21135

NATIONAL LIBRARY
OTTAWA

CANADA

BIBLIOTHÈQUE NATIONALE
OTTAWA

NAME OF AUTHOR.. Anthony R. Whitehurst

TITLE OF THESIS.. The Perceptual Role of Voice Onset Time

UNIVERSITY.. University of Alberta, Edmonton

DEGREE FOR WHICH THESIS WAS PRESENTED.. M. Sc.

YEAR THIS DEGREE GRANTED.. 1973

Permission is hereby granted to THE NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(Signed) Anthony R. Whitehurst

PERMANENT ADDRESS:

P.O. Box 1005
Los Banos, California
U.S.A. 93635

DATED. December 18 19 73

NL-91 (10-68)

THE UNIVERSITY OF ALBERTA

THE PERCEPTUAL ROLE OF VOICE ONSET TIME

by

© Anthony R. Whitehurst

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF SCIENCE

IN

EXPERIMENTAL PHONETICS

DEPARTMENT OF LINGUISTICS

EDMONTON, ALBERTA

SPRING, 1974

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled "The Perceptual Role of Voice Onset Time", submitted by Anthony R. Whitehurst in partial fulfilment of the requirements for the degree of Master of Science.

.....
Supervisor

.....

.....

.....

.....

Date 15 December 1973

ABSTRACT

The perceptual impact of Voice Onset Time (VOT) was examined in an experimental design including three factors: place of articulation, manner of articulation, and white noise masking. The perceived voiced/voiceless distinction was upheld throughout the entire range of masking conditions, while a significant place by masking interaction was observed.

High levels of white noise masking were also accompanied by the confusion of place of articulation. Such confusions gave rise to a re-evaluation of the present synthetic stimuli in acoustic (pre-linguistic) terms. Variations in a number of stimulus features in the frequency domain were observed to correspond closely to changes in VOT. As a result of this observation, in conjunction with earlier studies of both speech and non-speech auditory analysis, it was suggested that at least some of those associated stimulus features may be analyzed by the listener in a non-linguistic mode.

ACKNOWLEDGEMENTS

My sincere thanks are offered to all of the members of my thesis committee. They have been patient. Dr. J. T. Hogan has provided me with numerous comments and criticisms, all of which have aided me in directing my own thoughts. Professor Wm. J. Baker has once again offered invaluable guidance and understanding, especially in statistical matters. To be certain, Dr. A. J. Rozsypal and Mr. Allan Opperthauser have been more than helpful in instrumentation and the preparation of stimuli. I must also thank Miss Debbie Perkins for her conscientious job of typing the thesis.

This list of thanks due is by no means complete, since valuable commentary and assistance has arisen in many discussions with fellow students and members of the department. To all of them, and to the University of Alberta and the Department of Linguistics, who have supported me financially through departmental assistantships, I am grateful.

Above all, I must express my deepest gratitude to my wife, Jan, for her constant moral support, understanding, encouragement, and patience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Preliminary Considerations

The primary goal of this study was to examine in some detail certain aspects of the perceived 'voiced-voiceless' manner distinction in initial stop consonants. Although some consideration was given to several studies which were believed to be related to the issue on a general level, most serious attention was given to those which dealt more specifically with the notion of voice onset time (VOT), the time interval between consonant release and the onset of glottal pulsing. VOT is a recent addition to the list of features which are assumed to be in some way influential in the perceived difference between English 'voiced' (b-d-g) and 'voiceless' (p-t-k) categories in initial position.

The status of VOT has rapidly evolved from that of a descriptive ameter (Lisker & Abramson, 1964) to being indicative of a biologically significant aspect of human neural structure (Eimas & Corbit, 1973). In accordance with the logical and chronological development of the status of VOT, several major assertions have been offered:

Analytical claims

1a. For all languages which exhibit at least two perceived manner categories of initial stop

consonants, the perceptual manner categories can be differentiated along the single dimension of VOT. In this sense, VOT has been regarded as a universal phonetic dimension.

1b. Other proposed features are to be regarded simply as consequences of VOT, and therefore, need not be invoked in the categorization of initial stop consonants. Thus, VOT is a sufficient perceptual cue for the observed distinctions.

Biological claims

2a. All speakers must share some common means by which this universal dimension may be subdivided into perceptual classes.

2b. Experimental evidence interpreted as supportive of linguistic VOT distinctions in pre-verbal infants led to the hypothesis that the mechanism for VOT discrimination is an innate human neural complex, i.e., a specific 'linguistic' feature detector mechanism.

Examination of the few critical reports available has led to the conclusion that the above assertions may be premature. Furthermore, it has become clear that VOT, evidently treated as the sole necessary and sufficient cue in the perceived voiced-voiceless manner distinction has been

elevated to its present level of importance without having been observed under a sufficiently broad range of strictly defined experimental conditions.

Background of the Problem

The "Voiced/Voiceless" Distinction

English stop consonants are traditionally classed as either "voiced" (b-d-g) or "voiceless" (p-t-k). This descriptive classification was originally based on the presence or absence of low-amplitude glottal pulses during the interval of oral occlusion, just prior to the consonant release. Although, in initial position, the presence of this glottal "buzz" (or "voice bar", as it has been more recently termed in spectrographic analysis) would clearly indicate a voiced stop, it has been noted that this feature is not necessarily present in the case of certain English initial stop consonants which may nevertheless be perceived as voiced.

Because of such cases, it became necessary to conscript the use of another (previously secondary) feature, aspiration, to distinguish between the English voiced and voiceless stop categories in initial position. Aspiration can be seen in a spectrographic display as a noise component, appearing after the consonant release, whose frequency distribution is generally spread over the range of the second and third vowel formants. The presence of such a

noise component was cited as indication of a voiceless stop consonant. As a result, the added consideration of either voicing or aspiration was found to yield more consistent distinctions between the two manner categories. However, it became evident that for various languages, including English, the occurrence of these two key features was somewhat unpredictable, in that on the one hand, initial stops perceived as voiced might or might not exhibit a distinct voice bar, while on the other hand, those perceived as voiceless might or might not exhibit a distinct amount of aspiration in final position, or before unstressed vowels.

A third distinction between the stop consonant categories was based on acoustic qualities noted by Fletcher (1929). It was introduced as the 'tense/lax', or 'fortis/lenis' distinction. Fletcher discussed the two phonemic categories in terms of their relative amplitudes, and noted that the voiceless (tense) category (p-t-k) showed consistently higher audibility than the corresponding voiced (lax) category (b-d-g). In the course of his experimentation, he noted that the number of decibels of difference required for complete attenuation of a given fortis phoneme was consistently higher than that required for its lenis counterpart (Fletcher, 1929).

Jakobson, Fant, and Halle (1952) mentioned an apparent difference in the production of tense and lax consonants . . . : "Tense consonants are articulated with

greater distinctness and pressure than the corresponding lax phonemes (p. 38)." They proposed that previous distinctions could be viewed as largely redundant in light of this contrast. Thus, the tense/lax distinction alone was employed by Jakobson and his associates, in their phonological system, to separate the two groups of English phonemes. The proposed contrast was included in their set of "twelve binary oppositions" which they described as . . . "the inherent distinctive features which we detect in the languages of the world and which underlie their entire lexical and morphological stock . . . (p. 40)."

The nature of the tense/lax feature, however, must be questioned when referred to as dichotomous. This is especially true in light of the obviously continuous scales offered for the determination of tenseness/laxness, such as distinctness, pressure of articulation, muscular strain, tension, deformation of the vocal tract from neutral position, and segment duration. It should be mentioned that although this particular distinction bears a certain amount of interest in the present study, many have taken exception to the Jakobsonian phonological system, often solely on the basis of the highly arbitrary nature of features like tenseness/laxness, and on the grounds that such features may not truly be binary (see Fant, 1967).

Malecot (1970) argued in support of the tense/lax opposition, per se. Dealing strictly with physiological

evidence, he concluded that the opposition, based on a scale of "force of articulation", was apparently . . . "a linguistic reality but is primarily a synesthetic response to intrabuccal air pressure impulse, with closure duration perhaps playing a secondary role (p. 1591)."

If this distinction, still common in phonemic descriptions, were, in fact, independently capable of accounting for all English (p-t-k) : (b-d-g) contrasts, then reference to voicing (voice bar) or aspiration in the differentiation would be rendered clearly redundant. But, as noted by Lisker and Abramson (1964), . . . "it is too often the case to be accidental that voiceless and aspirated stops are discovered to be fortis, while voiced and unaspirated ones are at the same time lenis (p. 386)." Ultimately, it would seem only reasonable to ask what dimensions are actually in question in the perceived "voiced-voiceless" opposition, and further, what boundaries on those dimensions delimit the two perceived manner classes. With respect to the tense/lax opposition, neither of these questions has been unambiguously answered, leading one to conclude that the tense/lax distinction, so often co-operative with voicing and aspiration, lacks physical correlates which are truly independent of the other two features.

Questions regarding the isolation of relevant parameters, and notions of perceptual boundaries, might best be dealt with in terms of the so-called "categorical speech

perception" hypothesis as set forth by Liberman et al. (1957, 1958). Their notion was based on the premise that speech stimuli drawn from a physical continuum are perceived as members of discrete categories. Various subsequent experiments, conducted mainly at Haskins Laboratory, were aimed at the discovery of the relevant physical (physiological and acoustic) parameters upon which the language-user's perceptual processes rely.

The Concept of VOT

Lisker and Abramson (1964) began a series of studies in an attempt to find the dimension upon which the perceived voiced/voiceless distinction might be conclusively dependent. They proposed that Voice Onset Time (VOT) was the primary feature in the observed distinction, and defined VOT as the ". . . duration of the time inverval by which the onset of periodic pulsing either precedes or follows release (p. 387)." It was held by Lisker and Abramson that this fundamental timing relationship was the result of articulatory and glottal adjustments which were also responsible for the predictable co-occurrences of aspiration, voicing, and articulatory force (tenseness/laxness).

Implicit in Lisker and Abramson's position was the assumption that VOT was a primary unit of production, and that other features were simply its consequences. When stated in such terms, their interpretation of VOT could be easily introduced into the Motor Theory of Speech Perception

as a " . . . possible link between perception and articulation (p. 421)."

Evidence Proposed for the Universality of VOT

In a study which involved eleven languages, Lisker and Abramson (1964) elicited identificantions of stimuli drawn from natural speech. The languages observed were chosen from those which distinguished among two, three, or four consonant manner categories in initial position, before a vowel (Table 1). The fundamental issue in Lisker and Abramson's experiment was whether VOT alone could be used to determine phoneme class membership in any of the two, three, or four category languages. Their results would seem to indicate that this may not be the case.

The perceptual phoneme categories of the six two-cateogry languages were sufficiently distinguishable from one another on the basis of the features of voicing and aspiration, as shown in Table 1. Furthermore, the phonemes of the first two of the three-category languages, Eastern Armenian and Thai, were equally differentiated. However, according to Lisker and Abramson, a third feature, length, was required to differentiate between Korean's two classes of weakly and strongly aspirated voiceless initial stops. Finally, the initial stop consonants of both Hindi and Marathi, the four-category languages, were also classified satisfactorily on the basis of only the two features of voicing and aspiration.

TABLE 1

OCCURRENCE OF INITIAL STOP CONSONANTS

IN TWO-, THREE-, AND FOUR-CATEGORY LANGUAGES

Item	b-d-g	b^h-d^h-g^h	p-t-k	p^h-t^h-k^h
Voiced?	yes	yes	no	no
Aspirated?	no	yes	no	yes
Two-Category Languages				
English	x			x
Cantonese			x	x
Tamil	x		x	
Hungarian	x		x	
Spanish	x		x	
Dutch	x*		x	
Three-Category Languages				
E. Armenian	x		x	x
Thai	x*		x	x
Korean			x	x**
Four-Category Languages				
Hindi	x	x	x	x
Marathi	x	x	x	x

*Dutch and Thai do not exhibit an initial voiced velar.

**Korean bears a distinction between weakly- and strongly-aspirated voiceless initial stops.

VOT production values for initial stop consonants in all of the two category languages, plus Eastern Armenian and Thai, exhibited relatively distinct category divisions along the VOT continuum. Korean, however, presented a different case. The distribution of VOT's for the Korean initial unaspirated and "weakly-aspirated" stops showed a significant degree of VOT overlap, while the distribution for "strongly-aspirated" stops clearly stood alone. Lisker and Abramson defended VOT, stating that "'. . . while the distribution of values is thus somewhat anomalous, we cannot say with reasonable assurance that our measure of voice onset time fails to separate the three categories of Korean stops; it will certainly suffice to distinguish the aspirated set from the other two and it may still well be the single most important measure for separating the latter (1964, p. 403)."

Kim (1970) proposed, on the basis of cineradiographic evidence, that aspiration is an independent factor in the distinction of Korean initial stop categories. Kim found aspiration to be highly predictable, based on the size of the glottal opening at consonant release. Like Lisker and Abramson's approach to the articulatory foundations of VOT, Kim explained aspiration as a laryngeally controlled phonomenon. By showing a high correlation between size of glottal opening during consonant articulation and aspiration in the three Korean initial manner categories, Kim avoided the difficulty encountered by the Lisker and Abramson timing

relationship. In discussing their results, Lisker and Abramson (1964) did not attach any significance to the possibility that as in Korean, aspiration, or at least Kim's type of interpretation of it, might have been a more reliable index than VOT in distinctions among initial Korean voiceless stops.

A second observation in exception of Lisker and Abramson's hypothesis is not unlike the foregoing. Both of the four-category languages, Hindi and Marathi, presented overlapping VOT distributions in the production of aspirated and unaspirated voiced stops, while the voiceless unaspirated and voiceless aspirated stops composed the two remaining modes in the similar trimodal VOT distributions. As in the case of Korean, VOT may have been an insufficient measure in phoneme manner categorization.

With respect to the possibility of alternative explanations, their interpretation of results appears to be somewhat restricted: "To be sure, the voiced unaspirated and voiced aspirated stops show differences in average values that are almost systematic; nevertheless, they occupy ranges that are nearly co-extensive (p. 403)." The explanation offered in the case of Hindi and Marathi was as follows: "It seems very likely that the voiced aspirates are distinguished from the other voiced category by the presence of low amplitude buzz mixed with noise in the interval following release of the stop (p. 403)". Given that both

categories fall on the negative side of the VOT continuum (voicing precedes consonant release), it is difficult to justify such an explanation. If "low amplitude buzz" is to be equated, in this context, with the glottal waveform, then the distinction should have been reduced to one of aspiration. In any case, the ambiguity of this explanation is sufficient to suggest that, as in the case of Korean, additional measures, besides VOT, might have been more reliably enlisted in the observed distinction.

Additional counter-evidence has recently been presented by Caramazza, Yeni-Komshian, Zurif, and Carbone (1973). They examined VOT with respect to both the perception and the production of initial stop consonants by Canadian French, English, and bilingual French-English speakers. In both production and perception, the Canadian French speakers showed substantial VOT overlap for three classes (labial, apical, and velar) of initial stop consonants. Caramazza et al. concluded that these results " . . . strongly suggest that VOT is not a sufficient cue for the perception of voicing distinctions in Canadian French (p. 426)."

These authors also believed that their results warranted consideration with respect to the proposed universality of VOT. They reasoned that in the case of Canadian French, their data " . . . cast doubt on any theory assigning VOT a universal status in the total determination of the

phonetic dimensions of voicing, aspiration, and articulatory force (p. 426)", and offered the possibility that alternatives such as articulatory force or rate of formant transition may be more important in the Canadian French perceptual classification.

In light of the Korean, Hindi, and Marathi counter-evidence, Lisker and Abramson could not justify any general statement concerning the proposed universality of voice onset time as a perceptual cue. Nevertheless, the authors offered the following conclusion:

> " . . . this measure of voice onset time has been applied to word-initial stops in eleven languages and has been found to be highly effective as a means of separating phonemic categories, although these languages differ both in the number of those categories and in the phonetic features usually ascribed to them . . . It would seem that such features as voicing, aspiration and force of articulation are predictable consequences of differences in the relative timing of events at the glottis and at the place of oral occlusion (1964, p. 422)."

In addition, several authors (Eimas et al., 1971; Eimas and Corbit, 1973) maintained the universality of the VOT dimension, and continued to expand its range of applications.

Biological claims

In the first of two closely related studies, Eimas et al., (1971) were interested in possible biological foundations for the categorical perception of VOT. By

conditioning the non-nutritive sucking activity of one- and four-month old infants, when exposed to stimuli varying in VOT, it was hoped that the results would indicate first, a differential response to pairs of stimuli drawn from different adult English phoneme categories, and second, no differentiation between stimuli drawn from a single category. Stimuli were chosen from the set of synthetic CV syllables prepared by Lisker and Abramson for their own VOT experiments conducted earlier at Haskins Laboratory. When a subject had been habituated to a stimulus, a second stimulus was presented, and any changes in the rate of conditioned response (non-nutritive sucking) were measured.

The authors reported a significant increment in response rate when stimuli were drawn from different standard English perceptual manner categories, and no change when stimuli were drawn from a single manner category. They inferred from these results that " . . . the means by which the categorical perception of speech, that is, perception in a linguistic mode, is accomplished may well be part of the biological make-up of the organism (p. 306)." Offering equivocal conclusions, based on rather insufficiently defined evidence, this study was used as the basis for further extensions of the concept of VOT.

The Eimas and Corbit Model

Eimas and Corbit (1973) extended the notions of the assumed biological significance of VOT. Summarizing the

earlier works regarding VCT, such as Lisker and Abramson (1964), and Abramson and Lisker (1970), they presented the following characterization: "Of particular interest is the fact that the categorical nature of the perception of the VOT continuum appears to be universal (p. 101)." Moreover, the citation of Eimas et al. (1971) was presented as evidence of the pre-verbal occurrence of VOT discrimination, leading to the conclusion that "The apparent universality of this phenomenon suggests that it is a manifestation of the basic structure of the human brain (p. 101)." To propose that the foregoing assumptions are questionable on fundamental grounds should require no further justification than the equivocal results of those studies cited as supportive by Eimas and Corbit, since their subsequent experimental hypothesis was contingent upon the previously discussed analytical and biological claims.

Eimas and Corbit proposed that the categorical perception of VOT was mediated by a "selectively tuned linguistic feature detector" system. The system, they held, is composed of independent 'detectors', one of which is specifically tuned to a limited range of VOT values corresponding to voiced English initial stop consonants, and the other to the range of VOT values corresponding to voiceless initial stop consonants. Excitation of the former feature detector (FD_1) would induce the perception of a voiced consonant, while excitation of the latter

detector (FD_2) would induce the perception of a voiceless consonant.

Several important assumptions were required if Eimas and Corbit's model were to account for previous data from categorical perception experiments. First, the detectors had to be differentially sensitive, each to its own limited range of VOT values. Eimas and Corbit suggested that the sensitivity of a given detector " . . . might be measured, in principle, by the output signal of the detector (p. 108)." Next, since stimuli could presumably fall between the modal detection values of the two detectors, thereby exciting both simultaneously, it would have to be assumed that only the detector whose output signal was stronger would be recognized at higher levels of analysis in the auditory pathway. If, however, a stimulus were to excite both detectors equally, it was said to fall at the phonetic boundary (B).

Certain inferences can be drawn from the foregoing assumptions. For example, one might expect that the "-in principle-" measurement of the sensitivity of the feature detectors would provide (as projected in Figure 1) a graph of output signal strength, elicited by subjecting the detectors to a wide range of stimuli varying along the VOT continuum. The figure was constructed to agree with the properties discussed by Eimas and Corbit (p. 108). However, no such measurements were made by Eimas and Corbit.

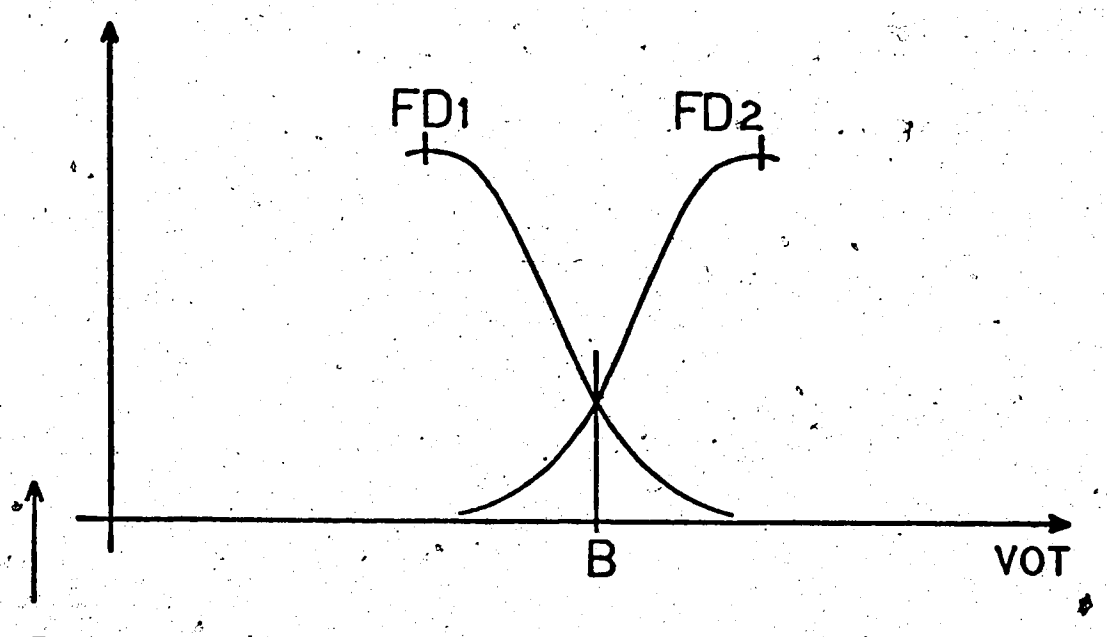


Figure 1: Normal output strength.

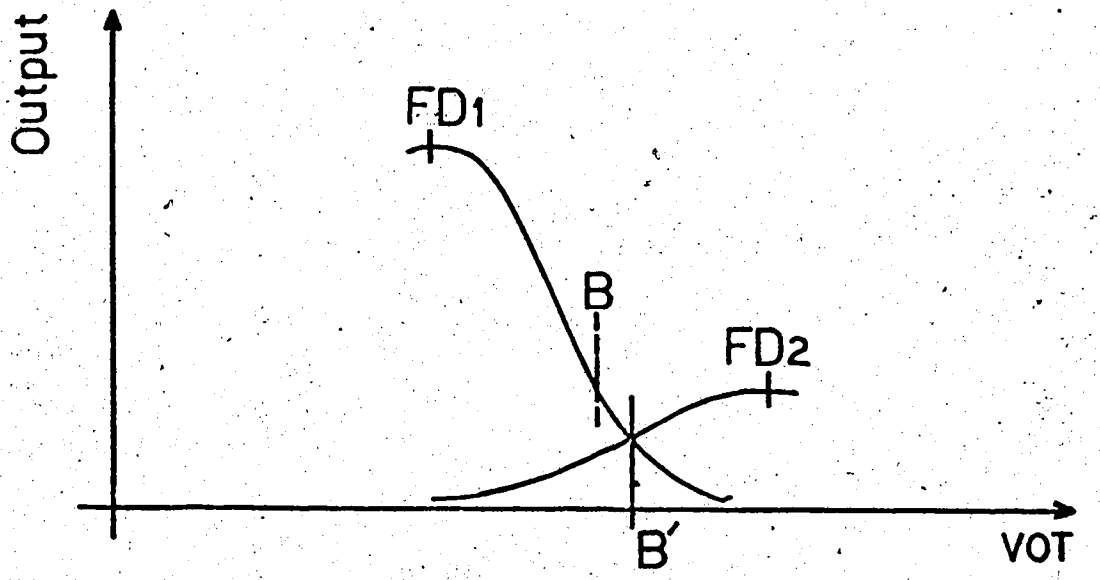


Figure 2: Output strength after adaptation of FD2 to optimum VOT.

Instead of pursuing the hypothesis that the feature detectors exist in the form described, Eimas and Corbit chose to take a step further and attempt to test the effects of adaptation on that mechanism. They postulated that repeated stimulation of either detector would cause fatigue, thereby lessening its sensitivity. The further assumed, "for purposes of simplicity (p. 108)", that the output signal strength of a fatigued detector would decrease equally for the entire range of VOT values to which it is normally sensitive. Figure 2 represents a projected graph of signal output strength after adaptation of FD_2. Note that Figure 2, like Figure 1, is qualified as this author's interpretation of the discussion by Eimas and Corbit (p. 108). They presented no such examples.

The effect that Eimas and Corbit hoped to observe in support of their proposed feature detector system was a shift in the phonetic boundary toward the adapted stimulus. Adaptation to a long VOT should have caused an upward shift in the phonetic boundary (as in Figure 2), while adaptation to a short VOT should have caused a downward shift in the phonetic boundary.

Here it is important to note certain shortcomings of Eimas and Corbit's model. First, given a binary response regardless of place of articulation, that is, a one-way classification of stimuli as either voiced or voiceless, a single "feature detector" would suffice. This is only one

area in which alternate models may have been considered. Second, in order to ensure accurate measurement of the phonetic boundary in the present model, it was necessary that the two feature detectors be 'specifically tuned' to overlapping ranges of VOT values. Although in this respect the efficiency of the proposed detectors falls somewhat short of ideal, it was on this basis alone that the authors could postulate changes in the phonetic boundary as a result of adaption procedures. In other words, had they assumed that the detectors were, in fact, sensitive to restricted, independent ranges of VOT values, then the adaption could not have been predicted.

Eimas and Corbit elicited identification functions for labial and dental stops (b-p, d-t), both before and after adaptation. Their experimental design employed a two- alternative forced choice (2AFC) paradigm, not unlike the forced choice methods used in earlier VOT studies, in which, for any given condition, both stimuli and responses were limited to a single place of articulation, and stimuli were to be identified as either voiced or voiceless. Using the Lisker and Abramson (1970) synthetic stimuli, initial identification functions were elicited for the labial and dental series, separately. Subjects were then asked to identify the same stimuli, after having been "adapted" to one of the other extreme of the VOT series for either the labial or dental stimuli. The obtained identification functions indicated a shift toward the adapting stimulus.

This shift was interpreted as a shift in the phonetic boundary, leading Eimas and Corbit to conclude that their model was, in fact, supported.

Eimas and Corbit attempted to rule out the possibility that any simple acoustic (as opposed to linguistic) variables might be held to account for the observed phenomenon. They showed that adaptation to an dental stop produced a shift in the identification function for labial stops. Conversely, adaptation to a labial stop produced a similar shift in the functions for dental stops. On the basis of this cross-series effect, they proposed that the detector system was not sensitive to "simple acoustic information", but only to "those complex aspects of the sound pattern that both series had in common, namely, voice onset time (p. 105)."

This assumption, however, may not be warranted, as evidenced by the fact that the proposed "phonetic boundary" for dental stimuli was several milliseconds greater than that for the labial stimuli (cf. p. 105), suggesting that VOT was dealt with in a somewhat different fashion for the two places of articulation. In order to deal with this particular phenomenon (to which Eimas and Corbit attached no apparent significance), and yet maintain the original notion of fixed VOT-detectors, it would appear necessary to postulate either a separate set of VOT-detectors for each place of articulation, or an independent mechanism,

capable of translating VOT's in different contexts, or places of articulation, into some unambiguous form amenable to the feature detector mechanism. Furthermore, it is not at all clear that the only feature shared by both series was VOT.

One might conclude that the mechanism described by Eimas and Corbit might have been more reasonably characterized as a highly complex mechanism, composed of numerous "feature detectors", all of which contribute to the gross function of the mechanism. For example, consider any complex mechanism, composed of smaller, functional units. The loss or alteration of the function of any of the compone s through adaptation procedures will, to a certain degree, affect the output function of the mechanism. Thus, experimental alterations of gross function (much like the results cited by Eimas and Corbit) would not be a clear indication of the specific properties of the mechanism, unless those lower-level functions had been anticipated. In light of possible alternative explanations, Eimas and Corbit's results are held to offer no unequivocal support of their particular model.

Statement of the Problem

The issue of greatest importance here is not simply whether the hypotheses put forward by Eimas and Corbit (1973) were supported. The main question to be offered is whether VOT alone, as originally defined by Lisker and Abramson (1964),

is truly a sufficient criterion for the assignment of phonemic status to initial stop consonants.

If one were to regard certain established 'cues', such as aspiration and force of articulation, as analytical consequences of articulatory and glottal gestures, then there would be no argument against the analytical necessity of VOT, since Lisker and Abramson have also discussed VOT in terms of essentially the same relative articulatory and glottal adjustments. On the other hand, to say that VOT alone is sufficient for the perceived distinction is to ignore other features which have been experimentally established as influential in the perceived voiced/voiceless distinction.

It would appear that a study of the effects of various treatments on the perceptual impact of VOT is in order. The acceptability of VOT as a universal perceptual cue should be contingent upon the replicability of earlier results under a broader range of well-defined experimental conditions. Furthermore, a critical re-evaluation of key assumptions underlying the concept of VOT and its acoustic consequences may provide a framework from which more reasonable accounts of observed phenomena may emerge.

The present study was designed to examine the nature of the effects of place of articulation, manner of articulation, and white noise masking on the distribution of stimulus identifications as measured on the VOT continuum. Using

the median VOT value of a given response distribution as a measure of performance, the various factors were introduced on the basis of several considerations.

First, since VOT is by definition totally independent of place of articulation, median VOT values for one place of articulation should not differ significantly from those of another place of articulation. By limiting stimuli and responses to a single place of articulation, the two alternative forced choice (2AFC) paradigm, as used in earlier VOT studies, could not have provided any avenue for the direct comparison of different places of articulation. The present study, however, employed a four alternative (4AFC) paradigm in which mixed labial and dental stimuli were not only to be identified as implicitly voiced or voiceless, but were also to be classified with respect to place of articulation.

Second, manner of articulation (voiced or voiceless) may, with certain qualifications, be regarded as the only factor in the present experiment in any way associated with the perceptual distinction between voiced and voiceless initial stops in English. If subjects are indeed capable of assigning the phonemic status of voiced or voiceless to stimuli which vary in VOT, then according to previous results, the distributions of voiced and voiceless responses should occupy relatively discrete ranges of VOT. That is, 'voiced' responses should be concentrated at the low end of

the VOT scale, while 'voiceless' responses should be concentrated at the high end. Under such circumstances, it follows that the median VOT values of those distributions should be consistently different from one another, as long as subjects maintain a discrete mode of response.

Third, white noise masking was chosen as a factor on the basis of its possible implications in two areas. First, if the temporal measure of VOT is, in fact, the sole perceptual cue in the voiced/voiceless distinction, then the masking of frequency information during the VOT interval should have no effect on observed patterns of stimulus identification. Moreover, with respect to the "linguistic feature detector" mechanism proposed by Eimas and Corbit (1973), additional motivation is provided by considerations in the domain of signal detection.

The separate effects of adaptation and of noise on a multiple detector mechanism are fundamentally different. Adaptation is defined as a decrease in the output of a single receptor (in this case, a 'tuned detector') in response to a constant stimulus. The addition of random noise to a signal should not have this effect. The effect of random noise on a receptor system has been described by Green and Swets (1966) in terms of 'interference', or uncertainty in sensory measurements. For fixed noise background levels, the amount of uncertainty is evenly distributed over the entire range of sensation of the particular modality.

Thus, random noise masking should not have a selective effect on a single receptor or output distribution, as would adaptation.

Finally, with respect to the notion of "fixed" neural detectors, subjects should not be expected to differ significantly in patterns of performance. One important aspect of the present study is the consideration of the possibility of differences among subjects.

CHAPTER II

METHOD

Stimuli

In nearly all respects, the synthetic CV syllables used as stimuli in the present experiment were identical to those used by Eimas and Corbit (1973). In order to achieve some understanding of the types of differences between the two sets of stimuli, it may be of assistance to first present a brief description of the latter.

The stimuli used by Eimas and Corbit were those generated by Lisker and Abramson in 1967 at the Haskins Laboratories for use in their early cross-language tests. They used a parallel resonance synthesizer, equipped with " . . . three formant resonators with variable frequencies and amplitudes, a choice of buzz or hiss excitation, or a mixture of the two, and control of the overall amplitude and fundamental frequency (Lisker & Abramson, 1967, pp. 563-564)." All of the synthesized initial consonants were followed by a three-formant approximation to the vowel [a], and the labial, apical, and velar place categories were achieved by the addition of appropriate release bursts and formant transitions at the beginning of each syllable. Voicing before consonant release (voicing lead) was represented by the insertion of low frequency harmonics of the buzz source. Voicing lag, voice onset after consonant release, was achieved by supression of the onset of the first formant, relative to the second and third formants, while filling

the interval between consonant release and voice onset with hiss excitation of the second and third formants to approximate aspiration. Of the broad range of 37 VOT variants in each place series (150 msec before release to 150 msec after release) synthesized by Lisker and Abramson, Eimas and Corbit used only 14 from the labial series (-10 msec to +60 msec) and 14 from the dental series (0 msec to +80 msec).

The present experiment employed a Parametric Artificial Talker (PAT) in the synthesis of the required stimuli (see Anthony & Lawrence, 1962, for original concepts in the development of the PAT). The PAT was driven and controlled by a PDP-12 digital computer. Included in PAT's eight control functions are: Frequencies of the three variable formants, frequency and amplitude of the glottal waveform, amplitude of the wide-band noise source for excitation of the formants, and finally, the central frequency and amplitude of the second wide-band noise source, which is independent of the formants. This second noise source, most commonly employed in the synthesis of fricatives, was not needed for the purposes of the present study. Control voltages for the remaining six parameters were graphed as a function of time (Figure 3), digitized by means of a Hewlett-Packard F-3B Line Follower, and stored in computer memory (see Hill, 1969, and Akitt, 1970, for further details of PAT programming in conjunction with the PDP system). A general overview of various speech synthesizers is also available in Cooper (1961).

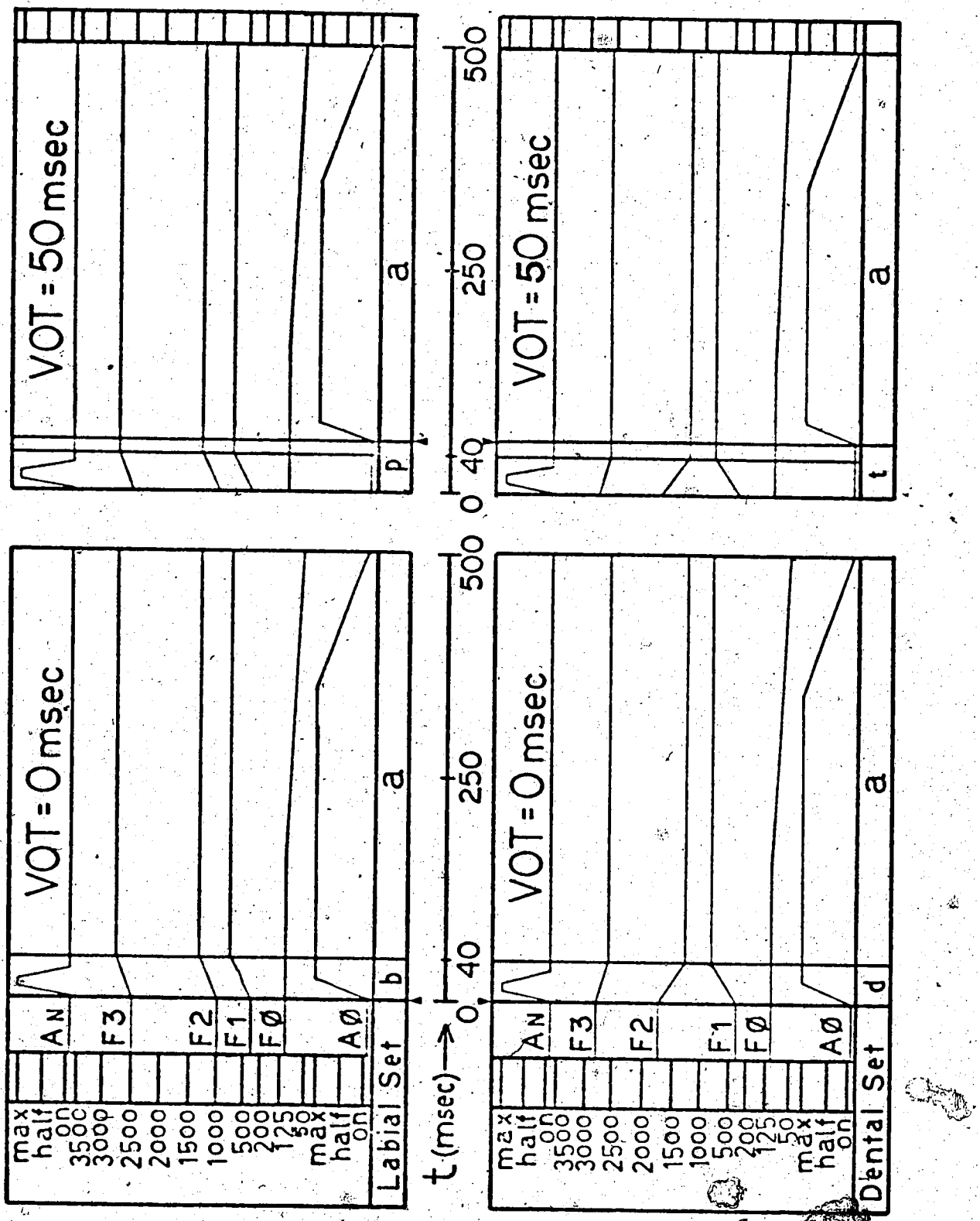


Figure 3: PAT control functions for the synthesis of two labial and two dental stimuli.

Two sets of stimuli, a labial series and a dental series, were prepared, each containing 14 VOT variants. The appropriate (Delattre, et al., 1955) formant loci and transitions were incorporated to identify initial consonants as either labial or dental. All of the 500 msec CV syllables ended with a three-formant approximation to the vowel [a], with a fundamental frequency of 110 Hz, dropping toward the end.

With duration of transition set at 40 msec, the three vowel formants for the labial series in ascending order, started at approximately 370 Hz, 1000 Hz, and 2500 Hz, rising to 760 Hz, 1280 Hz, and 2650 Hz, respectively. Formants for the dental series, in ascending order, started at approximately 370 Hz, 1800 Hz, and 2800 Hz, with F1 rising to approximately 760 Hz, while F2 and F3 fell to 1280 Hz and 2650 Hz, respectively. Variations in VOT were achieved by shifting the onset of the glottal waveform with relation to the release burst. VOT variants for the two series were the same as those used by Eimas and Corbit; the range of VOT for the labial set was from -10 msec to +60 msec and for the dentals, from 0 to +80 msec. In both series, VOT varied in 5 msec steps, up to +50 msec, after which VOT increased in 10 msec steps.

Recordings

Ten different randomizations of the entire set of 28 labial and dental stimuli were divided into two equal

Blocks of five different randomizations (140 presentations) each. The stimuli, separated by two seconds of silence, were then recorded on magnetic tape, in analog form, from PAT output, using a TEAC A-7030 tape-deck. The final recording consisted of five complete Blocks of stimuli (Block order 1-2-1-2-1), or 25 presentations of each stimulus.

During the experiment, the recorded stimuli were played back through the left channel of the tape-deck. Peak rms syllable value was held constant at optimum playback level, as monitored by the left channel VU-meter. As a source of the white-noise masking signal a GRC (General Radio Company) 1382 Random Noise Generator was used. Precise variations in signal-to-noise ratio were achieved with a Hewlett-Packard 350D Attenuator Set. The white noise output of the noise source passed through the attenuator, and finally, into the right channel of the tape-deck, where rms value could be monitored (see Figure 4). By regulating the white noise output level, signal and noise were balanced. For S/N=0 peak rms syllable value and rms white noise value were equal. Holding the signal constant, it was then possible to vary the attenuation of the white noise in 10 dB steps, from zero attenuation (S/N=-20 dB) to 40 dB attenuation (S/N=+20 dB, white noise barely audible).

Presentation

The two separate output channels from the tape-deck were mixed, using a Braun CSV 250 Power Amplifier, and the

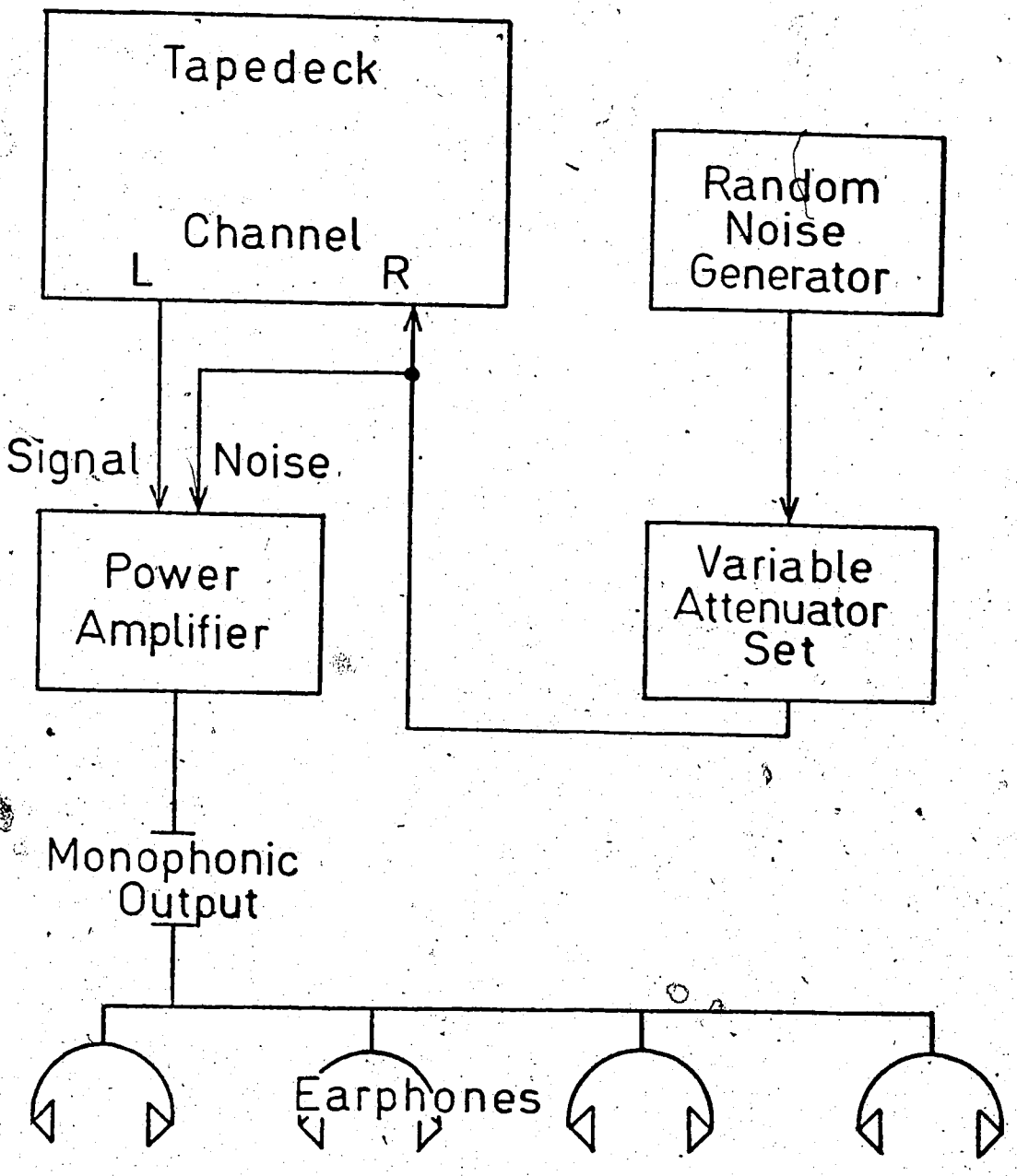


Figure 4: Shematic diagram of experimental apparatus.

resulting monophonic output was presented to subjects binaurally, at a comfortable listening level, using Telephonics TDH-49 earphones, with MX41/AR cushions. The level of white noise was changed after each Block presentation, in such a way as to prevent any two consecutive Blocks from having the same level of background noise. On the basis of S/N ratio, subjects were divided into two equal groups of five as indicated in Table 2. The order of white noise variation for Group 2 subjects was basically a reversal of that presented to Group 1 subjects. Group 1 and Group 2 subjects were tested separately, each normally in three separate one-hour sessions, of five, ten, and ten Blocks, respectively; the first session included a set of pre-recorded instructions (see Appendix A). Sessions were separated by at least 24 hours.

In what was essentially a 4AFC procedure, subjects were required to identify each stimulus as belonging to one of the four possible response categories, [p], [b], [t], or [d], and to record each response, according to presentation number and category, on a prepared IBM answer sheet. Responses were tabulated by an IBM optical scoring machine. The final set of raw data for each subject included 25 identifications of each of the 28 stimuli under each of the five signal-to-noise conditions (25 x 28 x 5 = 3500 responses for each subject). The basic experimental design for stimulus presentation is represented in Table 3.

TABLE 2

DISTRIBUTION OF S/N CONDITIONS BY SESSION

TO TWO GROUPS OF FIVE SUBJECTS

Group	Session	Block Number				
		1	2	1	2	1
1	1	+20	+10	0	-10	-20
	2	+10	0	-10	-20	+20
		0	-10	-20	+20	+10
	3	-10	-20	+20	+10	0
		-20	+20	+10	0	-10
2	1	-20	-10	0	+10	+20
	2	-10	0	+10	+20	-20
		0	+10	+20	-20	-10
	3	+10	+20	-20	-10	0
		+20	-20	-10	0	+10

TABLE 3

BASIC DESIGN FOR ELICITATION OF PSYCHOMETRIC FUNCTIONS FOR A SINGLE SUBJECT

SERIES	MASKING CONDITION	R	VOT			
			1	2	. . .	14
LABIAL	1	1 2 . . 25				
	2	1 2 . . 25				
	.	.				
	.	.				
	.	.				
	5	1 2 . . 25				
DENTAL	1	1 2 . . 25				
	2	1 2 . . 25				
	.	.				
	.	.				
	.	.				
	5	1 2 . . 25				

Subjects

Subjects were ten native English-speaking female student volunteers from the University of Alberta, whose ages ranged from 18 to 36 years. None of the subjects had had any experience with synthetic speech. All were subjected to standard Pure Tone Audiometric (sweep) Tests, with the use of a Beltone 10-D Audiometer. The mean results (both ears) for each subject are presented in Appendix B. With only one exception, subject HA, age 33, who showed a decrease in right ear sensitivity to frequencies above 6 kHz, all were found to be normally sensitive to frequencies in the speech range, with threshold criterion set at 15 dB.

CHAPTER III

RESULTS

Data Preparation

Previous experimenters have generally applied a Two Alternative Forced-Choice (2AFC) method to elicit responses. For a given place of articulation, say, labial or dental, a particular VOT variant was to be identified as either voiced or voiceless. Cumulative frequency distributions of category identifications by VOT were readily constructed, and the 50% cross-over point was taken as an overall index of the subject's performance under a given set of conditions.

Such a direct index was not readily available in the present experiment, which employed a Four Alternative Forced-Choice (4AFC) paradigm. Labial and dental stimuli were mixed, and presented at random, and subjects were required not only to implicitly identify stimuli as either voiced or voiceless, but also to categorize stimuli according to place of articulation (labial or dental). As a result of subsequent confusions in place of articulation, the cumulative frequency distributions of voiced and voiceless responses, based on total presentations for a given place of articulation, were asymmetrical. Thus, it became necessary to employ an alternate index of the subject's response.

It was believed that stimuli incorrectly classified

with respect to place of articulation (place errors), regardless of manner identification, could be treated as "errors". Subsequently, they were dealt with independently of the remaining "correct" responses. Discussion of errors appears in Chapter IV.

Dealing only with "correct" responses, it was apparent that the two cumulative frequency distributions of voiced and voiceless identifications would have to be represented by separate VOT values. That is, a single index was needed to represent each of the separate voiced and voiceless response distributions elicited for each of the two stimulus series (labial and dental). Given that the distributions under consideration in the present study are of an asymmetrical nature, the mean would present a biased estimate of central tendency. The asymmetry was sufficient to suggest the distribution mid-point as the preferred statistic to represent central tendency in the analysis of "correct" responses. The resulting VOT index values for voiced and voiceless identification distributions (Appendix C) were used as basic data points in the experimental design for the analysis of variance presented in Table 4.

Analysis of Variance

The analysis of variance can be described as a complete 2 (place of articulation) by 2 (manner of

TABLE 4

EXPERIMENTAL DESIGN FOR ANALYSIS OF VARIANCE

S/N	Ss	LABIAL		DENTAL	
		Voiced	Voiceless	Voiced	Voiceless
1	1 2 3 . . . 10				
. . .	. . .				
5	1 2 3 . . . 10				

articulation) by 5 (masking level) factorial design with repeated measures on all factors, for ten subjects. Each of the 20 cells in the present design (Table 4) contained the corresponding VOT index values computed for each of the ten subjects. The results of the analysis of variance are presented in Table 5.

Place by Masking Interaction

Although significant overall F-ratios were observed for both the Place and Masking factors, the interaction between them was also significant ($F=3.798$, $p<0.01$). Thus, the two factors cannot be discussed independently. The interaction is represented by Figure 5, in which VOT (ordinate) is plotted as a function of masking level (abcissa) for the two places of articulation.

Figure 5 indicates that, as a result of masking, VOT values for the dental series varied over a broader range than did VOT values for the labial series. An *a posteriori* Newman-Keuls test (Ferguson, p. 274) for significant differences among means was performed on VOT's for the labial and dental series separately. The results of the Newman-Keuls tests are presented in Tables 6(a) and 6(b).

Table 6(a) would seem to indicate that VOT's for the labial series did not undergo any significant change as a function of masking, while Table 6(b) reveals that at

TABLE 5

ANALYSIS OF VARIANCE

MEDIAN VOT VALUE AS A FUNCTION OF
PLACE OF ARTICULATION, MANNER OF ARTICULATION
AND MASKING LEVEL

SOURCE	SS	df	MS	F
Place Labial/Dental (A)	48.585	1	48.585	15.21***
Manner Voiced/Voiceless (B)	23358.87	1	23358.87	7315.65***
Masking Level (C)	179.049	4	44.762	14.001***
A x B	1.086	1	1.086	0.340
A x C	48.511	4	12.128	3.798**
B x C	16.708	4	4.177	1.301
A x B x C	13.970	4	3.492	1.093
Subjects R (ABC)	623.784	9	69.309	21.7065***
Error	545.9460	171	3.193	

**p 0.01

***p 0.001

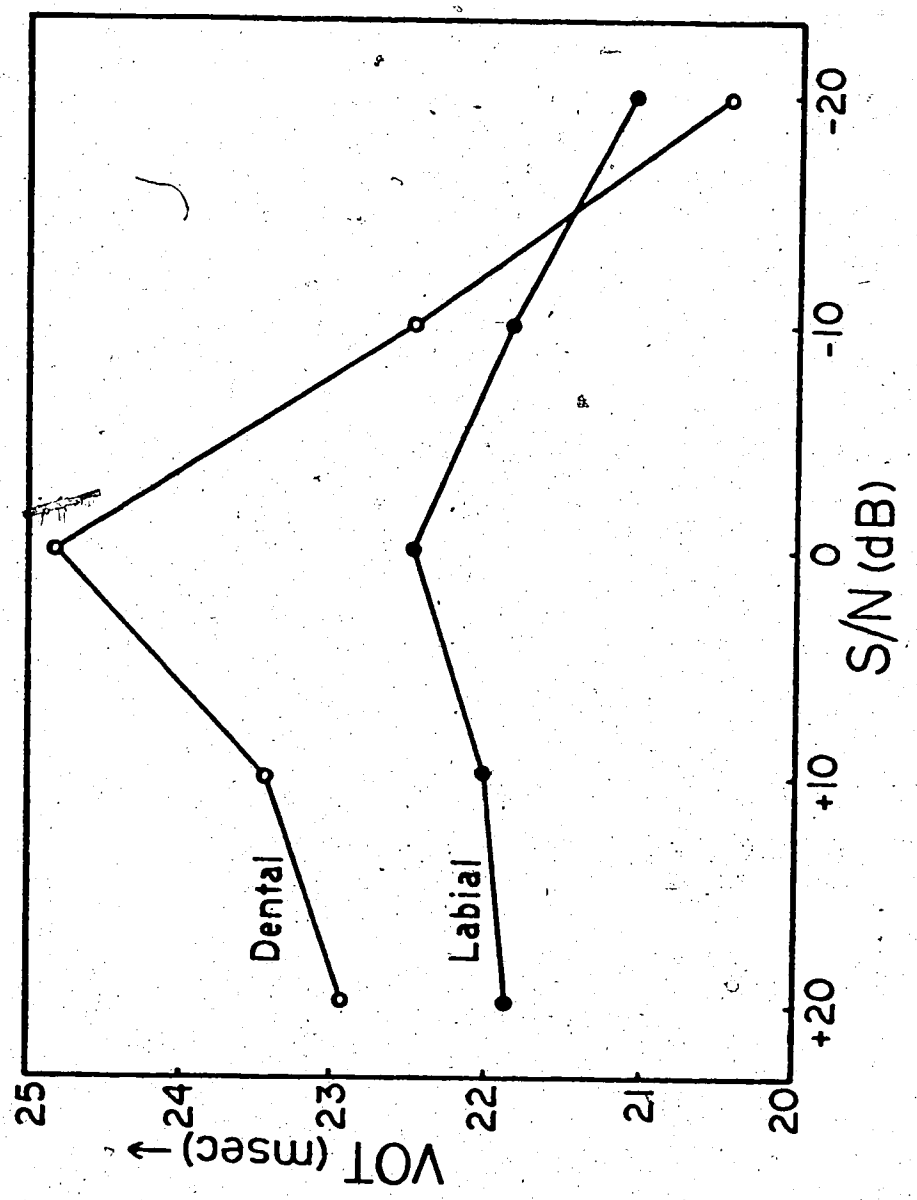


Figure 5: VOT as a function of S/N ratio.

TABLE 6

NEWMAN KEULS TESTS FOR SIGNIFICANT DIFFERENCES AMONG MEAN VOT VALUES BY S/N RATIO

(a) LABIAL SERIES

S/N ratio	Mean VOT's in order of magnitude	Differences Among Means			
		-10 dB	+20 dB	+10 dB	0 dB
-20 dB	21.05	0.80	0.82	0.96	1.42
-10 dB	21.85		0.02	0.16	0.62
+20 dB	21.87			0.14	0.60
+10 dB	22.01				0.46
0 dB	22.47				

(b) DENTAL SERIES

S/N ratio	Mean VOT's in order of magnitude	Differences Among Means			
		-10 dB	+20 dB	+10 dB	0 dB
-20 dB	20.43	2.04*	2.54**	3.03**	4.40**
-10 dB	22.47		0.50	0.99	2.36*
+20 dB	22.97			0.49	1.86
+10 dB	23.46				1.37
0 dB	24.83				

*p 0.05
**p 0.01

masking levels of 0 dB, -10 dB, and -20 dB, VOT's for the dental series were, in fact, significantly different. This would suggest that even though the labial and dental profiles in Figure 5 are in some respects visually similar, the effects of masking may only be realized in the case of S/N ratios below 0 dB for the dental series alone.

Why the dental series should be more sensitive in this respect to white noise masking is not a question which can be simply answered. This question will be discussed in Chapter IV. It will suffice at this point to state that VOT values for the labial series remained much more stable throughout the observed range of white noise masking than did those for the dental series.

Manner of Articulation

The exceptionally high F-ratio for manner categories ($F = 7315.65$; $p < 0.001$) was not altogether unexpected, and reflects the fact that this factor was by far the largest source of variation in the present analysis. VOT values for each of the four response distributions are plotted as a function of S/N ratio in Figure 6. Directly observable in Figure 6 is the fact that for either place of articulation the large difference (in msec) between VOT values for 'voiced' and 'voiceless' response distributions remained quite stable throughout the entire range of masking conditions. On the basis of the consistently separate VOT values, it may be argued that white noise masking had

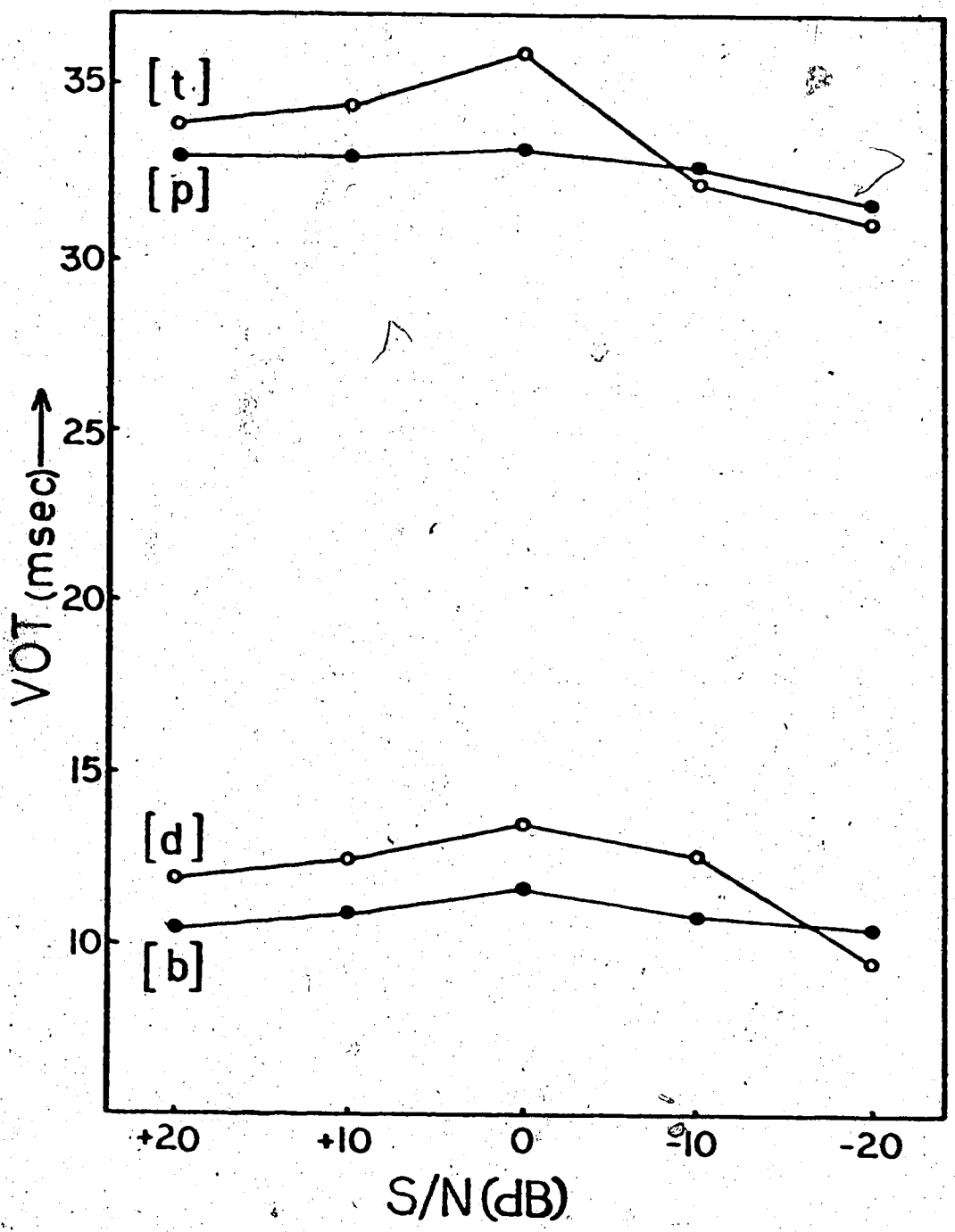


Figure 6: VOT's for each of the four response categories as a function of S/N ratio.

no effect on the ability of subjects to label stimuli as voiced or voiceless in a discrete fashion, on the basis of VOT.

Subjects

Differences among subjects were found to be significant ($F = 21.707$; $p < 0.001$). Although the specific nature of those differences is not directly considered in the present study, the significant F-ratio associated with inter-subject differences indicates that one must approach the notion of 'fixed' VOT detectors with some caution.

CHAPTER IV

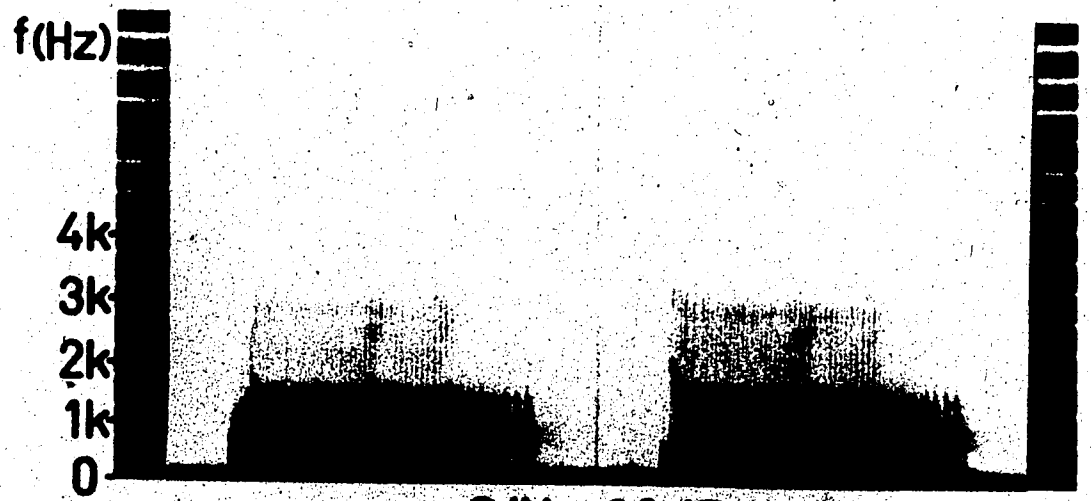
DISCUSSION

Although various interpretations can be offered to account for the results of the present study, above all, it is apparent that VOT is, in fact, an effective signal characteristic in the identification of voiced and voiceless initial stop consonants in English. However, in light of the interaction between the effects of place of articulation and masking on VOT values, complemented by the results of the *a posteriori* Newman-Keuls tests, it must be acknowledged that certain aspects of the dental stimuli rendered them more vulnerable than the labial stimuli to the effects of white noise masking.

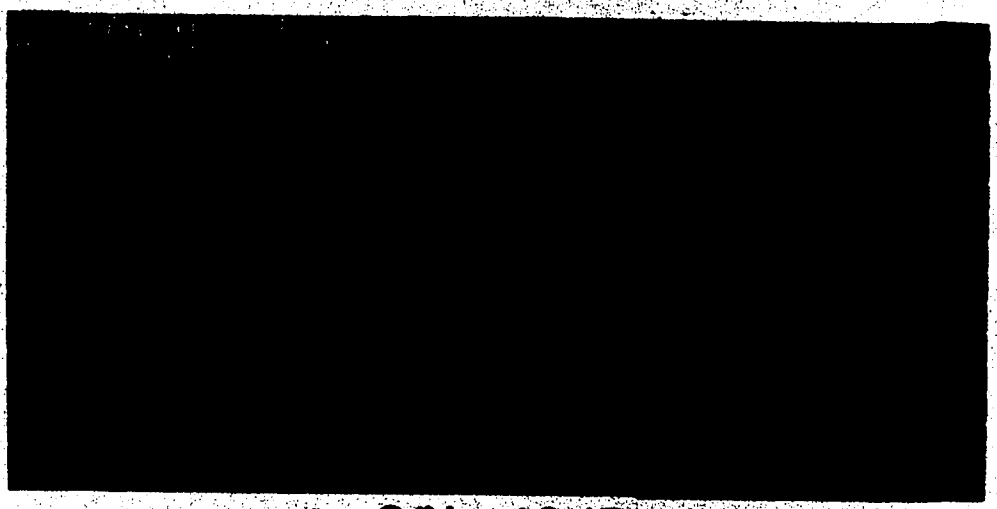
At high levels of masking, the effects of white noise on frequency domain properties of the stimuli was sufficient to cause a significant shift in the VOT values for the dental response distributions, while VOT values for the labial response distributions remained stable. This leads to the assumption that in addition to VOT, some concurrent frequency domain stimulus features may play an important role in the identification of voiced and voiceless initial stops, even though those same features might vary according to place of articulation.

Further motivation for the consideration of alternate stimulus features stems from the results of spectrographic analysis of the stimuli. The spectrograms in Figure 7 suggest that at S/N ratios of -10 dB and -20 dB, the

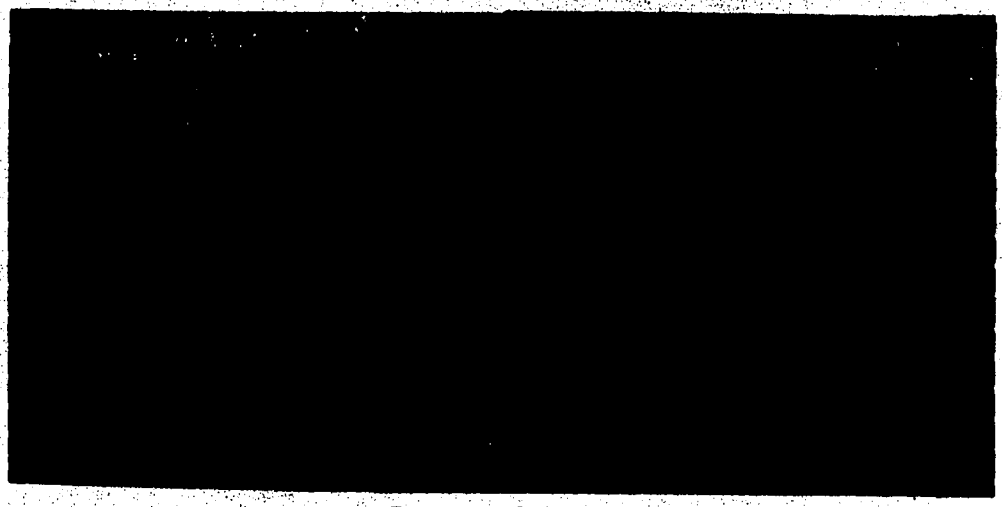
Figure 7(a): Sound spectrograms of a labial (left) and a dental (right) stimulus, at 0 msec VOT, under three S/N conditions. (time along abcissa)



S/N=+20dB



S/N=-10dB



S/N=-20dB

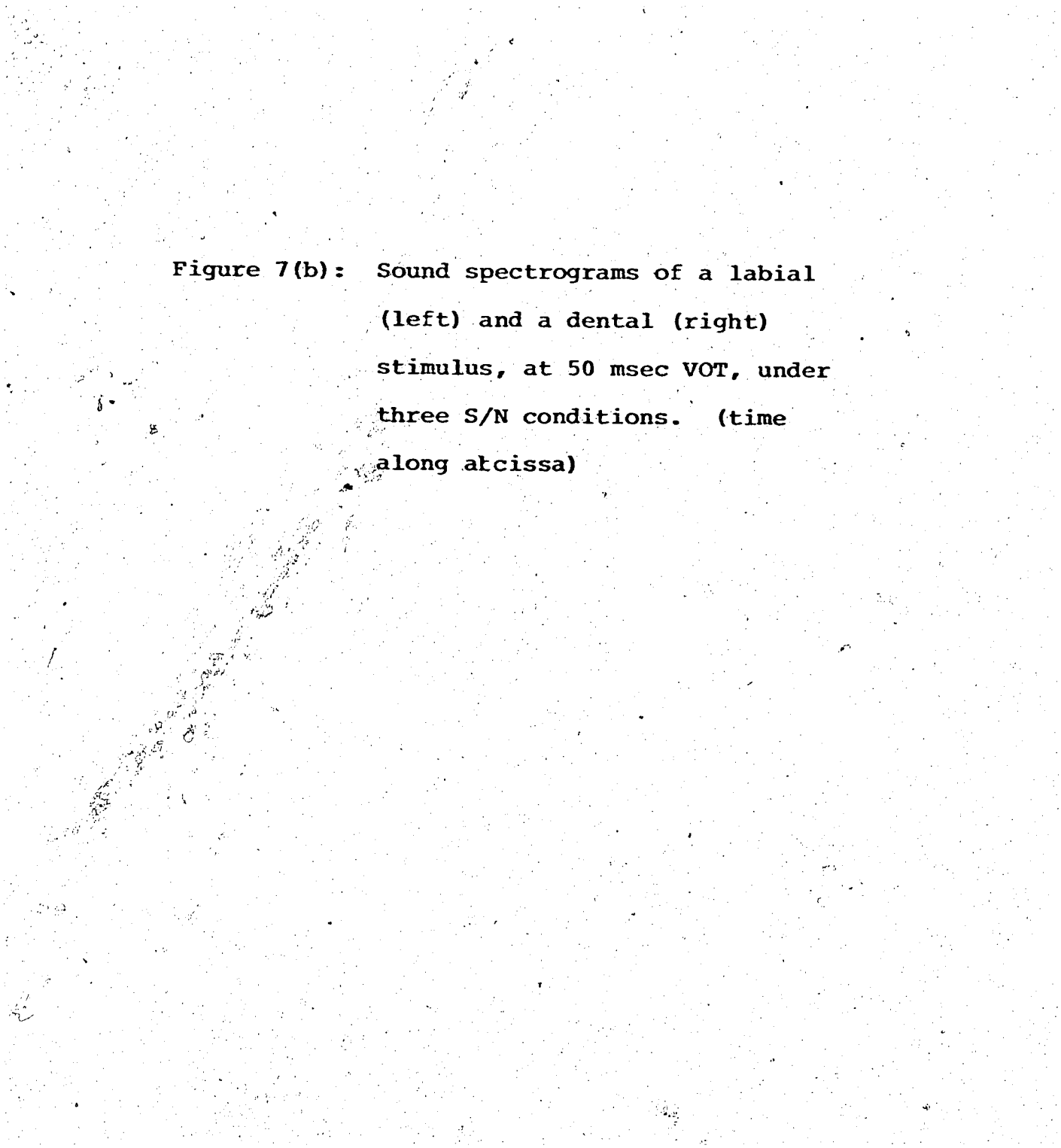
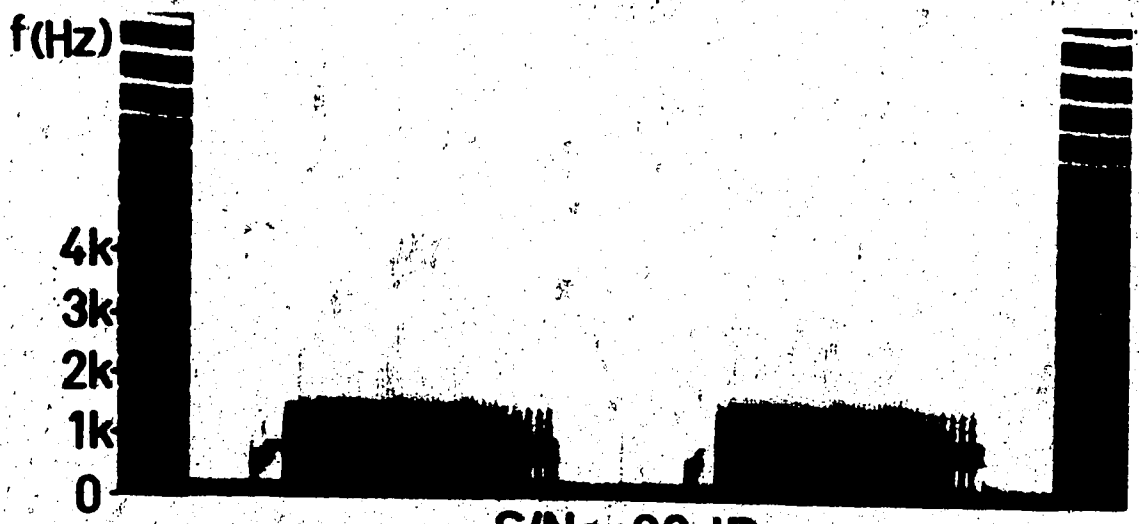
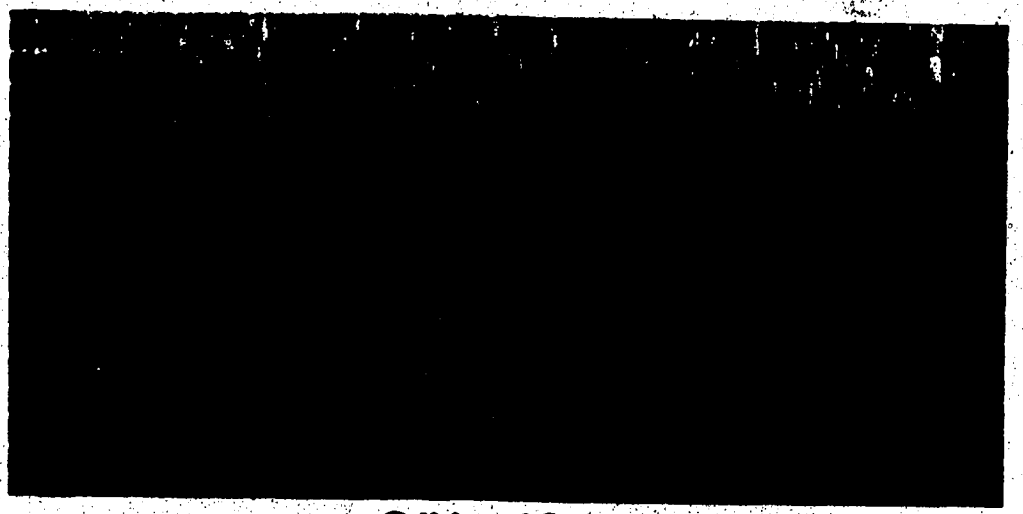


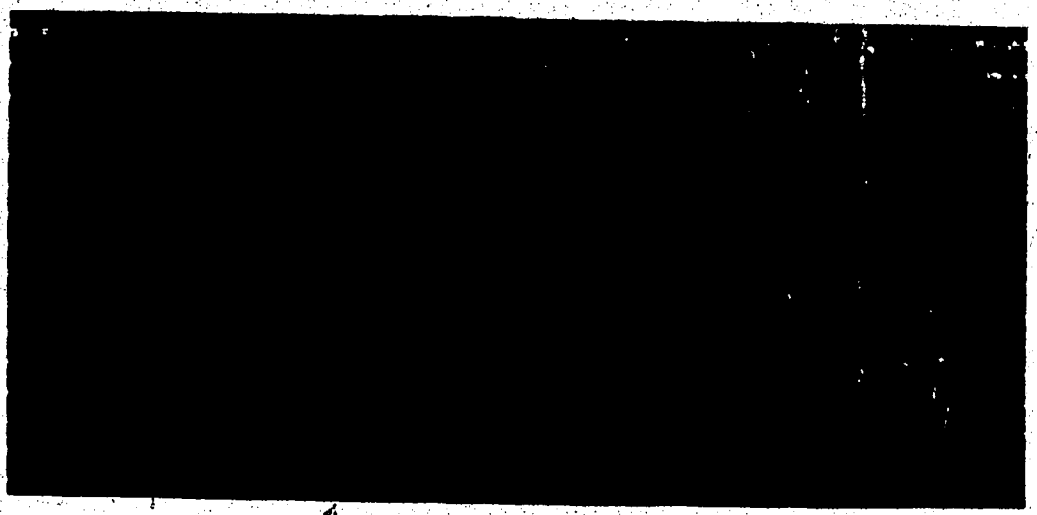
Figure 7(b): Sound spectrograms of a labial (left) and a dental (right) stimulus, at 50 msec VOT, under three S/N conditions. (time along abcissa)



S/N=+20dB



S/N=-10dB



S/N=-20dB

initial noise burst signalling consonant release may have been completely masked, rendering VOT virtually undetectable. Under such circumstances, it follows that the voiced/voiceless distinction may have been sustained entirely on the basis of features other than VOT.

By removing the initial noise burst from the present stimuli, and examining subsequent stimulus identifications, these contingencies could be more directly examined. Within the limits of the present study, however, one possibly beneficial approach to the relative perceptual importance of additional stimulus features might be provided by an examination of the confusions of place of articulation.

The Confusion of Place Categories

Figure 8 represents the number and percent of confusions of place of articulation under each masking condition, for each of the VOT variants from 0 msec to 60 msec (labial and dental series combined). This range covers all VOT values which overlapped both place series. As one would expect, place-errors were most pronounced at the highest level of background noise. Moreover, there appears to be a certain S/N ratio (0 dB) above which errors are negligible (maximum error $\leq$ 2%), and below which errors increase radically. The increase in errors is not equally spread over the entire range of VOT; at higher masking levels, relatively more errors resulted in the range of VOT above 25 msec.

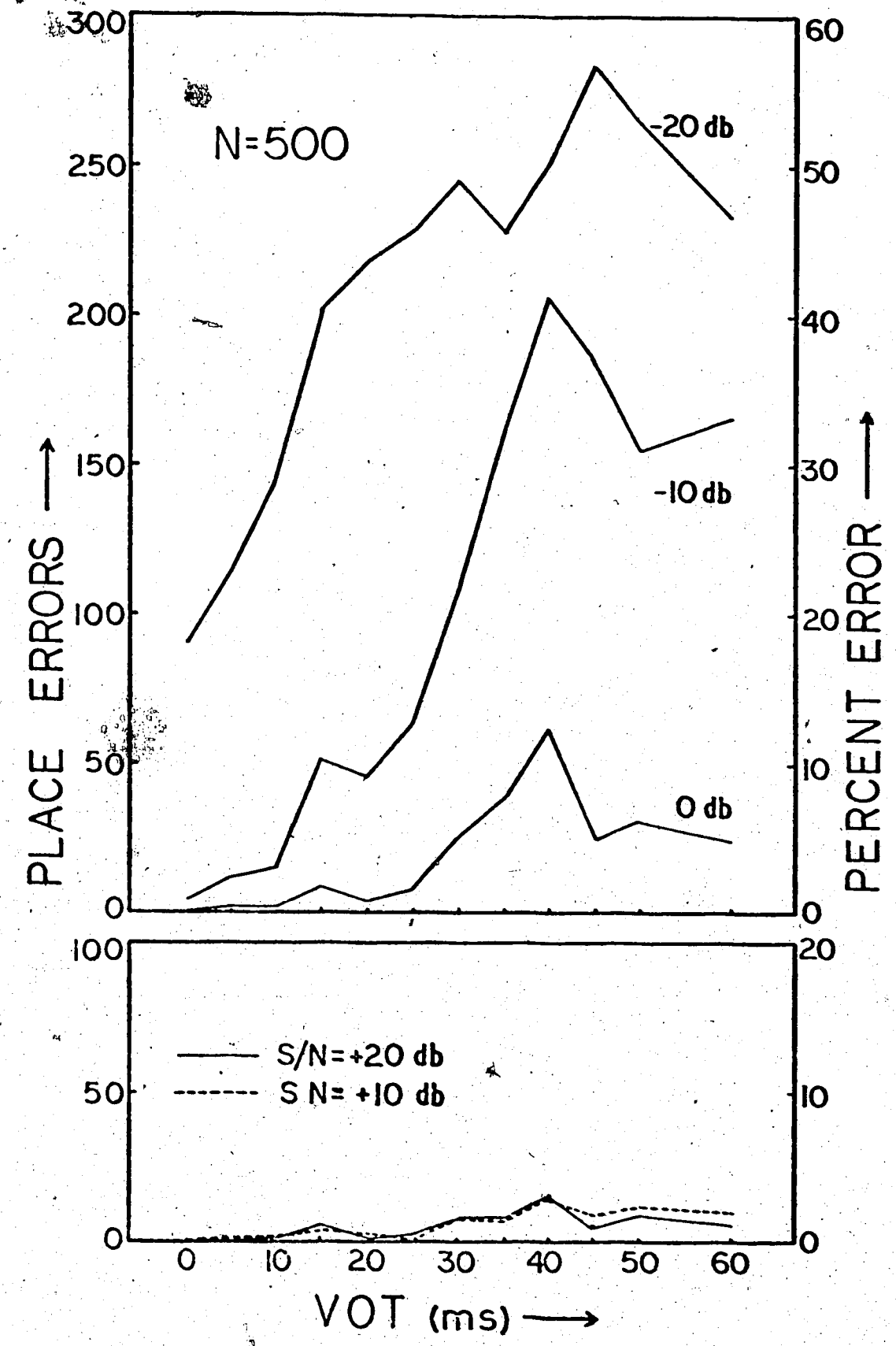


Figure 8: Place errors under each S/N condition.

The increase in error as a function of VOT is plotted in Figure 9. The frequency polygons in Figure 9(a) represent the number of times each stimulus from the dental series was identified as a labial, while the reverse holds for Figure 9(b). In both cases, place-errors are sub-divided according to manner category of response (voiced or voiceless). Once again, it is evident that dental and labial stimuli are treated differently by subjects. With respect to place of articulation, dental stimuli were more often mis-classified than labial stimuli.

In light of the above observations, it would not appear unwarranted to assume that as the masking level increased, so did the loss of acoustic information necessary to determine the place category of a particular stimulus, and furthermore that the loss was more substantial for dental stimuli than for labial stimuli. However, it is evident that at least minimal information about manner of articulation remained at even the highest masking level, in order to enable subjects to identify stimuli from the high end of the VOT scale as 'voiceless', and those from the low end as 'voiced'. This observation, in conjunction with the analysis of variance results, supports earlier claims that certain cues for place and manner of articulation are independent.

The relationship between S/N ratio and 'transmitted information' (in an information-theoretical sense) was

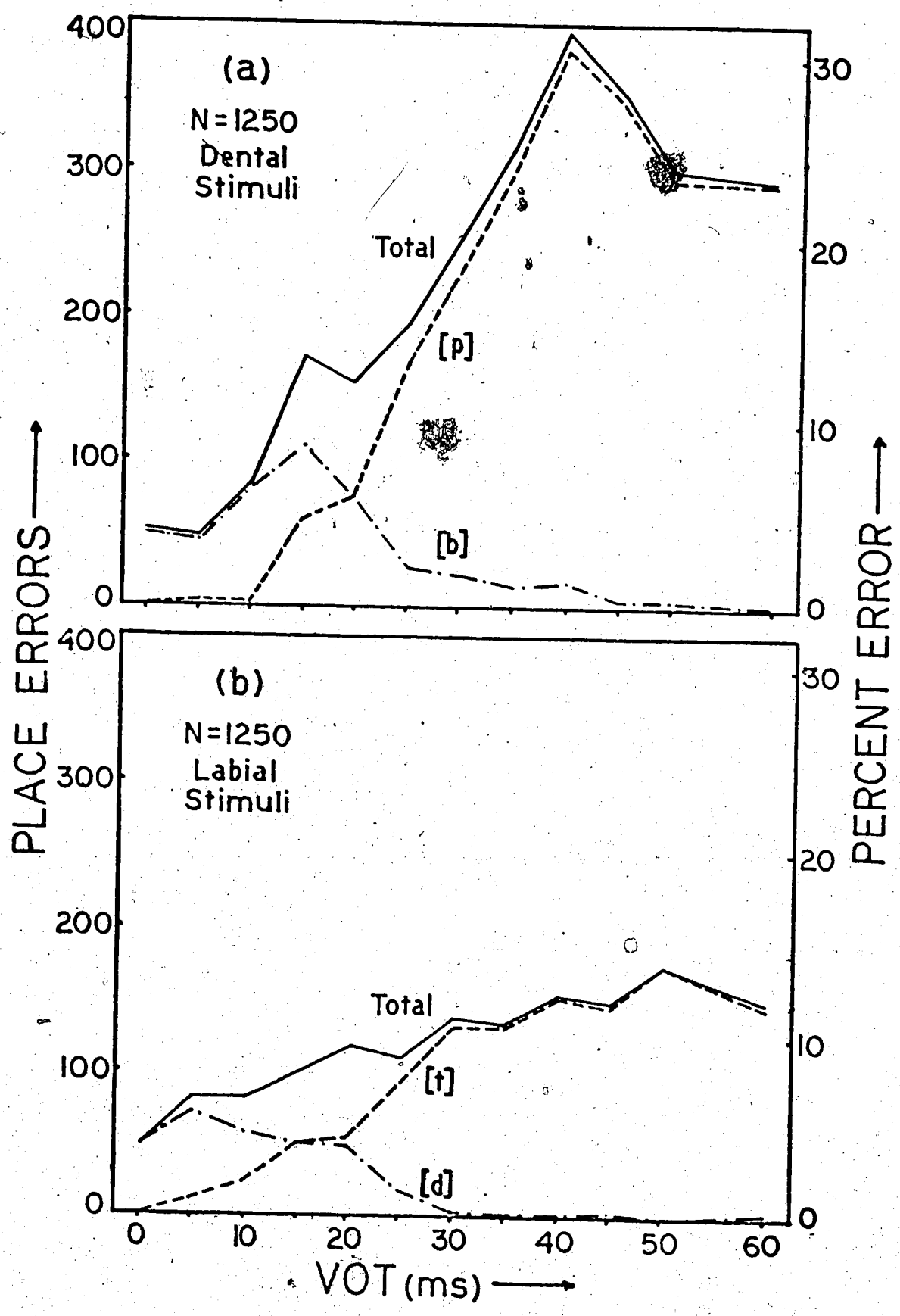


Figure 9: Total place errors for each stimulus series.

discussed by Miller and Nicely (1955, p. 348), and is generally supportive of this interpretation. They calculated that 'voicing information' can be transmitted at S/N levels 18 dB below those needed for the transmission of 'place information'. The problem remains, however, of accounting for the error patterns as they relate to VOT.

Interplay Between S/N Ratio and VOT as a Possible Cause of Place Errors

Numerous studies have reported experimental evidence to suggest that for both natural and synthetic speech stimuli, frequency characteristics of the release burst and adjacent vowel formant (F2) transitions are essential in the identification of stop consonants. The reader is directed to Miller and Nicely (1955), Delattre, Liberman, and Cooper (1955), Halle, Hughes, and Radley (1956), Liberman et al. (1957, 1958), and Kozhevnikov and Chistovich (1965) for further discussion.

For the present stimuli, these formant transitions can be precisely measured. In this case, all of the formant specifications of a given stimulus series were held constant (see Figure 3). Systematic variations in VOT were achieved by simply shifting the onset of the glottal waveform along the time scale, with respect to a fixed point (t=0) corresponding with consonant release. In other words, the temporal was identical for all stimuli.

As noted in Chapter II, the overall time alloted for formant transition was fixed at 40 msec. Labial and dental stimuli differed only in the direction and initial frequencies of the second and third formants. The 30 msec release noise burst was present in all stimuli as a source of initial short-duration formant excitation. Thus, for VOT's in the range 0 msec to 35 msec, all or some of the adjacent vowel formant transition was modulated by the glottal waveform, in conjunction with the initial 30 msec noise burst. For VOT's of 40 msec or more, the only source of formant excitation during the transition period was the 30 msec noise burst, since the onset of the glottal waveform did not occur until after the formant transitions had been completed, and the formants had reached appropriate steady-state values for the following vowel [a]. From this assessment, one would expect that using a white noise agent, it would be easier to mask the formant transitions in stimuli which exhibit long VOT's than in stimuli which exhibit short VOT's. Strong support for this claim resides in the observation that, for both stimulus series, VOT is highly correlated with place errors. For the labial series, $r = 0.9528$ ($p < 0.01$), and for the dental series, $r = 0.970$ ($p<0.01$).

On the other hand, this evidence could as easily be cited in support of the claim that another feature, namely, the duration of periodically excited formant transitions (dt), might also be offered to account for the same error

phenomena. From the perfect inverse relationship between VOT and dt in the range of VOT from 0 msec to 40 msec, it follows that dt is negatively correlated with errors, to exactly the same level of significance as is VOT. However, if the range is extended to include all VOT's from 0 msec to 60 msec, the correlations are no longer identical. In correlati g errors with VOT in the labial series, r = 0.9119, and in the case of errors and dt, (r = -0.9572). Similarly, for the error by VOT correlation in the dental series, r = 0.8351, while r = -0.9635 in the correlation between errors and dt. The purpose in pointing out these correlations is to suggest that over an extended range of VOT, place errors might be more closely associated with the duration of periodically-excited vowel formants.

This relationship might be clearer in Table 7, in which a more precise account of a number of correspondent changes in the structure of the F2 transitions for the labial and dental stimulus series are presented. From Table 7, it is noted that for VOT's of 40 msec or more, all stimuli are essentially the same, except for information conveyed by the 30 msec release noise burst. This might account for what appears in Figure 8 (S/N = -20 dB) to be the complete loss of ability to distinguish between labial and dental stimuli at over 40 msec VOT.

Presumably, the same parameters could be specified for stimuli used in earlier VOT studies, if a more precise

TABLE 7

CHANGES IN F2 AS A FUNCTION OF VOT
(Errors included for reference)

LABIAL SERIES				
VOT	f_i	df(Hz)	dt(ms)	Errors
0	1000	280	40	43
5	1040	240	35	82
10	1080	200	30	84
15	1120	160	25	102
20	1160	120	20	118
25	1190	90	15	110
30	1220	60	10	140
35	1250	30	5	131
40	1280	0	0	154
45	1280	0	0	151
50	1280	0	0	173
60	1280	0	0	146
DENTAL SERIES				
VOT	f_i	df(Hz)	dt(ms)	Errors
0	1800	520	40	51
5	1735	455	35	47
10	1670	390	30	81
15	1605	325	25	171
20	1540	260	20	153
25	1475	195	15	193
30	1410	130	10	255
35	1345	65	5	311
40	1280	0	0	397
45	1280	0	0	353
50	1280	0	0	299
60	1280	0	0	293

f_i = frequency of F2 at onset of glottal waveform.

df = interval (Hz) between initial and target frequency for the vowel [a] (1280 Hz).

dt = milliseconds of formant transition modulated by glottal waveform.

description of those stimuli were available. The possible perceptual importance of these stimulus features has been given very little attention by previous authors concerned with VOT. Those authors carried out their experiments under the assumption that VOT was the only property of the stimulus which was systematically manipulated, while in reality, the manipulation of VOT can be shown to result in a number of related changes in other stimulus features. On the basis of some of these features, it may be possible to re-evaluate the stimuli.

In general, periodic signals have been shown to be less vulnerable to the effects of white noise masking than aperiodic signals (Hirsh, 1952). With respect to 'place-information' conveyed by the release burst and adjacent formant transitions, stimuli exhibiting short VOT's can be said to represent the former type of signal, while stimuli with long VOT's resemble the latter. It is therefore likely that in order to cause a significant amount of place confusion, stimuli exhibiting a distinctive amount of periodic excitation of formant transitions (short VOT's) should require a higher level of random masking noise than stimuli which exhibit little or no periodic excitation of formant (long VOT's). It is important to note that this hypothesis contradicts the claims made by Fletcher (1929) to the effect that voiceless (tense) consonants are inherently more audible than voiced (lax) consonants.

The differences in error patterns might also be approached in terms of differential sensitivity to masking effects. Fischer-Jorgensen (1954) pointed out certain acoustic differences between [p] and [t] in natural speech. He described [t] as having higher inherent resonances than [p], and before a central vowel, [t] was indicated by a steep falling F2 transition. Because of the relative weakness of the F2 transitions associated with longer VOT's, one might expect the transition to be quite easily masked, leaving only the stronger, low-frequency energy associated with [p]. In such a case, to expect a [p] identification would not be unreasonable, since, as Fischer-Jorgenson stated, " . . . if there is no positive reason for hearing [k] or [t], there will be a majority for [p] (p.55)."

Within the present framework, the relationships which hold between confusions and VOT may be reconsidered, and possibly explained, in terms of the differential sensitivity of the present stimuli to white noise masking. This approach toward more complete stimulus description may prove beneficial in the explanation of observed VOT discriminability as well as confusions of place of articulation.

Formant Transitions as a Plausible Cue for Stimulus Discrimination

Liberman, Harris, Hoffman, and Griffith (1957) asserted that the listener can only discriminate between speech stimuli which he can identify as belonging to

different phonemic categories. Based on identification functions, the phonetic boundary between two categories was established by the point of subjective equality (PSE), that point on the stimulus continuum at which a stimulus was identified as belonging to either category with equal probability.

Lisker and Abramson (1970), Lazarus and Pisoni (1972), and Eimas and Corbit (1973) generally agreed that with respect to the VOT continuum, the phonetic boundary between the English voiced and voiceless categories was found in the region of 30 msec VOT.

Functions of discriminability along the VOT dimension have been reported by Abramson and Lisker (1970), Lazarus and Pisoni (1972), and Eimas and Corbit (1973). Their results were obtained by ABX triadic comparison procedures, in which the third stimulus in a series was to be judged as identical to either the first or the second. The authors reported a peak in discriminability at about 30 msec VOT. Performance at chance level was observed at about 15 msec in either direction from the peak. This consistency is not surprising in light of the fact that all three studies employed the same (Lisker & Abramson, 1970) synthetic stimuli.

Optimum VOT discriminability in the region of the phonetic boundary was offered by the above authors as evidence to support their claim that the listener

discriminates between stimuli on the perceptual basis of phoneme class membership. The only discussion of formant transitions associated with the stimuli used appears in Lazarus and Pisoni (1972, p. 4). Although very little information was presented about specific frequency parameters, it was stated that the duration of formant transitions was fixed at approximately 50 msec. An essentially complementary relationship has been established between VOT and the duration of periodic excitation of vowel formant transitions. Thus, earlier reports of optimum discriminability at just below 30 msec VOT may be restated as optimum discriminability for periodically-excited formant transitions of just over 20 msec in duration.

Under this interpretation, it is possible that peaks in observed discriminability along the VOT dimension may be due to limenal acoustic properties of associated formant transitions, rather than to abstract linguistic parameters. Such a proposal is by no means unwarranted. On the contrary, it would seem premature to assume the necessity of highly specialized mechanisms for the acoustic analysis of a particular type of speech signal without having demonstrated first, that the signal cannot be regarded as a complex of more fundamental acoustic variables, and second, that general auditory mechanisms are incapable of the same basic analysis.

Nabelek and Hirsh (1969) examined the ability of listeners to discriminate among various frequency transitions.

They computed relative difference limens (DL's) for frequency transitions varying in duration, frequency shift, and general frequency region. Three frequency regions were examined: 250 Hz, which the authors felt was related to pitch and intonation, and 1 kHz and 4 kHz, which they regarded important regions in the perception of formant transitions. In general terms, they reasoned that the listener relies heavily on "glide rate" (frequency shift per unit time) in the discrimination. In addition, their results indicated, with the exception of small frequency shifts in the pitch region, greatest discriminability (smallest DL's) obtained for transition durations of from 20 to 30 msec. Thus, since these measures agree well in magnitude with the range of durations of frequency transitions in normal speech, especially in the F2 - F3 regions, Nabelek and Hirsh concluded that the observation of peaks in discriminability reflects a " . . . general property of hearing and that it does not appear only in connection with speech sounds (p. 1518)."

Further general agreement with the re-evaluated results of VOT studies suggest that certain important acoustic aspects of VOT variants might also be analyzed by auditory mechanisms which are not specific to speech analysis. With respect to the "linguistic feature detector" model proposed by Eimas and Corbit (1973), it may be concluded that no such complex speech-specific mechanism (i.e., a phonetic-level processor) need be postulated, and

that various more general auditory models can be offered to account for the same phenomena.

The Need for Further Research

One point which is obvious is that a great deal of further research is necessary in areas related to VOT. Experiments should be designed for the examination of the nature of inter-subject differences. Although it is believed that the present data were elicited under a broader range of experimental conditions than data elicited in earlier studies, the experimental design itself leaves many important research questions out of reach. For example, in a number of different ways, the present results suggest that place of articulation is a very important factor in the consideration of VOT. Surmounting past methodological limitations, VOT studies should be expanded to include the perception of initial velar stop consonants and possibly the affricate series. Furthermore, the present shift of attention to such stimulus features as formant transitions necessitates the inspection of VOT and consonant perception in relation to a broader range of following vowels.

The apparent co-operation of VOT with other acoustic properties of the stimulus has given rise to several possible re-interpretations of experimental evidence which has been provided by earlier VOT studies. In this respect, agreement of certain aspects of the present results with

those of previous studies, dealing with both speech and non-speech stimuli, provides strong motivation for future postulations of more general auditory models.

CHAPTER V

SUMMARY

Voice onset time (VOT), defined as the time interval between consonant release and the onset of laryngeal pulsing, originated as an analytical measure in the determination of voicing in initial stop consonants. More recently, VOT has received increased attention as a possible 'cue' for the listener in the perceptual manner categorization of initial stop consonants. Based on the assumed perceptual universality of VOT, a "linguistic feature detector" model has been offered to account for the differentiation between English voiced and voiceless initial stops.

The viability of a language-specialized detector system whose sole input is VOT must be contingent upon empirical evidence of VOT's perceptual universality. A three-factory experiment with repeated measures was designed to observe the effects of place of articulation, manner of articulation, and white noise masking on the identification of stimuli in which VOT was systematically varied.

The voiced-voiceless distinction was apparently maintained throughout the observed range of masking conditions, suggesting that VOT is, in fact, an important cue for the distinction. However, under high masking conditions, median VOT values for dental response distributions changed significantly, while those for labial distributions did not. This result was interpreted as an indication that in addition to VOT, a certain amount of frequency domain information such

as aspiration and formant transitions may play an important role in the perceptual distinction between the voiced and voiceless categories.

High masking levels were also associated with the confusion of place of articulation. An attempt to determine the cause of such confusions gave rise to a re-evaluation of the present synthetic speech stimuli in terms of their primary acoustic (as opposed to linguistic) composition. A number of other stimulus characteristics, including the F2 transition, were found to vary concurrently with VOT.

When considering the results of stimulus re-evaluation in terms of earlier studies concerning both speech and non-speech auditory analysis, two important generalizations emerged. First, the concept of VOT as a universal perceptual cue may be inadequate, since associated theories place little perceptual significance on the dynamic acoustic components of the stimulus with which VOT is closely related. In turn, VOT might be regarded as a composite feature, subsuming a number of dynamic, perceptually relevant features, of which the interval between consonant release and the onset of laryngeal pulsing may be only one. Second, the fact that certain stimulus features which vary concurrently with VOT can be quantified at the acoustic, as opposed to the linguistic level, prompted the suggestion that VOT might be regarded as a primary acoustic feature rather than a primary linguistic feature of the stimulus.

Previous experimental results suggest that the auditory analysis and discrimination of frequency transitions may be a general auditory process, applicable to both speech and non-speech stimuli. The fact that formant transitions play an important role in initial consonant perception, and are also listed among stimulus features which vary concurrently with VOT, suggests that at least some of the properties of VOT stimuli can be analyzed or at least differentiated, on pre-phonetic grounds.

Such hypotheses present alternatives to the "linguistic feature detector" model and for that matter, any other auditory models requiring highly specialized mechanisms for the analysis of complex speech stimuli which could possibly be reduced to more fundamental acoustic components.

REFERENCES

Abramson, A.S., & Lisker, L. Discriminability along the voicing continuum: cross-cultural tests. In Proceedings of the Sixth International Congress of Phonetic Sciences, Prague, 1967. Prague: Academia, 1970, 569-573.

Akitt, B.J. Computer control system for a speech synthesizer (PAT) using the PDP 8/1. Unpublished manuscript, University of Calgary, 1970.

Anthony, J., & Lawrence, W. A resonance analogue speech synthesizer. In A.K. Neilsen (Ed.), Congress Report I, Fourth International Congress On Acoustics, Copenhagen, 1962.

Caramazza, A., Yeni-Komshian, G.H., Zurif, E.B., & Carbone,E. The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals JASA, 1973, 54, 2, 421-428.

Cooper, F.S. Speech synthesizers. Proceedings of the Fourth International Congress of Phonetic Sciences, Helsinki. 1961.

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M., & Gertsman, L.J. Some experiments on the perception of synthetic speech sounds. JASA, 1952, 24, 6, 597-606.

Delattre, P.C., Liberman, A.M., & Cooper, F.S. Acoustic loci and transitional cues for consonants. JASA, 1955, 27, 4, 769-773.

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

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

Fant, C.G.M. Auditory patterns of speech. In W. Wathen-Dunn (Ed.), Models for the Perception of Speech and Visual Form. Cambridge: M.I.T. Press, 1967.

Ferguson, G.A. Statistical Analysis in Psychology & Education. New York: McGraw-Hill, 1959.

Fischer-Jorgensen, E. Acoustic analysis of stop consonants. Miscellanea Phonetica, 1954, 2, 42-59.

Fletcher, H. Speech and Hearing. Princeton, New Jersey: D. Van Nostrand, 1929.

Green, S.M., & Swets, J.A. Signal Detection Theory and Psychophysics. New York: John Wiley and Sons, 1966.

Halle, M., Hughes, G.W., & Radley, J.P.A. Acoustic properties of stop consonants. JASA, 1957, 29, 1 107-116.

Hill, D.R. Some practical steps taken towards a man-machine interface using speech. Paper presented at the 2nd Hawaii International Conference on System Sciences, University of Hawaii, Honolulu, Hawaii, Jan. 22-24, 1969.

Hirsh, I.J. The Measurement of Hearing. New York: McGraw-Hill, 1952.

Jakobson, R., Fant, C.G.M., & Halle, M. Preliminaries to Speech Analysis. Cambridge: M.I.T. Press, 1963.

Kim, Chin-Wu A theory of aspiration. Phonetica, 1970, 21, 107-116.

Kozhevnikov, V.A., & Chistovich, L.A. Speech: Articulation and Perception. Washington, D.C.: Joint Publications Research Service, (10 June), 1965.

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

Lazarus, H.H., & Pisoni, D.B. Categorical and non-categorical modes of speech perception along the voice onset time continuum. Paper presented at 84th meeting of the Acoustical Society of America, Miami Beach, December 1972.

Liberman, A.M., Delattre, P.C., Gertsman, L.J., & Cooper, F.S. Tempo of frequency change as a cue for distinguishing classes of speech sounds. Journal of Experimental Psychology, 1956, 52, 2, 127-137.

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,

Liberman, A.M., Delattre, P.C., & Cooper, F.S. Some cues for the distinction between voiced and voiceless stops in the initial position. Language and Speech, 1958, 1, 159-166.

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

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

Lisker, L., & Abramson, A.S. Some effects of context on voice onset time in English stops. Language and Speech, 1967, 10, 1-28.

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, 563-567.

Malecot, A. The lenis-fortis opposition: Its physiological parameters. JASA, 1970, 47, 6, 1588-1592.

Miller, G.A., & Nicely, P.D. An analysis of perceptual confusions among some English consonants. JASA, 1955, 27, 2, 338-352.

Nabelek, I., & Hirsh, I.J. On the discriminability of formant transitions. JASA, 1969, 45, 6 1510-1519.

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

Wathen-Dunn, W. (Ed.) Models for the Perception of Speech and Visual Form. Cambridge: M.I.T. Press, 1967.

APPENDIX A

INSTRUCTIONS TO SUBJECTS

Hi. Thanks for helping me out by being a subject in this experiment. What I'm trying to do is to find out how English speakers distinguish between certain types of initial consonants. The tapes you are about to hear are made up of syllables in the form of a consonant followed by the vowel a. All I'd like you to do is to identify the consonant that you hear. I can tell you now that the only consonants you will hear will be p's, b's, t's, and d's, followed by a.

In the experiment, pa's, ba's, ta's, and da's will be mixed together, and presented in an irregular order for you to identify. Each individual presentation will be follwed by two seconds of silence, in which time you must identify the consonant you heard by simply filling in the appropriate location on the answer sheet. Obviously, some guessing will be necessary in cases where you aren't sure of the answer, but you have some idea. In these cases, I want you to guess. You must give a response for every item presented. In case you honestly miss an answer, simply leave it blank and go on to the next. We can correct that later.

Usually, answer sheets like these, which are analyzed by a computer, require that mistakes be completely erased.

However, in case you should accidentally fill in the wrong answer slot, time won't allow you to erase and fill another. So, just lightly cross out the mistake, and fill in what you think is the correct answer. Afterwards, I'll go back and erase the crossed-out answer for you. If you have any questions at this point, please raise your hand.

If you haven't already done so, please fill in the information for which blank spaces are provided at the top of the answer sheet. So that I can keep track of your answers, you will be required to repeat this information at the top of each answer sheet. Pay no attention to the space marked S.#; it refers to your 'Subject Number', and will be assigned later, during the analysis. Below the heading 'ID Number', please write your Student Identification Number, if you have one. The ten columns to the right of the boxes represent the numbers zero through nine, respectfully. Fill in the slot representing the number you have written in each corresponding row. Please be careful not to extend your pencil marks beyond the dashed guidelines; it presents problems in the computer analysis. Are there any questions?

What you are about to hear, then, are a few examples of the stimuli drawn from the p-b group:

(Four examples presented here).

Now listen to a few stimuli chosen from the t-d group:

(Four examples presented here).

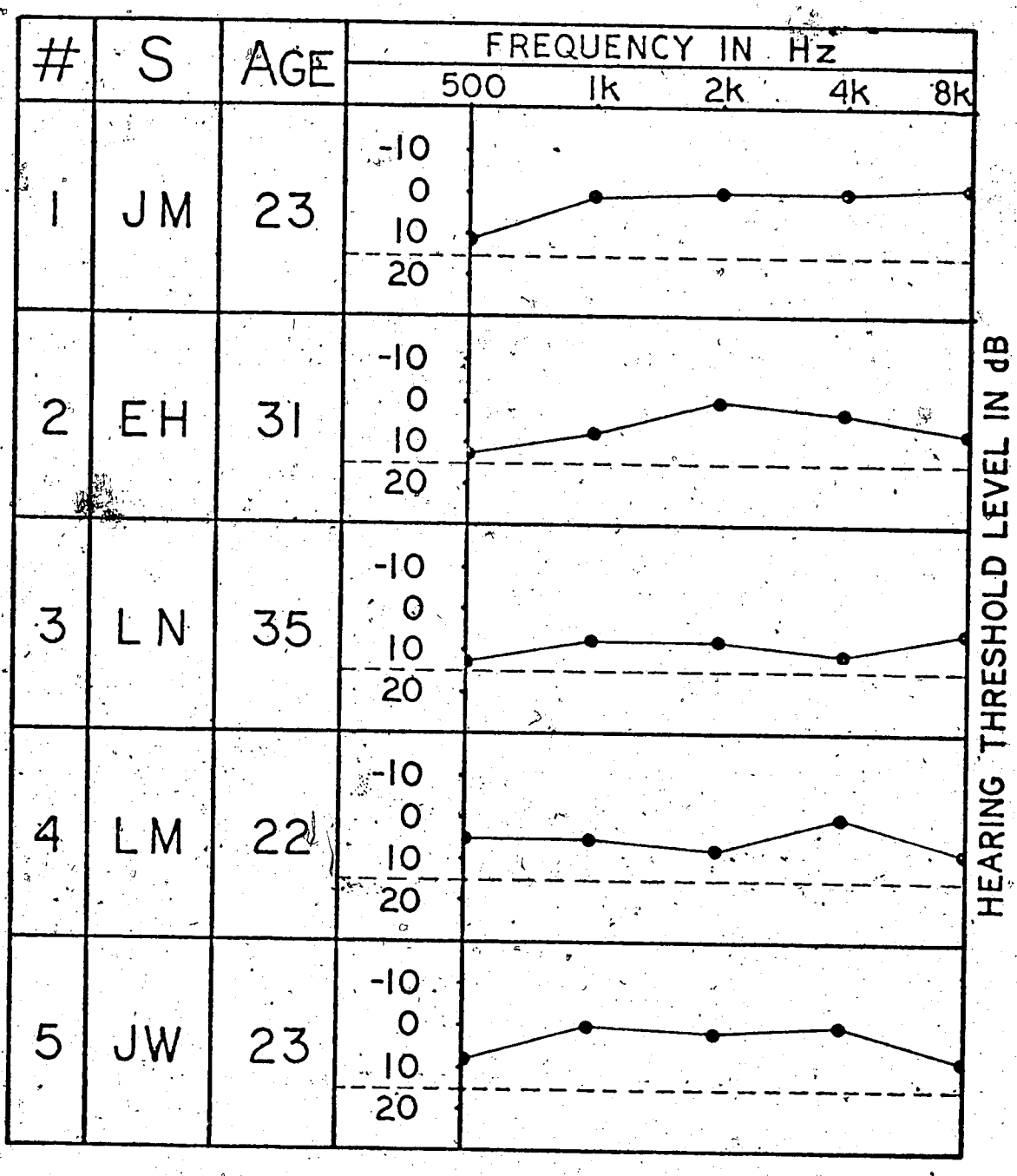
Now, see if you can distinguish between pa's and ba's, ta's and da's, when they've been mixed together:

(Minimum and maximum VOT representatives presented alternately)

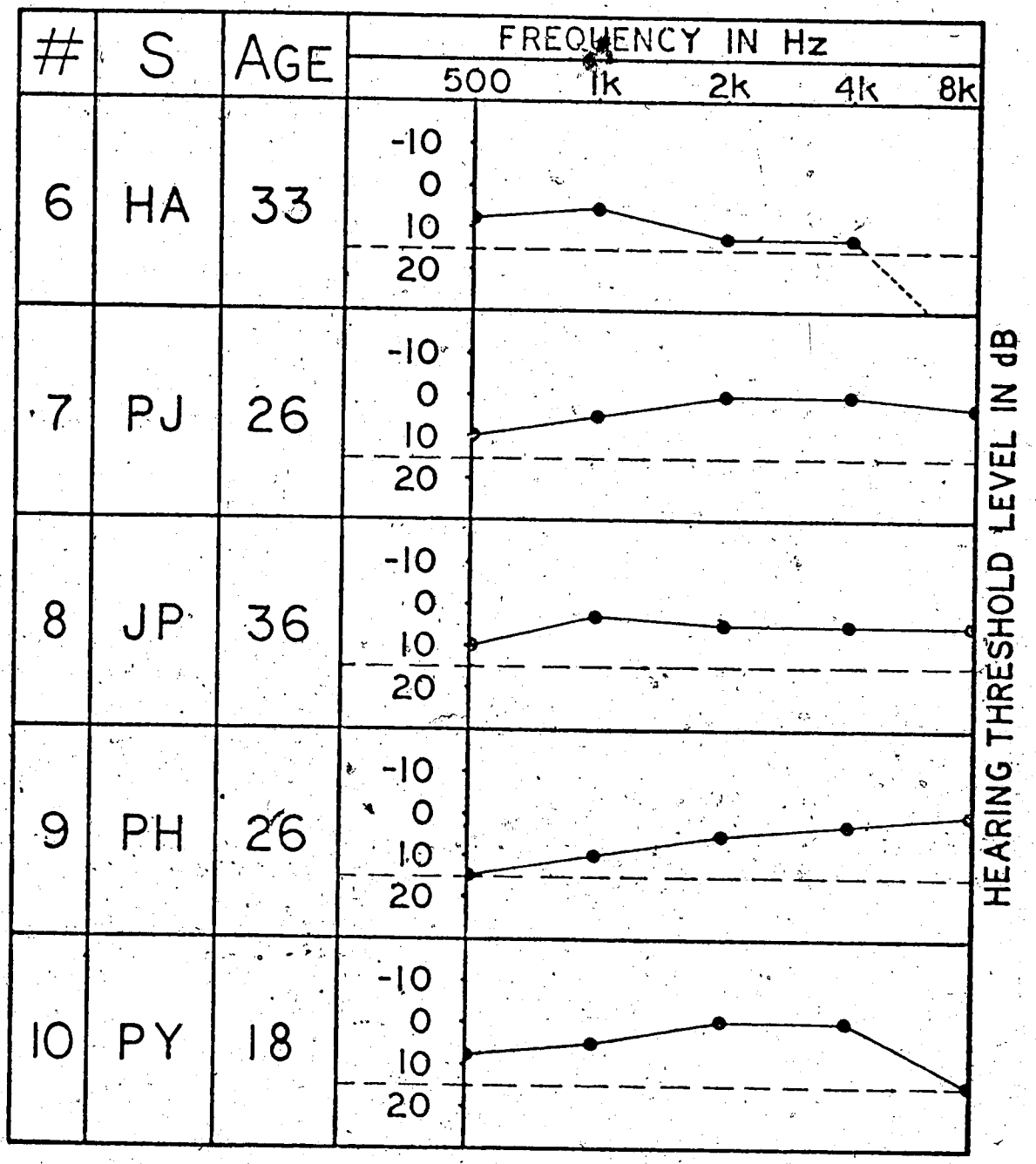
Since you will be making quite a few identifications, I've divided the presentations into small sets. After each column has been completed, there will be a short rest period, after we have completed the third page, we'll take a five or ten minute break. Are there any questions before we begin?

APPENDIX B

RESULTS OF PURE TONE AUDIOMETRIC SWEEPS TESTS



APPENDIX B (continued)



APPENDIX C

DATA POINTS FOR ANALYSIS OF VARIANCE MEDIAN VOT'S FOR THE LABIAL SERIES

S	Voiced				
	S/N Ratio				
	+20 dB	+10 dB	0 dB	-10 dB	-20 dB
1	6.73	8.65	9.67	9.05	9.75
2	12.40	12.39	12.90	11.67	9.43
3	11.41	10.96	10.24	11.45	13.75
4	10.42	11.00	11.10	10.69	11.13
5	8.07	8.23	8.65	8.48	7.84
6	10.13	10.00	10.55	9.44	7.39
7	13.10	15.42	18.88	15.00	12.05
8	13.00	12.90	13.50	12.50	11.85
9	9.28	9.14	9.38	9.60	11.00
10	11.35	11.40	11.85	11.88	11.03
Voiceless					
1	32.90	31.00	30.90	30.10	31.14
2	34.75	34.70	35.62	33.00	30.50
3	33.13	32.50	31.74	30.38	26.25
4	32.81	33.40	33.70	31.19	30.54
5	29.20	29.20	30.83	31.15	32.07
6	32.40	32.19	32.39	32.05	30.23
7	35.83	36.85	38.17	39.09	35.25
8	35.50	35.21	35.85	35.50	36.43
9	31.25	31.20	31.56	30.75	33.24
10	33.80	33.90	31.90	33.95	30.13

APPENDIX C (continued)

MEDIAN VOT'S FOR THE DENTAL SERIES

S	Voiced				
	S/N Ratio				
	+20 dB	+10 dB	0 dB	-10 dB	-20 dB
1	6.75	8.55	8.20	8.61	7.00
2	14.18	11.60	13.00	11.80	9.29
3	12.50	11.80	11.90	11.00	9.32
4	12.70	13.50	14.80	12.82	11.18
5	13.40	14.20	14.40	13.75	10.66
6	11.70	11.80	11.67	11.00	7.33
7	11.90	14.60	16.98	14.90	8.01
8	13.10	14.10	17.20	16.64	10.34
9	11.98	12.20	13.60	12.50	12.74
10	12.30	12.60	14.60	14.00	10.98
	Voiceless				
1	26.40	29.90	32.17	21.79	25.83
2	33.60	34.10	35.73	31.97	28.93
3	34.50	33.96	33.37	30.13	31.83
4	35.20	36.00	37.50	36.07	35.56
5	35.83	36.80	36.94	33.00	31.07
6	34.74	33.18	33.53	29.79	30.00
7	34.50	36.80	39.53	36.50	30.00
8	35.42	36.14	40.00	35.71	34.84
9	34.40	32.70	35.95	30.71	28.33
10	34.40	34.70	35.59	36.82	35.31