the questionnaire, subjects whose groupings followed the vowel-similarity pattern reported that pitch was the predominant factor in selecting their groups. Many described their classificarions as "high pitch" versus "low pitch." In light of these descriptions and the previous work with other acoustic patterns, it may be more appropriate to describe these classes as following a pitch-similarity rule, especially since the vowel-similarity rule appears to be a special subset of the former.[8]

Finally, the present results raise serious doubts about whether two-formant chirps lacking first formant cues are suitable for use as nonspeech controls in speech perception experiments. In the past, chirps lacking first formant information have often been used as control stimuli because it was claimed that they present the listener with the same critical information (i.e., acoustic cues) contained in syllables, but in a nonspeech context (e.g., Mattingly et al., 1971). First formant information was routinely deleted from such chirps to make them less speech-like. The justification offered for this procedure was that the first formant information did not help to distinguish the speech contrasts under investigation--e.g., [b] and [d]. However, this argument holds only under the assumption that the information content of each formant is independent of the other formants. If one assumes instead that it is the relationship which exists between the formants which is critical for perception, then the first formant information cannot be considered to be redundant. The present results support the latter view because they show that whereas the two formant chirps do not contain the information necessary to permit a grouping according to consonant-identity, such a grouping is easily achieved with patterns that do preserve first formant information--viz., the full-formant chirps. A similar view regarding the importance of first formant information for the perception of brief auditory patterns has been expressed by Pisoni (Note 3). He observed that the inclusion of first formant information in chirp patterns resulted in sharp discontinuities in discrimination functions in the vicinity of phonetic boundaries. Such discontinuities are not typically present for patterns lacking first formant information (e.g., Mattingly et al., 1971). Hence, it seems clear that acoustic energy in the region of the first formant plays a critical role in the perception of the overall acoustic pattern.

In summary, by employing two new methodologies--free classification and perceptual learning--the present study was able to demonstrate that there is invariant acoustic information available in the onsets of syllables which allows the listener to perceive the same consonant in different vowel contexts. Such findings suggest that previous studies have underestimated the degree of invariance present in the perceptual structure of speech stimuli.

Acknowledgements

This research was supported by an N.S.E.R.C. grant (A-0282) to the first author and an N.S.F. grant (BNS 78-13019) to the second author. Portions of this research were presented earlier in a paper given at the 99th meeting of the Acoustical Society of America, April 22, 1980 in Atlanta, Georgia. The authors would like to acknowledge the support of a number of persons including: Diane Kewley-Port for her help in deriving the onset spectra for the stimuli; Joel Katz and the late Frank Restle for their suggestions about the data analysis, Janice Murray for her help in scoring the results, and especially David Pisoni for his suggestions throughout this research project and for so graciously making available the facilities of the Speech Perception Laboratory at Indiana University. Finally we would like to thank Harris Savin who pointed out to one of us some years ago the fundamental problems which exist in determining what constitutes an appropriate nonspeech control.

Footnotes

[1]These findings were challenged some years later by Cole and Scott (1974) who argued that the burst information produced at the release of stop consonant occlusion serves as an invariant acoustic cue for the recognition of stop consonants in different vowel contexts. Subsequently, a more systematic investigation using a wider variety of vowel contexts and more precise controls of stimulus parameters indicated that the burst information by itself was not a sufficient cue for recognizing stops in all contexts (Dorman, Studdert-Kennedy, & Raphael, 1977). Rather, bursts and formant transitions appear to be functionally equivalent, context-dependent cues which are reciprocally related such that where the perceptual weight of one increases, the weight of the other declines.

[2]While initial burst information is typically included in the onset spectra of stop consonants, Blumstein and Stevens (1980) have noted that the onset spectra for stops generated without bursts do not differ appreciably. The main effect of including the burst is to enhance the global property which cues a particular place of articulation. As a check on the results of the present investigation, we have conducted a further series of experiments using stimuli containing burst information. The results of that study which will be reported in a forthcoming paper (English, Jusczyk, & Smith) are substantially the same as those reported here.

[3]We would like to thank Diane Kewley-Port for deriving the onset spectra for us.

[4]It is difficult to know just what to attribute the difference in consistency scores to. While one possibility is that the [bɛ] and [dɚ] stimuli may have been more ambiguous than the [bi] stimuli, this pattern did not recur in either

Experiment I or Experiment III. Thus, it seems preferable to treat it as just an anomaly in the data.

[5]There is some indication that this subject became confused and reversed the numbers he assigned to each category midway through the test series. In fact, if the category assignments are simply reversed at a point midway through the series, the data follow a vowel similarity pattern. It is also worth noting that this subject did sort the stimuli according to a vowel similarity rule during the first 80-item test block.

[6]The vowel similarity rule was chosen since many subjects seemed to regard it as the most natural way of grouping the stimuli. Thus, it provided a convenient upper bound for measuring the ease with which the consonant-identity rule was learned.

[7]Note that since only three training tapes had been prepared for each stimulus condition, any subject who required four training blocks heard one of the training tapes twice. The choice of which tape was to be repeated was counterbalanced across subjects.

[8]The notion that vowels might be classified with respect to pitch is a very old one in psychology traceable back at least to Kempelen (1791) and also investigated by such as Donders (1857), Helmholtz (1930), Koenig (1870), and Kohler (1910).

Reference Notes

1. Kewley-Port, D. KLTEXC: Executive program to implement the KLATT software synthesizer. Research on Speech Perception, 1978, Progress Report 4, Indiana University.

2. Kewley-Port, D. SPECTRUM: A program for analyzing the spectral properties of speech. Research on Speech Perception, 1979, Progress Report No. 5, Indiana University.

3. Pisoni, D. B. Discrimination of brief frequency glissandos. Research on Speech Perception, 1976, Progress Report No. 3, Indiana University.

References

Attneave, F. Applications of Information Theory to Psychology. New York: Holt, Rinehart & Winston, 1959.

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

Blumstein, S. E. & Stevens, K. N. Perceptual invariance and onset spectra for stop consonants in different vowel environments. Journal of the Acoustical Society of America, 1980, 67, 648-662.

Boring, E. G. The Physical Dimensions of Consciousness. New York: Century, 1933.

Boring, E. G. Auditory theory with special reference to intensity, volume, and localization. American Journal of Psychology, 1926, 37, 157-188.

Brady, P. T., House, A. S., & Stevens, K. N. Perception of sounds characterized by rapidly changing resonant frequency. Journal of the Acoustical Society of America, 1961, 33, 1357-1362.

Cole, R. A. & Scott, B. The phantom in the phoneme: Invariant cues for stop consonants. Perception & Psychophysics, 1974, 15, 101-107.

Donders, F. C. Ueber die Natur der Vocale. Beitrage Zur Nature - und - Heilkundes, 1857, 1, 157-162.

Dorman, M., Studdert-Kennedy, M., & Raphael, L. J. Stop consonant recognition: Release bursts and formant transitions as functionally equivalent context-dependent cues. Perception & Psychophysics, 1977, 22, 109-122.

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.

Garner, W. R. Uncertainty and Structure as Psychological Concepts. New York: Wiley, 1962.

Gibson, J. J. The Senses Considered as Perceptual Systems. Boston: Houghton Mifflin, 1966.

Harris, K. S. Cues for the discrimination of American English fricatives in spoken syllables. Language and Speech, 1958, 1(1), 1-7.

Grunke, M. E., & Pisoni, D. B. Some experiments on perceptual learning of mirror-image acoustic patterns. Paper presented at the Ninth International Congress of Phonetic Sciences, Copenhagen, 1979.

Helmholtz, H. Sensations of tone. Translated by A. J. Ellis 5th (ed.) New York: Longmans, Green, 1930.

Kempelen, W. Le mécanisme de la parole suivi de la description d'une machine parlante, Vienne: Imprimé chez B. Baurr et se trouvé chez J. V. Degen, 1791.

Kewley-Port, D. Representations of spectral change as cues to place of articulation in stop consonants. Unpublished Ph.D. dissertation. City University of New York, 1980.

Klatt, D. H. Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America, 1980, 67, 971-995.

Koenig, R. Sur les notes fixes caractéristiques des diverse voyelles. Comptes Rendus Hebdomadaires des Séances der l'académie des sciences Paris, 1870, 70 931-933.

Kohler, W. Akustiche Untersuchungen II. Zertschrift sür psychologie, 1910, 58, 59-140.

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

Liberman, A. M., DeLattre, P. D., & 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.

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

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

Searle, C. L., Jacobson, J. Z., & Rayment, S. G. Phoneme recognition based on human audition. Journal of the Acoustical Society of America, 1979, 65, 799-809.

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

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

Smith, L. B., & Kemler, D. G. Levels of experienced dimensionality in children and adults. Cognitive Psychology, 1978, 10, 502-532.

Stevens, K. N., & Blumstein, S. E. Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America, 1978, 64, 1358-1368.

Stevens, K. N., & Blumstein, S. E. The search for invariant acoustic correlates of phonetic features. In P. D. Eimas & J. L. Miller (Eds.), Perspectives on the Study of Speech. Hillsdale, N.J.: Lawrence Erlbaum Associates, (in press).

II. SHORT REPORTS AND WORK-IN-PROGRESS

SPEECH PERCEPTION WITHOUT TRADITIONAL SPEECH CUES

Abstract. A complex sinusoidal replica of a naturally produced utterance was judged by adults to be intelligible despite the unnatural speech quality of the signal. The dynamic properties of these highly artificial acoustic signals are apparently sufficient to support phonetic perception in the absence of traditional acoustic cues for phonetic segments.

A person listening to a continuously changing natural speech signal perceives a sequence of linguistic elements. Research has attempted to characterize this perceptual process by analyzing the acoustic properties of speech signals that specify its linguistic content (1). In the present study, however, listeners perceived linguistic significance in acoustic patterns with properties differing quite substantially from those traditionally held to underlie speech perception. And, although listeners accurately reported the linguistic content of these acoustic patterns, the results suggest that the signal was also perceived, simultaneously, to be nonspeech. These novel findings imply that the process of speech perception makes use of acoustic properties that are more abstract than the spectral templates and speech cues typically studied in speech research.

The stimuli used in our study consisted of time-varying sinusoidal patterns that followed the changing formant center-frequencies, or the natural resonances of the supralaryngeal vocal tract, of a naturally produced utterance. The sentence, "Where were you a year ago?" was spoken by an adult male talker, digitized at the rate of 10kHz, and analyzed in sampled data format. Frequency and amplitude values were derived every 15 msec. for the center frequencies of the first three formants by the method of linear prediction (2). These values were hand-smoothed and used as synthesis parameters for a digital sinewave synthesizer. Three time-varying sinusoids were then generated to match the center frequencies and amplitudes of the first three formants, respectively, of the natural speech utterance. Figure 1 shows narrow- and wideband spectrograms of the original spoken utterance and a narrowband spectrogram of its replica formed by the three time-varying sinusoids.

Figure 1 about here

Although our synthetic stimuli were designed to preserve the frequency and amplitude variation of natural speech formants, the three-tone patterns differ from natural speech in several prominent ways. First, the energy spectra of the tones differ greatly from those of natural and synthetic speech. Voiced speech sounds, produced by pulsed laryngeal excitation of the supralaryngeal cavities, exhibit a characteristic spectrum of harmonically related values (3). Because the frequencies of the individual tones in our stimuli follow the formant center frequencies, the components of the spectrum at any moment are not necessarily related as harmonics of a common fundamental. In essence, the three-tone pattern does not consist of harmonic spectra, although natural speech does.

Second, the short-time spectra of the tone stimuli lack the broadband formant structure that is also characteristic of speech. Because the resonant properties of the supralaryngeal vocal tract introduce short-time amplitude maxima and minima across the harmonic spectrum of sound generated at the larynx, some frequency regions contain harmonics with more energy than neighboring regions (4). Because our tone stimuli consist of no more than three sinusoids, there is no energy present in the spectrum except at the particular frequencies of each tone. Thus, the short-time spectra of the tone stimuli are also distinct in this way from the energy spectra of natural speech. There is literally no formant structure to the three-tone complexes, though the tones do exhibit acoustic energy at frequencies identical to the center frequencies of the formants of the original, natural utterance.

Third, the dynamic spectral properties of speech and tone stimuli are quite different. Across phonetic segments the relative energy of each of the harmonics of the speech spectrum changes. Formant center-frequencies may be computed by following the changes in amplitude maxima of the harmonic spectrum. However, natural speech signals do not exhibit continuous formant frequency variation. Rather, laryngeal activity in voiced speech creates distinct pulses characterized

by a formant structure. Thus, changes in formant structure, particularly when observed in wideband spectrograms, may erroneously appear to contain continuous formant variation over time. Figure 1b displays a wideband spectrogram, in which the finegrained amplitude differences are averaged over frequency to derive the formant pattern. In contrast to the case in speech, each tone in our stimuli continuously follows the computed peak of a changing resonance of the natural utterance. Overall, our three-tone pattern deliberately mirrors the abstract spectral changes of the naturally produced utterance, though in local detail it is unlike natural speech signals.

In the absence of fundamental period and formant structure, the complex tone signal therefore consists of none of those distinctive acoustic attributes that are assumed traditionally to underlie speech perception (5). None of the appropriate acoustic cues based on the acoustic events within speech signals is present in our stimuli, e.g., neither formant frequency transitions, which cue manner and place of articulation; nor formant target frequencies and steadystate durations, which cue vowel color and consonant voicing; nor fundamental frequency changes, which cue voicing and stress (6). Similarly, the short-time spectral cues, which depend on precise amplitude and frequency characteristics across the harmonic spectrum, are absent from these tonal stimuli, e.g., the onset spectra that are often claimed to underlie perception of place features (7). The absence of traditional acoustic cues to phonetic identity suggests that our sinusoidal replica of the sentence should be perceived to be three independently changing tones. However, if listeners are able to perceive the tones as speech, then we may conclude that traditional speech cues are themselves approximations of second-order signal properties to which listeners attend when they perceive speech.

Our perceptual test consisted of three conditions in which independent groups of listeners were informed to different degrees about the tonal stimuli that they

would hear. Within each instructional condition, different groups of eighteen listeners each were assigned to seven stimulus conditions: the three tones presented together (S1:T1+T2+T3); three pairwise tone combinations (S2:T1+T2; S3:T2+T3; S4:T1+T3); and each tone played separately (S5:T1; S6:T2; S7:T3). The three instructional conditions crossed with the seven stimulus conditions made twenty-one experimental conditions in all. In each condition a given sinusoidal pattern was presented four times in succession, at 85 dB SPL, by audiotape playback over matched and calibrated headsets.

In Instructional Condition A, listeners were asked simply to report their spontaneous impressions of the stimuli, having been told nothing in advance of the nature of the sounds. Multiple responses were permitted. The accumulated responses, organized by stimulus condition, are displayed in Table 1. Modal responses in each stimulus condition indicate that the majority of listeners did not hear the sinusoids as speech, though a small number of responses in several conditions favored human- or artificial-speech interpretations. Two listeners in the three-tone condition, however, responded that they heard the sentence, "Where were you a year ago?" This outcome might be anticipated only if there were stimulus support of some kind for perceiving the linguistic content of these patterns. Even as a response to a direct request to generate a sentence in English, the probability of producing this exact sentence is exceedingly small (8).

Table 1 about here

In Instructional Condition B, listeners were informed that they would hear a sentence produced by a computer, and were asked to transcribe the synthetic utterance as faithfully as possible. We scored the responses in each condition for correct number of syllables transcribed relative to the original utterance,

"Where were you a year ago?" Average transcription performance in each stimulus condition is presented in Figure 2a. It is clear that a large number of subjects can identify the sentence in Conditions S1 and S2. Nine of the listeners across these two conditions transcribed the entire sentence correctly, though ten others reported that they could hear no sentence at all in the tones. The remaining listeners transcribed various syllables correctly. We conclude from these first two instructional conditions that naive listeners may not automatically perceive sinusoidal replicas of natural speech as linguistic entities. When instructed to do so, however, they perform well presumably because the linguistic information, though not carried by acoustic elements producible by a vocal tract, is preserved in the dynamic structure of the stimulus pattern (9)

Figure 2 about here

In Instructional Condition C, listeners were asked directly to evaluate the speech quality of the stimuli. They were told that they would be presented with the sentence, "Where were you a year ago?" and they were asked to make three judgments. First, they reported whether the sentence was discernible in the tonal pattern by responding Yes or No; they also provided a confidence rating for their judgments using a dual five-point scale, as well. These responses were converted to a ten-point scale (1=confident Yes; 10=confident No) and are presented in Figure 2b grouped by stimulus condition. In five of the stimulus conditions, listeners were very confident that they did not hear the sentence in the tones. However, in Conditions S1 and S2, listeners were very confident that they recognized the intended sentence; the average confidence ratings in these two conditions did not differ significantly despite the absence of Tone 3 in Condition S2.

In the second task, listeners rated the number of words that could be identified in the particular pattern presented (1=all, 2=most, 3=a few, 4=almost none, 5=none). As shown in Figure 2c, for five of the stimulus conditions subjects indicated that they could not identify any of the words in the sentence. But, in the three-tone condition (S1), listeners reported that almost every word was clear. The omission of Tone 3 from the pattern in Condition S2 led subjects to report that significantly fewer words were intelligible, yet this condition remains significantly different from Conditions S3 through S7.

In the third task, listeners rated the voice quality of the sinusoidal stimuli [1=natural, 2=funny (peculiar), 3=unnatural, 4=nonspeech]. The average ratings appear in Figure 2d. The split between S1 and S2 and the other conditions is still quite evident, though these stimulus patterns were considered to have unnatural voice quality despite their clear intelligibility. In essence, listeners in these two conditions apprehended the linguistic significance of the patterns despite the radically unnatural, nonspeech quality of these signals (10). That is, they were able to perceive the sense of the utterance in the absence of acoustic patterns of the kind producible by the vocal tract.

The results of the present study cannot be explained within the framework of existing theories of speech perception (11), for our listeners could not have relied on an inventory of elemental speech cues in perceiving the linguistic message in the tones. Without question, sufficient stimulation must have been available in the tonal patterns to support phonetic perception, though it seems that the perceiver's attention--which can be directed to find the appropriate level of abstraction of the stimulus--ultimately determines whether a "science fiction sound" or a sentence is heard (12). We conclude, then, that speech perception can endure the absence of particular short-time acoustic spectra and traditional formant-based acoustic cues only insofar as certain of the dynamic relations in the natural signal are preserved over the transformation (13).

Further examples of nonspeech tonal analogs of natural speech utterances are needed to characterize these essential relations more precisely.

Robert E. Remez
Department of Psychology
Indiana University
Bloomington 47405

Philip E. Rubin
Haskins Laboratories
270 Crown Street
New Haven, Connecticut 06510

David B. Pisoni
Thomas D. Carrell
Department of Psychology
Indiana University

References and Notes

1. C.G.M. Fant, Logos, 5, 3 (1962); A.M. Liberman, F.S. Cooper, D.P. Shankweiler, and M. Studdert-Kennedy, Psychol. Rev., 74, 421 (1967); I.G. Mattingly, Am. Sci., 60, 327 (1972); K.N. Stevens and S.E. Blumstein, J. Acoust. Soc. Am., 64, 1358 (1978).

2. J.D. Markel and A.H. Gray, Jr., Linear Prediction of Speech (Springer, New York, 1976).

3. T. Chiba and M. Kajiyama, The Vowel: Its Nature and Structure (Tokyo-Kaiseikan, Tokyo, 1941); C.G.M. Fant, The Acoustic Theory of Speech Production (Mouton, The Hague, 1960). The closely spaced horizontal lines shown in Figure 1a are the harmonics of the fundamental frequency of phonation, and are typically revealed in narrowband spectrograms.

4. Typically, the amplitude of the valleys in the spectrum of natural speech ranges from 10-30 dB below the amplitude of the peaks [K.N. Stevens and S.E. Blumstein, in Perspectives in the Study of Speech, P.D Eimas and J.L. Miller, Eds. (Erlbaum, Hillsdale, N.J., in press)].

5. Descriptions of distinctive attributes of speech signals have been influenced significantly by theoretical models of sound production in the vocal tract. These models describe the speech signal as the product of a source and a filter [Chiba and Kajiyama, op. cit.; K.N. Stevens, in Handbook of Physiology-Respiration I, W.O. Fenn and H. Rohn, Eds. (Washington, D.C., American Physiological Society, 1964)]. Glottal pulsing provides a source in which energy is present at integral multiples of the fundamental frequency. The complex resonance of the pharyngeal, oral and nasal cavities is treated as a filter, in which the peaks in the transfer function represent the formants. The perceptual tests of potentially distinctive attributes, however, have typically employed electronic or digital analogs of the source-filter theory of speech acoustics to fabricate stimuli. In doing so, these tests had not questioned the neccessity of harmonic spectra or broadband formant structure in speech perception; neither had they raised the possibility empirically that listeners attend to higher-order relational properties of time-varying speech signals.

6. A.M. Liberman and M. Studdert-Kennedy, in Handbook of Sensory Physiology, Vol. VIII, "Perception" R. Held, H. Leibowitz and H.-L. Teuber, Eds. (Springer, New York, 1978).

7. K.N. Stevens and S.E. Blumstein, in Perspectives in the Study of Speech.

8. G.A. Miller and N. Chomsky, in Handbook of Mathematical Psychology, Vol. II, R.D. Luce, R.R. Bush, and E. Galanter, Eds. (Holt, New York, 1960).

9. It has often been emphasized that a variety of acoustic events may cue a single phonetic feature in the absence of other, redundant cues; experiments with synthetic speech in which phonetic distinctions were minimally cued indicate that listeners tolerate schematized speech signals with little loss of intelligibility [A.M. Liberman and F.S. Cooper, in Papers in Linguistics and Phonetics to the Memory of Pierre Delattre, A. Valdman, Ed. (Mouton, The Hague, 1972)]. For this reason, listeners probably do not require stimuli to display

the acoustic "stigmata" of speech to be candidates for phonetic interpretation [A.M. Liberman, I.G. Mattingly, and M.T. Turvey, in Coding Processes in Human Memory, A.W. Melton and E. Martin, Eds. (V.H. Winston, Washington, D.C., 1972). However, even schematized synthetic speech has consisted of acoustic cues which are utterable in principle as components of a speech signal; these cues enjoy specific articulatory rationales. This resemblence of schematized synthetic speech to natural speech may have led theorists to underestimate the abstractness of the stimulus properties relevant to perception. As acoustic signals, time-varying sinusoids are neither components of speech signals nor may their function as phonetic information receive a literal articulatory interpretation. For these reasons, the sinusoids may be said to specify phonetic identity abstractly.

10. Although much intelligible synthetic speech would also be judged unnatural, this may be ascribed to the practice of presenting the speech cues in contexts of minimal variation in the acoustic parameters indifferent to intelligibility--which affect speech quality nonetheless (Liberman and Cooper, op. cit.). A synthesizer which produces a harmonic spectrum, broadband formants and a fundamental period within the normal range will sound unnatural, and perhaps be unintelligible, despite the acoustic resemblence to natural speech if the synthesis of prosodic variation--of speech rhythm, meter, and melody--is defective [J. Allen, Proc. IEEE, 64, 433 (1976)]. The judgment that this kind of synthetic imitation of speech signals is unnatural is, therefore, quite different from the judgment of unnaturalness in the present case.

11. The proposal that listeners "track" formant frequency variations must be entertained as an explanation of our findings only if the meaning of the term "formant" is extended to mean "a peak in the spectrum." In its present sense the concept of the formant refers to a natural resonance of the vocal cavities [Hermann, Arch. ges. Physiol., 58, 264 (1894)]. Quite literally, then, there

are no vocal resonances in our tone complexes (though listeners who succeed in extracting the utterance probably do so because the tones preserve dynamic properties of vocally produced signals). Our preference is to retain the literal meaning of formant, and to conclude, therefore, that a difference between voiced speech signals and the tone signals is that one contains broadband formants and harmonic spectra, and the other merely inharmonic peaks with infinitely narrow bandwidths.

12. Our finding is related, in some sense, to early studies of "vowel pitch" in which simple steadystate tones were judged to possess "vocality," or speechlike qualities [W. Kohler, Z. Psychol., 58, 59 (1910); J.D. Modell and G.J. Rich, Am. J. Psychol., 26, 453 (1915); E.B. Titchener, described in E.G. Boring, Sensation and Perception in the History of Experimental Psychology (Holt, New York, 1942), p. 374]. More recent studies have shown that listeners may identify brief complex sinusoidal patterns as isolated syllables, and therefore as speech sounds, when they are supplied with restricted response alternatives in low uncertainty judgment tasks [P.J. Bailey, A.Q. Summerfield and M. Dorman, Haskins Laboratories Status Report on Speech Research, SR-51,52, 1 (1977); C.T. Best, B. Morrongiello and R. Robson, J. Acoust. Soc. Am., 66, S50 (1979); M.E. Grunke and D.B. Pisoni, Proc. Ninth Int. Congr. Phonet Sci., Vol. II, 461 (1979)]. The present study, however, makes use of neither a closed response set nor a low uncertainty task to obtain the effect of intelligibility.

13. We have recently synthesized the sentence, "A yellow lion roared," thereby extending the range of tone synthesis to nasal manner as well as the stop consonant, liquid consonant, and vowel phone classes represented here. Similar findings have been obtained with this sentence, indicating that the present results are not due to peculiarities specific to the sentence used in these tests.

14. We gratefully acknowledge the assistance of Charles Marshall, Joe Montelongo and Sue Gans. We also thank Dick Aslin, Alvin Liberman and Frank Restle, among many others, for very helpful advice and comments on an earlier version of the paper. This research was supported by NIMH Grant MH 32848 (to R.E.R), NIMH Grant MH 24027 (to D.B.P.) and NICHHD Grant HD-01994 to Haskins Laboratories.

RESPONSE CATEGORIES AND FREQUENCIES BY STIMULUS CONDITION

FOR GROUP A

STIMULUS CONDITION	RESPONSE CATEGORIES
S1 (T1+T2+T3)	Science Fiction Sounds (8), Computer bleeps (5), Music (4), Several simultaneous sounds (3), Human speech (3), Where were you a year ago (2), Radio interference (2), Human vocalizations (1), Artificial speech (1), Bird sounds (1), Reversed speech (1)
S2 (T1+T2)	Science fiction sounds (7), Computer bleeps (3), Sirens (2), Music (2), Radio interference (2), Tape recorder problems (1), Reversed speech (1), Whistles (1), Artificial speech (1), Human speech (1)
S3 (T2+T3)	Science fiction sounds (14), Radio interference (3), Music (2), Computer bleeps (2), Whistles (1), Several simultaneous sounds (1)
S4 (T1+T3)	Science fiction sounds (9), Artificial speech (5), Computer bleeps (4), Several simultaneous sounds(4), Whistles (3), Radio interference (2), Tape recorder problems (2), Human speech (1), Human vocalizations (1), Reversed speech (1), Music (1)
S5 (T1)	Science fiction sounds (5), Music (4), Reversed speech (4), Tape recorder problems (3), Human speech (2), Artificial speech (2), Animal cries (2), Bird sounds (2), Radio interference (2), Several simultaneous sounds (2), Human vocalizations (1)
S6 (T2)	Sirens (7), Bird sounds (4), Mechanical sound effects (4), Radio interference (4), Animal cries (3), Whistles (2), Computer bleeps (1)
S7 (T3)	Bird sounds (17), Whistles (6), Mechanical sounds effects (5), Human vocalizations (3), Human speech (1), Artificial speech (1), Computer bleeps (1), Animal cries (1), Music (1), Radio interference (1), Tape recorder problems (1)

FIGURE CAPTIONS

Figure 1. (a)Narrowband spectrogram of the natural utterance, "Where were you a year ago?" showing harmonic structure as narrow horizontal lines along the frequency scale. (b)Wideband spectrogram of the same utterance, showing formant pattern as dark bands along the time axis. Note that the vertical striations correspond to individual laryngeal pulses. (c)Narrowband spectrogram of the three-tone sinusoidal replica. The energy concentrations follow the time-varying pattern of the formants above, but there is no energy present except at the formant center frequencies.

Figure 2. (a)Transcription performance for Instructional Condition B. (b)Detection ratings for Instructional Condition C (1=Confident Yes, 10=Confident No); (c) Ratings of number of intelligible words in the tones (1=every, 2=most, 3=a few, 4=almost none, 5=none); (d)Naturalness ratings (1=natural, 2=peculiar, 3=unnatural, 4=nonspeech).

TABLE CAPTION

Table 1. Response categories and frequencies by stimulus conditions in Instructional Condition A.

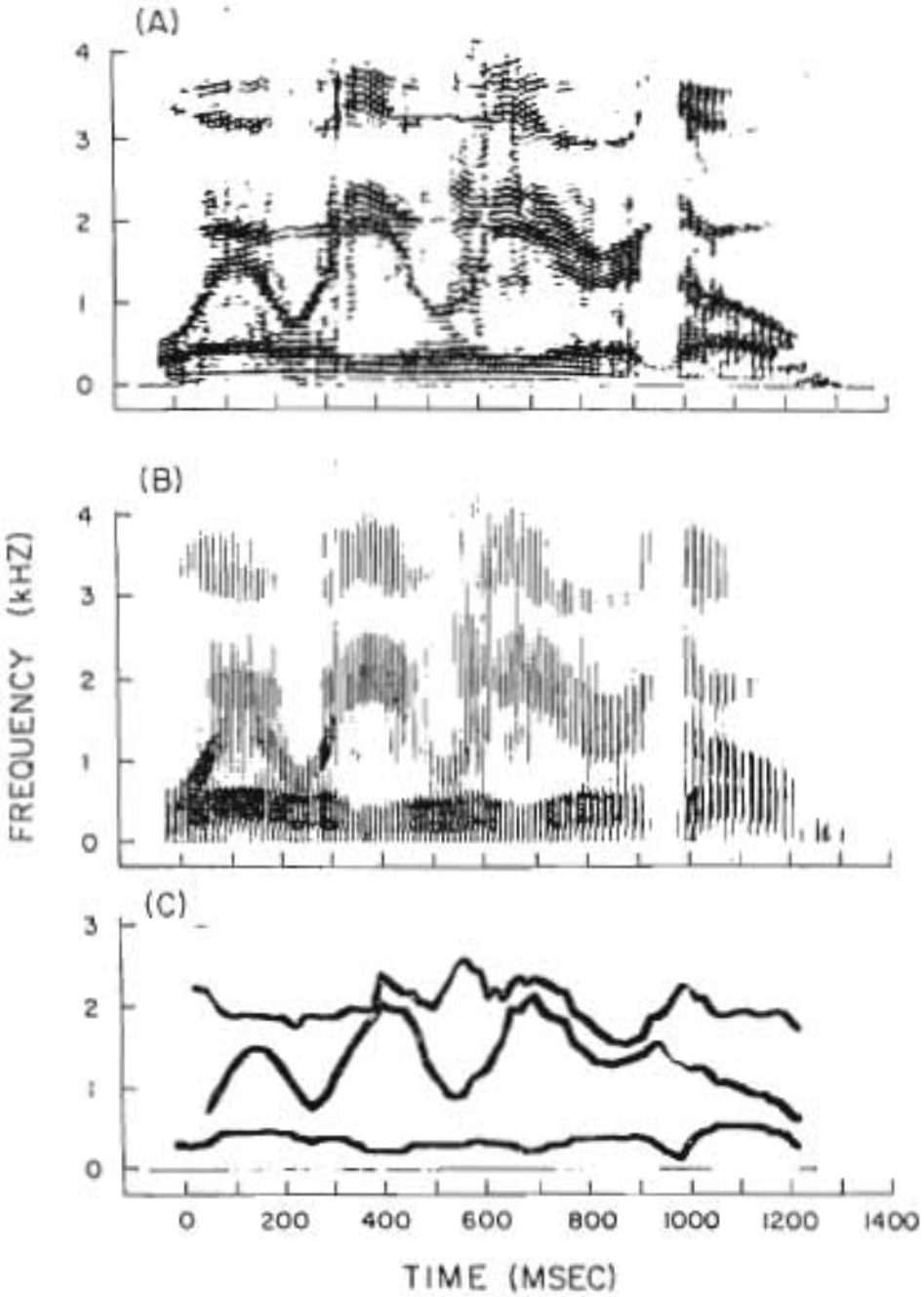

(A)
(B)
(C)
FREQUENCY (kHZ)
0
1
2
3
4
0
200
400
600
800
1000
1200
1400
TIME (MSEC)

FIGURE 1.

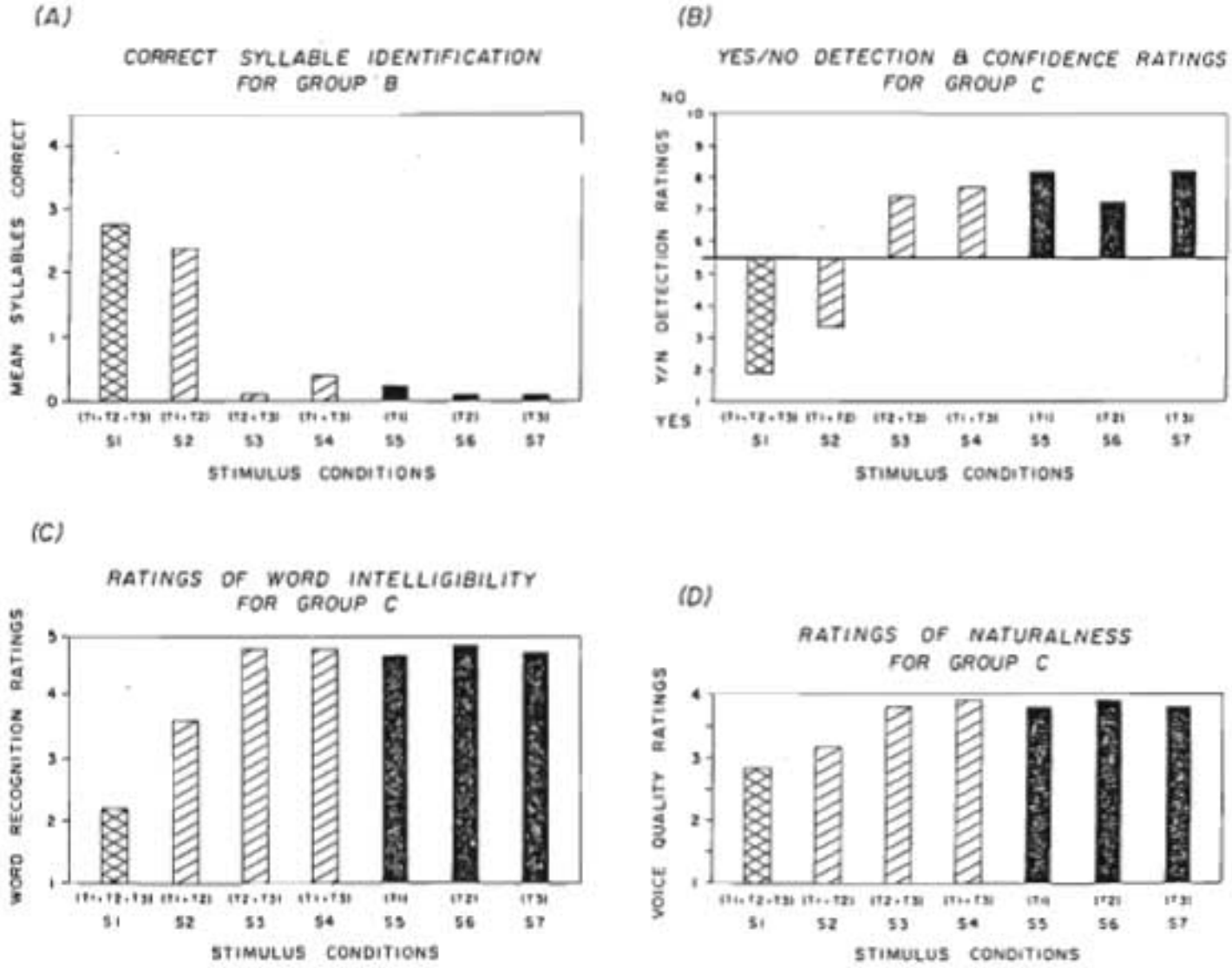

(A)
CORRECT SYLLABLE IDENTIFICATION FOR GROUP B
MEAN SYLLABLES CORRECT
STIMULUS CONDITIONS
(B)
YES/NO DETECTION & CONFIDENCE RATINGS FOR GROUP C
NO
YES
Y/N DETECTION RATINGS
STIMULUS CONDITIONS
(C)
RATINGS OF WORD INTELLIGIBILITY FOR GROUP C
WORD RECOGNITION RATINGS
STIMULUS CONDITIONS
(D)
RATINGS OF NATURALNESS FOR GROUP C
VOICE QUALITY RATINGS
STIMULUS CONDITIONS
S1
S2
S3
S4
S5
S6
S7

FIGURE 2.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 6 (1979-80) Indiana University]

Fundamental Frequency as a Cue to Postvocalic Consonant Voicing in Production: Developmental Data

Sue Ellen Krause
Department of Psychology
Indiana University
Bloomington, Indiana 47405

A paper presented at the 99th Meeting of the Acoustical Society of America, Atlanta, April, 1980.

INTRODUCTION

Several recent studies have demonstrated that fundamental frequency (F0) contour of a vowel affects perception of vowel length (e.g. Lehiste, 1976; Pisoni, 1976; Wang, Lehiste, Chang, and Darnovsky, 1976). Specifically, a vowel with changing F0 contour, either rising or falling, is perceived as longer than a monotone vowel. Moreover, it has been shown that vowel duration serves as a perceptual cue for postvocalic consonant voicing (Denes, 1955; Raphael, 1972; Krause and Fisher, 1980). In an identification paradigm, the probability of a voiced response increases as the vowel stimulus duration increases. Taking this one step further, Lehiste (1977) demonstrated that F0 contour of the vowel directly affects the segmental feature of voicing for apical stops in the postvocalic position. She suggested that changing F0 contour, through its mediating effect on vowel duration, can serve as a cue to the voicing feature of the postvocalic consonant.

Gruenenfelder and Pisoni (1980) replicated Lehiste's earlier findings and extended the effect to a fricative voicing contrast, as well. In addition, they examined F0 contours in speech production to determine whether the perceptual results could be explained by appeal to articulatory regularities in production. However, the majority of their talkers did not demonstrate significant differences in F0 contour. These results were in agreement with findings of two earlier studies (Mohr, 1971 and Lea, 1973). To explain the disparate findings between the perceptual and production data, Gruenenfelder and Pisoni

suggested that the effects of F0 contour in perception may simply reflect a more general psychophysical phenomenon that is not necessarily restricted to speech and probably not tied to any regularity in speech production.

The role of fundamental frequency contour in realizing differences in postvocalic consonantal voicing is also of interest from a developmental perspective. In identification tasks, I have demonstrated that children of 3- and 6-years of age require longer vowel duration values to shift their judgments from a voiceless consonant to a voiced consonant (Krause and Fisher, 1980). Similarly, in speech production, children demonstrate reliably longer vowels before voiced consonants than do adults (Krause, 1980). Thus, while children do use vowel duration as a cue to postvocalic consonant voicing, they seem to exaggerate the vowel duration difference between voicing contexts. It seems possible that suprasegmental features of production, such as F0 contour within a syllable, may be exaggerated, as well, perhaps reflecting less precise articulatory control over this phonetic contrast.

In addition, one might consider the progressively greater control over pitch that is demonstrated with increasing age. Although F0 stabilizes considerably by age 3-4 years, Kent (1976) has reported that young children do demonstrate more variability in their average F0 than do older children or adults. Thus, any indication of regular control over F0 contour with respect to the segmental feature, voicing, would be provocative. The specific purpose of the present study was to examine the F0 contours of

vowels in relation to the postvocalic voicing distinction from a developmental perspective.

METHOD

The subject sample consisted of five 3-year-olds, five 6-year-olds, and five adults. All subjects passed screening tests for normal language and articulation skills and normal hearing sensitivity.

The test stimuli consisted of six words, BIP, BIB, POT, POD, BACK, and BAG.

----------SLIDE 1----------

The postvocalic voicing distinction was tested in three different phonemic contexts: with the vowel /I/ followed by bilabial stops, with the vowel /ɑ/ followed by apical stops, and with the vowel /æ/ followed by velar stops.

----------SLIDE 2----------

Randomizations of the six test words, pictured as line drawings, were presented to the subjects in a sound-attenuated room. High fidelity audio-recordings were made of all productions. The 10 trials with the most consistent stress, intonation, and loudness levels were selected for analysis.

Spectrograms were made of each production with a Voice Identification Sound Spectrograph, using a filter bandwidth of 45 Hz, over an expanded scale of 50-4000 Hz.

----------SLIDE 3----------

Fundamental frequency was measured at the start of the vowel

(F0s) and at the end of the vowel (F0e). ΔF0, the change in F0, was calculated for each production of the test words for each subject. Those ΔF0 values were then averaged for each speaker's production of a test stimulus, yielding one ΔF0 value per word for each subject, which was used as the dependent measure for statistical analysis.

RESULTS

The next slide shows the effect for the Age factor, which was significant.

----------SLIDE 4----------

As shown here, ΔF0 decreases with increasing age. Only the contrast between 3-year-olds and adults was statistically significant. Moreover, a significant effect was found for the Voicing factor.

----------SLIDE 5----------

Specifically, ΔF0 was greater preceding voiced stop consonants than preceding voiceless stop consonants.

These two main effects are most meaningfully interpreted in terms of the Age x Voicing interaction. A display of the mean ΔF0 values for the three age groups by voicing feature is provided in the next slide.

----------SLIDE 6----------

The relation between the Age and Voicing factors was examined in two ways. The first examined the simple effect of voicing at each age level. The contrast in ΔF0 between the voiced and

voiceless contexts reached statistical significance only for the 3-year-olds. That is to say, 3-year-olds demonstrated reliably larger changes in F0 from the start to the end of the vowel before voiced stop consonants than before voiceless stop consonants. The finding that ΔF0 is essentially the same before voiced and voiceless consonants for the adult speakers in this study is consistent with the earlier adult studies (Mohr, 1971; Lea, 1973; Gruenenfelder and Pisoni, 1980). While no claims are being made here regarding F0 contour as an actual cue to postvocalic consonant voicing for the 3-year-old children, this finding does complement results of vowel duration measures made on the same set of data. This point will be discussed in more detail later.

I also examined the simple effect of Age for each of the voicing contexts. While a decrease in ΔF0 is shown with increasing age for vowels before voiceless stops, only the contrast between 3-year-olds and adults reached significance. For vowels preceding voiced consonants, values of ΔF0 differed reliably for all age contrasts. Again, this finding appears to be best interpreted in relation to the vowel duration data.

DISCUSSION

----------SLIDE 7----------

This slide displays the mean duration values for vowels preceding voiceless stop consonants and for vowels preceding the voiced stop consonants, averaged across the three phonemic

contexts tested in speech production (Krause, 1980). A significant interaction between the Age and Voicing factors was found for these data. While all age groups demonstrated significant differences in vowel duration before voiced and voiceless consonants, the 3-year-olds showed the greatest difference in vowel duration between the two voicing contexts, the 6-year-olds showed less difference, and the adults showed the least difference. The F0 contour data reflect a parallel effect. Specifically, the difference in F0 contour between voicing conditions decreased with an increase in age.

The interaction for vowel duration appears to be primarily due to decreasing values of duration before voiced consonants with increasing age. Age comparisons for the F0 contour data yielded comparable findings. The change in F0 from the start to the end of the vowel was most similar between age groups for vowels preceding voiceless consonants. Moreover, all age groups differed significantly on F0 contour of vowels preceding the voiced stops.

SUMMARY AND CONCLUSIONS

In summary, results of this study indicate a clear developmental effect of F0 contour with respect to the voicing contrast for postvocalic stop consonants. The adult data from this study are consistent with the adult data reported by other investigators. The F0 contours obtained for the children, particularly for the 3-year-olds, are different from those of

adults, although those data do complement earlier developmental findings on vowel duration as a function of postvocalic consonant voicing.

It is premature at this point to define either the mechanism(s) deployed by young children in realizing these F0 contour differences or why children differ from the adults' performance. However, we may consider certain factors that could contribute to the developmental effect. For example, the findings from this study could be, at least partially, a function of language-specific linguistic experience with English. As young children gain "knowledge" about the importance of various acoustic correlates of phonemic contrasts over time, they may try to maximize the differences in production between phoneme targets, as appears to be the case for the use of the vowel duration cue. Perhaps pitch differences are another means for realizing these same distinctions in voicing. That is, the developmental differences demonstrated in the present study may represent an early holistic stage in phonological development. Furthermore, since both pitch inflection and duration are correlates of stress, the larger pitch excursions observed in the children's speech may be a function of syllable stress.

As noted earlier, it is generally accepted that children's average fundamental frequency and formant frequency values are more variable than those of adults. The use of this F0 contour measure may represent another parameter that will stabilize with age. At this point, it is unclear as to whether F0 contour differences could be tied to poor control over vowel duration or

if they operate separately, but in a similar fashion to realize voicing in final stops.

Further data analysis is now in progress to assess other measures of F0 contour and to make quantitative comparisons between the F0 contour and the vowel duration data reported here. Moreover, additional developmental perceptual research is in progress to assess the use of multiple acoustic parameters that function as both segmental and suprasegmental features in adult speech, as well as how these parameters develop with age and experience in the language learning environment. The present results show that control over pitch has a developmental course much like my earlier findings dealing with regulation of vowel duration. The control over these distinctive acoustic correlates becomes more precise as the child's linguistic maturity approaches that of the adult.

REFERENCES

Denes, P. Effects of duration on the perception of voicing. Journal of the Acoustical Society of America, 1955, 27, 761-764.

Gruenenfelder, T.M. and Pisoni, D.B. Fundamental frequency as a cue to postvocalic consonantal voicing: Some data from speech perception and production. Manuscript submitted for publication, 1980.

Kent, R.D. Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies. Journal of Speech and Hearing Research, 1976, 19, 421-447.

Krause, S.E. Developmental use of vowel duration as a cue to postvocalic consonant voicing: Evidence from speech production with comparisons to speech perception. Manuscript submitted for publication, 1980.

Krause, S.E. and Fisher, H.B. Developmental use of vowel duration as a cue to postvocalic consonant voicing: A perception study. Manuscript submitted for publication, 1980.

Lea, W.A. Segmental and suprasegmental influences on fundamental frequency contours. In L.M. Hyman (Ed.), Consonant types and tones. Southern California Occasional Papers in Linguistics, 1973.

Lehiste, I. Influence of fundamental frequency pattern on the perception of duration. Journal of Phonetics, 1976, 4, 113-117.

Lehiste, I. Contribution of pitch to the perception of segmental quality. Paper presented at the 9th International Congress on Acoustics, Madrid, 1977.

Mohr, B. Intrinsic variations in the speech signal. Phonetica, 1971, 23, 65-93.

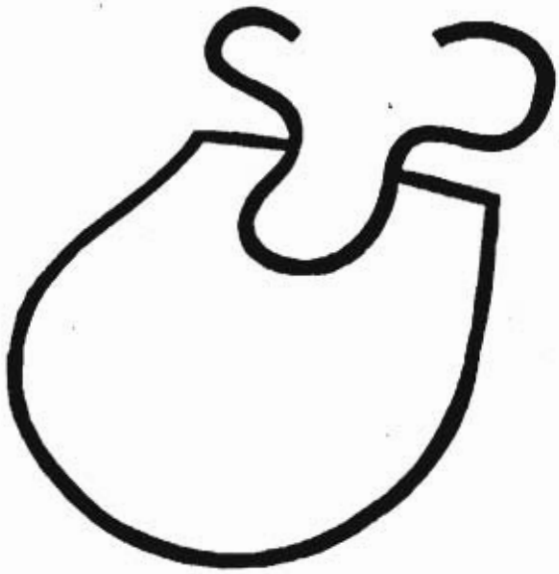
Pisoni, D.B. Fundamental frequency and perceived vowel duration. Paper presented at the 91st meeting of the Acoustical Society of America, Washington, D.C., April, 1976.

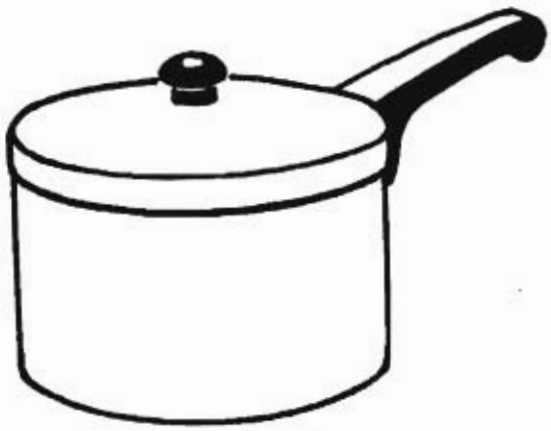
Raphael, L.J. Preceding vowel duration as a cue to the perception of voicing of American English consonants in word-final position. Journal of the Acoustical Society of America, 1972, 51, 1298-1303.

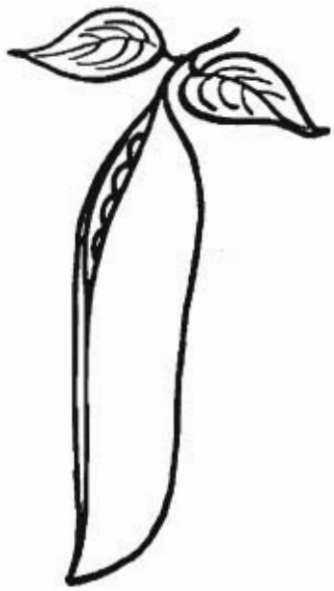
Wang, W.S-Y., Lehiste, I., Chuang, C-K., and Darnovsky, M. Perception of vowel duration. Paper presented at the 92nd meeting of the Acoustical Society of America, San Diego, November, 1976.

	BIP	BIB	POT	POD	BACK	BAG
3 YR						
6 YR						
ADULT						

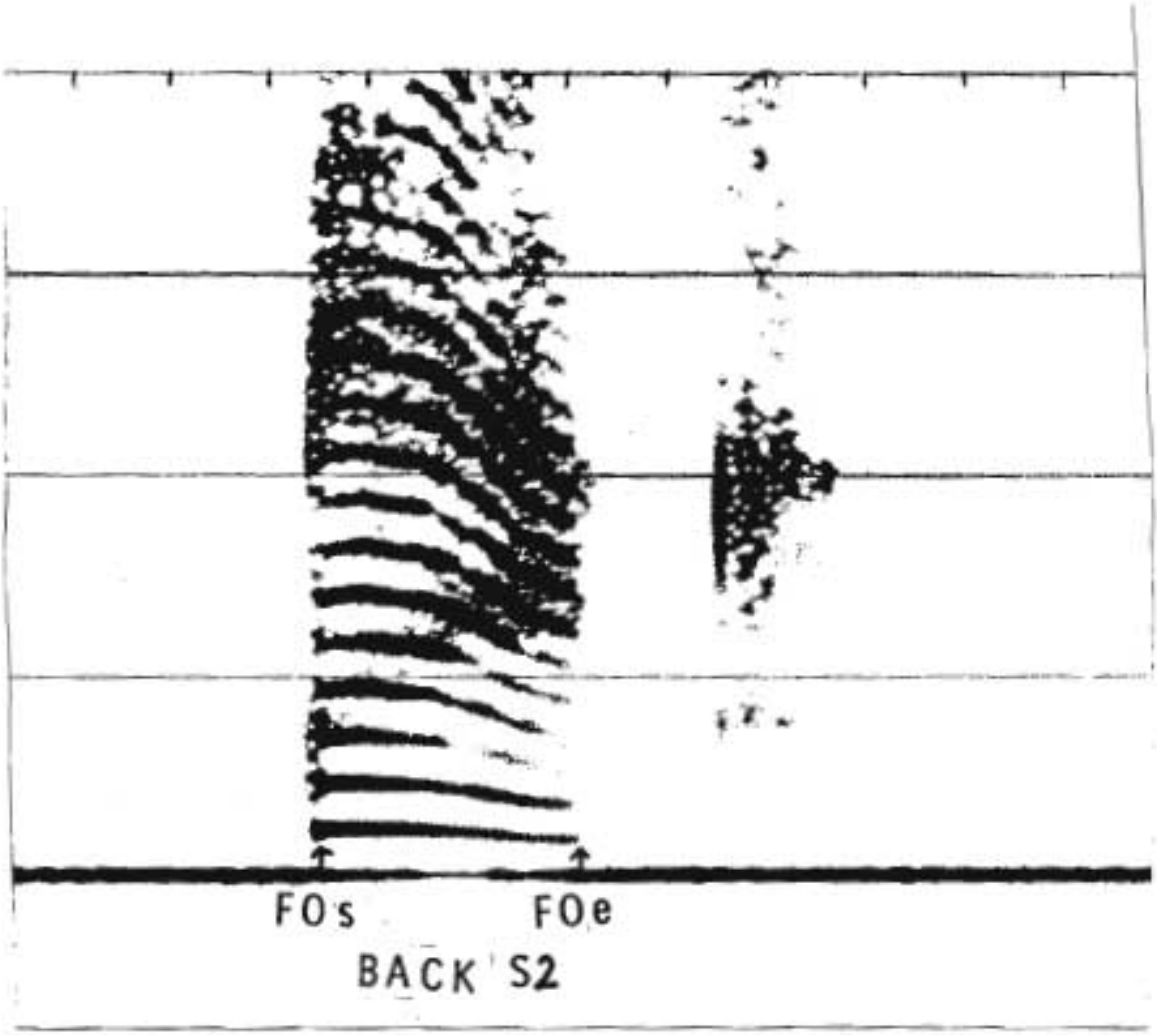
Slide 1



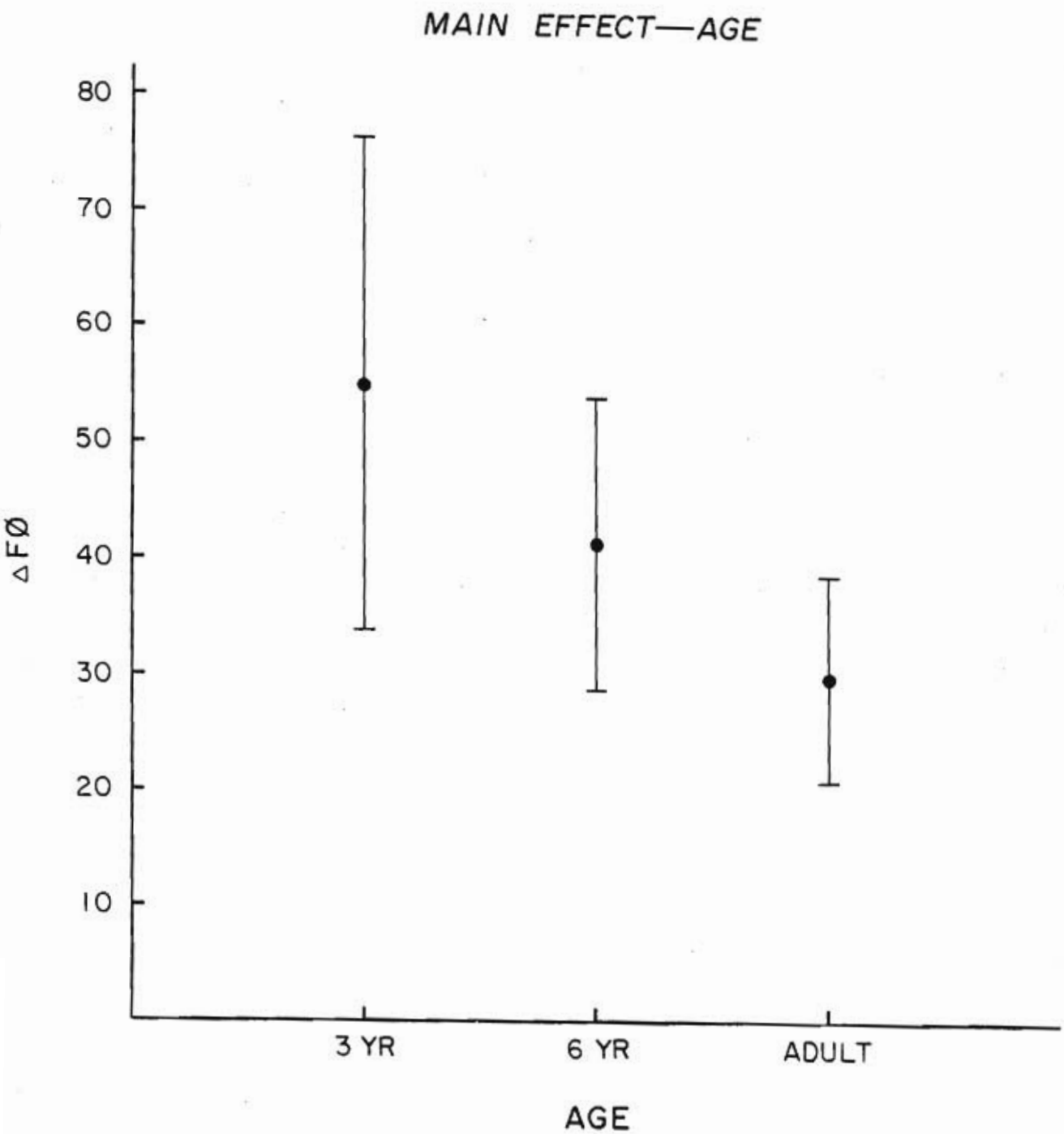




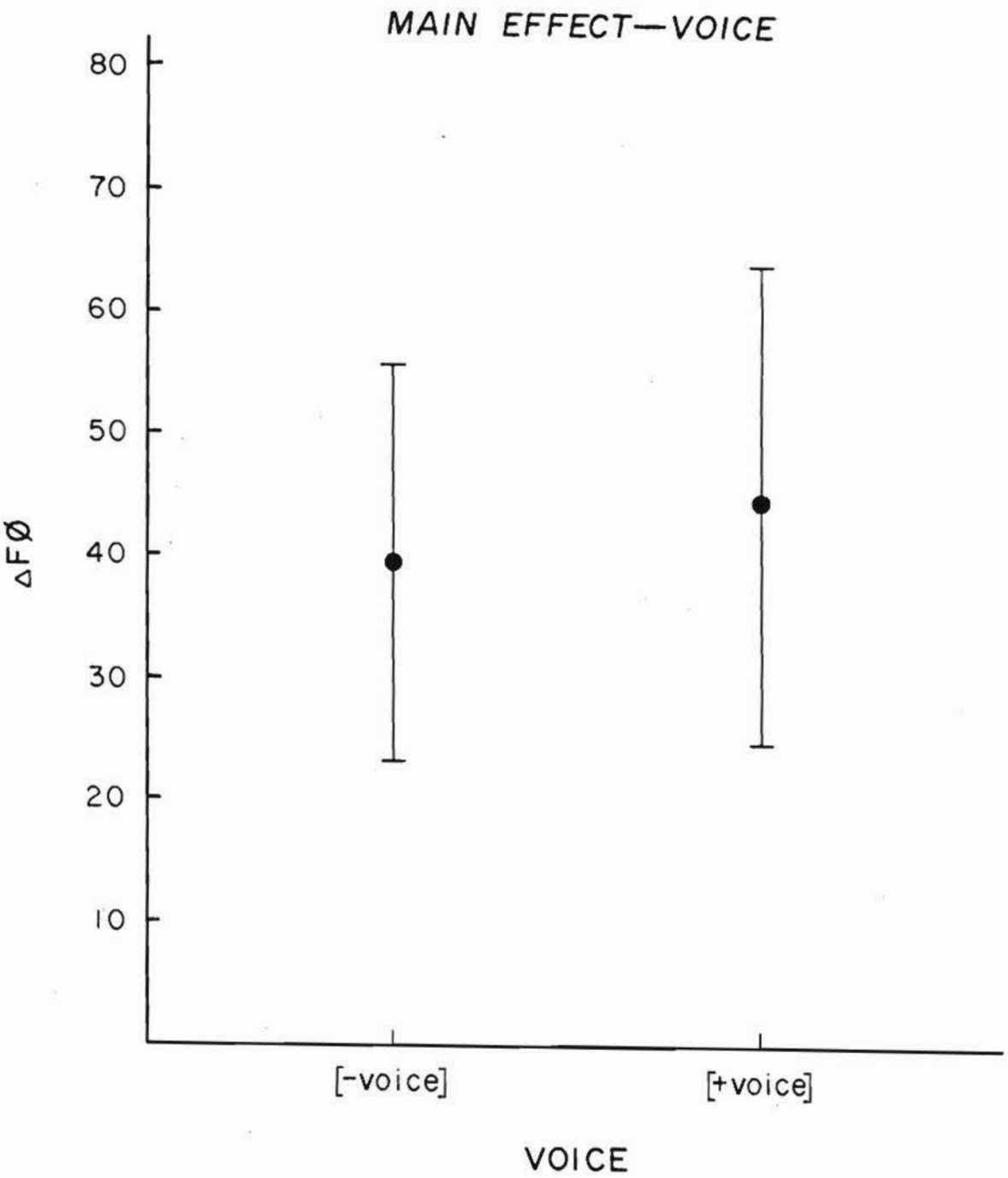
Slide 2



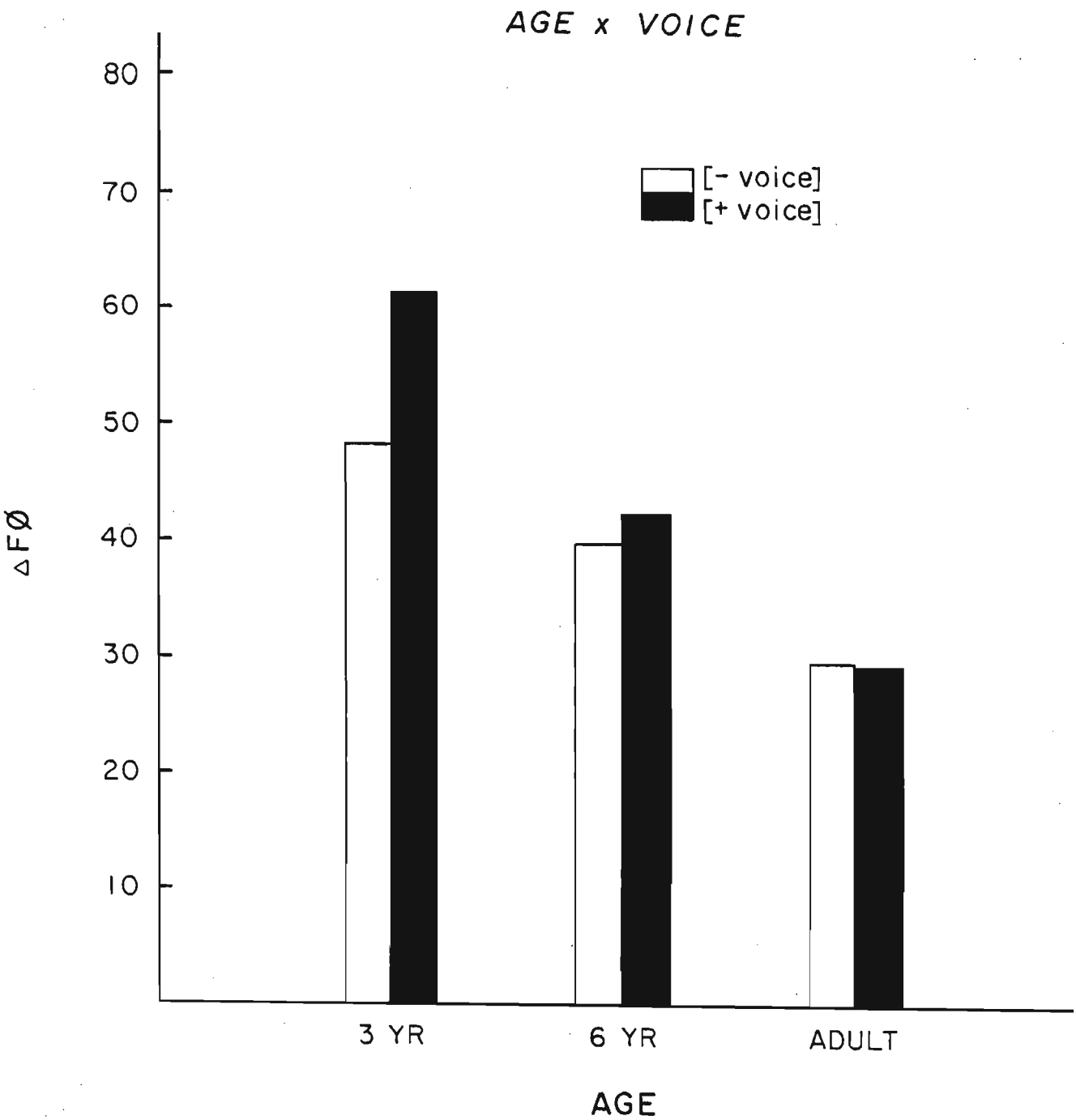
Slide 3



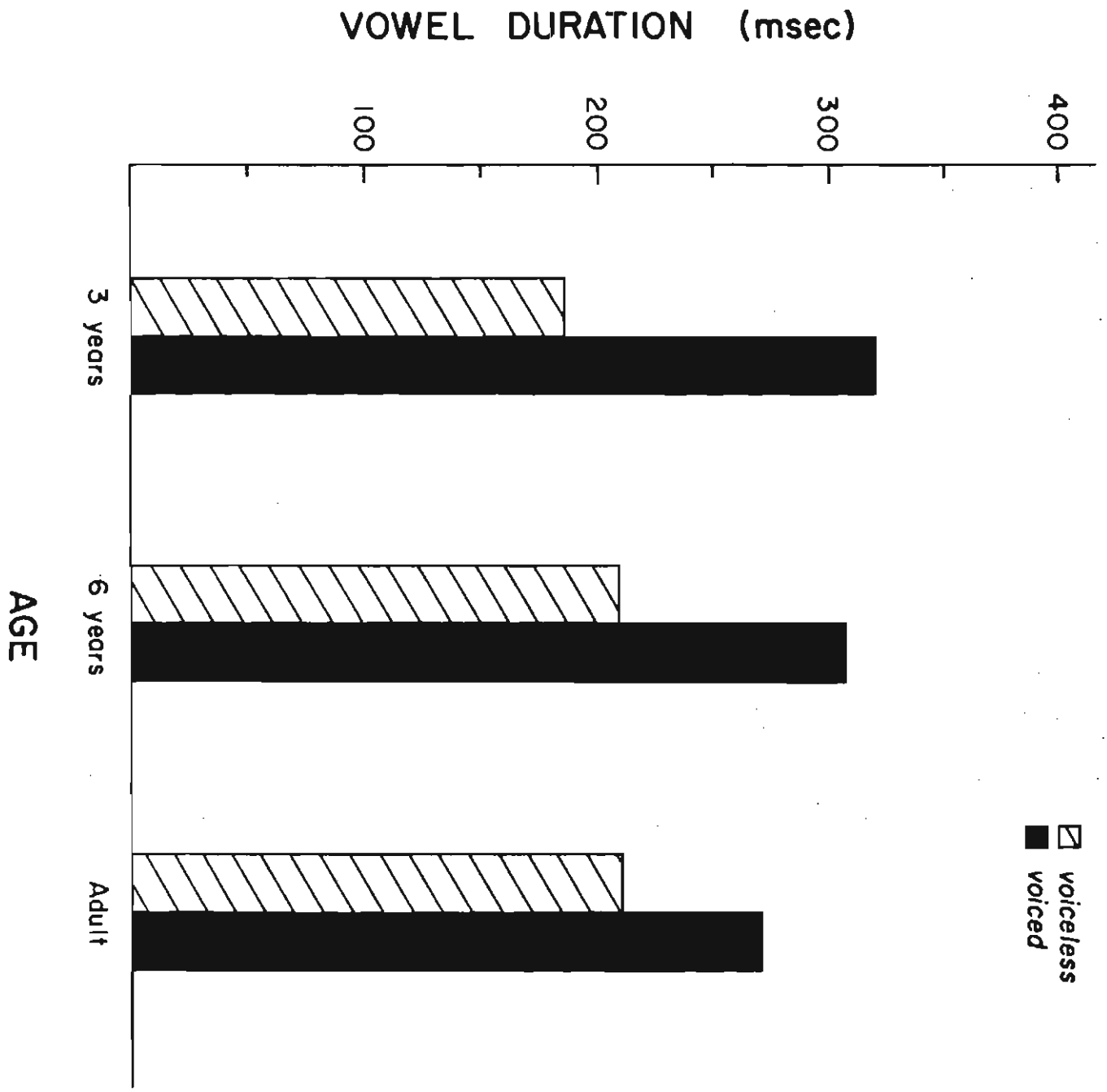
Slide 4



Slide 5


AGE x VOICE
80
70
60
50
40
30
20
10
ΔFØ
[- voice]
[+ voice]
3 YR
6 YR
ADULT
AGE

Slide 6


VOWEL DURATION (msec)
100
200
300
400
AGE
3 years
6 years
Adult
voiceless
voiced

Slide 7

Classification of CV Syllables by Readers and Prereaders

Amanda C. Walley, Linda B. Smith and Peter W. Jusczyk

Indiana University
Bloomington, Indiana 47405

This is a version of a paper that was presented at the 88th Annual Convention of the American Psychological Association, September 1, 1980, Montreal, Canada. The research reported here was supported by grants from NIRC, NSF and the Social Sciences and Humanities Research Council of Canada. We thank Melanie Lockyear for her help in the data collection.

Abstract

The present experiment investigated two hypotheses that have been advanced to account for the young child's difficulty in performing explicit phonemic judgements and manipulations. According to one hypothesis, this difficulty results from the absence of phonemic components in the child's underlying representation of speech. According to the other, it reflects an inability to access these units. In the classification task employed in the present experiment, prereading children and more experienced readers were asked to group CV syllables with either of two pretrained syllables. Classification by either consonant or vowel identity was possible, although the stimulus sets were designed to bias children towards consonantal classifications. Even very young children made a substantial number of consonant and vowel classifications indicating that they represent speech in terms of phonemic segments and can access this information. Because the classification results were not strongly related to reading ability, it was proposed that the classification task taps sensitivity to phonemic information at a level of processing that is different from that used by explicit segmentation tasks. Such sensitivity may, in fact, serve as the basis for the development of segmentation abilities.

Classification of CV Syllables by Readers and Prereaders

The difficulty young children 4 or 5 years of age experience in tasks requiring explicit phonemic judgements and conscious manipulations, such as judging the number of phonemic segments in words, making same-different judgements about phonemic segments, and rearranging and deleting specified segments (e.g., Bruce, 1964; Calfee, Chapman & Venezky, 1972; Liberman, Shankweiler, Fischer & Carter, 1974; Savin, 1972) has been interpreted in two different ways. The first interpretation holds that although phonemic units form part of the child's representation of speech, they are not accessible for conscious inspection (Gleitman & Rozin, 1977; Liberman, Shankweiler, Liberman, Fowler & Fischer, 1977; Rozin & Gleitman, 1977). This position is consistent with current interpretations of developmental differences in cognitive abilities, particularly those related to memory (see Brown, 1978). The second view assumes that the child's difficulty indicates that phonemes do not constitute perceptually real entities for the child - i.e., that they are not a part of the child's underlying representation of speech (Treiman & Baron, 1980). Indeed, the 'psychological reality' of the phoneme constitutes a topic of debate even with respect to adult speech perception. The controversy centers around the question of whether or not this information is necessarily computed in fluent speech processing prior to lexical access (e.g., see Klatt, 1980; Foss & Blank,

1980) since such information is available to adults at least after word identification. These two accounts of the child's difficulty differ then in their assumptions about how speech is represented in the child's mental lexicon. They differ further with respect to the implied relations between the perception of speech and learning to read. We will briefly consider these two views and their implications for learning to read.

The view that phonemes are "real" for the child, but only become accessible with age, is motivated, in part, by the young child's rather sophisticated ability to produce and understand fluent speech. Some researchers (Gleitman & Rozin, 1977; Liberman et al., 1977; Rozin & Gleitman, 1977) have maintained that perceptual constancy of the phoneme and phonemic segmentation are achieved implicitly by a mechanism that is, following the position of the Haskins Laboratory group, specialized for speech processing. The coarticulation of speech sounds gives rise to a continuously varying acoustic waveform in which reliable cues for phoneme segmentation are absent and in which sounds are encoded in a context-dependent manner, such that one portion of the waveform cannot be interpreted without reference to adjacent portions. The speech perception mechanism deals with this encodedness problem by reference to the motor-articulatory patterns which initially gave rise to these contextual dependencies. However, the relations between the articulatory and auditory systems which support this decoding are not simple; they are presumed to be "hard-wired" into the speech perception mechanism and apparently do not form a part of the child's explicit knowledge of speech. At the level of perception, at least as perception is viewed by Gleitman and

Rozin, and Liberman and her associates, segmentation of the waveform is not possible for the child and phonemic units are not consciously realized. These researchers propose that the syllable represents the 'lower bound as a real-time (accessible) unit' of speech processing, which is the lowest level of linguistic structure that maps linearly onto the sound stream and thus the syllable lends itself more readily to conscious awareness.

The contrasting view is that the difficulty young children have in making phonemic judgements reflects a fundamental difference between children and adults in the underlying representation of speech (see Treiman and Baron, 1980). The child's difficulty, according to this view, reflects a more general tendency of the immature perceptual system to perceive what are for adults separable dimensions of stimuli (i.e., phonemes) as integral - i.e., as undifferentiated wholes (Smith & Kemler, 1977; Shepp, 1978). The child's poor performance is not, therefore, specific to the encodedness problem of speech or a result of a mere inability to access phonemic information.

Treiman and Baron (1980) have reported several findings which support their position. First, they have shown that young children classify speech sounds by their wholistic similarity rather than by phonemic identities. For example, children tend to judge syllables such as /bi/ and /vɛ/ (which are wholistically alike) as belonging in one category, while adults tend to judge the syllables /bi/ and /bo/ (which share a constituent phoneme) as belonging in the same category. Second, in a free classification task, no significant difference in the percentage of consonant vs. vowel classifications was found, and in a constrained

classification task, children learned to make common phoneme classifications equally well for stops and fricatives. These findings would seem to indicate that it is not, as the inaccessibility hypothesis contends, simply acoustic considerations which determine the difficulty of phonemic analysis.

Not only do these two groups of researchers differ in how they view the child's representation of speech, but they also differ in how they characterize the emergence of segmentation skills. As Read (1978) has put it, these abilities appear at a decidedly 'suspicious' point in development - at a time when the child begins to learn to read (an alphabetic orthography). According to Gleitman and Rozin, and Liberman and her associates, learning to read an alphabet is difficult (in comparison to a syllabary, for example) because, in order to take advantage of the alphabetic principle, it requires that the user learn letter-sound correspondences which are not based on an easily accessible perceptual category. Thus, it is suggested that learning to read is a prerequisite for the development of phonemic perception or awareness and, more generally, that metalinguistic insights which are fostered by learning to read causally precede these cognitive categories. Stated another way, phonemic perception may emerge specifically as a consequence of learning to read an alphabetic orthography. If this characterization is an accurate one, it implies an interesting relationship between speech and reading; i.e., the latter is generally viewed either as equivalent or parasitic to the primary linguistic process of speech, yet here is an instance of the secondary linguistic process exerting a profound influence on the primary process (see also Miller, 1972,

for some interesting comments on the relationship between writing and the evolution of Western logic).

The claim that learning to read an alphabetic orthography leads to the formation of abstract perceptual/cognitive categories and metalinguistic insights (i.e., access to phonemes) may be contrasted with the opposing view which holds that phonemic categories are a prerequisite for reading and the growth of metalinguistic knowledge - i.e., phonological awareness. To some extent, this position is embodied in Treiman and Baron's view of the development of phonemic perception. According to these investigators, the emergence of the ability to perform segmentation tasks represents a fundamental change in the way speech is represented by underlying perceptual mechanisms and, indeed, a change that occurs throughout the perceptual system. Integral perception of speech becomes separable with general development at a time when this change occurs for other stimuli (e.g., for the dimensions of visual stimuli, such as color and form). Although Treiman and Baron might not disagree that this change may be prompted by learning to read, to the extent that this change occurs in other realms of perceptual processing, their position is more that it is the result of general cognitive growth. It may be noted that although segmentation abilities improve markedly around the time the child begins to learn to read, this is also the time that the child begins a program of formal education in which a variety of explicit analytic skills are taught. It need not be the case, therefore, that dimensions that are initially perceived as integral by the child become separable simply according to some internal developmental

schedule. Rather, various forms of experience, including that of reading, might foster this development. The critical claim, however, is that the progression from the "integral" to "separable" perception of phonemes is part of a general developmental trend which is likely to emerge without specific training. This is consistent with the finding (see Liberman et al., 1977) that the relative performance of prereaders in a phoneme counting task is correlated with their later reading ability.

Because it is impossible to systematically prevent reading instruction for children, disentangling these two explanations of the development of phonemic perception - i.e., learning to read an alphabetic orthography vs. more general cognitive development - is difficult. There is a host of evidence suggesting that some kind of relationship exists between children's perception of phonemes and learning to read an alphabetic orthography (e.g., Liberman et al., 1977; Makita, 1968; Rozin, Poritsky & Sotsky, 1971; Rozin & Gleitman, 1977), but the causal direction of this relationship remains unclear. There is, nevertheless, some support for the notion that access to phonemic structure requires reading instruction with an alphabetic orthography. For example, Morais, Cary, Alegria and Bertelson (1979) studied segmentation abilities in illiterates and found that, like children, these adults had no problem producing or understanding fluent speech, but, relative to a group of literates, they had substantial difficulty on segmentation tasks (adding and deleting specified segments). This result, which is presumably not predicted by the general cognitive growth hypothesis, suggests that the failure to develop these

segmentation skills was a result of the absence of reading instruction, rather than an arrest in cognitive growth. (One serious problem with accepting this interpretation, however, is that these investigators failed to control for the illiterates' level of general intelligence - i.e., it is not clear whether the illiterates did not learn to read through lack of opportunity or lack of ability to do so.)

While young children apparently do not have easy access to phonemic information or sophisticated abilities for segmenting and manipulating this information, there are several reasons for suspecting that phonemes do actually form part of the child's underlying representation of speech and that these are to some extent accessible. First, recent attempts to model the initial stages of speech processing (Kewley-Port, 1979; Searle, Jacobson & Rayment, 1979; Stevens & Blumstein, 1978, 1980), which employ novel (vs. traditional spectrographic) methods of speech analysis, have met with considerable success in identifying invariant acoustic correlates for place of articulation - a feature that has been particularly resistant to any satisfactory acoustic definition (see Liberman et al., 1967). The existence of such invariants does not preclude the notion that the developing child is faced with the task of extracting these invariants from the waveform and that this may be difficult, but it does at least allow for the possibility that these invariants are present in the child's representation of speech before reading is undertaken. Second, these possible acoustic invariants may provide the basis for the constancy of the phonemic/phonetic percept. At least three studies of infant speech perception have shown that such

perceptual constancy may exist even at a prelinguistic point in development (Eilers, 1977; Fodor, Garrett & Brill, 1975; Kuhl, 1976). Such findings would seem to constitute a further indication that phonemes or the perceptual categories on which they are based are eventually a part of the prereader's representation of speech. Third, there is additional empirical evidence, which is largely overlooked in discussions of phonemic perception and its development, indicating that the prereader is sensitive to phonetic similarities between different vowels (Read, 1973). While such sensitivities may not indicate the existence of phonemic categories per se, they do suggest that the child possesses some ability to attend to the component segments of speech sounds. This ability may, in turn, provide the basis for the child's rather sophisticated knowledge of phonological rules and knowledge about what constitute relevant and irrelevant phonetic variations in speech (i.e., phonemes) - knowledge which is not easily accounted for by the wholistic similarity view of the child's perception of speech.

How can the discrepancy between Read's results, together with those from the infant speech research, and the wealth of studies demonstrating the difficulty of phonemic tasks for young children be resolved? Read (1978) attributes his success in demonstrating sensitivity to phonetic relationships between different vowels to the nature of the task that he employed. He has proposed that accessibility of linguistic structures is possible to varying extents - i.e., as sensitivity to these structures is manifested in the ability to perform tasks of different complexity. Thus, the ability to make similarity judgements in his task is indicative of

access to these structures and is easier than the categorization and segmentation studies typically used to study speech perception in children. These tasks are at a level of complexity very close to the skills required in reading. Whereas Read's task simply required that children judge whether a sound like /fɛd/ was more similar to /ɛd/ than was a sound like /fʊd/, segmentation tasks require explicit, conscious manipulation of phonemes (for example, counting the number of phonemes in the word "dog", pronouncing the word "monkey" without the "k" sound, etc.). Children's performances in these two types of tasks may tap quite different levels of awareness of the sound structure of language; Read's results may be indicative of some initial sensitivity to this structure, whereas the results of segmentation studies may define the limits of this sensitivity. Thus, one reason for previous failures in demonstrating the accessibility of phonemic categories may be due to the general assumption that segmentation, which is apparently very difficult, but which has usually been required in these studies, is the only indicator of accessibility of this information. One could hypothesize, however, that there exists in the child an initial sensitivity to phonemic components that matures into explicit awareness of segments. Read's results can, nevertheless, be interpreted in terms of the child's sensitivity to the wholistic similarity of his test items (i.e., /ɛd/ and /fɛd/ are more wholistically alike than /ɛd/ and /fʊd/). Furthermore, even if there does exist this developmental trend from initial sensitivity to explicit awareness of phonemic segments, what fosters it - general cognitive growth or learning to read?

We addressed these two issues - how sensitive the young child is to phonemic segments and how this sensitivity is related to developmental level and learning to read - with a classification task. Specifically, we investigated the ability of prereaders (kindergarteners) and beginning and more experienced readers (second and fifth graders) to classify speech sounds (CV syllables) by their constituent phonemes. It may be noted that in Treiman and Baron's (1980) free classification task, wholistic similarity was pitted against the phonemic classification. Therefore, their results may indicate that children merely prefer the wholistic similarity classification, as opposed to having no representation in terms of phonemes. We attempted to structure our classification task such that it would be optimal for the extraction of consonantal information from the stimulus array. According to Gibson's (1969) view of perceptual development and learning, it should be easier for the child to extract invariants from a stimulus array in which there is a considerable number of exemplars of that invariant. Furthermore, this extraction process may also be facilitated when identity on the relevant dimension is preserved, but when the stimuli do not possess a large degree of wholistic similarity (Kemler & Smith, 1979). Specifically, the stimulus sets used in the present experiment were designed in such a way that the stimuli could be grouped on the basis of shared consonantal information (+labial and + dental) or on the basis of shared vocalic information (+ and - front vowel). However, the opportunity for grouping according to initial consonant was made optimal by presenting stimuli which shared consonantal information more often than vocalic information and by using rather dissimilar vowels in the syllables.

Method

Subjects.

The subjects for the present experiment were children enrolled in an elementary school in the Bloomington area and paid volunteers whose parents responded to an advertisement. Twenty-five kindergarteners (mean age = 5,11 years), 26 second graders (mean age = 8,1 years) and 27 fifth graders (mean age = 10,2 years) participated in the experiment. In addition, 10 kindergarteners and 1 fifth grader contributed data, but failed either to meet the pretraining criterion (10 correct consecutive responses) or to meet an acceptable level of correct responding (75%) on pretrained items in testing and their data were not analyzed. No gross speech or hearing disorder was reported for any child by parents at the time of testing and all subjects were native speakers of English.

Procedure.

Each subject was tested individually in a single session lasting no more than 45 minutes. The session included a classification task (pretraining and testing), a picture-letter matching task, and a reading test where applicable (see below). The subject was seated at a table facing two puppets. Experimenter 1 (E1) sat facing the child and experimenter 2 (E2) sat to one side of the table.

The experimental session began with pretraining. In this phase of the experiment, the child was informed that each of two puppets made a special sound that the child was to learn. The

child was asked to pat the correct puppet on the head whenever he heard that puppet's sound. E1 then read the two pretraining items from a randomized list. Ex2 recorded each response from the child and informed him whether or not he was correct. After a minimum criterion of 10 correct consecutive responses, the testing phase of the task was initiated.

In the testing phase of the procedure, the child was told that he was going to listen to several more sounds made by either of the puppets and that he was to choose which one of the puppets had made the new sound and indicate his choice by patting one of the puppets on the head. E1 then read each of the 8 test items within a block. Each test item was read in the carrier sentence "if puppet A says (pretraining item 1) and puppet B says (pretraining item 2), who would say (test item)?" E2 recorded the child's response on each trial. The order of the puppets and their associated sounds in the carrier sentence was alternated between blocks. In addition, at the beginning of each block, E2 asked the child what sound each puppet had started out saying (the pretrained items) and if he did not remember, the child was prompted. No feedback was given on individual test trials, but general encouragement was given at the end of each block.

In the picture-letter matching task, the child was shown 3 pictures and asked to identify each object. If the child did not respond with the intended object name, E2 corrected him. The child was then asked to point to the picture that "went best" with a letter shown to the side of the pictures. E1 recorded each child's response. Only those subjects who responded correctly on 8 out of the 10 trials advanced to the reading test.

In the reading test, E1 held up cards with the test items one at a time in random order. E2 transcribed all readings of each word offered by the child. Each child received the same one-syllable real and nonsense (1SR, 1SN), and two-syllable real and nonsense (2SR, 2SN) word lists (in this order). If a child could not read any of the first four words shown in either of the 1SR or 2SR lists, the session was terminated. Before testing on the nonsense words, the child was told that he was going to see some "pretend" words that E1 and E2 had made up and to try to say these aloud the way he thought they should sound even though they might sound silly.

Stimulus and Design.

Pretraining and test items were drawn from the oral stop consonant stimulus set /bi, bɛ, ber, bo, di, dɛ, der, do/ or from the nasal stop consonant set /mi, mɛ, mer, mo, ni, nɛ, ner, no/. Subjects in the oral stop condition (OS) and the nasal stop condition (NS) were randomly assigned to 1 of 8 possible pretraining contrasts that pitted both place of stop consonant articulation and vowel frontness against each other. Thus, for example, the pretraining constrasts for OS were /bi/ vs. /der/, /bi/ vs. /do/, /bɛ/ vs. /der/, /bɛ/ vs. /do/, /ber/ vs. /di/, /ber/ vs. /dɛ/, /bo/ vs. /di/ and /bo/ vs. /de/. Puppet assignment to a pretraining item in a contrast was counterbalanced. The testing phase of the classification task consisted of nine test blocks. Within each block, each item of a stimulus set, including pretrained items, appeared once in random order. E1, who spoke the local dialect of the children tested, read all the sounds aloud from a prepared list.

In the picture-letter matching task, 10 groups of 3 hand-drawn pictures were presented (on paper) together with the orthographic symbol for a consonant, which began the name of only 1 of the 3 pictures.

Test items in the reading task were handprinted on 3 x 5 " cards. Nine items appeared in the 1SR and 1SN in the word lists and 10 items in the 2SR and 2SN word lists.

Results

Figure 1 shows the mean percentages of correct consonant

Insert Figure 1 about here

---

classifications made by the kindergarteners, second and fifth graders. The first two test blocks were regarded as an introduction to the classification task and the mean percentages shown in this figure are, therefore, based on performance on the last 7 test blocks. The scores for each child were submitted to an analysis of variance for a 2 (Grade) x 2 (Stimulus set) design. This analysis revealed only a main effect of Grade ($F(2,73) = 9.52$, $p < .01$). Planned comparisons further revealed that the mean percentages of consonantal classifications by kindergarteners differed significantly from those of second and fifth graders ($t(75) = 4.31$, $p < .01$; $t(75) = 3.39$, $p < .01$, respectively). The mean percentage of classifications by second and fifth graders did not differ significantly from one another ($t(75) < 1.00$). These results point very clearly to an increasing ability to attend to

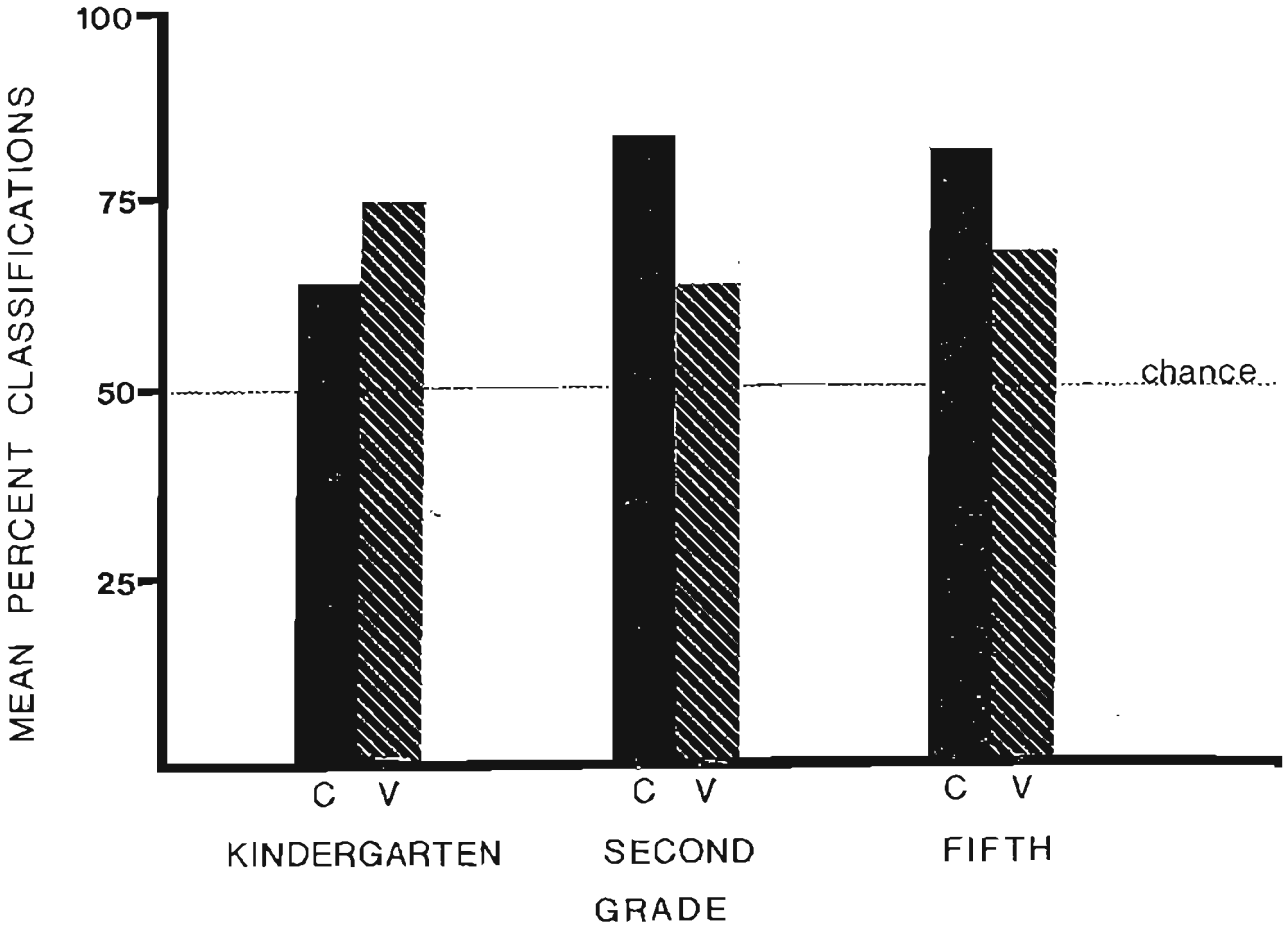


Figure 1. Mean percentages of consonant and vowel classifications in relation to grade level.

consonantal information in CV syllables between kindergarten and grade 2, at which point this ability appears to asymptote. Nevertheless, the kindergarteners ability to make consonantal classifications is greater than chance ($t(24) = 4.44$, $p < .01$), indicating that they can and do attend to this type of component segment.

Also shown in Figure 1 are the mean percentages of correct vowel classifications for each grade. A vowel classification refers to a classification made on the basis of vowel identity. Thus, for example, if the test item /di/ was grouped with the pretrained item /bi/, this was scored as a correct vowel classification. Subjects' vowel scores were based on the pretrained vowels. Thus, if a subject was pretrained on /bi/ vs. /do/, his vowel score was based on only the 4 sounds /bi/, /bo/, /di/ and /do/. A two-way analysis of variance showed only a marginally significant main effect of grade level ($F(2,73) = 2.97$, $p < .10$). Planned comparisons indicated that the mean percentages of vowel classifications by kindergarteners differed significantly from those of second graders ($t(75) = 2.58$, $p < .01$), but not from those of fifth graders ($t(75) < 1.00$). The mean percentages of vowel classifications by second and fifth graders fell just short of being significantly different ($t(75) = 1.65$, $p > .05$). Selective attention to vowel segments appears, therefore, to follow a curvilinear trend which drops between kindergarten and second grade and increases again between second and fifth grade. It is the case, however, that the mean percentage of vowel classifications made by the second graders is still greater than chance ($t(25) = 4.20$, $p < .01$).

In addition to examining the mean percentages of consonant and vowel classifications for each grade, we also determined the classification "rule" used by each subject. A rule was defined as classification (by consonant, vowel or some other means) on 80% or more of the test trials. The percentages of subjects for each grade using these three rules are given in Table 1. As one can

---

Insert Table 1 about here

---

see, the pattern of rule use by individual subjects in each grade conforms well with the results derived from the group analysis. Specifically, classification by consonant increases between kindergarten and second grade where it plateaus, whereas classification by vowel is better described by a u-shaped function. Finally, it may be noted that there is a decline in the use of the "other" rule with increasing age.

Since virtually all of the fifth graders performed at a ceiling level in the reading test, their data do not illuminate the relationship between classification and reading and their reading data are not shown here. Therefore, only the reading performances of the kindergarteners and the second graders will be considered below.

Table 2 shows the results of the picture-letter matching task

---

Insert Table 2 about here

---

in relation to rule use for the kindergarteners. Subjects were divided into two groups - those who scored 80% or better in the

Table 1

Percentage of Subjects Using a
Consonant, Vowel or "Other" Classification Rule

	Grade		
Rule Use	K	2	5
C	20	69	44
V	40	16	37
O	40	19	19

Table 2

Rule Use by Kindergarteners
in Relation to Reading Ability

	Matching Task		Reader	
Rule Use	Pass	Fail	Yes	No
C	4	1	2	3
V	3	7	0	10
O	7	3	1	9
Total	14	11	3	22

matching task and advanced to the reading test (Pass) and those who did not (Fail). As can be seen in the table, over half (14/25) of the kindergarteners (most notably those who used a consonant or an "other" rule) passed the matching task by this criterion. Very few of these children (3/14) could actually read more than 4 of the 9 1SR words - a result which would seem to argue against there being any strong or clear relationship between rule use and reading ability. The three subjects who were able to read (S1, S9, S18) scored at least 7/9 on the 1SR words and 3/9 on the 1SN words. Two of these subjects were consonant rule users, the third an "other" rule user.

The second graders performed at ceiling for the picture-letter matching task and the 1SR and 2SR word lists and therefore these results are not shown. Two levels of reading proficiency (Pass and Fail) were defined by 70% or better performance on the 1SN and 2SN word lists. The results are given in Table 3. As can be seen from the table the ability to read

---

Insert Table 3 about here

---

these simple nonsense words is not clearly related to rule use since roughly half of the subjects in each rule group met this criterion. A consideration of the results of the two syllable nonsense words in relation to rule use is somewhat more informative (see Table 3). It appears that all subjects who met the pass criterion tended to use either a consonant or vowel rule. However, use of either of these rules did not guarantee reading success (as is also indicated by the absence of any relationship

Table 3

Rule Use by Second Graders
in Relation to Reading Ability
for 1 and 2 Syllable Nonsense Words

	1 Syllable		2 Syllable	
Rule Use	Fail	Pass	Fail	Pass
C	9	9	13	5
V	1	2	2	1
O	3	2	5	0

between classification and reading ability in the kindergarteners data and the ability of fifth graders to make both consonant and vowel classifications and to read). Admittedly, the small number of second graders who used a vowel rule makes it difficult to draw any firm conclusions.

Discussion

How aware are children of the phonemic segments of speech? Our classification results suggest that even very young children are sensitive to these components. The experiment did, nevertheless, reveal an increasing tendency with age (up to second grade) to classify CV syllables by identity of their initial consonants. However, it was also found that even kindergarteners were able to make this type of classification at a level of performance above chance. Furthermore, a rather substantial ability to classify the same sounds according to vowel identity was also evident in children in all grades and this ability appeared to exhibit a slight curvilinear trend - i.e., classification by vowel decreased between kindergarten and second grade and tended to rise again in the fifth graders. Note that classification by consonant does require attention to phonemic constituents as the child must group wholistically dissimilar sounds together (e.g., /bi/ and /ber/) and wholistically similar sounds apart (e.g., /bi/ and /di/). Thus, the finding that kindergarteners are able to make consonantal classifications would seem to indicate that these constituent phonemes do, in fact, form a part of the young child's underlying representation of speech

(in contrast to Treiman and Baron's, 1980, claim) and that the structure of this representation is to some extent accessible (in contrast to the position taken, for example, by Gleitman and Rozin, 1977, Liberman et al., 1977, and Rozin and Gleitman, 1977).

The answer to the question of how this sensitivity is influenced by learning to read is less clear. This may be due, in part, to the insensitivity of our measure of reading ability. Nevertheless, it was the case that although very few (3/25) of the kindergarteners could read even very simple words, they did demonstrate a substantial ability to classify the test items according to their constituent consonants and vowels and to consistently use either a consonant or vowel rule. This may be interpreted as support for the idea that these constituents, which form a part of their representation of speech, are accessible to some extent prior to reading. The relationship between rule use by second graders in classification and reading performance also does not appear to be a strong one. Although all of the second graders who succeeded in reading the two syllable nonsense word list were consonant or vowel rule users, many of the children who used these rules in classification did not succeed in the reading task. This finding indicates that classification skills, which are assumed to reflect access to phonemic structure can exist somewhat independently and (as the kindergarten data suggests) prior to learning to read.

The decrease in vowel classifications around second grade is particularly intriguing in view of the fact that vowels, in comparison to consonants, have relatively more invariant acoustic representations, although in English, their orthographic

representations are more varied than those of consonants. The second graders apparently decided, instead, to attend primarily to consonants - a decision which probably does not reflect a change in the underlying representation of speech (since the kindergarteners can make consonant classifications also), but rather the assumption that, having just begun to learn to read, the initial sounds of words are particularly important for lexical access (see Cole & Jakimik, 1980; Marslen-Wilson and Welsh, 1978). This notion and the general pattern of our results are consistent with the hypothesis that children at all age levels are sensitive to both consonant and vowel segments in syllables, but that the latter, presumably because of their acoustic properties are more perceptually salient. Thus, it may be that the kindergarteners' classifications are determined primarily by perceptual salience, but that the second graders must learn to ignore this saliency and selectively attend to initial consonants in order to accomplish one of the major tasks that they are faced with - namely, learning to read. Finally, the fifth graders appear to be influenced in their classifications by both the perceptual salience of the vowel segments and the knowledge that initial segments (here, consonants) are important. In any event, their performance in our task has the effect of making that of the kindergarteners appear more sophisticated than would otherwise be judged if fifth graders had not made vowel classifications. Indeed, in an experiment almost identical to the one reported here, but using synthetic speech, Jusczyk, Smith and Murphy (1980) found that even though adults can make consonant classifications, they prefer to make classifications on the basis of vowel information.

Our results indicate that young children even prior to knowing how to read are sensitive to the internal structure of syllables in terms of their phonemic segments. Thus, the underlying representation of speech at this age may be the same as that for older children. However, the sensitivity tapped in our task is not equivalent to the explicit and conscious manipulations required by the segmentation tasks used in previous studies and this could account for the stronger relationship of the latter task to reading ability. It seems important, nevertheless, to determine the nature and extent of the initial foundations of segmentation abilities. The initial sensitivity which eventually gives rise to segmentation abilities obviously places constraints on these abilities and the existence of this sensitivity serves to emphasize that segmentation abilities do not emerge from out of nowhere.

Our suggestion then is that young children, like older children, represent speech in terms of its component segments and that what changes in development is the accessibility of these segments for conscious manipulation. However, this developmental trend may not be motivated exclusively by learning to read and this could explain why there was no strong relationship between classification and reading performance in our task. According to Treiman and Baron, the development of phonemic perception is one manifestation of a more general trend that occurs in perceptual/cognitive development. However, some investigators of perceptual development do maintain that dimensions, such as color and form, become separable with age because of an increasing ability to access these underlying dimensions; i.e., the

dimensions of color and form are represented as such within the perceptual system and what develops is the ability to access and manipulate these dimensions separately. (See, for example, Smith & Kemler, 1978, as opposed to Shepp, 1978.) Our view of phonemic perception is in accord with this account of general cognitive development and in this respect it differs from the account offered by Treiman and Baron.

In summary, we have provided evidence for the existence of phonemic structure in the young child's internal representation of speech. Our results are, therefore, also consistent with the assumptions made by other researchers concerning the presence of phoneme-like units in the speech processing mechanism (e.g., Gleitman & Rozin, 1977; Liberman et al., 1977). However, unlike the position taken by these investigators, we have also shown that even young prereaders have some access to this structure. With respect to the causal relationship between phonemic perception and reading, this suggests the viability of a third view of their relationship. Rather than phonemic perception providing the basis for reading or vice versa, it may be that the rudimentary sensitivity to the component segments of speech which we observed provides part of the initial basis for learning to read. This in turn may support an increased ability to overtly access and manipulate phonemes. Phonemic segmentation abilitites and the ability to read may, therefore, develop in a highly interactive manner (see Brown, 1978, for a discussion of such interactions between and the increasing co-ordination of metaknowledge and knowledge representations).

References

Brown, A. L. Knowing when, where, and how to remember: A problem of metacognition. In R. Glasser (Ed.), Advances in Instructional Psychology, Vol. I. Hillsdale, N. J.: Lawrence Erlbaum Associates, Inc., 1978.

Bruce, D. J. The analysis of word sounds by young children. British Journal of Educational Psychology, 1964, 34, 158-170.

Calfee, R., Chapman, R. & Venezky, R. How a child needs to think to learn to read. In L. W. Gregg (Ed.), Cognition in Learning and Memory. New York: John Wiley & Sons, Inc., 1972.

Cole, R. A. & Jakimik, J. A model of speech perception. In R. Cole (Ed.), Perception and Production of Fluent Speech. Hillsdale, N. J.: Lawrence Erlbaum Associates, Inc., 1980.

Eilers, R. E. Context sensitive perception of naturally produced stop and fricative consonants by infants. Journal of the Acoustical Society of America, 1977, 61 (5), 1321-1336.

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

Foss, D. J. & Blank, M. A. Identifying the speech codes. Cognitive Psychology, 1980, 1, 1-31.

Gibson, E. J. Principles of Perceptual Learning and Development. New York: Appleton-Century-Crofts, 1969.

Gleitman, L. R. & Rozin, P. The structure and acquisition of reading I: Relations between orthographies and the structure of language. In A. S. Reber & D. L. Scarborough (Eds.), Toward a Psychology of Reading. Hillsdale, N. J.: Lawrence Erlbaum Associates, Inc., 1977.

Jusczyk, P. W., Smith, L. B. & Murphy, C. The perceptual classification of speech and nonspeech sounds. Paper presented at the 99th Meeting of the Acoustical Society of America in Atlanta, Georgia on April 22, 1980.

Kemler, D. G. & Smith, L. B. Accessing similarity and dimensional relations: Effects of integrality and separability on the discovery of complex concepts. Journal of Experimental Psychology: General, 1979, 108 (2), 133-150.

Kewley-Port, D. Continuous spectral change as acoustic cues to place of articulation. Research on Speech Perception Progress Report No. 5, Indiana University, 1979, 327-346.

Klatt, D. H. Speech perception: A model of acoustic-phonetic analysis and lexical access. In R. Cole (Ed.), Perception and Production of Fluent Speech. Hillsdale, N. J.: Lawrence Erlbaum Associates, Inc., 1980.

Kuhl, P. K. Speech perception in early infancy: The acquisition of speech-sound categories. In S. K. Hirsh, D. H. Eldredge, I. J. Hirsh & S. R. Silverman (Eds.), Hearing & Davis: Essays Honoring Hallowell Davis. St. Louis, Mo.: Washington University Press, 1976.

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

Liberman, I. Y., Shankweiler, D., Fischer, F. W. & Carter, B. Explicit syllable and phoneme segmentation in the young child. Journal of Experimental Psychology, 1974, 18, 201-212.

Liberman, I. Y., Shankweiler, D., Liberman, A. M., Fowler, C. & Fischer, F. W. Phonetic segmentation and recoding in the beginning reader. In A. S. Reber & D. L. Scarborough (Eds.), Toward a Psychology of Reading. Hillsdale, N. J.: Lawrence Erlbaum Associates, Inc., 1977.

Makita, K. The rarity of reading disability in Japanese children. American Journal of Orthopsychiatry, 1968, 38, 599-614.

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

Miller, G. A. Reflections on the conference. In J. F. Kavanagh & I. G. Mattingly (Eds.), Language by Ear and by Eye. Cambridge, Mass.: M.I.T. Press, 1972.

Morais, J., Cary, L., Alegria, J. & Bertelson, P. Does awareness of speech as a sequence of phones arise spontaneously? Cognition, 1979, 7, 323-331.

Read, C. Children's awareness of language, with emphasis on sound systems. In A. Sinclair, R. J. Jarvella & W. J. M. Levelt (Eds.), The Child's Conception of Language. New York: Springer-Verlag, 1978.

Read, C. Children's judgements of phonetic similarities in relation to English spelling. Language Learning, 1973, 23, 17-38.

Rozin, P. & Gleitman, L. R. The structure and acquisition of reading II: The reading process and the acquisition of the alphabetic principle. In A. S. Reber & D. L. Scarborough (Eds.), Toward a Psychology of Reading. Hillsdale, N. J.: Lawrence Erlbaum Associates, Inc., 1977.

Rozin, P., Poritsky, S. & Sotsky, R. American children with reading problems can easily learn to read English represented by Chinese characters. Science, 1971, 171, 1264-1267.

Savin, H. B. What the child knows about speech when he starts to learn to read. In J. F. Kavanagh & I. G. Mattingly (Eds.), Language by Ear and by Eye. Cambridge, Mass.: M.I.T. Press, 1972.

Searle, C. L., Jacobson, J. E. & Rayment, S. G. Phoneme recognition based on human audition. Journal of the Acoustical Society of America, 1979, 65, 799-809.

Shepp, B. E. From perceived similarity to dimensional structure: A new hypothesis about perceptual development. In E. Rosch and B. B. Lloyd (Eds.), Cognition and Categorization. Hillsdale, N. J.: Lawrence Erlbaum Associates, Inc., 1978.

Smith, L. B. & Kemler, D. G. Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. Journal of Experimental Child Psychology, 1977, 24, 279-298.

Smith, L. B. & Kemler, D. G. Levels of experienced dimensionality in children and adults. Cognitive Psychology, 1978, 10, 502-532.

Stevens, K. N. & Blusmtein, S. E. Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America, 1978, 64 (5), 1358-1368.

Stevens, K. N. & Blumstein, S. E. Perceptual invariance and onset spectra for stop consonants in different vowel environments. Journal of the Acoustical Society of America, 1980, 67 (2), 648-662.

Treiman, R. & Baron, J. Segmental analysis ability: Development and relation to reading ability. In T. G. Waller & G. E. MacKinnon (Eds.), Reading Research: Advances in Theory and Practice, Vol. 2. New York: Academic Press, 1980.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 6 (1979-80) Indiana University]

Young Children's Understanding of Ambiguous Sentences

Beth G. Greene and David B. Pisoni

Department of Psychology
Indiana University
Bloomington, Indiana 47405

Short Title: Ambiguous Sentences

Abstract

Several earlier studies have reported that children appear to be unable to detect surface structure and deep structure ambiguities in sentences before the age of twelve. Since school age children (6 - 12 years old) routinely deploy their linguistic knowledge in complex linguistic tasks such as reading and writing, it seems quite likely that they should be able to detect the presence of ambiguity in sentences. The present study was designed to assess the linguistic abilities of young children to recognize sentence ambiguity by means of a forced-choice picture pointing task. Three types of sentence ambiguity were examined: lexical ambiguity, surface structure ambiguity and deep structure ambiguity. Fifty-four subjects ages 3.5 to 5.6 years of age were instructed to point to the two pictures that showed what the machine was talking about. For each test sentence, two pictures in the display were correct and two were not. Overall performance in detecting both interpretations of an ambiguous sentence was 47% correct (chance = 16.7%). Older subjects performed better than younger subjects. These findings suggest that earlier conclusions regarding young children's understanding of ambiguities at these three different levels of linguistic structure have substantially underestimated their linguistic abilities. Much more attention should be directed to the specific methodologies used in studies assessing the linguistic competence of young children and the types of linguistic and metalinguistic knowledge needed to perform the task.

Young Children's Understanding of Ambiguous Sentences

Beth G. Greene and David B. Pisoni

Every native speaker of a language has the ability to determine the "sentencehood" of a potentially infinite number of novel sentences; that is they can recognize the grammatical sentences of the language and discriminate these from the ungrammatical sentences. Moreover, adult native speakers are able to consciously identify and explain such linguistic phenomena as paraphrase, ambiguity and semantic anomaly and they can make numerous other sophisticated metalinguistic judgments about the sentences in their language. Unfortunately, current understanding of the metalinguistic abilities of young children is much less extensive at the present time primarily because it has been very difficult to interrogate and probe the linguistic intuitions of young children. Moreover, young children are often unable to consciously express introspective judgments about language that require access to their metalinguistic knowledge. In order to gain insights into the linguistic abilities of young children, special care must be devoted to developing appropriate paradigms and experimental techniques to reveal the nature of their linguistic competence.

One class of linguistic judgments that has been studied extensively in the adult psycholinguistic literature is sentence ambiguity (Garrett, 1970; Foss, Bever & Silver, 1968; Bever,

Garrett & Hurtig, 1973; Mistler- Lachman, 1972). Interest in the study of sentence ambiguity in adults derives primarily from the assumption that recognition of different types of ambiguities may reveal something about the order or sequence of processing stages in normal sentence comprehension (Foss, 1970; MacKay, 1966; MacKay & Bever, 1967). Unfortunately, the results of many of these adult studies on processing ambiguous sentences have not always led to the same conclusions primarily because of differences in experimental procedures and design.

Over the last decade several developmental studies of the perception of sentence ambiguity in young children have been reported (Kessel, 1970; Shultz & Pilon, 1973; Wiig, Gilbert & Christian, 1978). Like the earlier adult studies on processing ambiguous sentences, these studies have also yielded inconsistent and somewhat contradictory results.

In an early study, Kessel (1970) examined children's comprehension of linguistic ambiguity using a picture selection task followed by a series of free report questions designed to elicit additional information about the child's underlying thought processes (Piagetian interview). Lexical ambiguities were correctly interpreted by most six year olds (over 50%). Kessel stressed that this conclusion may be true only for the "lexical ambiguities used here" (p.51). The four lexical items used in his study were relatively easy to detect. Among the youngest subjects, both surface and underlying syntactic

ambiguities were detected equally often (about 30% corrrect). However, detection of syntactic ambiguity did not approach adult proficiency until age 12. Kessel concluded that interpretation of these types of ambiguities requires reflecting on or thinking about one's own comprehension, a cognitive process that marks the transition from the concrete operational stage to the stage of formal operations (cf. Piaget, 1967).

In several adult studies response latency data has been used as evidence for a hierarchy of processing stages. Lexical ambiguity is typically detected most rapidly followed in turn by surface structure ambiguity; deep structure ambiguities generally show the longest latencies (MacKay, 1966; MacKay & Bever, 1967). Longer response latencies are interpreted as reflecting more complex internal processing. Shultz & Pilon (1973) suggest that the pattern of latency results (lexical < surface < deep) may also reflect a developmental trend. In other words, more complex psychological processes may require more time to develop in children acquiring their language.

To examine this hypothesis, Shultz & Pilon (1973) presented four types of linguistically ambiguous sentences to children ranging in age from 6 to 15 years old. A free report paraphrase task was used followed by a picture selection (verification) task. In addition, the child was required to restate the sentence in his or her own words and justify his or her choice of the pictures. Shultz & Pilon felt that "mere pointing" was not a

sensitive enough measure of the child's understanding of the meaning(s) of the sentence. Therefore they required explicit verbal justification as described above. Results of their study indicated that detection of phonological ambiguity showed the greatest improvement between age 6 and 9. There was a steady linear increase with age in detecting lexical ambiguity. Deep and surface syntactic ambiguities were virtually not detected until age 12. Furthermore, the predicted ordering of detection abilities phonological < lexical < syntactic (surface and deep) was observed. Shultz & Pilon did not find any differences between the two types of syntactic ambiguity.

In a more recent study, Wiig et al. (1978) presented lexically and syntactically ambiguous sentences to kindergarten, second grade, fourth grade, sixth grade and college age students. Using a forced choice picture selection technique, they required subjects to choose two pictures for every sentence. An examination of the responses for the youngest subjects (kindergarten and second grade) who correctly chose both interpretations of an ambiguous sentence, showed that the lexically ambiguous sentences were correctly identified only 18% of the time whereas the syntactically ambiguous sentences were correctly identified 26% of the time. One explanation for the relatively poor performance of Wiig et al's subjects may lie in the experimental materials themselves. Many of their sentences were abstract and, even with the pictorial analogs, they appeared

very difficult for young children to interpret (e.g., The bill is large (15% correct); The duck is ready to eat (5% correct)). However, the forced-choice procedure used in this study demonstrated that young children could detect syntactic ambiguity at performance levels that were higher than previously reported.

Taken together, these results indicate that children do not appear to be able to detect sentence ambiguities until somewhere around the age of twelve. Such experimental results were surprising to us given the fact that young children are quite capable of recognizing and understanding a very large number of novel sentences in their linguistic environment. Moreover, school age children are often called upon routinely to deploy their linguistic knowledge in complex linguistic tasks such as reading and writing. While the experimental literature suggests that children are not capable of using their linguistic intuitions to recognize and understand ambiguous sentences, these findings may simply be due to the specific experimental methodology employed and may not be due to basic limitations on the linguistic abilities of young children or their underlying linguistic competence.

The present study was designed to assess the linguistic abilities of young children to recognize sentence ambiguity with a forced-choice pointing technique. The forced-choice procedure provides an opportunity to make a more quantitative assessment of the young child's linguistic intuitions without the effects of

bias that are present in typical paraphrase or free report procedures that have been used in past studies of ambiguity.

Method

Subjects. A total of 54 children served as subjects. The children attended a private cooperative nursery school in Bloomington, Indiana. Subjects ranged in age from 3.5 to 5.6 years of age. There were 30 males and 24 females. Subjects were divided into three groups of 18 children each: youngest (3.5 - 4.0), middle (4.1 - 4.8), and oldest (4.9 - 5.6).

Materials. A total of eighteen test sentences was developed using vocabulary items appropriate for this age group. Three types of sentence ambiguity were examined: (1) lexical ambiguity, (2) surface structure ambiguity, and (3) deep structure ambiguity. A sentence was considered lexically ambiguous if a word or sequence of words had two distinct meanings (e.g., The man is holding a pipe). Surface structure ambiguity involved the possibility of two distinct groupings or parsings of adjacent words in a sentence (e.g., They fed her dog biscuits). Deep structure ambiguities involved a change in the logical relations between words but not a change in meaning of individual words nor a change in the groupings of words (e.g., The mayor will ask the police to stop drinking).

The eighteen test sentences were recorded in two different random orders on audiotape by a female talker (BGG) in a flat monotone intonation using a high-quality microphone and tape recorder (Uhrer 4000 Report-L). For each ambiguous sentence, a set of four colored drawings were prepared. Two of the pictures represented the two alternate interpretations of the sentence; the two other pictures were designed as distractors. Each distractor picture contained appropriate visual content but the meaning of the picture was different from the two meanings of the sentence under test. Each drawing was mounted on a 5" X 8" (12.5 cm x 22 cm) piece of cardboard and labeled appropriately on the back to facilitate randomization during the experiment. For the test sentence, "The boy broke the glasses", the pictures showed: (1) a boy with a pair of broken eyeglasses, (2) a boy standing in the kitchen with broken drinking glasses on the floor at his feet, (3) a boy drinking from a glass, and (4) a boy with a broken lamp. The former two pictures were correct while the latter two were distractors.

Insert Figure 1 about here

Initially, a pool of eight sentences for each type of ambiguity was evaluated. The two alternate interpretations of each sentence and the four pictorial choices (two correct visual representations of the sentence and two distractor pictures) were

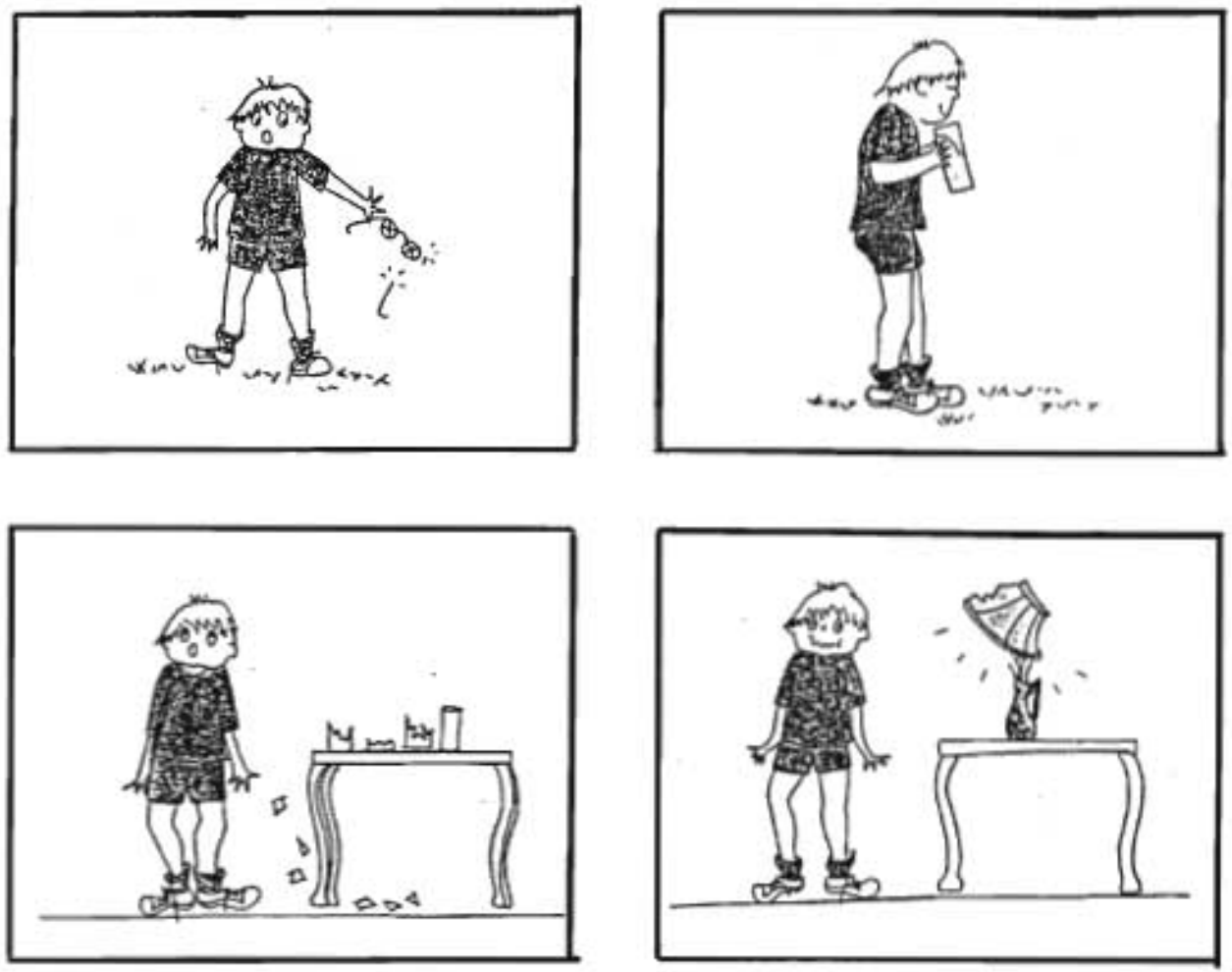


Figure 1

Example of placement of drawings on display board for the sentence "The boy broke the glasses".

evaluated by two independent raters. Both raters, graduate students in linguistics, assessed the content of the sentences in relation to the meaning expected and the drawings used to represent the sentences. Suggested modifications of the drawings were made and raters asked to reevaluate. Any remaining differences of opinion were resolved by mutual agreement. The best six sentences of each of the three types of ambiguity formed the eighteen experimental sentences used in this study.

Procedure. Each child met with the experimenter (BGG) alone in a large quiet classroom at the nursery school. Both were seated on the floor with a 36" X 36" (91.5 cm x 91.5 cm) display board between them and a high quality tape recorder (Uhrer 4000). Specific instructions to the children were as follows:

> We're going to play a game. I am going to show you some pictures. This machine is going to say some things to you. I want you to listen to what it says. Then I want you to show me the pictures it is talking about. I want you to show me two pictures every time. The machine will be talking about two pictures. Are you ready? Here are the pictures. Now listen to the machine. Ready? Listen.

Sentences were then presented one at a time via tape recorder and its internal loudspeaker. Each sentence was repeated twice. Prior to the test phase of the experiment, six practice trials were conducted. Four pictures were placed on the display board. Then the tape recorder was turned on and the word "ball" was presented twice. The child was told "Point to the two

pictures..." Three individual words and three simple unambiguous sentences were presented as test items. Each child was prompted if needed to insure that he/she fully comprehended the experimental tasks in the practice session which included listening to a stimulus, responding with a pointing response and then pointing to two pictures on every test trial. Each set of four pictures was arranged in a different random order on the display board for each subject prior to the presentation of each test sentence. If the child made only one pointing response, the experimenter gave the prompt, "Show me the two pictures the machine is talking about." Each subject's first and second response was recorded by letter/location on prepared response sheets. The experimental session lasted about 15-20 minutes. No time constraints were imposed nor was the child pressured to continue or complete the task. Three children had to be eliminated from the study after the six practice trials because they could not or would not perform the required picture pointing task. Two additional children who simply stopped performing the task during the course of the experiment were also eliminated from the study. Testing sessions took place over a period of two weeks.

Results. Performance in the forced-choice task was analyzed in two ways. First, we summarized the data on all trials in which subjects gave two correct responses to each test sentence. Summed over all ages and all test sentences, performance for two

correct responses was .47, a value that is significantly above chance expectation ($Z=15.37$, $p<.0001$). With four alternatives taken two at a time, there are six possible pairwise arrangements of which only one is correct for a chance probability of 1/6 or .167. When the two correct responses were subsequently broken down by experimental conditions and analyzed by an analysis of variance, we found a significant age effect ($F(2,51)= 5.76$, $p<.01$). Older subjects performed better than younger subjects. However, there were no other overall statistically significant differences between the three types of ambiguities nor was there a significant interaction between age and ambiguity type for the measure involving two correct responses.

In the second analysis, we examined trials on which only the first response was correct; that is, responses indicating that the child perceived only one interpretation for each test sentence. An analysis of variance on these first correct data showed a main effect for age ($F(2,51)= 3.83$, $p<.05$). In this case, however, the younger children showed more first correct response than the older children. This apparent reversal is due entirely to the greater number of two correct responses observed for the older group of subjects. The main effect for ambiguity type and its interaction with age were also not significant in this analysis.

Insert Table 1 about here

The eighteen sentences used in this study are shown in Table 1 along with the percentage of subjects who selected both correct pictures for each test sentence. It can be seen from Table 1 that two of the experimental sentences were responded to quite differently from the remaining eighteen sentences. The sentence, "They gave Mother the flower(flour)", elicited very few both correct responses. On the other hand, almost all subjects selected two correct responses for the sentence, "They are visiting firemen." One straightforward reason for these two anomolies lies in the set of drawings we used to represent the sentences. In the case of "flower-flour", the drawing representing the "flour" interpretation may not have been understood by the children since it was rarely selected. For the sentence, "They are visiting firemen", neither distractor picture portrayed firemen so very few children selected the distractors. The distractor pictures showed children with clowns and with baseball players and therefore could be discriminated quite easily from the two correct pictures.

Discussion

In the early studies of detection of linguistic ambiguity by young children investigators focused on uncovering the parameters

Table 1

Percentage of Subjects Who Selected Both Correct Pictures for Each Sentence

	SENTENCE	PERCENTAGE
Lexical		
	The book was read(red).	37
	They gave Mother the flower(flour).	11
	Bill is drawing his gun.	31
	John paints a house.	69
	The boy broke the glasses.	57
	The man is holding a pipe.	57
Surface Structure		
	He fed her dog biscuits.	43
	The man saw the big boy and girl.	37
	There is a hot dog.	39
	Big cats and dogs chase birds.	56
	He told her baby stories.	48
	The boy saw the tree in the yard.	57
Underlying Structure		
	Father chased the boy in his pajamas.	46
	The boy hit the girl with the books.	33
	The boy chased the girl on the bicycle.	41
	They are visiting firemen.	93
	They are flying planes.	46
	The lady is washing.	52

of linguistic competence and performance. These studies were based on previous findings obtained with adult subjects (Kessel, 1970; Shultz & Pilon, 1973; Wiig et al., 1977). The study carried out by Kessel in 1970 is frequently cited in the literature as the first investigation of young children's understanding of linguistic ambiguity. Subsequent investigations of ambiguity have been undertaken to explain or elaborate on Kessel's results or to provide some theoretically based hypotheses that can be tested with children at various points in the language acquisition process. Since we reviewed these early results in the introduction, we will simply summarize the major conclusion of these studies here, namely, that young children do not appear to be able to detect syntactic ambiguities until the age of twelve. The youngest subjects that were able to detect lexical ambiguity with any degree of confidence appear to be about age six. The results of the present study demonstrate clearly that children between three and six can and do detect both types of ambiguities well above chance expectation and well above levels previously reported. The reasons for these relatively high levels of performance in this study will be discussed below since we believe them to be methodological in nature. However, there are several other important issues that need to be considered in research dealing with the comprehension of ambiguity by young children.

Several more recent studies have examined linguistic ambiguity in a different manner (Evans, 1977; Hirsh-Pasek, Gleitman & Gleitman, 1978; Hakes, 1980). Unlike the earlier studies cited above which focused on ambiguity simply as a class of sentences to be studied, these more recent investigations examined ambiguity within the framework of the child's developing metalinguistic abilities. That is, ambiguous sounds, words and sentences have been studied recently because these types of linguistic materials may reveal information about how the child comes to gain conscious access to his/her knowledge of the language in tasks requiring more than simple forced-choice responses.

In a study of children's understanding of sentence ambiguity, Evans (1977) proposed that understanding sentence ambiguity is one of a number of cognitive, verbal and perceptual changes that occur between ages six and eight. This particular point in development corresponds to Piaget's transition from the preoperational to the concrete operational stage of cognitive development. The ability to understand riddles is one of these changes. Preoperational children do not "get the joke" in verbal riddles that rely on ambiguity while concrete operational children do. For example, Evans notes that the riddle,

Question: What has 18 legs and catches flies?
Answer: A baseball team.

relies on the lexically ambiguous word flies. Since the listener is led to interpret the word flies as meaning "insects", the riddle becomes funny because the meaning of flies must be shifted to "batted balls" to get the joke (Evans, 1977).

Hirsh-Pasek, Gleitman and Gleitman (1978) also examined children's detection and conscious report of ambiguity. In this study, verbal riddles and jokes served as the experimental materials to examine understanding of various types of ambiguity in children from grade one through grade six. Their results indicated that the ability to detect ambiguity emerges much earlier than the ability to explain the ambiguity in explicit terms. The youngest children, between six and seven years of age, were unable to provide adequate verbal responses for most types of ambiguity while the oldest group, between ten and eleven years of age, performed adequately. While these findings are not surprising nor are they different from earlier results, Hirsh-Pasek et al. attribute the developmental changes to growth in metalinguistic abilities rather than linguistic ability. That is, the child shows both competence and performance for a particular linguistic structure yet is unable to consciously explain the basis of the ambiguities.

A more recent series of studies carried out by Hakes and his colleagues (Hakes, 1980) further explores the development of metalinguistic ability in children. These studies focused on the period of middle childhood. Hakes views this time as one of

transition where there is an emergence of general metalinguistic ability (p. 100). Children ranging in age from four to eight years of age performed tasks involving conservation, comprehension, synonymy, acceptability, and phonemic segmentation. As expected, developmental differences were found between the younger children and older children and the division seemed to be ordered along the preoperational- concrete operational dimension. Examination of the interrelationships among the various task results were interpreted as support for the notion that there is a single, underlying developmental change that results in the emergence of general metalinguistic abilities. Furthermore, these changes appear to correspond to the major shift observed in the level of cognitive development.

Taken together the recent studies by Evans, Hirsh-Pasek et al., and Hakes reveal that young children have a substantial capacity to use metalinguistic knowledge. Children acquire language over a period of several years and throughout the time course of development many related cognitive and perceptual abilities develop as well. It is of interest and importance particularly in terms of assessing linguistic competence in young children to recognize that metalinguistic abilities also follow a developmental time course. Such abilities may therefore interact with the particular types of experimental tasks and procedures used to assess the child's linguistic abilities. Thus, it is not surprising to us that the earlier results and conclusions about

young children's understanding of ambiguity were so inconsistent with each other. Moreover, the differences in experimental materials and specific tasks across studies also contributed to failure to find clear cut agreement among investigators.

Another aspect of children's understanding of ambiguous sentences that must be addressed is the methodology used in studies of this kind. As mentioned above, the earlier studies used a variety of procedures as well as a variety of stimulus materials (see Evans, 1977, for a detailed summary of these procedures as well as a critical evaluation). Initially, the present study was undertaken to eliminate the response biases present in unstructured free report techniques. Young children had been shown to be unable to detect more than one interpretation of an ambiguous sentence (Kessel, 1970; Shultz & Pilon, 1973; Wiig et al., 1977). Even the recent studies of Brause (1977), Evans (1977), Hirsh-Pasek et al. (1978), and Hakes (1980) report that children age six and seven are generally unable to detect linguistic ambiguity. The procedure we used was a forced-choice pointing response that required that the child make two responses on every trial. It is important to emphasize here that the present study was carried out with children who were much younger than most of the children studied in earlier investigations. With the exception of Hakes (1980) the subjects in all of the earlier studies of ambiguity were of kindergarten age and above. Our results indicate that very young children are

able to detect linguistic ambiguity when placed in a forced-choice situation. The children were provided with an experimental task that limited the response possibilities and therefore reduced the uncertainty and complexity of the problem to be solved. Since the children were given very explicit instructions and some practice with the task, they were quite willing and able to respond to more than one picture on each trial. In other words, the task demands were well within their perceptual and cognitive abilities. Thus, given these considerations, it is not surprising that our results were considerably better than previous reports and considerably above chance expectation in a task such as this.

The nature of the experimental materials employed in studies of this type must also be considered. As noted earlier, the first studies of detection of linguistic ambiguity were modeled after studies conducted with adult subjects. The experimental materials used in the developmental studies were derived from those used with the adult subjects, often with little regard to the lexical knowledge of the children. Therefore, it is not very surprising to find that young children were unable to detect linguistic ambiguity. When the experimental materials were selected to be more appropriate for testing children, the level of performance improved (Evans, 1977; Hirsh-Pasek et al., 1978). Thus, performance on the task interacts with the kind of knowledge the child needs to use to perform the task. This

knowledge can be linguistic or metalinguistic.

There can be no question that metalinguistic knowledge plays an important role in young children's understanding of linguistic ambiguity, particularly in tasks that require explicit and conscious explanation of the ambiguity under consderation. A child who is at a certain level of cognitive development, i.e., preoperational, cannot be expected to succeed in a task that requires cognitive abilities that have not yet developed. Thus, young children age six and under can detect linguistic ambiguities but they are not yet able to provide explicit descriptions of their reasons for choosing one response over another. Therefore they do poorly in free report and interrogation tasks which demand such abilities and they fail to understand jokes and riddles that rely on ambiguity. When the task requirements limited them to only detecting the ambiguity using the forced-choice pointing response, the level of performance increased dramatically. After the child has made the transition from preoperational to concrete operational, improved performance in free report tasks can be expected because the child has the necessary cognitive abilities to explain why the sentence is ambiguous or why the joke is funny. That is, the child can dissociate himself from control of language and consider it as a formal object of study. Detecting the ambiguity relies on tacit knowledge of the language much like the knowledge utilized in speaking and listening, something young children do

with relative ease. On the other hand, explaining ambiguity relies on a more controlled, explicit and conscious process, a process that relies heavily on metalinguistic ability (see Hirsh-Patek et al.). Thus, as metalinguistic ability develops the ability to explain ambiguity improves. And at some point, children reach the stage in development of metalinguistic ability that allows them to reflect on the structure of language and how it is used. That is, they need to gain control over these strategies and essentially be able to step back from a situation and think about it (Hakes, 1980).

In the specific case of understanding ambiguity, the child must be able to switch from one interpretation of an ambiguous sentence to the other in order to fully appreciate the different meanings involved. Younger children appear to be unable to do this and therefore cannot explain their own responses when asked to in free report or interview tasks of the kind used in earlier studies of ambiguity. As metalinguistic abilities develop, the capacity to switch from one meaning to another also develops and the child is able to both detect the ambiguity and explain it in explicit terms to the experimenter.

Summary

In summary, the main results of this study demonstrate quite clearly that very young children between the ages of three and six years are capable of understanding ambiguous sentences at

levels well above chance expectation when provided with appropriate alternative responses in a forced-choice testing situation. The results of this study indicate that very young children of this age have quite sophisticated linguistic abilities which enable them to detect ambiguities in sentences at three distinct levels of analysis. Our findings suggest that young children's linguistic abilities have been substantially underestimated by previous reports in the developmental literature primarily because of a failure to recognize the differences between linguistic and metalinguistic abilities involved in carrying out the required experimental tasks.

References

Bever, T.G., Garrett, M.F., & Hurtig, R. The interaction of perceptual processses and ambiguous sentences. Memory and Cognition, 1973, 1, 277-286.

Brause, R.S. Developmental aspects of the ability to understand semantic ambiguity. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

Brodzinski, D.M., Feuer, V., & Owens, J. Detection of linguistic ambiguity by reflective, impulsive, fast/accurate, and slow/inaccurate children. Journal of Educational Psychology, 1977, 69, 3, 237-243.

Evans, J.R. Children's comprehension and processing of ambiguous words in sentences. (Doctoral dissertation, The University of Texas at Austin, 1976) Dissertation Abstracts International, 1977, 37, 6365-B. (University Microfilms No. 77-11,511)

Foss, D.J. Some effects of ambiguity upon sentence comprehension. Journal of Verbal Learning and Verbal Behavior, 1970, 9, 699-706.

Foss, D.J., Bever, T.G., & Silver, M. The comprehension and verification of ambiguous sentences. Perception and Psychophysics, 1968, 4, 304-306.

Garrett, M.F. Does ambiguity complicate the perception of sentences? In G.B. Flores d'Arcais & W.J.M. Levelt (Eds.), Advances in psycholinguistics. Amsterdam: North-Holland, 1970.

Hakes, D.T. The development of metalinguistic abilities in children. Berlin: Springer-Verlag, 1980.

Hirsh-Pasek, K., Gleitman, L.R., & Gleitman, H. What does the brain say to the mind? A study of the detection and report of ambiguity by young children. In A. Sinclair, R.J. Jarvella, & W.J.M. Levelt (Eds.) The child's conception of language. Berlin: Springer-Verlag, 1978.

Kessel, F.S. The role of syntax in children's comprehension from ages six to twelve. Monographs of the Society for Research in Child Development , 1970, 35 (Whole No. 6).

MacKay, D.G. To end ambiguous sentences. Perception and Psychophysics, 1966, 1, 426-236.

Mistler-Lachman, J.L. Levels of comprehension processing of normal and ambiguous sentences. Journal of Verbal Learning and Verbal Behavior, 1972, 11, 614-623.

Piaget, J. Six psychological studies. New York: Random House, 1967.

Shultz, T.R. & Pilon, R. Development of the ability to detect linguistic ambiguity. Child Development, 1973, 44, 728-733.

Wiig, E.H., Gilbert, M.F., & Christian, S.H. Developmental sequences in perception and interpretation of ambiguous sentences. Perceptual and Motor Skills, 1978, 46, 959-969.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 6 (1979-80) Indiana University]

Perception of the duration of rapid spectrum changes: Evidence for context effects with speech and nonspeech signals*

T. D. Carrell, D. B. Pisoni and S. J. Gans

Department of Psychology
Indiana University
Bloomington, Indiana 47405

Abstract

For a number of years investigators have focused attention on the effects one acoustic segment has on the perception of other acoustic segments. In one recent study, Miller and Liberman [Percept. & Psychophy. 25 (6), 457-465 (1979)] reported that overall syllable duration influences the location of the labeling boundary between the stop [b] and the semivowel [w]. They claim that this "context effect" reflects a form of perceptual normalization whereby the listener somehow readjusts his perceptual apparatus to take account of the differences in rate of articulation of the talker. More recently, Eimas and Miller [Science, 209 (5), 1140-1141 (1980)] have reported that prelinguistic infants also show similar context effects in discrimination of these stimuli. The inference to be drawn from the latter study is that infants perceive these speech sounds like adults and that such context effects reflect the operation of perceptual mechanisms that underlie a phonetic-like mode of processing specific to speech signals. In the present paper, we report the results of several critical comparisons between speech and nonspeech signals. We observed comparable context effects for perception of the duration of rapid spectrum changes as a function of overall duration of the stimulus. Our results with these nonspeech control signals therefore falsify both of the earlier claims by demonstrating clearly that context effects of the kind described by Miller and Liberman are not peculiar to the perception of speech signals or to normalization of speaking rate by the listener. Rather, such context effects may simply reflect general psychophysical principles that influence the perceptual categorization and discrimination of all acoustic signals whether speech or nonspeech.

*This is a draft of a paper to be presented at the 100th meeting of the Acoustical Society of America, November 19, 1980, Los Angeles, California. The research reported here was supported in part by NIMH research grant MH-24027 to Indiana University. We thank Paul Luce for his assistance in running subjects and Tom Jonas for his help in programming.

For many years investigators have been interested in how one phonetic segment affects the perception of adjacent segments in the speech signal. This form of "context conditioned variability" has been of major theoretical interest in the past and, despite some thirty years of research, it still continues to occupy the attention of many researchers even today. While some formal attempts have been made to deal with this problem by offering theoretical accounts of speech perception framed in terms of motor theories, analysis-by-synthesis or feature detector models of recognition, there is still no satisfactory solution to the problem of how listeners compensate for the extensive context-conditioned variability observed in the speech signal. These various forms of variability arise from several sources including: the effects of the immediately surrounding phonetic context, differences in speaking rate, differences in talkers and the variability of segmental durations that is conditioned by the syntactic and semantic structure of sentences. Indeed, this particular problem has been so elusive that some investigators such as Klatt have proposed to solve it by completely denying that the problem exists at all. Instead, Klatt has proposed a set of context-sensitive spectral templates that directly encode the acoustic-phonetic variability of phonemes in different phonetic environments. Words are recognized directly from sequences of these spectral templates without an intermediate level corresponding to a phonetic or phonemic representation.

In this paper we are interested in the following question: To what extent is the observed perceptual compensation carried out by perceptual mechanisms that are specific to the processing of speech signals? We became interested in this problem after a report by Miller and Liberman (1979) who found that the perceptual boundary between [b] and [w] was influenced by the nature of the context immediately following the critical acoustic cues to the stop vs. semi-vowel distinction -- namely the duration of the formant transitions at stimulus onset. They reported that the duration of the vowel in a CV syllable systematically influences the perception of the formant transition cues. With short syllables, subjects required shorter transition durations to perceive a [w] than with longer syllables. Miller and Liberman presented these results as a demonstration of perceptual normalization for speaking rate -- listeners adjust their decision criteria to compensate for the differences in vowel length that are conditioned by the talker's speaking rate. It is well known, for example, that vowels become much shorter when speaking rate increases. Thus, according to Miller and Liberman's account, the listener interprets a particular set of acoustic cues, say for a [b] or a [w], in relation to the talkers speaking rate rather than by reference to some absolute set of context invariant acoustic attributes in the stimulus itself. In their study, the location of the boundary for a syllable-initial [b-w] contrast was determined by the duration of the syllable containing the target phoneme.

The particular claims surrounding perceptual compensation or adjustment for speaking rate have taken on even more significance with the very recent report in Science by Eimas and Miller (1980).

They found that young prelinguistic infants also perceive these same [b-w] stimuli in a "relational manner" that is similar to the earlier data obtained with adult listeners by Miller and Liberman. Moreover, these results, like the ones obtained with adults, were interpreted as support for the argument that speech is not processed in a strictly left-to-right linear fashion one phoneme at a time. Rather, phonetic perception requires the integration of numerous widely distributed acoustic cues. As Miller and Liberman (1979) have put it:

> "The duration and structure of the syllable provide information about rate, and the listener uses this information when making a phonetic judgement of [b] and [w]."

We wanted to know to what extent these two sets of findings were a consequence of perceptual mechanisms that are specific to processing speech signals. That is, are these findings with adults and infants a result of phonetic categorization per se or are they due to more general factors related to auditory perception of both speech and nonspeech signals? In order to answer this question, we generated several sets of nonspeech control stimuli. These stimuli preserved all of the durational cues contained in the speech stimuli but, nevertheless, they did not sound like speech. We also generated several sets of synthetic speech stimuli which were modelled after the specifications provided by Miller and Liberman to see if we could replicate their earlier findings with adults. Experiments using both speech and nonspeech signals are now underway in our laboratory using young infants as well, although we will not report on these findings here today.

May I have Slide 1 please?

Slide 1 shows schematic representations of the formant motions of the endpoint stimuli we used. In the top panel two sets of stimuli are shown, long CVs and short CVs. The stimuli in each set contained identical formant transitions and differed only in the overall duration of the syllable. The stimuli were based on the values provided in Miller and Liberman's report. The long CV stimuli were 295 msec in duration; the short stimuli were 80 msec. We also generated several other sets of CVC stimuli which contained formant transitions appropriate for a /d/ in syllable-final position. These stimuli are shown in the bottom panel of this figure.

We then synthesized four sets of speech stimuli corresponding, in turn, to each of the four cells in the figure. For each set there were eleven test stimuli in which the duration of the formant transitions was varied from 15 msec to 65 msec in 5 msec steps. We also generated four additional sets of nonspeech stimuli by digitally manipulating three sinusoids to follow the same durations and formant trajectories of the speech patterns shown in this figure.

May I have Slide 2 please?

Slide 2 shows the results of a replication of the Miller and Liberman findings that the overall duration of the syllable affects the locus of the boundary for the [b-w] contrast. The

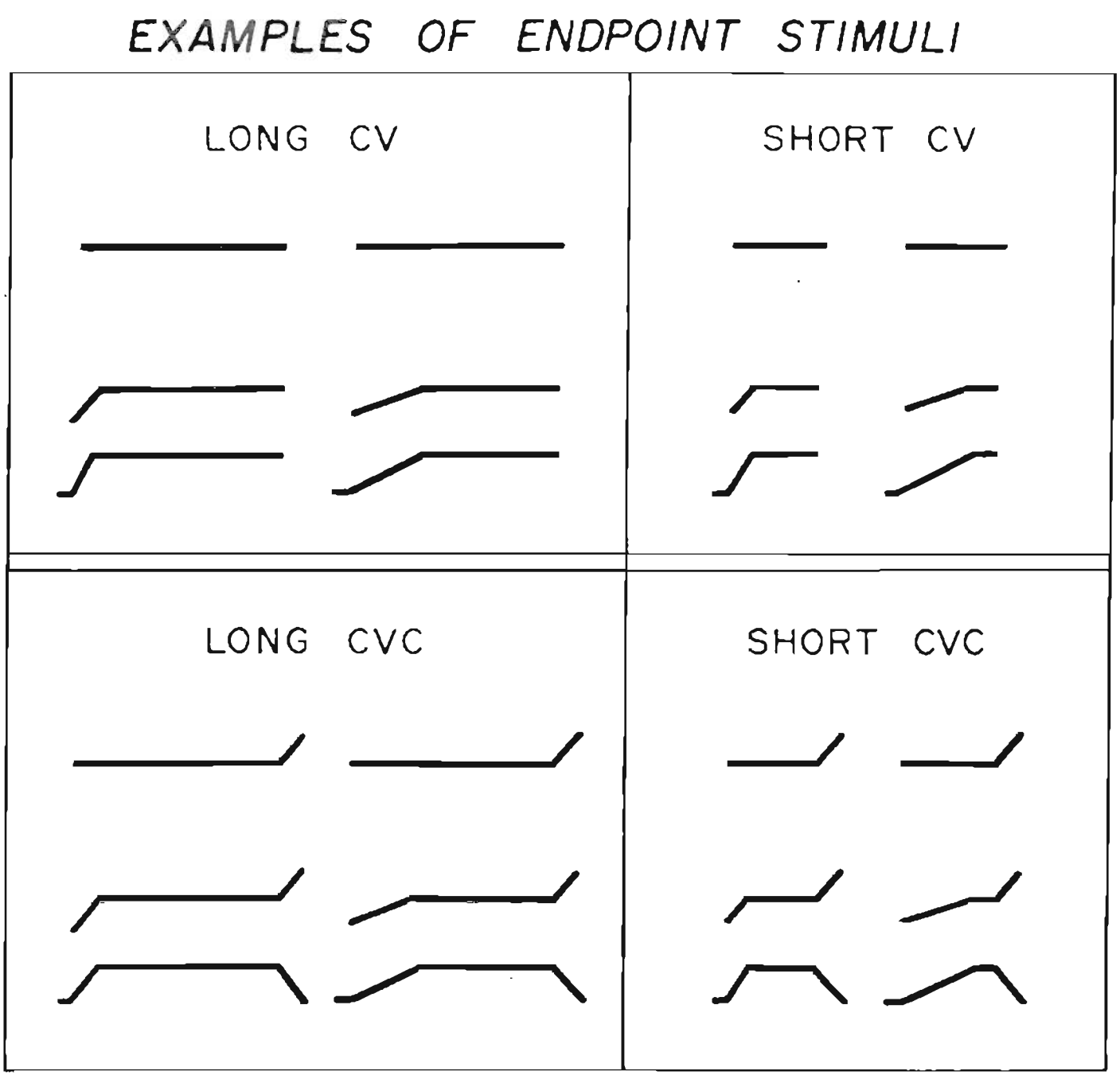

EXAMPLES OF ENDPOINT STIMULI
LONG CV
SHORT CV
LONG CVC
SHORT CVC

FIGURE 1.

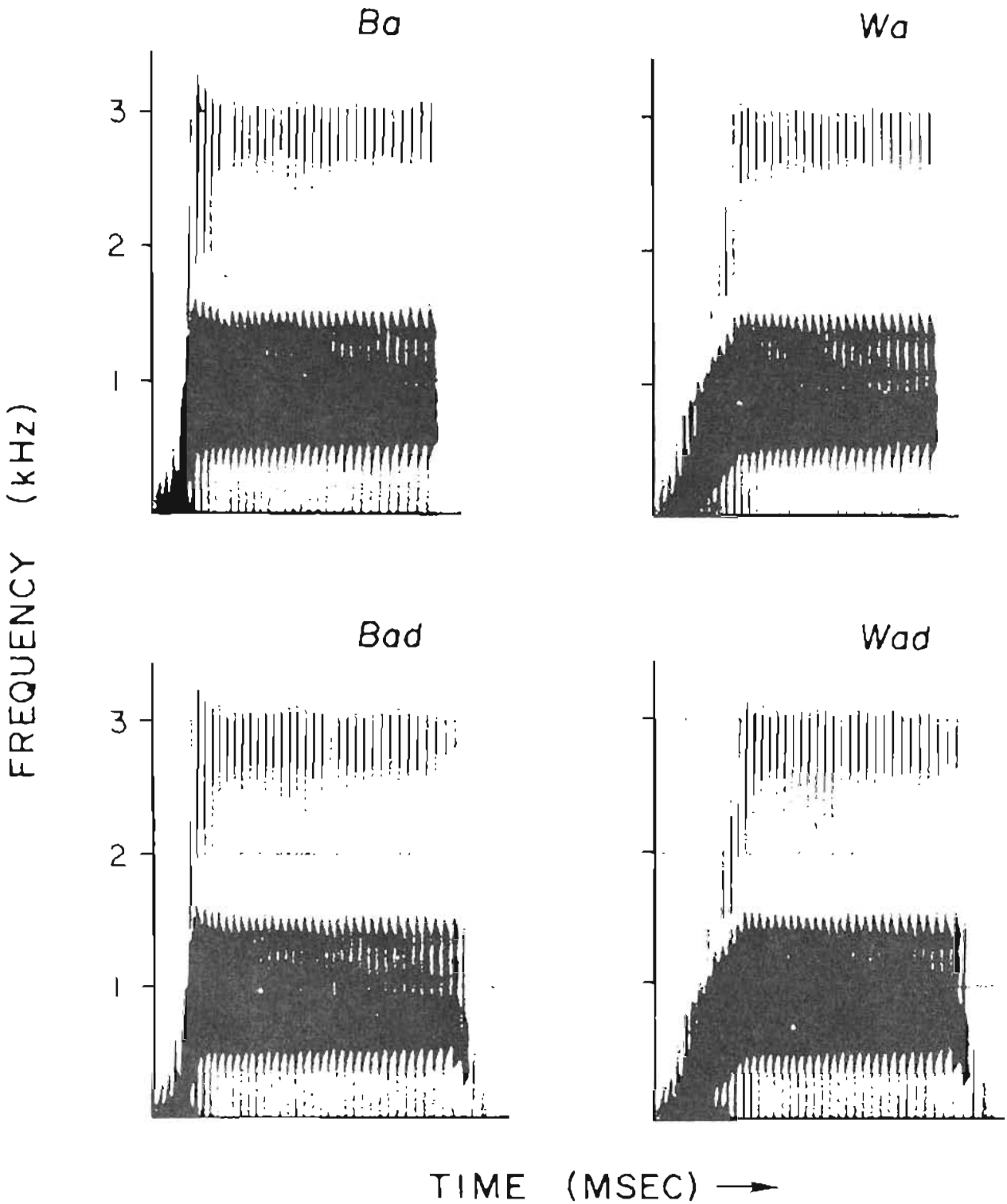


FIGURE 1A.

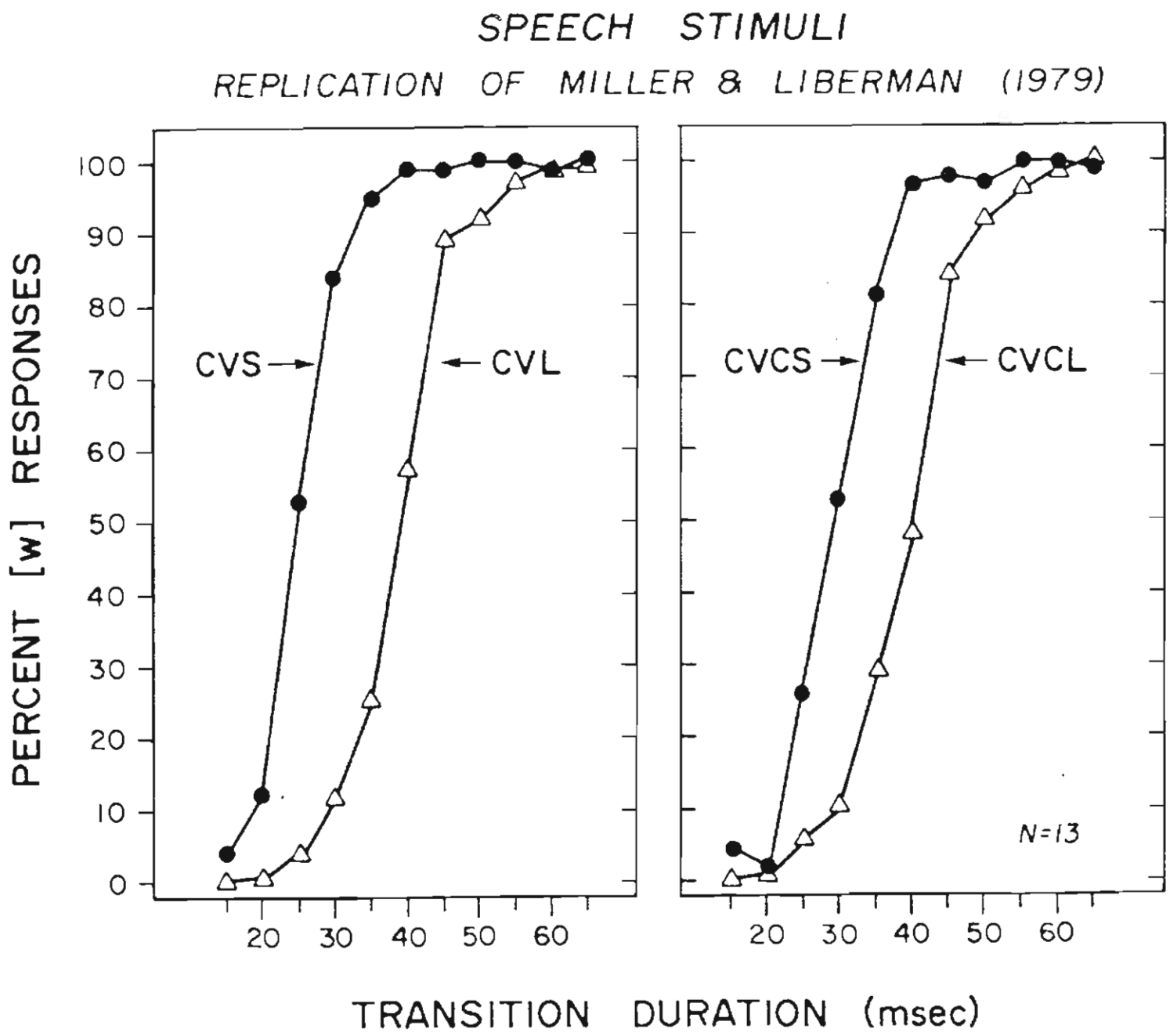

SPEECH STIMULI
REPLICATION OF MILLER & LIBERMAN (1979)
PERCENT [w] RESPONSES
100
90
80
70
60
50
40
30
20
10
0
CVS
CVL
CVCS
CVCL
N=13
20
30
40
50
60
TRANSITION DURATION (msec)

FIGURE 2.

results for the CV stimuli are shown in the left panel, the results for the CVC stimuli are shown in the right panel. In both cases, the boundary for the short stimuli in each set is displaced toward the left relative to the corresponding long stimuli. These findings are highly significant and very consistent across the same group of thirteen subjects who showed the effect in both CV and CVC conditions.

May I have Slide 3 please?

Slide 3 shows another set of data collected with the same speech stimuli but now with separate groups of subjects in each condition. Panels A and B replicate the findings we saw in the previous slide. That is, a consistent shift can be observed in the [b-w] labelling boundary as a function of syllable duration.

Panel C in this figure also displays the results for cross-series comparisons between CV and CVC syllables, of equal duration. Miller and Liberman reported that the context effects are not only due to the duration of the syllable but also to the internal structure of the syllable. When Miller and Liberman added formant transitions to their CV syllables, the perceptual boundary shifted in the direction opposite to that observed when the steady-state portion of the vowel was lengthened in a CV syllable. We also replicated this same effect in Panel C of this figure although the result is somewhat weaker than the main effects shown in panels A and B.

Having replicated the basic effect reported by Miller and Liberman with speech stimuli, we turned our attention to the

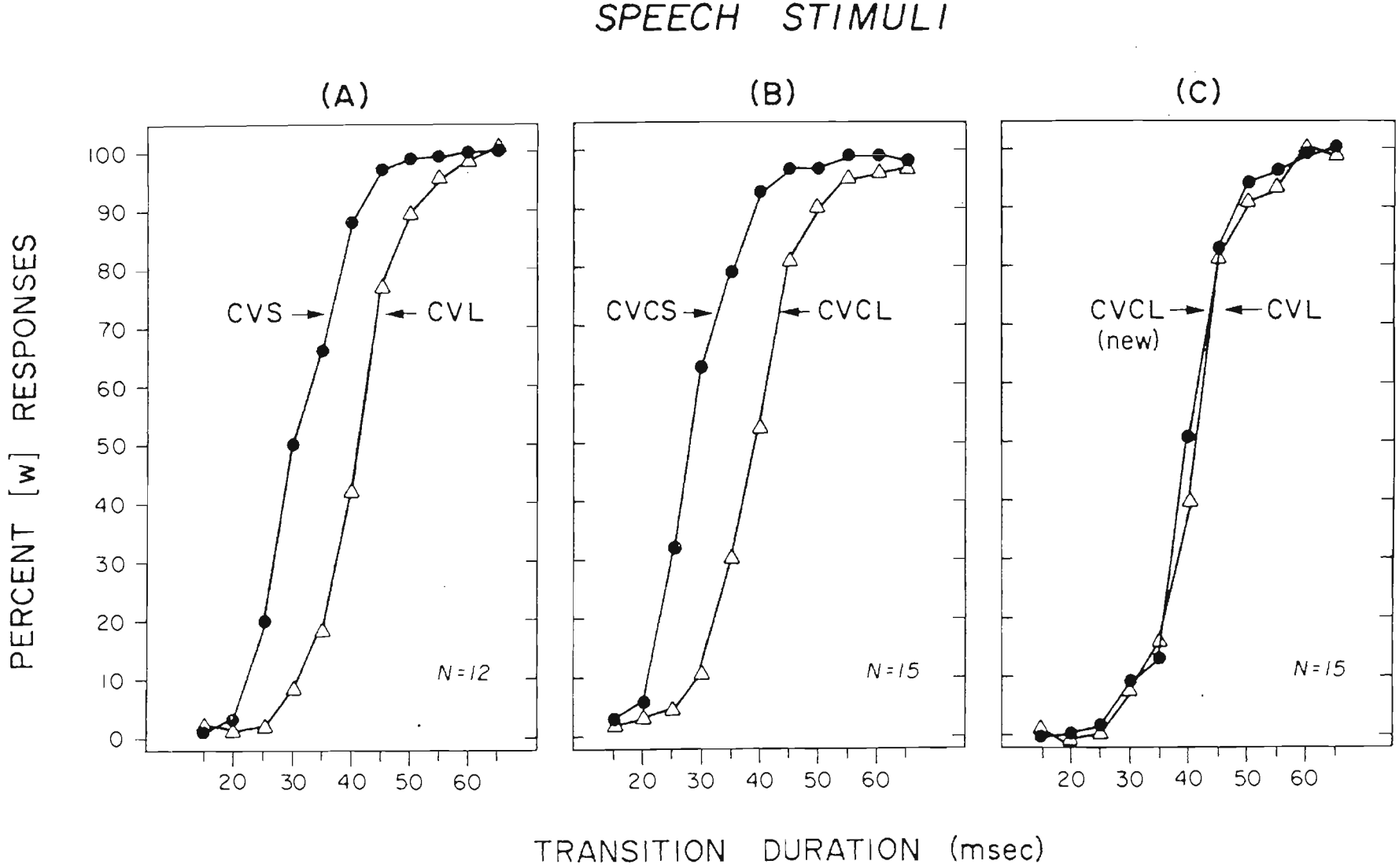

SPEECH STIMULI
(A)
(B)
(C)
PERCENT [w] RESPONSES
100
90
80
70
60
50
40
30
20
10
0
CVS
CVL
N=12
CVCS
CVCL
N=15
CVCL
(new)
CVL
N=15
20
30
40
50
60
TRANSITION DURATION (msec)

FIGURE 3.

nonspeech control stimuli. In this experiment we first trained our subjects to identify the endpoint nonspeech stimuli as beginning with either an "abrupt onset" or a "gradual onset" using a disjunctive conditioning procedure. After training on the endpoints was completed, subjects who met a predetermined criterion of 90 percent were asked to return to the lab for generalization testing. In this phase, all eleven test stimuli in a given series were presented in a random order for identification.

May I have Slide 4 please?

The labelling results for the nonspeech control stimuli are shown in Slide 4. Panel A on the left shows the CV stimuli, Panel B in the middle shows the CVC stimuli and Panel C shows the cross-series comparisons between CVs and CVCs. In Panels A and B there is a very substantial shift in the labelling boundary for "abrupt" and "gradual" onsets as the duration of the overall stimulus is decreased from 295 msec to 80 msec. The effects were statistically reliable and very consistent over subjects. Thus, these two sets of data show comparable context effects for the perception of the duration of a rapid spectrum change at the onset of nonspeech control stimuli.

The results shown in Panel C, on the right of this figure, also replicate the effects reported by Miller and Liberman for stimuli containing formant transitions in syllable-final position. In this case, the boundary for the CVC nonspeech condition is shifted to the left away from the CV stimuli as if they were

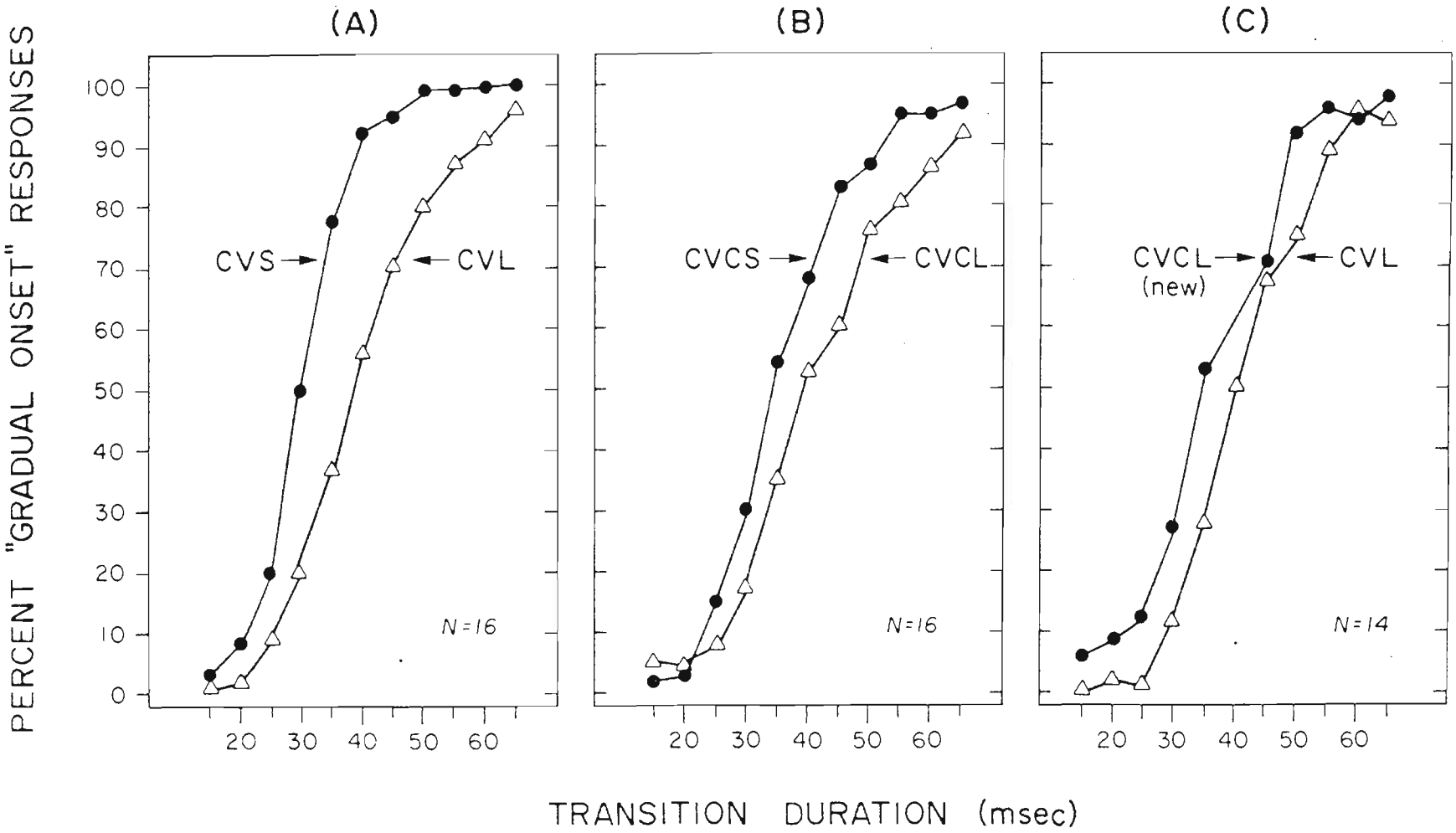

NONSPEECH CONTROL STIMULI
(A)
(B)
(C)
PERCENT "GRADUAL ONSET" RESPONSES
100
90
80
70
60
50
40
30
20
10
0
CVS
CVL
CVCS
CVCL
CVCL
(new)
CVL
N=16
N=16
N=14
20
30
40
50
60
TRANSITION DURATION (msec)

FIGURE 4.

perceived as having a shorter overall stimulus duration. This is true despite the fact that both series were of equal physical duration.

Taken together, all three nonspeech stimulus conditions replicate the context effects obtained by Miller and Liberman with speech stimuli. These nonspeech stimuli were designed to model, as closely as possible, the four sets of speech stimuli that differed in the cues for the [b-w] distinction. As we showed in the previous two experiments with speech stimuli, the major effects of syllable duration could be replicated quite easily with stimuli that differed in the duration of formant transitions. More importantly, however, the present results also demonstrate that comparable context effects can also be obtained with nonspeech stimuli as well.

We believe that our findings with nonspeech stimuli undermine the earlier conclusions of Miller and Liberman that such context effects reflect a form of "perceptual compensation" that is due to the specification of articulatory rate by overall syllable duration. It seems very unlikely that listeners in our nonspeech control conditions carried out an appropriate "adjustment" for changes in articulatory rate since these stimuli were not perceived as speech signals at all in our experiments. Moreover, the present findings demonstrate rather clearly that the perceptual categorization of stimulus onsets as either "abrupt" or "gradual" is also influenced by later occurring events in the stimulus configuration and that nonspeech signals may also be processed in a "relational" and nonlinear fashion, that is, in a manner that is comparable to the perception of speech signals.

Finally, these results seriously question the implications that Eimas and Miller have tried to draw from their recent study with young infants which demonstrated comparable context effects for discrimination of the duration of formant transitions for the [b-w] distinction. While 2-4 month old infants may show context-conditioned sensitivity in discriminating differences in formant transition duration, it seems very unlikely to us that these context effects reflect the operation of perceptual mechanisms that are involved in the phonetic coding of speech signals or the interpretation of speech signals as linguistic entities such as phonemes, phonetic segments or distinctive features. As in the case of the perception of voice onset time, we suspect that infants are responding to the basic psychophysical or sensory properties of these acoustic signals without reference to their linguistic significance. The results of our infant experiments currently underway should provide further data on this important issue.

In summary, we have carried out several critical comparisons between speech and comparable nonspeech control signals that differed in terms of the duration of a rapid spectrum change at stimulus onset, an acoustic cue that has been shown to be sufficient to distinguish between the stop [b] and the semi-vowel [w]. Our findings with speech stimuli replicated the earlier context effects reported by Miller and Liberman. Overall syllable duration clearly affects the location of the [b-w] boundary. However, we also found similar effects for the perception of nonspeech stimuli. We interpret these new results as evidence that context effects are not peculiar to perception of speech signals

or to the listener's perceptual normalization or adjustment to articulatory rate. Our findings therefore call into question the major conclusions of Miller and Liberman that listeners somehow monitor or extract estimates of the talker's articulatory rate and then subsequently carry out operations which adjust their perceptual criteria for interpreting a particular set of acoustic cues as a specific phonetic segment.

Our findings with nonspeech stimuli also raise serious questions concerning the recent interpretations of infant data made by Eimas and Miller. They have argued that the presence of similar context effects in discrimination of [b-w] contrasts implies that infants are perceiving speech signals in a "relational manner" like the mature adult listener. We believe that such conclusions are also unwarranted given the results of the present study demonstrating comparable context effects for nonspeech stimuli.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 6 (1979-80) Indiana University]

Infants' Discrimination of Cues to Place of Articulation in Stop Consonants*

R. N. Aslin and A. C. Walley

Department of Psychology
Indiana University
Bloomington, IN 47405

Abstract

Stevens and Blumstein (1978; 1980) have proposed that the shape of the onset spectrum provides contextually invariant information about place of stop consonant articulation for CV syllables and that sensitivity to properties of the onset spectrum underlies the infant's ability to discriminate place differences. According to this view, contextually variable formant transitions constitute only secondary, learned cues to place of articulation. Prelinguistic infants are, therefore, assumed to be incapable of discriminating place differences in two-formant stimuli which supposedly lack invariant spectral attributes. Our analysis revealed, however, that the two-formant labial and velar CV stimuli constructed for the present study were spectrally similar to their full-formant counterparts at stimulus onset. Therefore, if spectral shape does provide cues to place of articulation, infants would actually be expected to discriminate these stimuli. An operant head-turning paradigm was employed to test this hypothesis with 6-9 month old infants. The results indicated that infants can discriminate two-formant stimuli -- a finding which renders the distinction between primary vs. secondary cues invalid. Moreover, because several infants discriminated the two-formant alveolar and velar stimuli, whose onset spectra are very similar, it is probably not the gross shape of the spectrum at stimulus onset which mediates infants' discrimination of place of articulation differences in stop consonants.

*A version of this paper was presented at the 100th meeting of the Acoustical Society of America, November 18, 1980, Los Angeles, California. The research reported here was supported in part by research grants from NICHHD (HD-11915-03) and NIMH (MH-24027--06) and by a doctoral fellowship awarded to ACW by the Social Sciences and Humanities Research Council of Canada. We would like to thank Diane Kewley-Port for her help in stimulus construction and Kathy Mitchell and Martha-Lyn Wayne for their assistance in data collection.

A major problem in the study of speech perception is to provide an account of the constancy of the phonetic percept given the apparent lack of acoustic-phonetic invariances in the speech waveform. This noncorrespondence has been particularly evident in spectrographic analyses of stop consonants (e.g., Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967). This problem is, however, viewed by one class of speech perception theories (see, for example, Cole and Scott, 1974; Fant, 1960; Stevens and Blumstein, 1978; 1980) as the result of a failure to achieve an appropriate psychological description of the speech signal, rather than an inherent characteristic of the speech signal that requires the assumption that active, high-level perceptual mechanisms act to interprete it appropriately (e.g., Chomsky and Halle, 1968; Liberman et al., 1967; Stevens and House, 1972). In fact, recent attempts to model the initial stages of speech processing, which use new digital methods of speech analysis, have met with some success in identifying acoustic-phonetic invariances (e.g., Blumstein and Stevens, 1979; 1980; Kewley-Port, 1980; Searle, Jacobson and Rayment, 1979; Stevens and Blumstein, 1978).

Stevens and Blumstein (1978; 1980) have, for example, maintained that a specific form of linear predictive coding (LPC) of CV syllables, which is performed at stimulus onset and integrates energy over the first 25.6 msec. of the syllable, provides a distinctive and contextually invariant description of different places of stop consonant articulation. Moreover, they have suggested that a similar analysis may also be adequate for characterizing place of articulation for stops in VC syllables and

for syllable-initial nasals (Blumstein and Stevens, 1979). Specifically, the context-independent properties of the onset spectrum proposed by Stevens and Blumstein include the relative location and diffuseness of spectral energy. These properties were derived from theoretical considerations from the acoustic theory of speech production (Fant, 1960) and from an empirical examination of many natural stop productions. As can be seen in Figure 1, which contains theoretical spectra for the voiced stop

---

Insert Figure 1 about here

---

consonants /b/, /d/ and /g/, labials may be characterized by a diffuse flat or falling spectrum, alveolars by a diffuse rising one, and velars by a prominent mid-frequency spectral peak. These global properties of the onset spectrum are determined primarily by the initial portions of the formant transitions and are enhanced by the presence of the burst which follows consonantal release in stop CV syllables.

According to Stevens and Blumstein, it is sensitivity to these global properties which allows listeners to differentially identify place of articulation for stop CV syllables in the face of the contextual variability inherent in formant transition information. Moreover, it is claimed that the peripheral auditory system is innately endowed with feature detectors which are sensitive to the acoustic correlates of the onset spectrum. Listeners do not, so Stevens and Blumstein maintain, typically use formant transition information (e.g., formant starting frequencies and formant trajectories) to identify place of articulation. Such

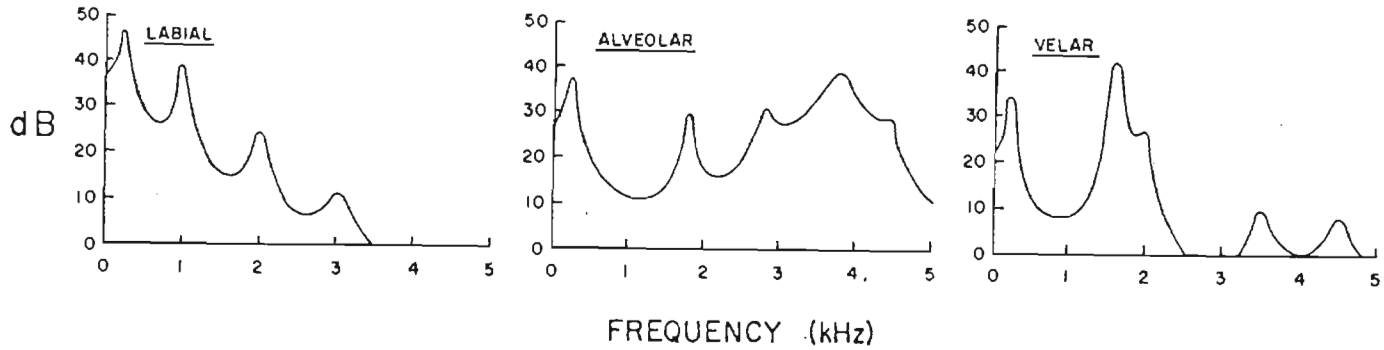


Figure 1. Theoretical onset spectra of labial, alveolar and velar stop consonants (after Stevens and Blumstein, 1978).

information is context-dependent (e.g., the second formant falls in /da/, but rises in /di/) and constitutes only a secondary cue to place of articulation. This secondary cue can be used by adults in the absence or distortion of the primary, context-independent properties of the onset spectrum (but see Walley and Carrell, 1980). This ability is one which is, according to Stevens and Blumstein, acquired in development by virtue of the co-occurrence of the primary and secondary cues. Thus, the properties of the onset spectrum are claimed to be primary for the perception of place of articulation in the sense that: i) for the adult, they constitute the normal or preferred basis for perception and ii) developmentally, they are used prior to formant transition information and sensitivity to these properties is, in fact, innate.

Such sensitivity may, Stevens and Blumstein suggest, account for the ability of prelinguistic infants to discriminate synthetic three-formant and three-formant + burst stimuli with different places of articulation (e.g., Bush and Williams, 1978; Leavitt, Brown, Morse and Graham, 1976; Miller and Morse, 1975; Moffitt, 1971; Morse, 1972). Presumably, the onset spectra of the stimuli employed in these studies had the primary, context-independent properties proposed by Stevens and Blumstein. If the results of these studies of infants' discrimination of place of articulation are to be explained in this way, then an additional developmental claim follows from Stevens and Blumstein's theory: infants, who cannot yet have learned the co-occurrrences of the primary, context- independent properties of the onset spectrum and the secondary, context- dependent cues provided by formant

transitions, should not be able to discriminate stimuli in which place of articulation differences are represented only by the secondary, formant transition cues. More specifically, infants should be unable to discriminate place of articulation differences in two-formant stimuli, which, according to Stevens and Blumstein, lack distinctive, contextually invariant onset spectra.

In order to investigate this prediction of Stevens and Blumstein's theory, we synthesized two sets of three CV syllables corresponding to /ba/, /da/ and /ga/. The spectro-temporal specifications of these stimuli are shown in Figure 2. The "full

---

Insert Figure 2 about here

---

cue" set of stimuli contained bursts and formant transitions and were modelled very closely after the best exemplars of each place of articulation category from Stevens and Blumstein's (1978) transition+burst /ba-da-ga/ continuum with a modified version of the Klatt (1980) synthesizer (see Kewley-Port, 1978). The "partial cue" or two-formant set of stimuli contained only formant transition cues (i.e., no bursts and no higher formants) and were modelled, in turn, after the parameters of the first two formants of the full cue stimuli.

Figure 3 shows the onset spectra of the full and partial cue

Insert Figure 3 about here

---

stimuli. These onset spectra were obtained using an LPC analysis very similar to that used by Stevens and Blumstein. Each of the

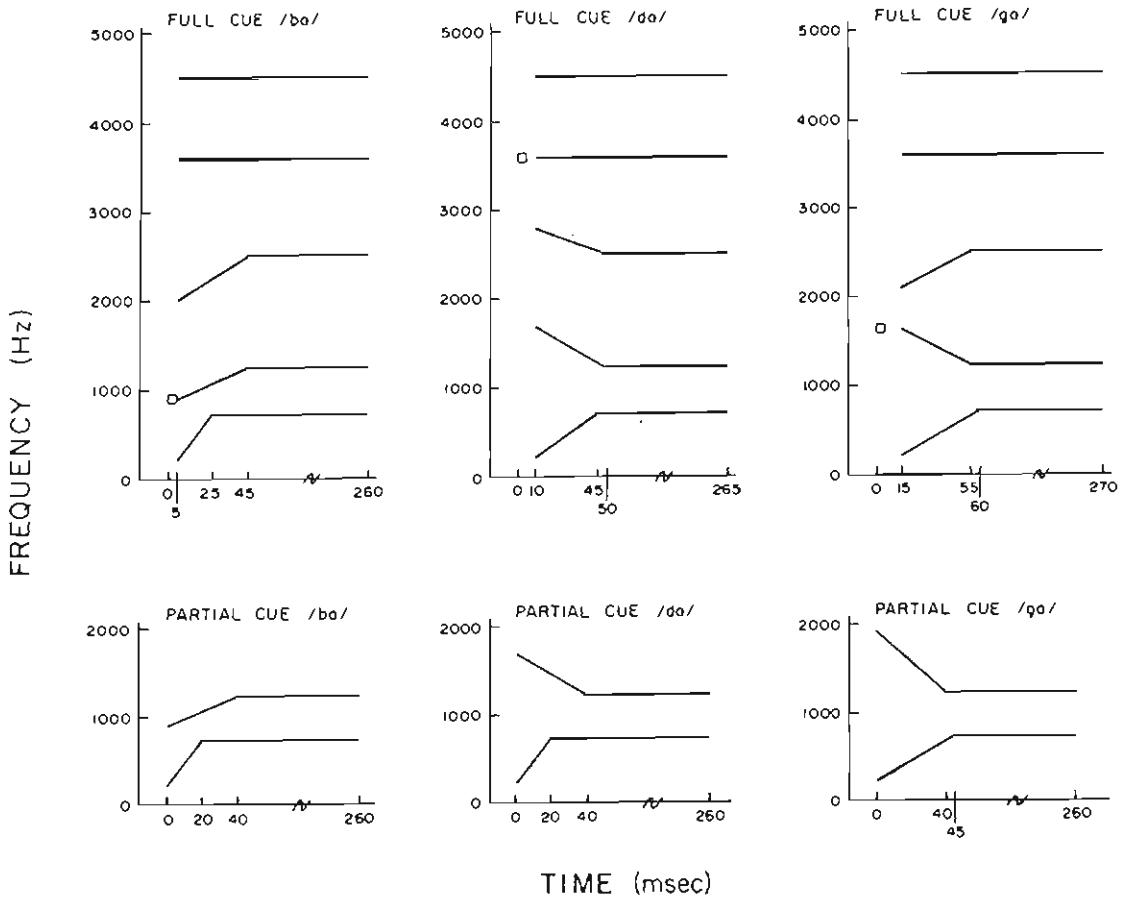


Figure 2. The spectro-temporal specifications of the full and partial cue stimuli.

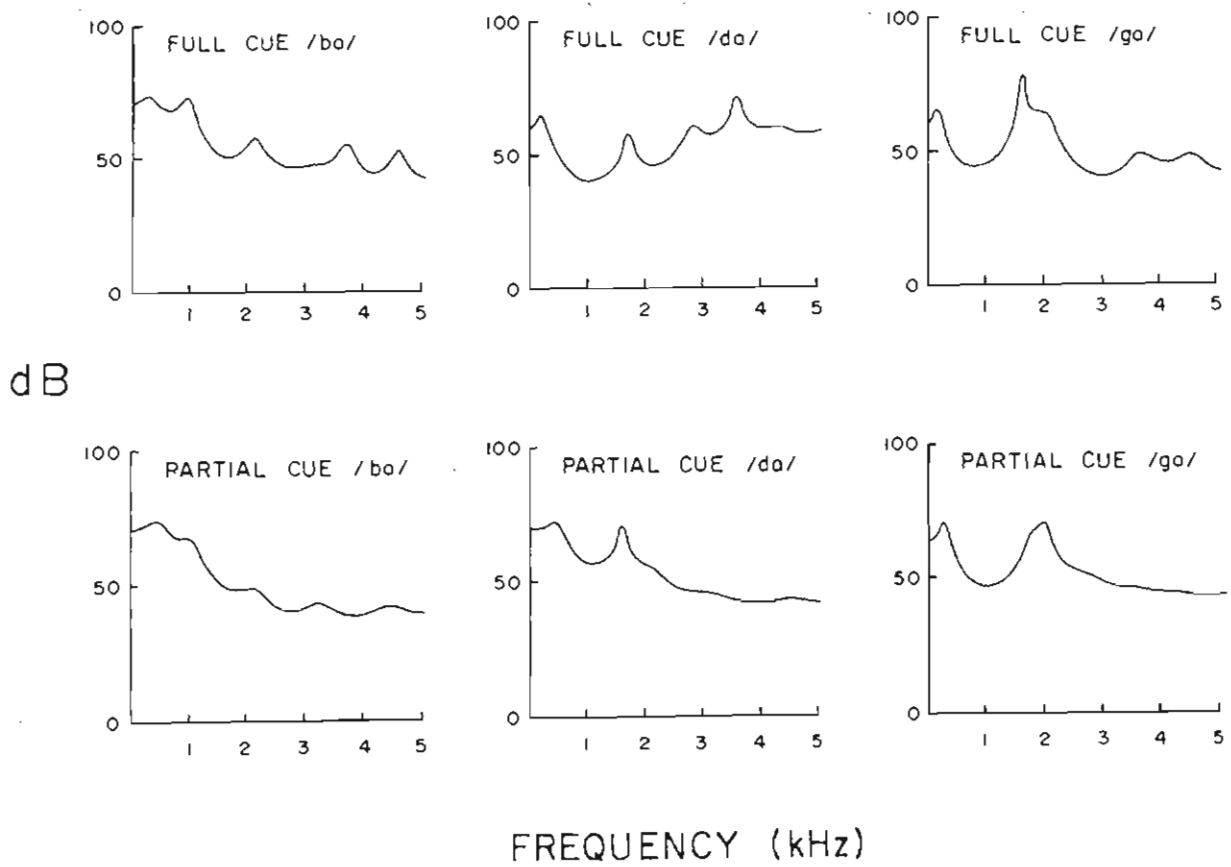


Figure 3. The onset spectra of the full and partial cue stimuli.

full cue stimuli (shown in the top of the figure) fits the appropriate Stevens and Blumstein place of articulation template, and is rejected by the other two templates (for a description of these templates, see Blumstein and Stevens, 1979). The partial cue or two-formant /da/ and /ga/ both fit the velar template and are rejected by the others. The two- formant /ba/ fits the labial template and is rejected by the other two. The results of this inspection of the onset spectra for the two- formant stimuli would seem, therefore, to motivate a revision in the prediction of Stevens and Blumstein's theory. That is, if sensitivity to the properties of the onset spectrum mediates place of articulation discrimination, then infants might afterall be expected to discriminate the partial cue or two-formant contrasts /ba vs. da/ and /ba vs. ga/, but not the contrast /da vs. ga/, since the onset spectra of these latter two stimuli are almost identical.

Sixty-nine infants, ranging in age from 5.5 to 8 months, were assigned to 1 of 3 full and partial cue contrasts and their discrimination of these contrasts assessed using an operant head-turning procedure. Figure 4 shows the experimental apparatus

---

Insert Figure 4 about here

---

used for this procedure. Inside a sound-attenuated IAC booth, an assistant attracts the gaze of the infant seated on the parent's lap. A repeating background stimulus (one member of the contrast being tested) is presented to the infant from a loudspeaker, while the assistant and parent listen to masking music over headphones. In an initial shaping phase which is controlled by an experimenter

Figure 4. The experimental apparatus used in the operant head-turning procedure.

outside of the booth, the infant is trained to make a 90 degree head-turn in anticipation of a visual stimulus (a mechanical toy) which serves as a reinforcer whenever a change in the speech stimulus occurs (i.e., 3 repetitions of the target stimulus or other member of the contrast being tested). The target stimulus is initially 10 dB above the background stimulus to facilitate shaping of the head-turn response. This difference is attenuated in 5 dB steps until a 0 dB difference is reached. In testing, the experimenter initiates trials (which consist of a change or no-change in the speech stimulus) and records any head-turns within the 4 second trial interval. During this phase, the experimenter is unaware of the particular stimulus conditions and reinforcement conditions which are controlled by a computer. A head-turn response on a change-trial (i.e., an experimental trial) constitutes a correct response, a head-turn response on a no-change trial (i.e., a control trial) constitutes an incorrect response.

Infants were required to achieve 80 % correct responding on a minimum of five change and five no-change trials in a row ($p<.05$, by the binomial expansion test) in order to be considered as having met discrimination criterion. As shown in Table 1,

---

Insert Table 1 about here

---

approximately one half of the infants who began testing on a contrast met this criterion in both the full and partial cue conditions. At least four subjects met discrimination criterion for each of the full and partial cue contrasts. In fact, out of

PROPORTION OF SUBJECTS MEETING DISCRIMINATION CRITERION

	FULL CUE N=31	PARTIAL CUE N=38
SHAPING	.94 (n=29)	.66 (n=25)
TESTING	.52 (n=16)	.53 (n=20)

Table 1. The proportion of subjects who passed the shaping and the testing phase of the procedure.

the 13 subjects who began the experiment on the partial cue contrast /da vs. ga/ (a starting N which includes subjects who failed to respond to the intensity difference in shaping and might, therefore, be classified as inattentive), six met discrimination criterion.

An examination of the mean proportions of correct responses on all trials prior to, but within the same session that discrimination criterion was met (see Figure 5), indicated that

---

Insert Figure 5 about here

---

performance was above chance for all the full and partial cue contrasts, but that there were no differences within or across cue condition. The mean number of trials required to meet criterion within the session that the criterion was met (see Figure 6) also

---

Insert Figure 6 about here

---

did not differ between any of the contrasts either within or across cue condition. Nor were there any differences between any of the contrasts within or across cue condition in the mean number of sessions required to meet discrimination criterion. Finally, performance on any of the full or partial cue contrasts by these three measures did not differ from a third group of infants who were tested more recently in an identical manner on the same contrasts, but with digitized natural productions of /ba/, /da/ and /ga/.

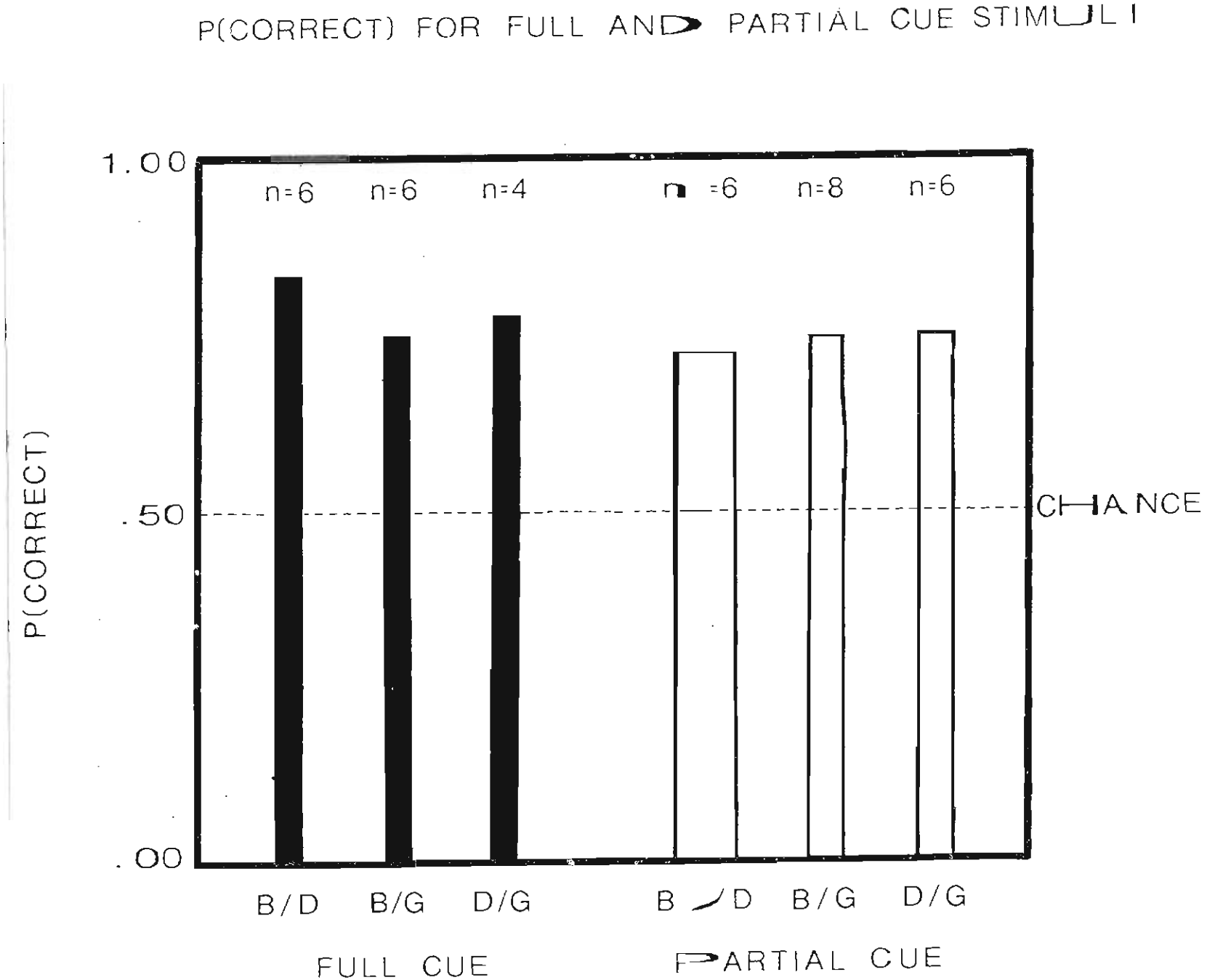


Figure 5. The mean proportions of correct responses for each of the full and partial cue contrasts.

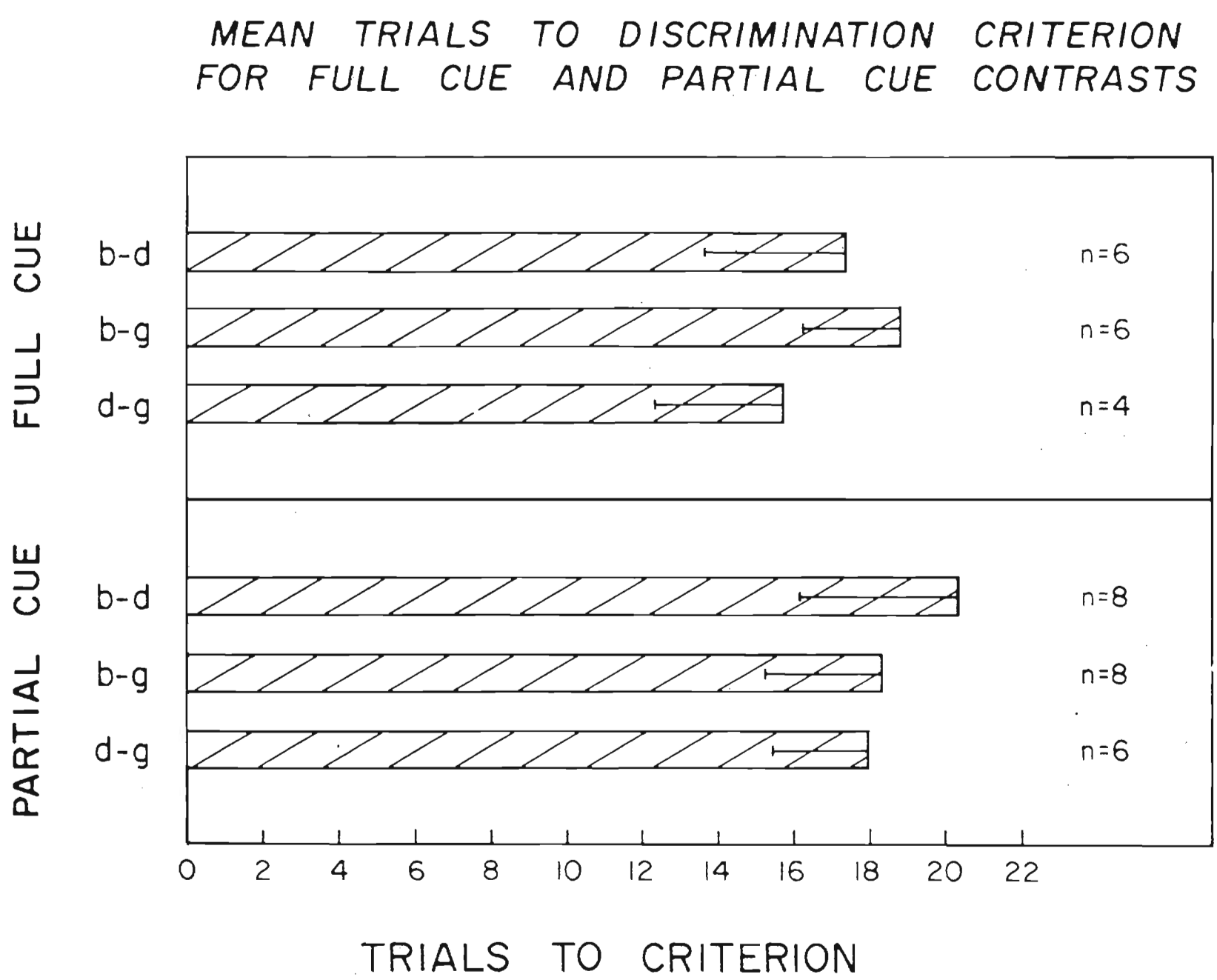


Figure 6. The mean number of trials required to meet discrimination criterion for each of the full and partial cue contrasts.

The major conclusion to be drawn from these results is that young infants are capable of discriminating partial cue or two-formant stimuli differing in place of articulation. This finding is, in fact, consistent with Eimas' (1974) earlier demonstration (using the high-amplitude sucking discrimination paradigm) that infants can discriminate two-formant synthetic labial and alveolar stops. Since an inspection of the onset spectra of our two-formant stimuli revealed that these stimuli fit two of Stevens and Blumstein's templates, one might expect to observe such discrimination -- at least for the /ba vs. da/ and /ba vs. ga/ contrasts. However, this observation and the finding that prelinguistic infants actually discriminate these contrasts argue against Stevens and Blumstein's proposal that two-formant stimuli only possess "secondary", context-dependent cues to place of articulation which cannot be used by infants in perception. In fact, in the case of the two "degraded" two-formant stimuli /ba/ and /ga/, which possess the appropriate Stevens and Blumstein onset spectra, there would seem to little need for making the distinction between primary, innate cues and secondary, learned cues.

It might be concluded, nevertheless, that Stevens and Blumstein are still correct in their claim that it is the gross shape of the spectrum at onset which enables infants to discriminate place of articulation differences. However, the finding that infants discriminated the two-formant /da/ and /ga/, which possessed very similar onset spectra, and the finding that performance on this contrast did not differ across or within cue condition do not support Stevens and Blumstein's claim that it is

the gross shape of the onset spectrum which is used for place of articulation perception. Moreover, the results of this study do not provide any empirical support for the distinction they make between primary and secondary cues to place of articulation in stop consonants.

References

Blumstein, S. E. and Stevens, K. N. Acoustic invariance in speech production: Evidence from tl characteristics of stop consonants. Journal the Acoustical Society of America, 1979, 66, 1001-1017.

Blumstein, S. E. and Stevens, K. N. Perceptual invariance and onset spectra for stop consonants in different vowel environments. Journal of the Acoustical Society of America, 1980, 67, 648-662.

Bush, L. and Williams, M. Discrimination by young infants of voiced stop consonants with and without release bursts. Journal of the Acoustical Society of America, 1978, 63 (4), 1223-1226.

Chomsky, N. and Halle, M. The Sound Pattern of English. New York: Harper and Row, 1968.

Cole, R. A. and Scott, B. Towards a theory of speech perception. Psychological Review, 1974, 81, 348-374.

Eimas, P. D. Auditory and linguistic processing of cues for place of articulation by infants. Perception and Psychophysics, 1974, 16 (3), 513-521.

Fant, G. Acoustic Theory of Speech Production. The Hague: Mouton, 1960.

Kewley-Port, D. KLTEXC: Executive program to implement the KLATT software speech synthesizer. Research on Speech Perception Progress Report No. 4, Indiana University, 1978, 235-246.

Kewley-Port, D. Representations of spectral change as cues to place of articulation in stop consonants. Research on Speech Perception, Technical Report No. 3, December, 1980.

Klatt, D. H. Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America, 1980, 67, 971-995.

Leavitt, L. A., Brown, J. A., Morse, P. A. and Graham, F. K. Cardiac orienting and auditory discrimination in 6-week infants. Developmental Psychology, 1976, 12, 514-523.

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

Miller, C. L. and Morse, P. A. The "heart" of categorical speech discrimination in young infants. Journal of Speech and Hearing Research, 1976, 19, 578-589.

Moffitt, A. R. Consonant cue perception by twenty-four-week-old infants. Child Development, 1971, 42, 717-731.

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

Searle, C. L., Jacobson, J. Z. and Rayment, S. G. Stop consonant discrimination based on human audition. Journal of the Acoustical Society of America, 1979, 65, 799-809.

Stevens, K. N. and Blumstein, S. E. Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America, 1978, 64, 1358-1368.

Stevens, K. N. and Blumstein, S. E. The search for invariant acoustic correlates of phonetic features. In P. D. Eimas and J. Miller (Eds.), Perspectives on the Study of Speech, 1980.

Stevens, K. N. and House, A. S. Speech perception. In J. Tobias (Ed.), Foundations of Modern Auditory Theory: Volume II. New York: Academic Press, 1972.

Walley, A. C. and Carrell, T. D. Onset spectra vs. formant transitions as cues to place of articulation. Paper presented at the 100th meeting of the Acoustical Society of America, November 19, 1980, Los Angeles, California. (See also this report.)

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 6 (1979-80) Indiana University]

Onset Spectra vs. Formant Transitions as Cues to Place of Articulation*

A. C. Walley and T. D. Carrell

Department of Psychology
Indiana University
Bloomington, Indiana 47405

Abstract

Stevens and Blumstein (1978; 1980) have proposed that the gross shape of the onset spectrum of a CV syllable provides listeners with a primary and contextually invariant cue to place of stop-consonant articulation. Contextually variable formant transitions are, on the other hand, claimed to constitute only secondary cues to place of articulation that are learned through their co-occurrence with the primary spectral ones. The present experiment assessed this claim about the relative importance of these two cues by obtaining listeners' identifications of synthetic stimuli whose onset spectra and formant transitions specified different places of articulation. The results indicated that listeners use formant transition information more often than the overall shape of the spectrum at stimulus onset - a result which argues against Stevens and Blumstein's claim that the onset spectrum (at least by their description) is the primary cue to place of articulation. However, because some listeners did, in some instances, consistently use the onset spectra cues for identification of place of articulation, it would seem that different perceptual strategies are used in phoneme identification.

*A version of this paper was presented at the 100th meeting of the Acoustical Society of America, November 19, 1980, Los Angeles, California. The research reported here was supported in part by NIH research grant NS-12179-05 and NIMH research grant MH-24027-06 to Indiana University and by a doctoral fellowship awarded to ACW by the Social Sciences and Humanities Research Council of Canada. We would like to thank L. B. Smith and D. B. Pisoni for their many helpful comments on this work.

A classic problem in the study of speech perception has been to account for the constancy of the phonetic percept given the apparent lack of acoustic-phonetic invariances in the speech waveform. This noncorrespondence has been particularly evident in spectrographic analyses of stop consonants (e.g., Liberman, Cooper, Shankweiler and Studdert- Kennedy, 1967). However, according to one class of speech perception theories (e.g., Cole and Scott, 1974; Fant, 1960; Stevens and Blumstein, 1978; 1980), this problem is indicative of a failure to find the appropriate psychological description of the speech signal, rather than an inherent characteristic of the speech signal which necessitates the postulation of active, high-level perceptual mechanisms to perform initial sensory analyses (e.g., Chomsky and Halle, 1968; Liberman et al., 1967; Stevens and House, 1972). Indeed, the advent of new digital methods of speech analysis has supported the possibility of identifying acoustic-phonetic invariances (see, for example, Blumstein and Stevens, 1979; 1980; Kewley-Port, 1980; Searle, Jacobson and Rayment, 1979; Stevens and Blumstein, 1978).

Stevens and Blumstein (1978; 1980) have, for example, maintained that the onset spectrum of a CV syllable provides a distinctive and contextually invariant description of place of stop-consonant articulation. This onset spectrum is obtained using a linear prediction analysis that integrates energy over the first 25.6 msec. of the syllable. In addition, they have suggested that a similar analysis may be an appropriate one for characterizing place of articulation for stops in VC syllables and for syllable-initial nasals (Blumstein and Stevens, 1979).

Specifically, the context-independent properties of the onset spectrum proposed by Stevens and Blumstein include the relative location and diffuseness of spectral energy. These properties were arrived at on the basis of theoretical considerations from the acoustic theory of speech production (Fant, 1960) and from an empirical examination of many natural stop productions. As one can see from Figure 1, which contains theoretical spectra for the

---

Insert Figure 1 about here

---

voiced stop consonants /b/, /d/ and /g/, labials may be characterized by a diffuse flat or falling onset spectrum, alveolars by a diffuse rising spectrum, and velars by a prominent mid-frequency spectral peak. These properties are determined primarily by the initial portions of the formant transitions and are enhanced by the presence of the burst which follows consonantal release in stop CV syllables.

According to Stevens and Blumstein, it is sensitivity to these global, contextually invariant properties of the onset spectrum which allows listeners to differentially identify place of articulation for stop CV syllables in the face of the contextual variability inherent in formant transition information. Moreover, it is claimed that the peripheral auditory system is innately endowed with detectors which are sensitive to these properties of the onset spectrum. Listeners do not, so Stevens and Blumstein maintain, <u>typically</u> use formant transition information (for example, formant starting frequencies and formant trajectories) to identify place of articulation. Such information

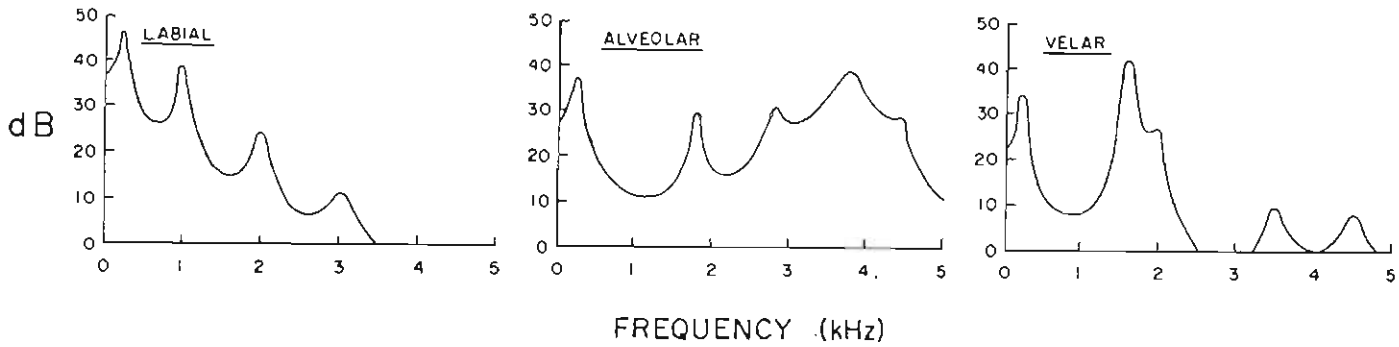


Figure 1. Theoretical onset spectra of labial, alveolar and velar stop consonants (after Stevens and Blumstein, 1978).

is contextually variable (e.g., the second formant falls in the syllable /da/, but rises in /di/) and constitutes a secondary cue to place of articulation. This secondary cue can be used by adults in the absence or distortion of the primary, invariant properties of the onset spectrum -- an ability which, according to Stevens and Blumstein, is acquired in development by virtue of the co-occurrence of the primary and secondary cues (for work which addresses this particular claim, see Aslin and Walley, 1980). Thus, the properties of the onset spectrum are asserted to be primary for the perception of place of articulation in that: i) for the adult, they constitute the normal or preferred basis for perception and ii) developmentally, they are used prior to formant transition information and sensitivity to these properties is actually innate.

In support of their theory that it is sensitivity to the gross properties of the onset spectrum which mediates place of articulation perception, Stevens and Blumstein (1978) report the results of a perceptual experiment in which adult listeners identified synthetic CV syllables from several continua. These continua varied in place of articulation (b-d-g or b-g) in one of three possible vowel contexts (/i/, /a/ or /u/) and contained either bursts and transitions, transitions only or bursts only. It was found that the best exemplars (according to subjects' identification functions) of each place of articulation category for the burst and transition and transition-only continua had onset spectra which possessed the proposed primary, context-independent cues.

As Stevens and Blumstein themselves note, this study did not really constitute a strong test of their theory since formant transition information and onset spectra properties were confounded in their stimuli. Therefore, optimal onset spectra may have co-occurred with optimal formant transition information and poor, nondistinctive onset spectra may have co-occurred with ambiguous formant transition information (for other criticisms of this work and a new analysis, see Kewley-Port, 1980). In an attempt to conduct a strong test of Stevens and Blumstein's theory, we obtained listeners' identifications of synthetic speech stimuli in which formant transition information specified one place of articulation, but in which the onset spectrum specified a different place of articulation -- a manipulation which was achieved by varying the relative amplitudes of the formants. In this way, we hoped to assess the relative contribution of spectral information at stimulus onset and transition information to the perception of place of articulation. Stevens and Blumstein's theory would, of course, predict that since onset spectra cues are primary, a listener's perception should agree with place of articulation as specified by the onset spectrum.

The three control stimuli for the present experiment, whose spectro- temporal specifications are shown in Figure 2, were

---

Insert Figure 2 about here

---

modelled after the best exemplars of each place of articulation category from Stevens and Blumstein's (1978) transition-only /ba-da-ga/ continuum. The stimuli were synthesized in the parallel

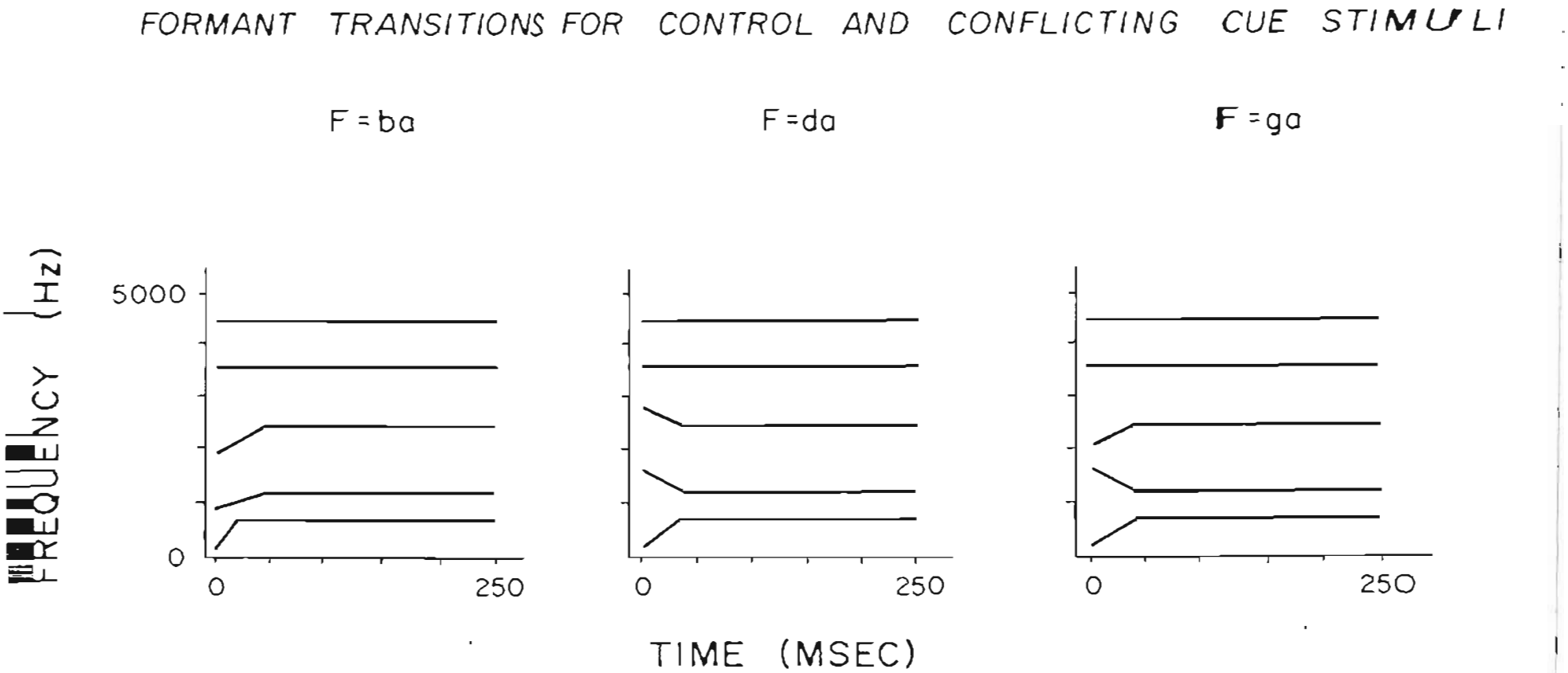


Figure 2. The spectro-temporal specifications of the control and conflicting cue stimuli.

branch of a modified version of the Klatt (1980) synthesizer (see Kewley-Port, 1978). The onset spectra of these stimuli, which were obtained using an LPC analysis similar to Stevens and Blumstein's, are shown in the top of Figure 3. The onset spectrum of each

---

Insert Figure 3 about here

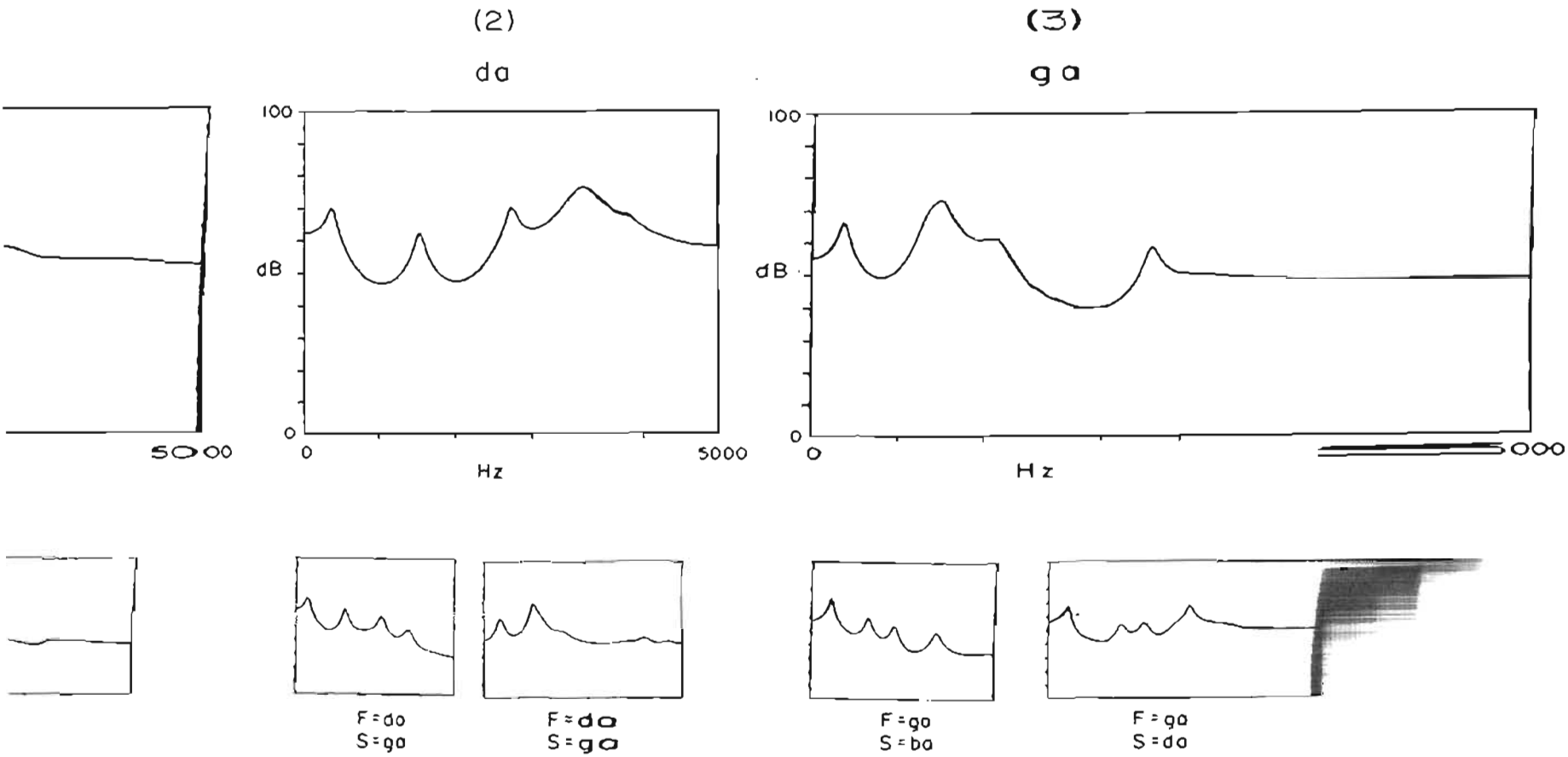
---

control stimulus is consistent with the place of articulation specified by the starting frequency and direction of its formant transitions and is accepted by the appropriate Stevens and Blumstein place of articulation template, and is rejected by the other two templates (for a description of these templates, see Blumstein and Stevens, 1979).

Two experimental or "conflicting cue" stimuli (shown in the bottom of Figure 3) were derived from each control stimulus. A conflicting cue stimulus differed from the original control signal only in that the formant amplitudes were manipulated such that its onset spectrum specified a place of articulation different from that specified by its formant transitions. Each conflicting cue stimulus had an onset spectrum which is accepted by one of the Stevens and Blumstein templates, but which is rejected by the other two templates.

Eighteen Indiana University undergraduates were presented with 24 repetitions of each of the 6 conflicting cue stimuli and 48 repetitions of each of the 3 control stimuli (such that they heard an equal proportion of control and experimental stimuli) in random order. The subjects were asked to identify these computer-generated versions of the syllables /ba/, /da/ and /ga/

SPECTRA OF CONTROL AND CONFLICTING CUE STIMULI



The onset spectra of the control and conflicting cue stimuli.

as quickly as possible by pressing one of three labelled response buttons. Both identification and reaction-time responses were collected. Since the reaction-time data has not yet been analyzed, only the identification data will be presented in this report.

Figure 4 shows the results of the group analysis of the

Insert Figure 4 about here

---

identification data. First, note that the proportion of hits for each control stimulus (shown on the left side of each panel) is very high. These identification responses could be based (among other things) on the gross properties of the onset spectrum, as Stevens and Blumstein have claimed or on information in the formant transitions. In order to determine which of these two cues subjects might have been using, we compared the proportion of hits for control stimuli to the proportion of identifications according to formant transitions for conflicting cue stimuli. In Figure 4, the proportion of identifications by formant transitions denoted by (F) in the figure, by onset spectrum (S), and the proportion of "other" responses (O) for each conflicting cue stimulus is shown to the right of the control stimulus from which it was derived.1

Overall, there <u>was</u> a decrement in the proportion of formant-based responses for the conflicting cue stimuli compared to the proportion of hits for control stimuli, but, as one can see, this decline was stimulus-dependent. Although for three of the conflicting cue stimului there was a decline in the proportion of formant-based responses, there was no such decline for the remaining three conflicting cue stimuli -- a result which does not

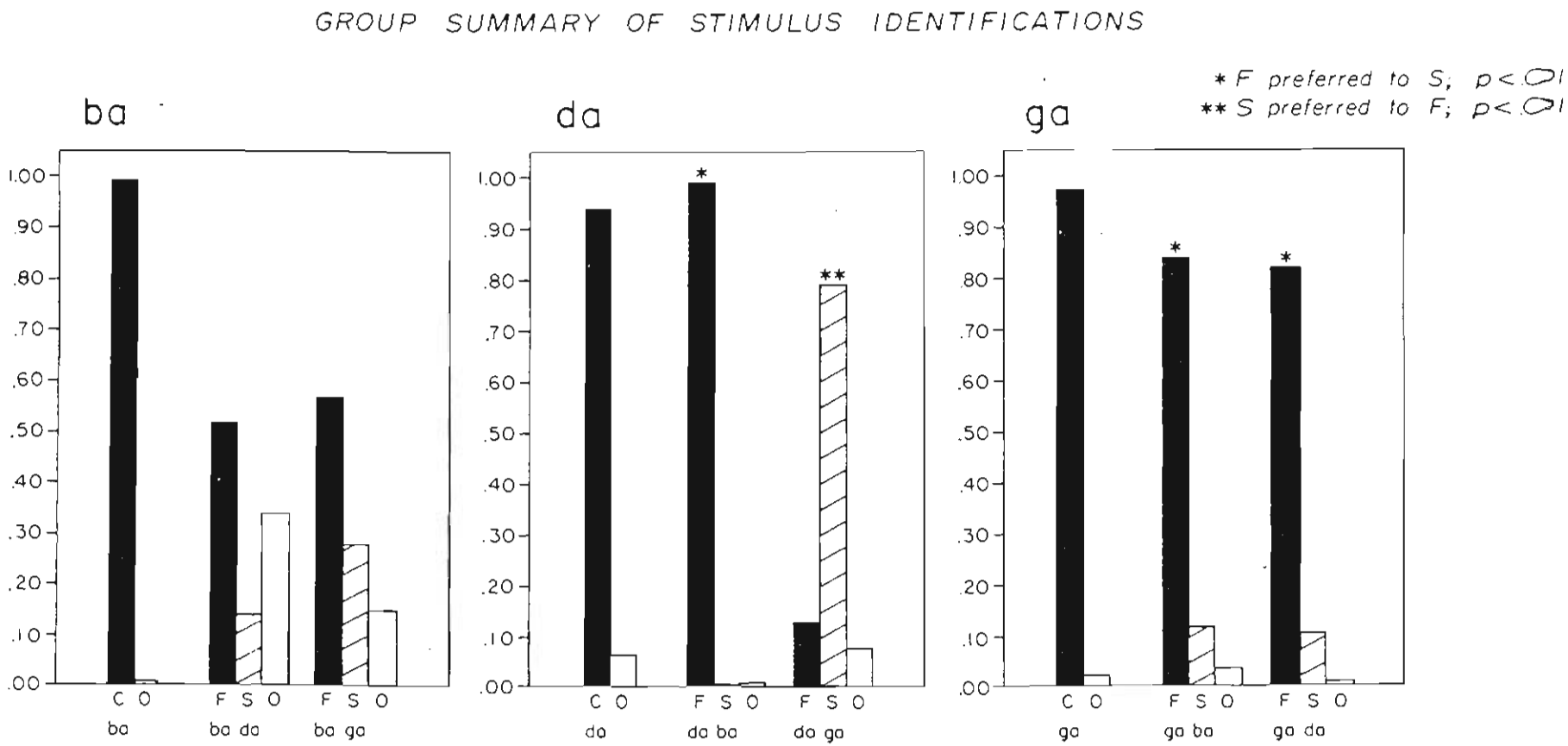


Figure 4. The results of the analysis of the grouped identification data.

support the prediction of Stevens and Blumstein's theory that perception depends on the gross shape of the spectrum at onset.

The decline in formant-based responses observed for three of the conflicting cue stimuli could have resulted from listeners' use of onset spectra in making their identifications. However, this decrement would also be expected if subjects are able to use both cues independently of one another or if putting onset spectra in conflict with formant transitions makes the stimuli ambiguous or poor. The results of two additional analyses suggest very strongly that the conflicting cue stimuli are not simply poor ones.2 First, when "other" responses are ignored, an onset spectrum- based response was preferred for one of the conflicting cue stimuli (identified by ** in Figure 4). Thus, for this particular stimulus, Stevens and Blumstein's prediction was, in fact, confirmed. However, in the three remaining cases where a preference was observed, this preference was formant-based (the stimuli are identified by * in Figure 4). Although no preference was observed for the two conflicting cue stimuli in the first panel of Figure 4, and the decrease in formant-based responses here could, therefore, still be attributed to the distorted nature of these stimuli, a second analysis of individual subjects' data argues against such an interpretation.

In Figure 5, it can be seen that although placing onset

Insert Figure 5 about here

---

spectra and formant transitions in conflict with one another did increase inconsistencies or "other" responses for some subjects,

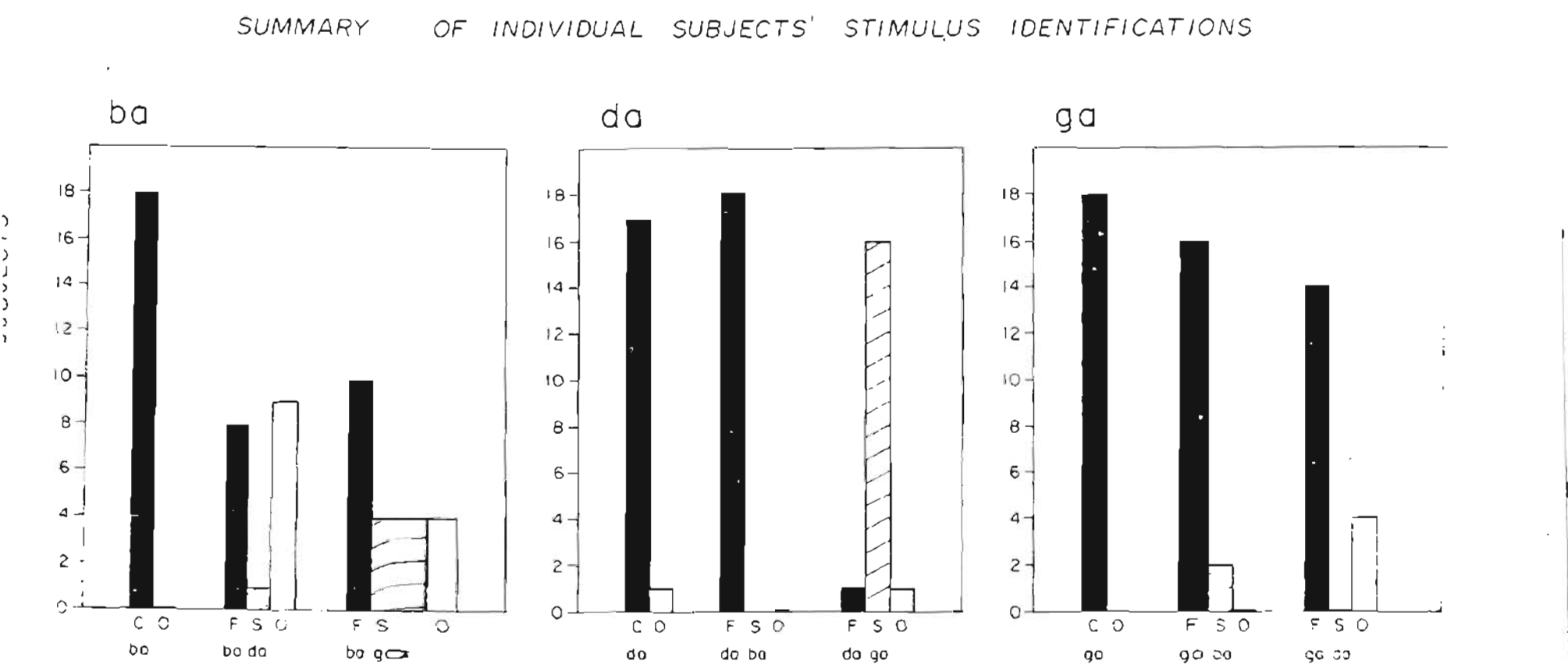


Figure 5. The results of the analysis of individual subjects' stimulus identifications.

many subjects were very consistent in using either a formant - or an onset spectrum -based response. The three panels in the figure show the number of subjects that correctly identified each control stimulus greater than 60% of the time and the number of subjects that used a formant, onset spectrum or "other" response greater than 60% of the time to identify the conflicting cue stimuli derived from the control stimuli. The probability of doing so by chance is less than .01 by a chi-square test.

As can be seen from the individual subjects' data, there were large individual differences in the type of response used and these responses were also stimulus-dependent. For example, eight subjects consistently identified the first conflicting cue stimulus in the first panel of Figure 5 according to its formant transitions (i.e., as a /ba/), whereas one subject consistently identified this same stimulus according to its onset spectrum (i.e., as a /da/). The data for the two subjects shown in Table 1

Insert Table 1 about here

---

illustrates the within-subject consistency and the between-subject differences and similarities observed in the individual subjects' analysis. The results obtained by this analysis suggest that both types of cues may be adequately specified at least in some of the stimuli and that both can be used by adult subjects. Nevertheless, it can be concluded that, as in the group analysis, formant-based responses predominated.

In conclusion, it is apparent from the analysis of the individual subjects' data collected in this study that both

AN EXAMPLE OF INDIVIDUAL DIFFERENCES IN STIMULUS IDENTIFICATIONS

SUBJECT NO.	CONTROL	CONFLICTING CUE	
	Proportion Correct	Proportion Formant Responses	Proportion Spectrum Responses
	ba	*ba* ⟶ *ga*	
3	1.00	.04	.96
4	1.00	1.00	.00
	da	*da* ⟶ *ga*	
3	1.00	.00	1.00
4	.90	.00	1.00
	ga	*ga* ⟶ *ba*	
3	1.00	1.00	.00
4	1.00	1.00	.00

Table 1. The pattern of stimulus identifications observed for two subjects

formant transition and onset spectrum cues were consistently used in the identification of the conflicting cue stimuli (a result which was both stimulus and subject dependent). To this extent, our results support Stevens and Blumstein's notion that the two cues may co-exist. However, formant-based responses predominated in both the group and the individual subjects' analyses -- a finding which argues strongly against Stevens and Blumstein's claim that properties of the onset spectrum are the primary cue to place of articulation for syllable-initial stop consonants. Although it may be that properties of the onset spectrum were not specified optimally and that Stevens and Blumstein's templates are merely wrong in their local details, the finding that subjects used two types of cues in identifying place of articulation does not appear to lend any empirical support to Stevens and Blumstein's distinction of primary vs. secondary cues. Further research is currently underway in our laboratory to elaborate on these perceptual processes in adults and infants.

Footnotes

1. A response was classified as "other" when a subject chose a response which did not conform to place of articulation as specified by either the formant transitions or onset spectrum of a stimulus. For example, if the formant transitions of a conflicting cue stimulus were appropriate for a /ba/ and its onset spectrum was appropriate for a /da/, then if a subject identified this stimulus as /ga/, the response was classified as "other". For present purposes, such responses are essentially considered as random (however, see Footnote 2).

2. In one case only (for the first conflicting cue stimulus in the first panel of Figure 5) are there really a large number of "other" responses. In fact, the individual subjects' analysis indicated that these were consistent responses. It is, therefore, interesting to note that by a post-hoc inspection of this stimulus according to Kewley-Port's (1980) feature classification system, which examines changes in spectral energy over time, that this stimulus might, in fact, be classified as a /ga/.

References

Aslin, R. N. and Walley, A. C. Infants' discrimination of cues to place of articulation in stop consonants. Paper presented at the 100th meeting of the Acoustical Society of America, November 18, 1980, Los Angeles, California. (See also this report.)

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

Blumstein, S. E. and Stevens, K. N. Perceptual invariance and onset spectra for stop consonants in different vowel environments. Journal of the Acoustical Society of America, 1980, 67, 648-662.

Chomsky, N. and Halle, M. The Sound Pattern of English. New York: Harper and Row, 1968.

Cole, R. A. and Scott, B. Towards a theory of speech perception. Psychological Review, 1974, 81, 348-374.

Fant, G. Acoustic Theory of Speech Production. The Hague: Mouton, 1960.

Kewley-Port, D. KLTEXC: Executive program to implement the KLATT software speech synthesizer. Research on Speech Perception Progress Report No. 4, Indiana University, 1978, 235-246.

Kewley-Port, D. Representations of spectral change as cues to place of articulation in stop consonants. Research on Speech Perception, Technical Report No. 3, December, 1980.

Klatt, D. H. Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America, 1980, 67, 971-995.

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

Searle, C. L., Jacobson, J. Z. and Rayment, S. G. Stop consonant discrimination based on human audition. Journal of the Acoustical Society of America, 1979, 65, 799-809.

Stevens, K. N. and Blumstein, S. E. Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America, 1978, 64, 1358-1368.

Stevens, K. N. and Blumstein, S. E. The search for invariant acoustic correlates of phoentic features. In P. D. Eimas and J. Miller (Eds.), Perspectives on the Study of Speech, 1980.

Stevens, K. N. and House, A. S. Speech perception. In J. Tobias (Ed.), Foundations of Modern Auditory Theory: Volume II. New York: Academic Press, 1972.

III. INSTRUMENTATION AND SOFTWARE DEVELOPMENT

Infant Speech Perception Laboratory: Current Computer Resources*

Jerry C. Forshee

Indiana University

Introduction

The current laboratory computer system was purchased from Digital Equipment Corporation as a packaged system with funds made available through NSF research grant BNS 77-04580 to Indiana University. The computer system was delivered in the Fall of 1977 and has undergone almost continuous hardware and software development to achieve its current hardware configuration and operational status. The primary instrumentation objective of the computer system for this grant (HD-11915) is to support the presentation to infant subjects of relatively short (300 to 500 msec.) speech signals in real-time and the collection of subsequent responses made by a "blind" adult observer. In the following sections of this report we will describe in some detail the design goals, the current configuration and the typical functions supported by the computer system. This report is not intended as an exhaustive description of these systems but is intended to summarize the current capabilities of the laboratory at the present time for carrying out a fairly wide range of activities associated with speech processing tasks including analysis, synthesis and perceptual experimentation with infant and adult observers.

System Design Goals

Our primary objective for the PDP 11/34 was to develop a computer system capable of conducting on-line perceptual experiments. Our definition of conducting perceptual experiments includes not only the real-time presentation of auditory signals to infants and the collection of observer's responses, but also the development of all required programs and the data recovery tasks, i.e., data summarization and analysis. Although considerable effort has been directed toward the design of this system, it is obvious that the current configuration is not completely permanent; as new research needs are encountered modifications will need to be made to meet these needs. What we are certain of, however, is that the PDP 11/34 system is an excellent choice around which to build a flexible laboratory of this type for it provides a very adequate set of facilities in its basic configuration and allows for more add-on capability than required by nearly any small-sized research laboratory. The current configuration of the PDP 11/34 computer system is displayed in Figure 1, and will be described below under the following major headings: (1) basic system resources, (2) analog input/output system, (3) subject I/O, and (4) operator control station.

Insert Figure 1 about here

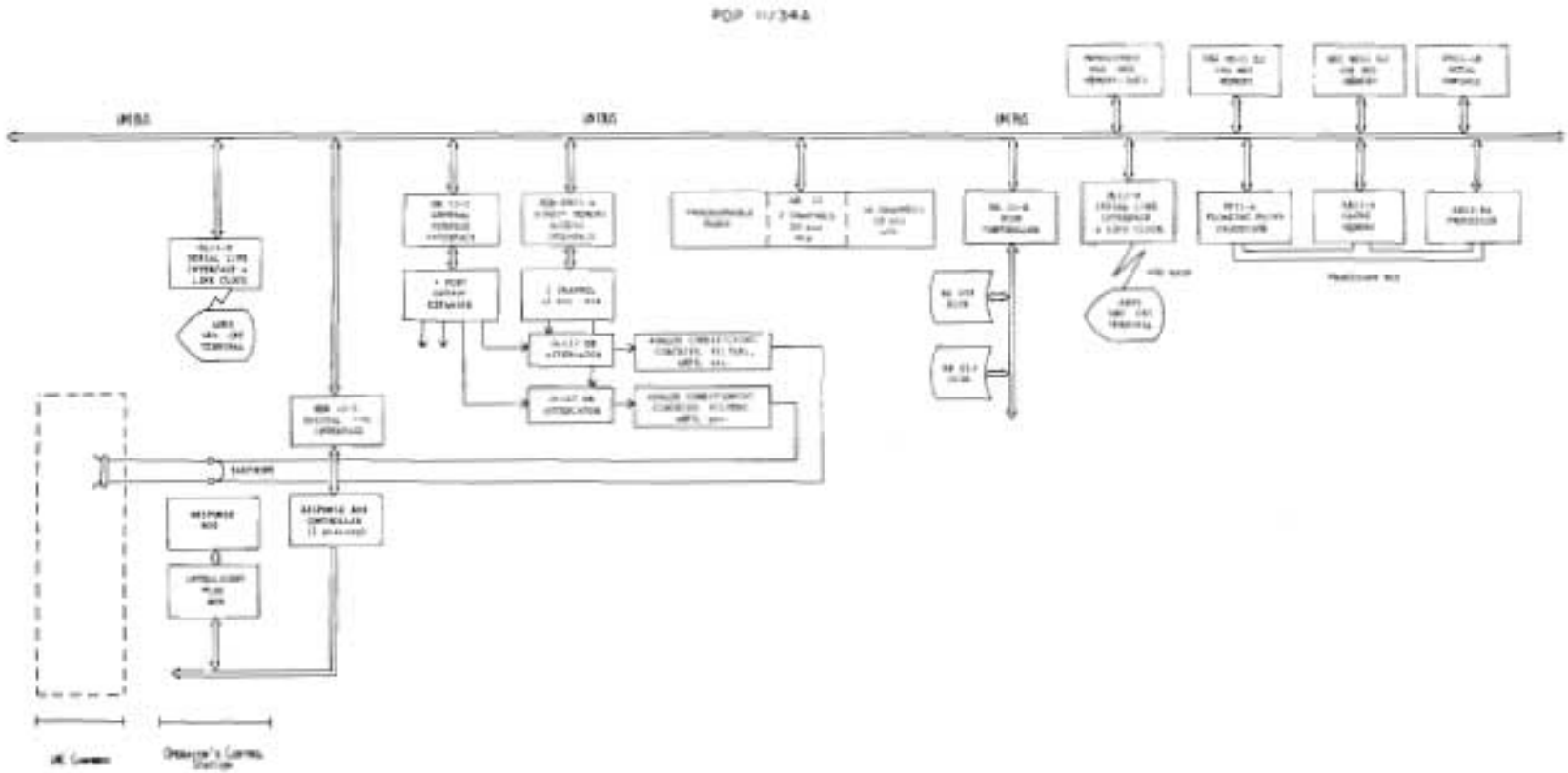


Figure 1. PDP 11/34 computer system in the Infant Perception Laboratory

Basic System Resources

The basic processor in this system has been complemented by two add-on options which combine to make the PDP 11/34 a very fast computational machine. These include the KD11-EA processor itself, the FP11-A floating point processor and the KK11-A cache memory. This later device acts as a high speed buffer between memory and the processor, providing a 40% to 60% increase in throughput depending on the organization of the program being executed. The total memory on the system is now 96K words which consists of the original DEC 32K word MS11-EJ and a Monolithic Systems 64K word MOS 3603 system. The processor is equipped with the optional Programmers Octal Keypad Console, the KY11-LB subsystem. The system programming console is an ADDS 580 CRT terminal, operating at 9600 baud, and connected to the Unibus by a DEC DL11-W serial line interface which also provides the hardware line frequency clock function. The system storage device is an RK05 disk system consisting of the DEC RK11-D disk controller, an RK05-J removable media 2.5 Mbyte drive and a fixed media 5.0 Mbyte RK05-F drive. The system hardcopy device is a DEC LA180 DECprinter which operates at 180 cps print speed. The programmable real-time clock on this system is one of the functional units of a DEC AR-11 interface. This clock is used to drive the various software timers used throughout the system.

Analog Output System

Experience gained in developing other systems at Indiana University led us to develop an analog output system that is modular and adaptable, thus allowing us to meet our instrumentation demands as they change from experiment to experiment. As such dynamic growth is the normal expectation in research of the kind carried out in our laboratory, we felt that choosing modularity and flexibility as one of the major design criteria in our analog system interface was a sound decision.

The first link in our chain of analog interface components originates with an MDB DR11-B Direct Memory Access (DMA) controller. This device is an embedded controller which provides all necessary interfacing with the PDP 11 Unibus to provide DMA capability for the system. This DMA controller leaves space for additional circuitry to be installed by the user to tailor the controller more exactly to his particular application. We used this feature of the DMA controller to add our own logic circuits to establish a unique configuration of control bits in the control and status register of the DMA D/A controller. These control bits allow for program selection of different hardware features. Among other features, the program may select: sampling rate of either 10K or 20K; inclusion or exclusion of pre-emphasis and de-emphasis circuits; signal routing to channel-A D/A or channel-B D/A, or to both channels for dichotic output of two independent signals simultaneously. Adjacent to the DMA controller in the computer is the D/A interface which contains the two 12 bit D/A converters, Datel Model DAC-VR12B2D.

One of our design goals in developing this system was to completely isolate the digital and analog systems. The analog power for the D/As comes from a power supply external to the computer and the logic and analog power systems are therefore kept totally isolated. The only two signals leaving the computer are the two D/A analog outputs which use isolated grounds and differential line drivers for maximum isolation and noise immunity. These two analog signals are connected to a separate equipment rack containing all the analog signal conditioning circuits. The ultimate design criteria for the analog rack was one of modularity and signal standardization. Each module has a standard input/output impedence and voltage protocol. This design scheme allows for modules to be patched in or out of the signal path quickly to satisfy the need of a particular experiment or related task. This modular standardization also allows for quick repairs, usually by exchanging the defective module with a spare one on hand. Such standardization also reduces the effort of incorporating new devices in the system; all that is needed is to convert the new device to our standard, if required, before adding it to the analog equipment rack. Another advantage is the great flexibility it permits; existing signal conditioning functions may be removed and new ones added in the chain with no effort required to break the chain or rearrange the existing modules. The standard units currently employed in this rack include: (1) the differential line receiver module which receives the signals from the D/A interface in the computer and establishes the electrical protocol for the analog signals; (2) low pass filters, both 5K and 10K are available; (3) programmable

attenuators, one for each analog channel, which has 9 stages with attenuation step sizes of 1/4 dB, 1/2 dB, or 1 dB; our default configuration is 1 dB steps with a range of 0 to -127 dB; (4) auxiliary output panel for driving experimenter earphones or speakers and other output devices such as tape recorders; (5) the final link in the analog interface chain is the power amplifier which drives the speakers providing sound to the subject in the experimental sound chamber.

The digital outputs required to operate the attenuators come from a simple yet extremely useful device we designed and constructed to provide an extended number of output bits while imposing a minimal load on the Unibus. This interface, the DR11-C Output Expander, produces four words of digital output with a single load presented to the Unibus by using a standard DR11-C module. The one word output of the DR11-C is multiplexed to four 16 bit holding registers controlled by two bits of the DR11-C control and status register. One of these expanded output words is used for controlling each of the programmable attenuators.

Subject I/O

All interaction with our subjects and observers is provided by an interface we designed specifically for this purpose. The Response Box Controller (RBC) system is basically a data multiplexer which is connected to the computer via a MDB11-C digital interface. Multiplexing is provided to allow an output buss of eight data bits to each of two subject stations, each one being independently latched in the RBC. Likewise, on the input, the RBC acts as a holding register for eight bits of input data

for each of these subject stations. Digital logic is provided in the RBC to handle the stacking of input data in the case of both subject stations responding simultaneously. When multiple responses occur, the data is held in the RBC and an output interrupt to the computer is generated each time an input is received until all pending responses have been read by the computer. In this effort of interfacing the human observer, a versatile and modular approach was taken with the organization of electrical signals available at the subject station. A device called the Intelligent Plug Box (IPB) was also designed which greatly facilitates the frequent changes between particular response manipulanda demanded by different experimental paradigms in use in the laboratory.

The IPB provides the interface between the RBC and the several types of response boxes which may be used. The IPB provides four different jacks that connect subject I/O devices using two different electrical protocols. The first connector jack provides points to connect the eight input and eight output lines using the "event" protocol. This protocol supports switches, pushbuttons, small relays, incandescent lamps and similar devices. The second output connector repeats the output lines so that input and output functions can be built in separate enclosures and then mixed and matched to suit a particular experiment. For "event" devices no additional logic is required beyond that contained in the IPB itself. Thus, it is quite economical to have multiple sets of subject I/O devices as only the enclosure, buttons, and cue lamps need be duplicated. This organization also makes it quite possible for a laboratory user

to construct a unique subject I/O device for his particular needs with a minimal amount of technical knowledge and skill and with minimal time requirements.

A second signal protocol is used on the third and fourth connectors on the IPB. The standard TTL logic family protocol is used with both the eight input and eight output lines appearing on the third connector while only the eight output lines are repeated on the fourth, for the same reason of combining I/O boxes as in the "event" protocol. This second mode, the "logic" protocol, is used to directly connect keyboards, numeric keypads, LEDs, seven segment displays or custom devices that use conventional TTL logic signal protocol.

Operator Control Station

Experimenter/Observer interaction with the computer during experimental sessions is carried out by a high speed CRT terminal located in the subject room. The experimenter/observer uses this terminal to start, pause and run the actual on-line experiment. When appropriate, the experimenter can enter new parameter values as required in conducting an experimental session. This terminal is also used extensively by the particular program conducting the experiment to provide data to the experimenter/operator regarding the progress of the infant subject participating in the experiment. The experimenter/ observer also has available a special response box used to interact directly with the experiment program. This facility permits the experimenter/observer to present stimuli and feedback during training trials and enter observational responses during test

trials. Also available to the experimenter/observer is an earphone set to allow listening to the actual stimuli being presented to the infant subject. During testing stages of the experiment the experimenter/operator's earphones are switched to a tone generator which the computer gates on and off in synchronization with the speech stimuli presented to the infant subject. This allows the experimenter/observer to be informed of the observation interval while remaining blind to the actual stimuli presented on a given trial.

Summary and Conclusions

It has been our intent in this report to present the reader with the design criteria, research goals and instrumentation strategies that have led to the development and implementation of the current computer facilities of the Infant Speech Perception Laboratory. We view the current laboratory as just one step in a continual process of evolution; each completed research project giving birth to the next idea and its own unique set of instrumentation needs. The attributes of modularity, adaptability and functional transportability have been central to our thinking, planning and implementation of each laboratory facility. We have also placed high priority on developing a well organized and efficiently operating laboratory with readily available software support as we believe these are fundamental prerequisites for conducting high quality research and collecting reliable experimental data under highly controlled conditions.

Footnotes

*The development and implementation of the computer facilities described in this report was supported by funds from NICHD research grant HD-11915 and NSF research grant BNS 77-04580 to Indiana University in Bloomington, IN 47405.

IV. Publications:

Aslin, R. N. & Pisoni, D. B. Some developmental processes in speech perception. In G. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), Child Phonology: Perception and Production. New York: Academic Press, 1980, Pp. 67-96.

Aslin, R. N. & Pisoni, D. B. Effects of Early Linguistic Experience on Speech Discrimination by Infants: A Critique of Eilers, Gavin and Wilson (1979). Child Development, 1980, 51, 107-112.

Pisoni, D. B. Adaptation of the Relative Onset Time of Two-Component Tones. Perception & Psychophysics, 1980, 28, (4), 337-346.

Jusczyk, P. W., Pisoni, D. B., Walley, A. and Murray, J. Discrimination of relative onset time on two-component tones by infants. Journal of the Acoustical Society of America, 1980, 67, (1), 262-270.

Remez, R. E. Susceptibility of a stop consonant to adaptation on a speech-nonspeech continuum: Further evidence against feature detectors in speech perception. Perception & Psychophysics, 1980, 27, 17-23.

Remez, R. E., Cutting, J. E. & Studdert-Kennedy, M. Cross-series adaptation using song and string. Perception & Psychophysics, 1980, 27, 524-530.

Gruenenfelder, T. M. and Pisoni, D. B. Fundamental Frequency as a Cue to Postvocalic Consonantal Voicing: Some Data from Perception and Production. Perception & Psychophysics, 1980, 28, 514-520.

Pisoni, D. B. Some Remarks on the Perception of Speech and Nonspeech Signals. Proceedings of the Ninth International Congress of Phonetic Sciences, Copenhagen, August, 1979. Copenhagen: Institute of Phonetics, University of Copenhagen, 1980. Pp. 301-312.

Pisoni, D. B. Summary of Symposium No. 8: Perception of Speech versus Nonspeech. Proceedings of the Ninth International Congress of Phonetic Sciences, Copenhagen, August, 1979. Copenhagen: Institute of Phonetics, University of Copenhagen, 1980. Pp. 312-320.

Pisoni, D. B. and Hunnicutt, S. Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. 1980 IEEE International Conference Record on Acoustics Speech and Signal Processing, April, 1980. Pp. 572-575.

Bernstein, J. and Pisoni, D. B. Unlimited text-to-speech device: Description and evaluation of a microprocessor-based system. 1980 IEEE International Conference Record on Acoustics, Speech and Signal Processing, April, 1980. Pp. 576-579.

Manuscripts to be published:

Pisoni, D. B. Some Measures of Intelligibility and Comprehension. In J. Allen (Ed.) Conversion of Unrestricted English Text to Speech. 1980 (In Press).

Pisoni, D. B. Variability of vowel formant frequencies and the Quantal Theory of Speech: A first report. Phonetica, 1980 (In Press).

Carrell, T. D., Smith, L. B. and Pisoni, D. B. Some Perceptual Dependencies in Speeded Classification of Vowel Color and Pitch. Perception & Psychophysics, 1980, (In Press).

Remez, R. E., Rubin, P. E., Pisoni, D. B. and Carrell, T. D. Speech Perception without Traditional Speech Cues. Science, 1981 (In Press).

Walley, A. C., Pisoni, D. B. and Aslin, R. N. The Role of Early Experience in the Development of Speech Perception. In R. N. Aslin, J. Alberts and M. R. Petersen (Eds.), The Development of Perception: Psychobiological Perspectives. New York: Academic Press, 1981 (In Press).

Blank, M. A., Pisoni, D. B. and McClasky, C. Effects of Target Monitoring on Comprehension of Fluent Speech. Perception & Psychophysics, 1981 (In Press).

Grunke, M. E. and Pisoni, D. B. Some Experiments on Perceptual Learning of Mirror-Image Acoustic Patterns. Perception & Psychophysics, 1981 (In Press).

Aslin, R. N., Pisoni, D. B., Hennessy, B. L. and Perey, A. J. Discrimination of Voice-onset Time by Human Infants: New Findings Concerning Phonetic Development. Child Development, 1981 (In Press).

Sinnott, J. M. and Pisoni, D. B. Pure Tone Thresholds in the Human Infant and Adult. Infant Behavior and Development, 1981 (In Press).

V. Speech Perception Laboratory Staff, Associated Faculty and Personnel:

Research Personnel:

David B. Pisoni, Ph.D. --------- Professor of Psychology
Richard N. Aslin, Ph.D. -------- Associate Professor of Psychology
Beth G. Greene, Ph.D. ---------- Research Associate
Diane Kewley-Port, Ph.D. ------- Research Associate
Michael R. Petersen, Ph.D. ----- Assistant Professor of Psychology
Robert E. Remez, Ph.D. --------- Visiting Assistant Professor of Psychology*
Joan M. Sinnott, Ph.D. --------- Research Associate
Linda B. Smith, Ph.D. ---------- Assistant Professor of Psychology
Rebecca Treiman, Ph.D. --------- Assistant Professor of Psychology

Michelle A. Blank, Ph.D. ------- NIH Post-doctoral Fellow**
Hans Brunner, Ph.D. ------------ NIH Post-doctoral Fellow
Barry Green, Ph.D. ------------- NIH Post-doctoral Fellow
Sue Ellen Krause, Ph.D. -------- NIH Post-doctoral Fellow***

Thomas D. Carrell, M.A. -------- Graduate Research Assistant
Paul A. Luce, B.A. ------------- Graduate Research Assistant
Christopher Murphy, B.A. ------- Graduate Research Assistant
Aita Salasoo, B.A. ------------- Graduate Research Assistant
Marthalyn Wayne, B.A. ---------- Graduate Research Assistant

Technical Support Personnel:

Jerry C. Forshee, M.A. --------- Computer Systems Analyst
Nancy J. Layman ---------------- Administrative Secretary
David Link --------------------- Electronics Engineer

Susan J. Gans ------------------ Undergraduate Research Assistant
Thomas Jonas ------------------- Undergraduate Research Assistant
Esti Koen ---------------------- Undergraduate Research Assistant

Special Consultants:

Peter W. Jusczyk, Ph.D. -------- University of Oregon

---

*Now at Barnard College, New York, N.Y.
**Now at Bell Laboratories, Whippany, N.J.
***Now at Rush Presbyterian Medical Center, Chicago, IL.