ACOUSTIC CHARACTERISTICS OF STOP CONSONANTS: A CONTROLLED STUDY

by

VICTOR WAITO ZUE

BSEE, University of Florida
(1968)

MSE, University of Florida
(1969)

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENT FOR THE DEGREE OF

DOCTOR OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May, 1976

Signature of Author Victor W. Zue
Department of Electrical Engineering and Computer Science,
May 14, 1976

Certified by Kenneth N. Stevens
Thesis Supervisor

Accepted by _____
Chairman, Departmental Committee on Graduate Students

ACOUSTIC CHARACTERISTICS OF STOP CONSONANTS:
A CONTROLLED STUDY

by
VICTOR WAITO ZUE

Submitted to the Department of Electrical Engineering and Computer Science on May 14, 1976 in partial fulfillment of the requirements for the degree of Doctor of Science.

ABSTRACT

The research reported in this thesis has two distinct and integral parts. The first part is directed towards the development of a highly interactive computer facility where controlled studies of the acoustic characteristics of selected consonants, consonant clusters, and vowels in a prescribed phonetic environment can be carried out. In conjuction with the development of the data-base facility, a large corpus of acoustic data has been colleced. The format of the data is a nonsense hə´CVC utterance embedded in a carrier sentence "Say _____ again", where the consonants and vowels are systematically varied. 15 vowels and diphthongs were used to formed the syllable nuclei and 51 word-initial consonants and consonant clusters were included.

The second half of the thesis utilizes the collected data and the developed facility to study the acoustic characteristics of English stops, both in singleton and in clusters. The data included 1,728 utterances spoken by 3 male speakers. Various aspects of the temporal and spectral characteristics of these stops were quantified and discussed in detail. The findings in general suggest the presence of context independent acoustic properties for these stops. The exact nature of the acoustic invariance, however, still remains a topic of further investigation.

Thesis Supervisor: Kenneth N. Stevens

Title: Professor of Electrical Engineering

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Professor Kenneth Stevens for his time and effort in supervizing this thesis. During my past six years at MIT I have, on numerous occasions, benefited tremendously from his constant guidance, his penetrating questions, his insightful remarks, and his infinite patience. My association with him represents the high point of my educational carrier. Besides thanking the readers for reading and offering valuable comments on the manuscript, I would also like to thank each one of them individually: to Dennis Klatt for his constant tutorage in acoustic phonetics; to Alan Oppenheim for teaching me what digital signal processing is all about; to Ben Gold for his friendship, and his generosity in resources and manpower.

Most of the research reported in this thesis was conducted at the MIT Lincoln Laboratory. To all of those at Lincoln Laboratory who had helped me along the way, I acknowledge their assistance with appreciation.

Last but not least, I would like to thank my parents for their undying patience. To Stephanie, I express my deepest gratitude for her constant encouragement, for our many discussions on speech, and, above all, for happiness.

TABLE OF CONTENTS

FIGURES AND TABLES

CHAPTER 1

INTRODUCTION

In the process of human communication by spoken language, the speech signal, i.e., the acoustic waveform, plays a very unique role. On one hand it is the final result of the complex encoding that transpires at various stages of the speech production process. At the receiving end, however, the speech signal is the principal information carrier upon which the perceptual decoding process must operate.

Because of the relative ease of access and manipulation of the speech signal, it has been the focus of many past research efforts that seek to better understand the nature of language. Although much has been learned about the acoustic events of speech, our understanding of the relationship between the acoustic characteristics of speech sounds and their underlying linguistic units still remains, for the most part, vague.

The purpose of this thesis is to probe further into this relationship for a subset of the English speech sounds, namely the stop consonants. The study is carried out under a controlled linguistic environment, using a data-base

facility designed for acoustic phonetic research.

Before reviewing past research on the subject and pointing out the problem areas, it is appropriate to first provide a brief account of the physiology and the acoustics of speech production, as well as to summarize the linguistic framework on which this research is based.

1.1 The Physiology and the Acoustics of Speech Production

Speech is generated by closely coordinated movements of several groups of human anatomical structures. One such group of structures consists of those that enclose the air passage below the larynx. Through control of the muscles and through forces generated by the elastic recoil of the lungs, pressure can be built up below the larynx. This pressure eventually provides energy for the speech signal.

Immediately above the trachea is the larynx, which constitutes the second group of structures essential to the production of speech. The vocal cords in the larynx can be positioned in many ways so that air can flow through the glottis either with or without setting the vocal cords into vibration. When the vocal cords are set into vibration, the airflow through the glottis is interrupted quasi-periodically, thus creating the effect of modulation.

The third set of structures consists of the tongue, jaw, lips, velum, and other components that form the vocal and nasal cavities. By changing the configuration of the vocal tract, one can shape the detailed characteristics of the speech sounds being produced.

It is convenient to describe the acoustics of speech production in terms of three distinct stages. First, through interaction between airflow from the lungs and the laryngeal and supraglottal structures, a source of acoustic energy is created. This acoustic source may be one of several types, and may have several possible positions. The source acts as the excitation for the cavities above and below it. The filtering that is imposed on the source by these cavities is the second stage in the generation of speech sounds. Finally, sound is radiated from the lips and/or the nostrils.

1.2 Linguistic Framework

Studies of the way language is organized have produced overwhelming evidence that underlying the production and perception of speech there exists a sequence of basic discrete segments that are concatenated in time. These segments, called "phonemes", are assumed to have unique articulatory and acoustic characteristics. It has been

proposed by Jakobson, Fant and Halle [Jakobson, Fant, and Halle 1952] that the phonemes can be characterized by a set of invariant attributes called distinctive features. The distinctive features bear a direct relationship to the articulatory gesture from which the speech sound is produced, and they have certain well-defined acoustic correlates. Therefore, at the phonemic level, the linguistic structure of an utterance can be represented by a two dimensional matrix with columns representing the phonemes, rows listing the distinctive features, and the matrix entries indicating the presence or absence of a feature for a given phoneme.

It should be noted that at the phonemic level, the distinctive feature theory necessitates a discrete (or even binary) selection, whereas at the articulatory and acoustic levels the feature correlates appear to take on a continuum of values.

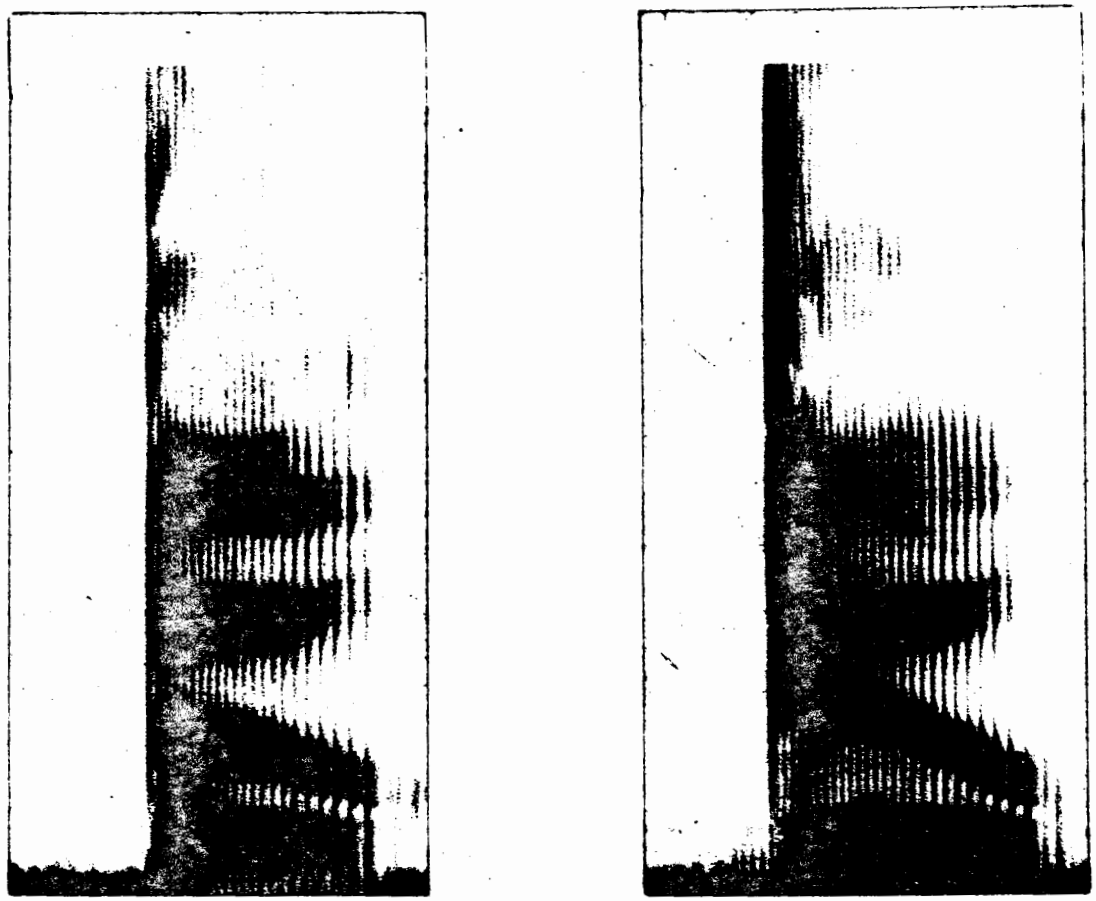
1.3 Problem Areas

During the production of speech, the linguistic contents of the feature matrix are transformed into actual neuro-muscular commands that set the articulators (lips, jaw, tongue, etc.) into motion. Although the commands may be discrete, or step-wise, the actual motions of the

articulators and the resulting acoustic signals are continuous, due to the interaction among various structures and their different degrees of sluggishness. The result is an overlap of phonemic information from one segment to another. In other words, although the features have certain well defined acoustic correlates, there is hardly a one-to-one correspondence between a given feature and its correlates. More precisely, the acoustic manifestation of a given feature appears to depend on the presence or absence of other features. Furthermore, when phonemes are concatenated to form an utterance, the acoustic correlates of the underlying features will undergo modification and distortion as a consequence of the phonetic environments. For example, Figure 1.1 shows spectrograms of the two words "do" and "boo". Although the vowels in the two words are phonemically identical, the acoustic characteristics of the vowels can be seen to be quite different.

Similarly, one can clearly observe the differences, both in temporal and in spectral characteristics, of the phoneme /t/ in three diffeent words "tea", "steep", and "tree", as shown in Figure 1.2.

These examples illustrate the important fact that in any study of the acoustic properties of speech sounds, the influence of the phonetic environment must be taken into



BOO DO

Figure 1.1 Spectrograms of the words "boo" and "do"
(formant frequencies of the vowel are modified by the consonant)

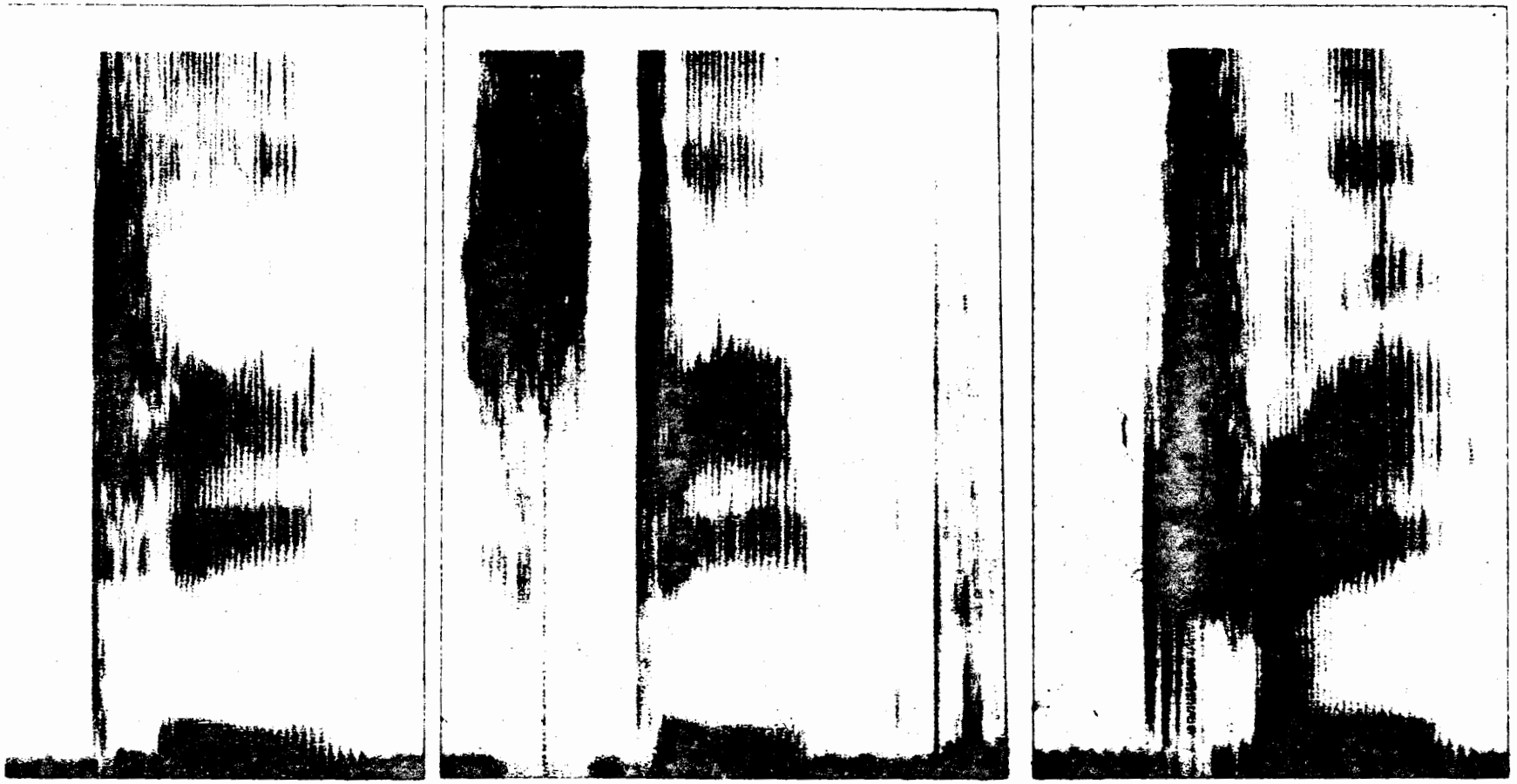


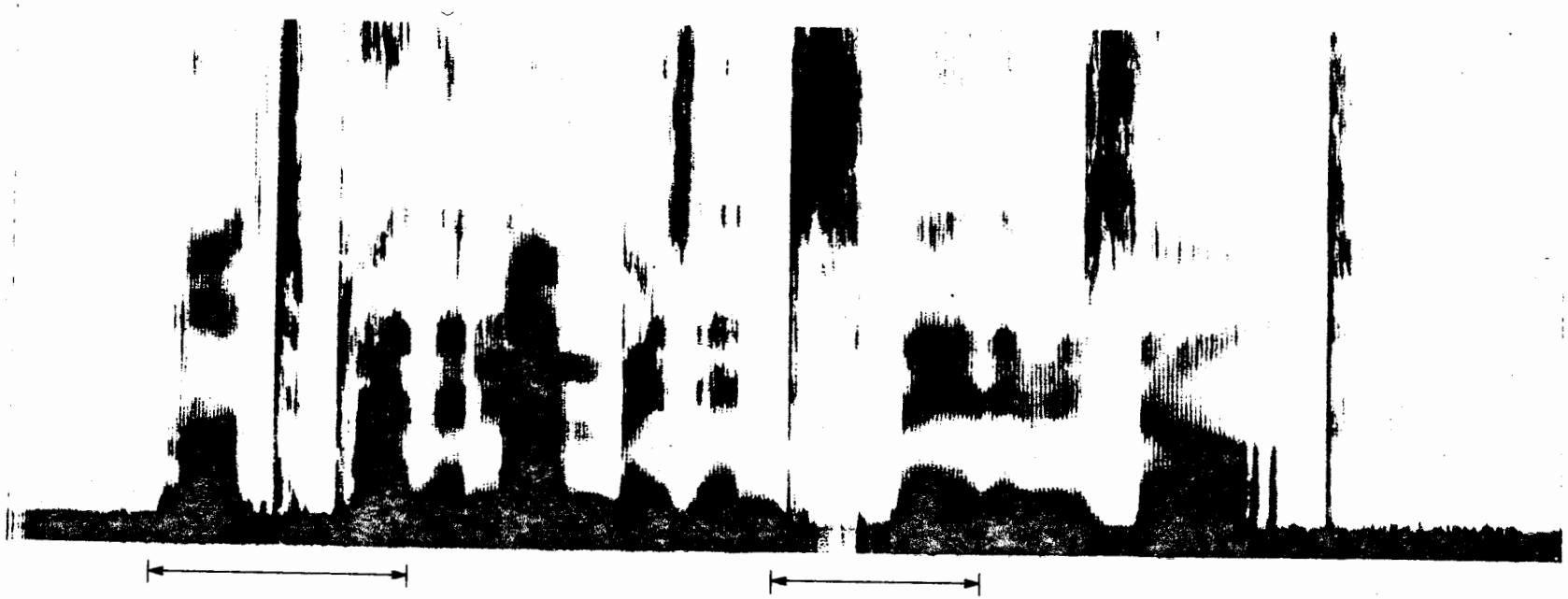
Figure 1.2 Spectrograms of the words "tea", "steep", and "tree" (spectral and temporal characteristics of the stop release are modified by the phonetic environment)

account.

Another important feature of speech communication is that sometimes a speaker can distort the acoustic properties of speech sounds so severely that even the environment will provide no acoustic cues to the identity of the phoneme. Figure 1.3 provides examples of such acoustic distortion. The schwa in the second syllable of the word "multiply" and the first syllable of the word "display" can be devoiced such that it exhibits no acoustic characteristics commonly associated with vowels. Such distortion is possible because a listener is capable of decoding an utterance not only from the acoustic signal, but also from his familiarity with the syntactic and semantic constraints, and with rules governing allowable phoneme seguences of his language.

The above examples serve to emphasize tha fact that although there might be invariant acoustic cues for a distinctive feature, the surface realization of such an acoustic cue is very much intermingled with other cues. Only through a very careful and controlled study can one satisfactorily answer the question of the acoustic invariance, if any, of phonetic features.

Another important factor contributing to our lack of success in relating acoustic characteristics to phonetic features is the variability from one speaker to another.



MULTIPLY THE NUMBERS AND DISPLAY THE RESULT

Figure 1.3 Spectrogram of the sentence "multiply the numbers and display the result" (example of schwa deletion as a consequence of phonological effect)

The acoustic properties of speech sounds depend on the physiological structure of the vocal apparatus, which varies from speaker to speaker. Furthermore, given a single speaker, the same utterance pronounced on two separate occasions could vary considerably. In any study of the acoustic properties of speech sounds, these inter- and intra-speaker variations will have to be accounted for. This requirement usually translates into multiple speaker and multiple session analysis, which in turn suggests a large corpus of data.

1.4 Literature Review

The acoustic characteristics of stops and the effect of coarticulation have been studied by many in the past, and a number of the studies date back some twenty years. Most of these studies are directed towards the search for perceptually important acoustic cues, and each study achieved a varying degree of success. Due to various technical difficulties involving the processing and the storage of a large amount of data, most of these studies have been rather limited in scope.

The pioneering work at Haskins Laboratory represents the early research for perceptually important acoustic cues of stops [Cooper et al. 1952, Delattre et al. 1954]. Using

the pattern playback machine that converts hand-painted spectrograms into sounds, the investigators were able to vary independently the frequency location of the stop burst and the amount and direction of the second formant transition into the following vowel. Their major finding has been that burst frequency as well as direction and degree of formant transitions are perceptually important cues to the identification of the stop consonants. These findings prompted the subsequent proposition of the existence of acoustic loci for these consonants. Delattre et al. speculated on the association of different formant loci with the place and manner of articulation of these consonants.

Concurrent with the Haskins studies, Fischer-Jorgensen [Fischer-Jorgensen 1954] reported a study of the acoustic properties of the six Danish stops. Similar but different results were found with regard to the formant loci of these consonants. The author noted the significant role played by the aspiration which serves to differentiate the Danish /p,t,k/ from /b,d,g/. It was also found that vowel environment tended to alter the acoustic properties of stops.

Spectral characteristics of English stop consonants were studied by Halle et al. [Halle et al. 1957].

Quantitative data were gathered and possible criteria for identification were proposed and tested. The study asserts that burst and transitions are the two major cues for the identity of these consonants, although the authors took issue with the "locus theory" proposed by the Haskins group. It was felt that a set of more complex rules seemed to operate on the formant transitions.

Lisker and Abramson [Lisker and Abramson 1964] reported a cross-language study of voicing in initial stops. The major finding of this study is that the features of voicing, aspiration and force of articulation could be plausible consequences of a single variable -- voice-onset time (VOT). VOT was found to be not only a basis for separating the voicing categories, but also sensitive to the place of stop closure.

Subsequent studies by Lisker and Abramson [Lisker and Abramson 1967] extended the results to other phonetic environments. The study by Klatt [Klatt 1975] was the first reported investigation of the variation of VOT in clusters.

In summary, a good deal of information with regard to the acoustic characteristics of stop consonants has been published in the literature. Although each of these studies had contributed individually to our understanding of the acoustic events of speech, the results are fragmental and

lack continuity. They all suffer, in one way or another, from some of the problems stated in the previous section.

1.5 Overview

In Chapter 2 we define the research goal of this thesis, and describe the data collection procedure. Chapter 3 is devoted to the description of the analysis system as well as the data-base facility. Chapter 4 presents results and discussion of the temporal characteristics of English stops, both in isolation and in clusters. The spectral characteristics of these stops are presented in Chapter 5.

CHAPTER 2

DEFINING THE RESEARCH GOAL AND THE DATA COLLECTION PROCESS

One long-range goal of the research in acoustic phonetics is to determine the relevant acoustic properties of all speech sounds and to relate these properties to the underlying features that characterize the sounds. As was pointed out in the previous chapter, the nature of speech as a communication medium makes this goal a formidable one. This thesis deals with a problem of much reduced magnitude in that the number of speech sounds studied is rather limited, and the phonetic environment as well as linguistic and phonological influences are carefully controlled.

This chapter first describes the scope of the study in detail. Considerations that went into the design of the data format are then presented, followed by a description of the acquisition of acoustic data.

2.1 Scope of the Study

The study reported here has two distinct and integral parts. The first part involves the collection of a large

corpus of acoustic data, as well as the development of a highly interactive computer facility where a substantial amount of acoustic data can be stored, examined and analyzed. The second part utilizes the facility to study a subset of the collected data, namely the English stops, both in singleton and in clusters, preceding stressed vowels.

It was decided at the outset that this study would be directed primarily towards prestressed consonants and consonant clusters. Data collection as well as the design of the data format consequently reflect such a constraint. We have restricted ourselves to the study of prestressed consonants (and clusters) for several reasons. First, a stressed consonant-vowel sequence is universal among all known languages. Studies of prestressed consonants can provide a common ground on the basis of which cross-language differences can be compared. Secondly, stressed syllables in an utterance are probably articulated with greater care and effort, thereby resulting in a robust acoustic signal where parameters can be extracted more reliably. Furthermore, based on some early studies [Stevens and Klatt 1968, Stevens 1969], it may be hypothesized that the intrinsic acoustic properties of consonants are least distorted by the environment when they appear in stressed C-V syllables. Acoustic invariants are more likely to be observed in such an environment. Therefore this phonetic

environment might provide, in some sense, the clearest indication of the ideal relationship between the underlying features and their acoustic correlates.

Although inter- and intra-speaker variabilities are not the major concerns of this study, the data-base was to include a number of talkers, each recorded on several occasions. The inclusion of several speakers and sessions presumably minimizes the probability of an individual speaker or recording session introducing a bias in the results.

In conjunction with the collection of acoustic data, an acoustic analysis system was developed to enable the preliminary processing of acoustic data. Digitization of the speech waveform, computation of the short-time spectra, and the hand-marking of utterances with phonetic labels can all be done conveniently. The facility was also developed such that utterances were stored in a data-base where retrieval and analysis could be done with ease. Various aspects of the analysis system and data-base will be discussed in great detail in Chapter 3.

Using the facility developed, the acoustic characteristics of prestressed English plosives, both in singleton and in clusters, have been studied in detail. Durational and spectral characteristics of the plosives were

examined as a function of the phonetic environments, and possible interpretations of the data in connection with production and perception were suggested. These results are presented in Chapters 4 and 5.

2.2 Data Format

Several considerations have been weighed in the design of the data format. We have thus far limited ourselves to stressed consonant-vowel syllables where the consonant (or consonant cluster) and vowel are varied in some systematic way. Experience has shown that stressed C-V syllables in isolation can be articulated in a rather unnatural manner. It is therefore advantageous to frame the C-V syllable in a carrier sentences so as to simulate a more natural, continuous-speech-like environment. Since certain English vowels do not appear in syllable final positions, a final consonant was added to the C-V syllable. Finally, we would like to eliminate as much as possible the linguistic and phonological influences. The last criterion is important because a speaker can sometimes distort the acoustic properties of a phonetic segment so severely that even the environment will provide no cues to the identity of the segment. Such a distortion is possible because a listener is capable of decoding such an utterance not only from the acoustic signal, but also from his familiarity with the

syntactic and semantic constraints, and with the rules governing allowable phoneme sequences of his language. In order to study the acoustic characteristics of speech sounds, it is certainly desirable to minimize such higher-level influences.

The format of the data was finally decided to be a nonsense word hə ˊCVC embedded in a carrier sentence, "Say _____ again". The prestressed consonants and consonant clusters, the vowels, and the poststressed consonants that were included in our data collection are listed in Table 2.I. They included the 15 vowels and diphthongs in English and essentially all allowable word initial consonants and consonant clusters in English. The hə ˊCVC nonsense word format, incidentally, had been used previously by other researchers [House and Fairbanks 1953, Stevens and House 1963, Stevens et al. 1966] to study the acoustic properties of vowels and consonants.

2.3 Acquisition of Acoustic Data

All the acoustic data were recorded in a sound-proof room where the signal-to-noise ratio is above 50 dB. An Altec 684B microphone was used in conjunction with a Presto model 800 tape recorder for the recording. The microphone was suspended from the ceiling and was placed approximately

SINGLETON CONSONANT	2-ELEMENT CLUSTERS	3-ELEMENT CLUSTERS	VOWEL
p	pl	spr	i
t	pr	spl	I
k	tr	str	e
b	tw	skr	ɛ
d	kl		æ
g	kr		ɑ
m	kw		ʌ
n	bl		o
s	br		ɔ
š	dr		u
f	dw		U
θ	gl		3
z	gr		ɑy
ž	gw		ɔy
v	fl		aw
ð	fr		
l	sp		
r	st		
w	sk		
y	sm		
	sn		
	sl		
	sw		

Table 2.I
A list of all the consonants, clusters, and vowels used

2 inches above the subject's upper lip and 10 to 12 inches in front of the subject. After being seated in front of the microphone, the subject was then asked to frame the utterance in a carrier sentence and read out loud. A sample of the list of utterances is included in Table 2.II. Prior to the actual recording, the subject was asked to read aloud for approximately one minute so as to allow adjustment of recording level and reading speed. Speaking rate was roughly maintained at 5 syllables per second within each sentence-like utterance. The subject was asked to speed up or slow down without explicit knowledge of the fact that speaking rate was being controlled. During the recording session, the person operating the recorder monitored the utterances through a set of headphones. Communication between him and the subject was achieved via signs displayed on the large window between the sound-proof room and the room where the recorder was situated. The subject was asked to pause after each list at which time he was asked to repeat those utterances that were erroneous. An entire recording session lasted approximately 45 minutes, marked by one or two rest periods. The subject preceded each utterance with a pre-assigned identification number as shown in Table 2.II. This number contains complete information of the phonetic context. For example, the number 2311 can be decoded to be a cluster /pr/, to be followed by the vowel

2201	hə'yɪt	2301	hə'pɻɪt	2401	hə'prɪt
2202	hə'yɪt	2302	hə'pɻɪt	2402	hə'prɪt
2203	hə'yɛt	2303	hə'pɻɛt	2403	hə'prɛt
2204	hə'yɛt	2304	hə'pɻɛt	2404	hə'prɛt
2205	hə'yæt	2305	hə'pɻæt	2405	hə'præt
2206	hə'yɑt	2306	hə'pɻɑt	2406	hə'prɑt
2207	hə'yʌt	2307	hə'pɻʌt	2407	hə'prʌt
2208	hə'yɔt	2308	hə'pɻɔt	2408	hə'prɔt
2209	hə'yɔt	2309	hə'pɻɔt	2409	hə'prɔt
2210	hə'yʊt	2310	hə'pɻʊt	2410	hə'prʊt
2211	hə'yʊt	2311	hə'pɻʊt	2411	hə'prʊt
2212	hə'yɝt	2312	hə'pɻɝt	2412	
2213	hə'yɑyt	2313	hə'pɻɑyt	2413	hə'prɑyt
2214	hə'yɔyt	2314	hə'pɻɔyt	2414	hə'prɔyt
2215	hə'yɑwt	2315	hə'pɻɑwt	2415	hə'prɑwt

Table 2.II A sample utterance list

/U/. Therefore, at the time when phonetic context is actually entered into the data-base, the user needs only to refer to this number which the computer will decode automatically. This method eliminates the confusing and sometimes difficult task of deciding the phonetic context from human perception of the utterance.

CHAPTER 3

ANALYSIS SYSTEM AND DATA-BASE FACILITY

One of the major aspects of this study is the development of a good acoustic analysis system. It is very desirable to have an analysis procedure where acoustic changes can be monitored as closely as possible, since acoustic characteristics of speech sounds can change significantly within a few milliseconds. To monitor the rapid spectrum changes at the onset of the release of a plosive, for example, would require an analysis technique more sophisticated than a conventional filter bank or sound spectrogram. Furthermore, in order to process such a large corpus of data in a reasonable amount of time, and to provide the user with relative ease of data access and retrieval, the processing and storage capabilities of the analysis system must be taken into consideration.

This chapter first gives an account of the signal processing aspects of our present study, with emphasis on a particular signal processing technique called linear prediction. The computer facilities are discussed next. We then describe the data-base facility in detail, giving examples of its capability. Procedures through which

acoustic data are entered into the data-base are also included in this chapter.

3.1 Acoustic Analysis Procedure

Digital computers and digital signal processing techniques have been chosen to perform our acoustic analysis. This choice offers many advantages, such as great flexibility, large data storage capability, and accuracy.

While certain acoustic events can be monitored conveniently in the time domain, experience has shown that frequency-domain representation of the speech signal often provides greater insights into the relationship between the articulatory and the acoustic realization of speech. For example, spectral peaks in non-nasalized vowels can be quite reliably correlated with the resonances of the vocal tract, and the frequency location of the major energy concentration in a plosive release gives good indications about the location of the constriction in the vocal tract. It is therefore often desirable to compute short-time spectra of the signal.

After experimenting with various methods of computing and smoothing the short-time spectra, we have chosen to compute the spectra via a speech analysis procedure known as linear prediction. The theory and limitations of linear

prediction analysis will now be presented.

3.1.1 Linear Prediction

Detailed treatment of the various formulations of linear prediction analysis can be found in the literature [Atal and Hanauer 1971, Markel 1971, Makhoul and Wolf 1972]. We shall elaborate on one of these formulations, which is commonly referred to as the covariance formulation. We have chosen to discuss this formulation primarily because the relevance of linear prediction analysis to the speech signal is most apparent this way.

Linear prediction analysis is based on the speech production model shown in Figure 3.1. The all-pole digital filter H(z) represents the combined effect of the glottal source, the vocal tract, and the radiation losses. In this idealized model the filter is excited either by a periodic impulse train for voiced speech, or random noise for unvoiced speech.

The speech production model can be equivalently characterized by the difference equation

$$s(n) = \sum_{k=1}^{p} a(k)s(n-k) + x(n) \qquad [3.1]$$

where s(n) and x(n) are the n-th samples of the output

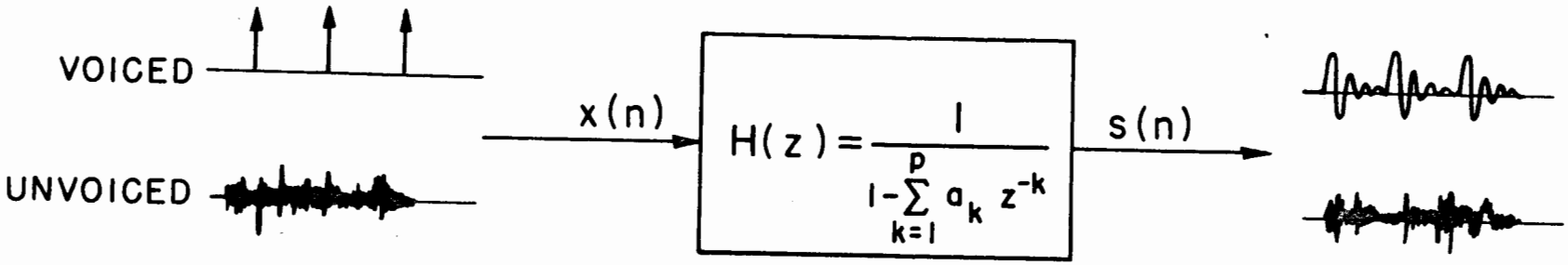


Figure 3.1 All-pole model of speech production

speech wave and the excitation, respectively. The a(k)'s are the coefficients characterizing the filter H(z), and henceforth will be referred to as the predictor coefficients.

From Equation [3.1] it is clear that one can determine the a(k)'s if the input and 2p consecutive values of s(n) are known, with the first p of these values serving as initial conditions. We shall restrict the following discussion to voiced speech for which the input is a periodic impulse train. In this case the a(k)'s can be determined with the knowledge of only 2p consecutive values of s(n) and the position of the impulse. For this idealized model, we can define the predicted value of s(n) as

$$\hat{s}(n) = \sum_{k=1}^{p} a(k)s(n-k) \qquad [3.2]$$

The difference between s(n) and $\hat{s}(n)$ will be zero except for one sample at the beginning of each period.

In reality, however, s(n) is not produced by this highly idealized model and therefore prediction of s(n) based on Equation [3.2] will introduce error. If we are to approximate s(n) by $\hat{s}(n)$ as defined in Equation [3.2], the a(k)'s can be determined only with the specification of the error criterion.

We can choose to determine the predictor coefficients by minimizing the sum of the squared-difference between s(n) and $\hat{s}(n)$, that is, by minimizing

$$E = \sum_{n=0}^{N-1} [s(n) - \hat{s}(n)]^2 \qquad [3.3]$$

where the minimization is to be carried out over a section of s(n) of length N.

The minimum mean-squared error criterion is chosen over other criteria because the determination of the a(k)'s now reduces to the solution of the following set of linear equations

$$\sum_{k=1}^{p} a(k)\, \phi(i,k) = \phi\,(i,0) \qquad i=1,2,..,p \qquad [3.4]$$

where

$$\phi(i,j) = \sum_{n=0}^{N-1} s(n-i)s(n-j) \qquad i,j=1,2,..,p \qquad [3.5]$$

Equation [3.4] can be written in matrix form as

$$\Phi\, \underline{a} = \underline{x} \qquad [3.6]$$

where Φ is a p by p matrix with typical element $\phi(i,j)$;

$\underline{a}$ and $\underline{x}$ are p-dimensional vectors with the i-th component given by a(i) and $\phi(i,0)$, respectively. The solution of the matrix equation is greatly simplified by the fact that the matrix is symmetric and hence recursive procedures are applicable [Faddeeva 1959].

It is of interest to compare the analysis procedure outlined above for two different cases. If the fundamental frequency of voicing (FØ) is known in advance, the analysis can be carried out directly in the sense that Equation [3.5] can be evaluated exactly. In practice, however, it is very desirable to carry out the analysis without a priori knowledge of FØ. In this case an approximation has to be made and additional error is introduced. We shall illustrate this point by a simple example. The argument can easily be generalized to include more complicated situations.

Let us assume that there is only one pitch pulse present in the data and that it occurs at n=m. If m is known, then Equation [3.5} can be evaluated as

$$\phi(i,j) = \sum_{\substack{n=0 \\ n\neq m}}^{N-1} s(n-i)s(n-j) \qquad [3.7]$$

Equation [3.5] can not be evaluated explicitly, however, if m is unknown.

If we choose to approximate $\phi(i,j)$ by

$$\hat{\phi}(i,j) = \sum_{n=0}^{N-1} s(n-i)s(n-j) \qquad [3.8]$$

Comparing Equations [3.7] and [3.8], we find that the error in $\phi(i,j)$ is given by

$$\varepsilon(i,j) = \hat{\phi}(i,j) - \phi(i,j) = s(m-i)s(m-j) \qquad [3.9]$$

By the nature of the speech signal, $s(m-i)$ and $s(m-j)$ are small compared with samples at the beginning of each period. Therefore the error $\varepsilon(i,j)$ is small compared with $\phi(i,j)$ for any reasonable N. Results comparing the two analysis procedure will be presented in a later section.

The theory of linear prediction has also been formulated in a slightly different way. Let e(n) denote the output of the inverse filter $H^{-1}(z)$ when it is excited by s(n). If we choose to determine the a(k)'s by minimizing the total energy in e(n), the set of equations obtained can be shown to be almost identical to Equations [3.4] and [3.5] [Markel 1971]. The major difference in the result is that the matrix Φ in this second formulation is of Toeplitz form

$$\phi(i,j) = \phi(|i-j|,0) \qquad [3.10]$$

so that the matrix equation can be solved by still more efficient algorithms [Levinson 1966]. The second formulation is sometimes referred to as the autocorrelation formulation.

From the predictor coefficients, the approximated spectral envelope of s(n) can be computed as $|H(e^{j\omega})|$. Note that the unit sample response of the inverse filter $H^{-1}(z)$ is given by

$$h(n) = \begin{cases} 1/a(0) & n=0 \\ -a(n)/a(0) & n=1,2,..,p \\ 0 & \text{otherwise} \end{cases} \qquad [3.11]$$

Therefore $|H(e^{j\omega})|$ can be obtained efficiently by computing the discrete Fourier transform of h(n) with a fast Fourier transform algorithm, and then inverting the result.

Linear prediction analysis assumes a specific speech production model where the combined effect of the glottal excitation, the vocal tract, and the radiation losses is represented by an all pole filter. Whether such a model is adequate for speech analysis leaves room for controversy. It is well known that for all non-nasalized sonorants the transfer function of the vocal tract has only poles. For

fricatives and nasals, however, the transfer functions have zeros as well as poles. Added to the problem is the fact that glottal peculiarities could sometimes introduce zeros. Therefore, from a theoretical standpoint, the all pole model is not always adequate. However, Atal has shown [Atal 1971] that these zeros are inside the unit circle, ans thus can be approximated by a number of poles via Taylor series expansion. Makhoul and Wolf [1972] have shown that linear prediction analysis can be viewed as a method of analysis-by-synthesis where the number of poles is specified and the result is a good fit to the envelope of the short-time spectrum. Therefore, within the context of spectrum analysis, the all pole model can be quite adequate. It is when one attempts to associate these poles with the genuine transfer function of the vocal tract that the all pole assumption needs to be reexamined. Experimental findings, with both synthetic and natural speech, have shown [Zue 1972] that linear prediction analysis captures the essential spectral characteristics of nasals and fricatives.

The number of poles, p, is determined by the sampling frequency as well as our knowledge of the speech production mechanism. For example, there exist in general 5 to 6 complex pole pairs for the vocal tract transfer function up to 5 kHz. Adding two real poles for the combined effect of the glottal source and the radiation losses, we arrived at a

value of p between 12 and 14 for the speech signal sampled at 10 kHz. For those with zeros in the transfer function, a higher value of p would be desirable.

As mentioned earlier, there are several formulations of the linear prediction technique that are closely related but have important theoretical differences. The set of issues has been explored in great detail elsewhere [Portnoff 1972, Makhoul and Wolf 1972]. However, when applied to a complete speech analysis-synthesis system, comparable results have been reported in the literature [Atal and Hanauer 1971, Itakura and Saito 1968]. We have implemented both the autocorrelation and the covariance methods of linear prediction, and the quantitative differences between the two methods were found to be minimum. The autocorrelation method has the advantage of smoother spectral variation from frame to frame, a direct consequence of windowing the speech. Also, computationally the autocorrelation method is very efficient when increasing the order of the predictor from p to p+1, as is probably required for nasals and nasalized vowels.

3.1.2 Results Comparing various Spectrum Analysis Techniques

Figure 3.2 compares spectra of a synthetic vowel /ɑ/ obtained by various spectral smoothing techniques: (a) and

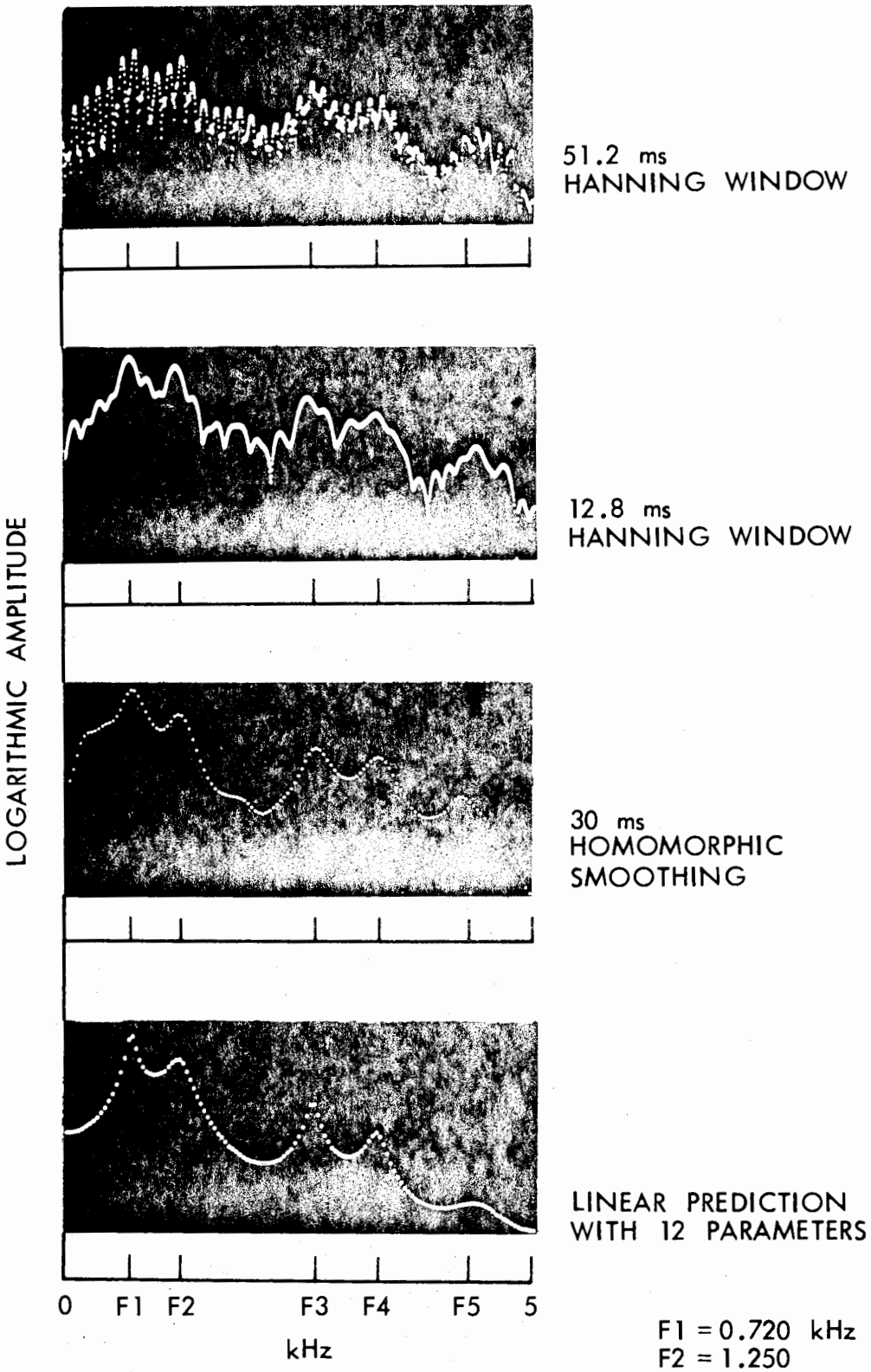


Figure 3.2 Spectra of a synthetic /a/ (obtained by various analysis techniques)

(b) by windowing (with different window widths) and Fourier transforming the waveform, (c) by cepstral smoothing [Oppenheim and Schafer 1968], and (d) by linear prediction. In Figure 3.2(a) the effect of glottal periodicities can be seen as the ripples superimposed on the spectral envelope. These ripples are greatly reduced in Figure 3.2(b) because of the spectral smearing of the wide frequency window. In Figure 3.2(c) the effect of the glottal excitation is removed by a homomorphic technique. This effect is also removed in Figure 3.2(d). However, since the linear prediction analysis is based on a specific speech production model and thus limits the number of spectral peaks, there are no extraneous peaks in Figure 3.2(d). If we compare the locations of the spectral peaks with the actual values of the five formants, it is clear that, for this example, the spectrum derived from linear prediction provides accurate formant information.

Figure 3.3 shows the spectra for the same vowel obtained by linear prediction, except in this case the analysis is carried out pitch-synchronously. Comparing Figures 3.2(d) and 3.3, except for the bandwidth of the second spectral peak, we find that the qualitative difference between the two spectra is quite small, as discussed in the previous section.

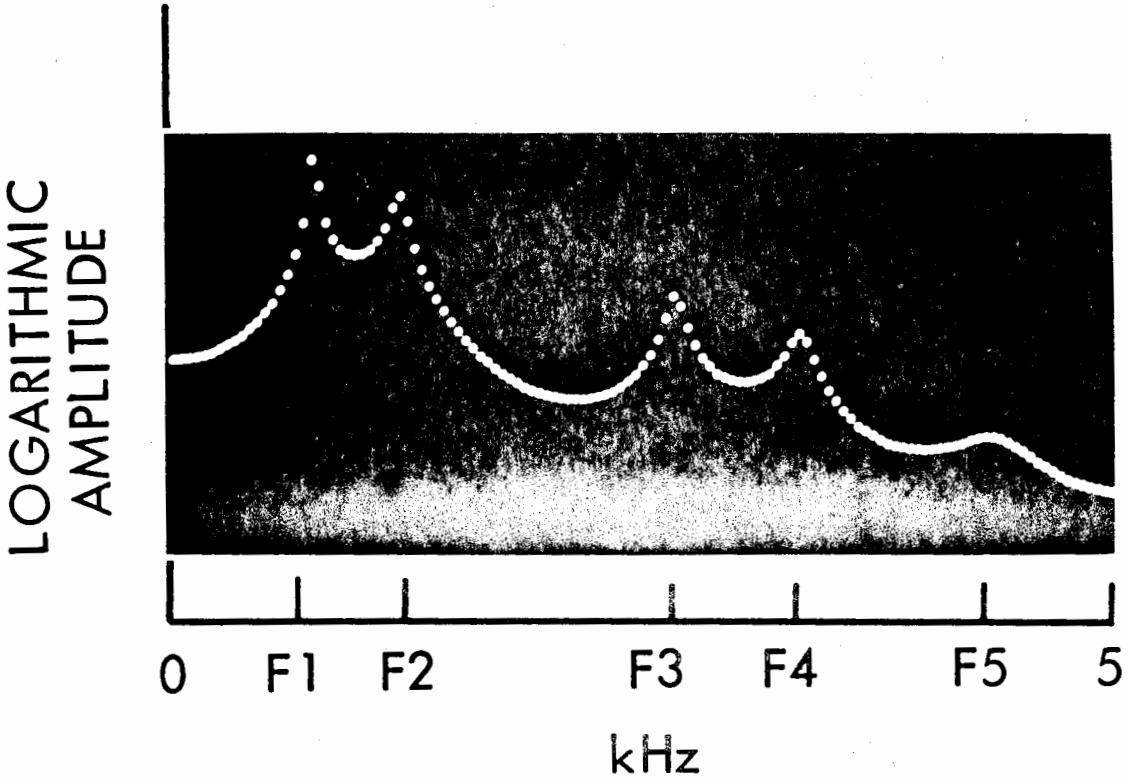


Figure 3.3 Spectrum of a synthetic /ɑ/ (obtained by pitch-synchronous linear prediction analysis)

Figure 3.4 displays a linear prediction spectrogram and three time functions derived from the same speech material. It can be seen that the RMS amplitude measurements derived from the speech signal and from linear prediction smoothed spectra are practically identical. The linear prediction spectrogram also bears a distinct resemblance to conventional spectrograms.

One property of linear prediction analysis is that the technique is relatively insensitive to pitch variations [Atal and Hanauer 1971]. Figures 3.5 and 3.6 illustrate this point with synthetic speech of different fundamental frequency contours. When the fundamental frequency (FØ) is held constant and is less than, say, 200 Hz, the analysis technique is practically insensitive to the value of FØ. When FØ is time varying, the technique shows various degree of deterioration, depending on the rate of FØ change and the length of the analysis window, N. Atal and Schroeder [1974] have outlined the necessary modifications to the linear prediction analysis precedure when FØ becomes extremely high.

Figures 3.7-8-9 compare spectra obtained by linear prediction with those obtained by discrete Fourier transform for some vowel and fricative sounds. The spectral matching property of the linear prediction technique is illustrated

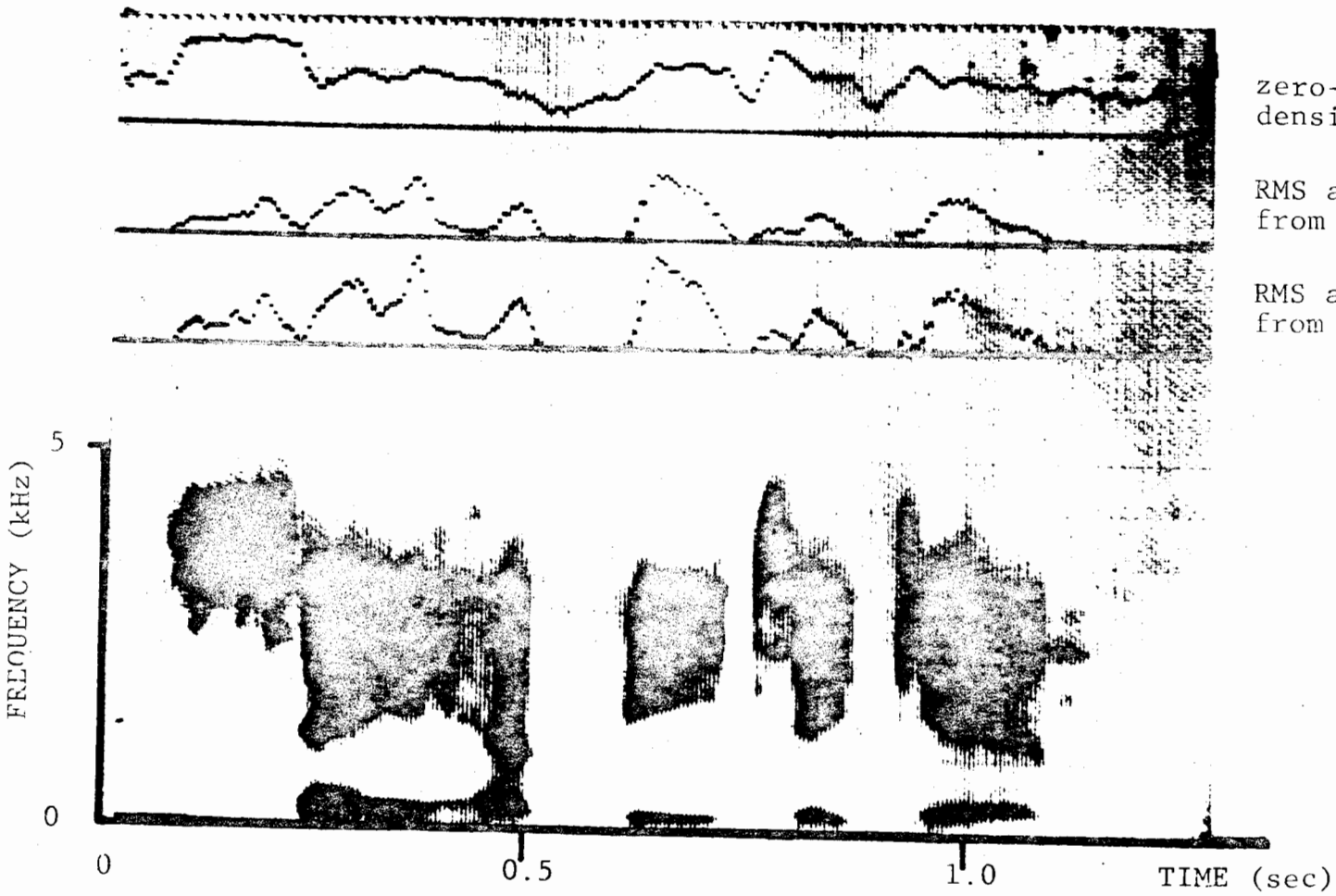
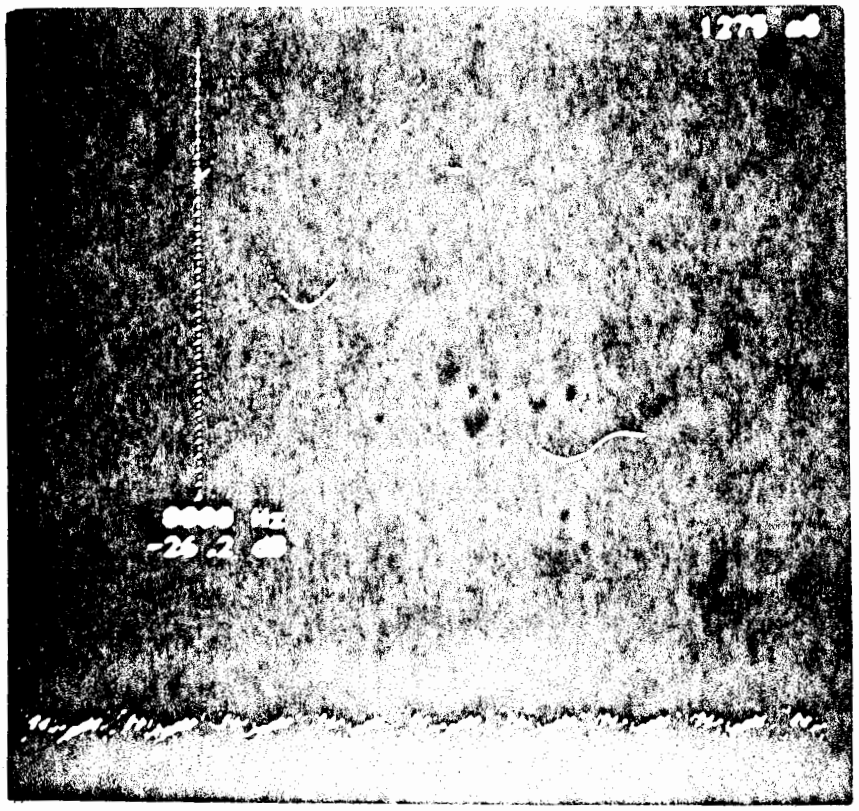
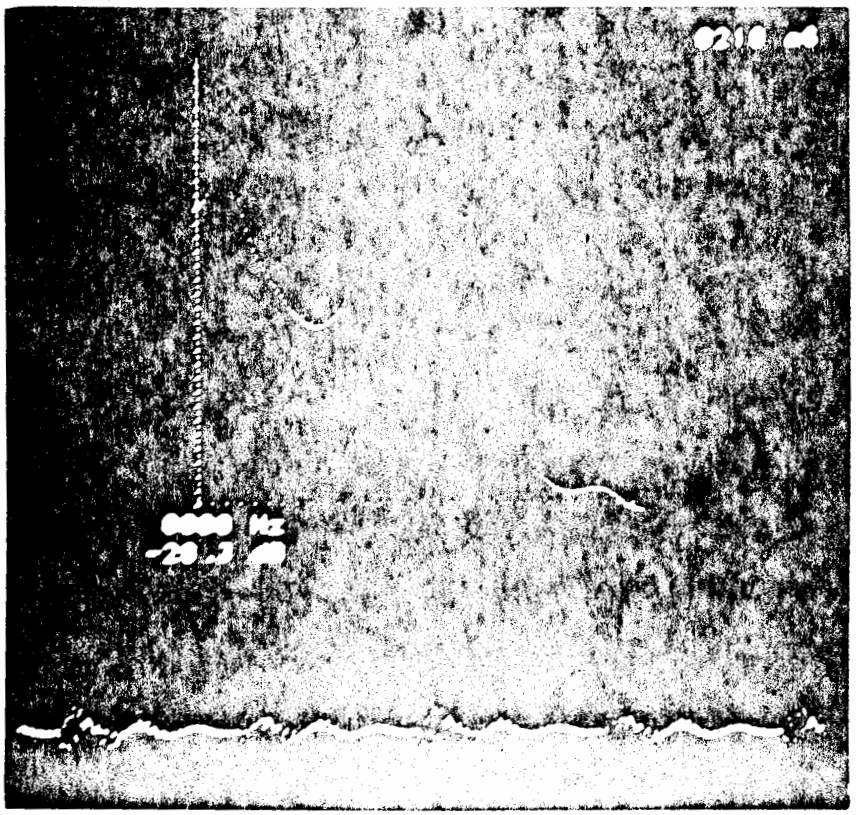


Figure 3.4 A linear prediction spectrogram
(the utterance is "say hə'dit again")

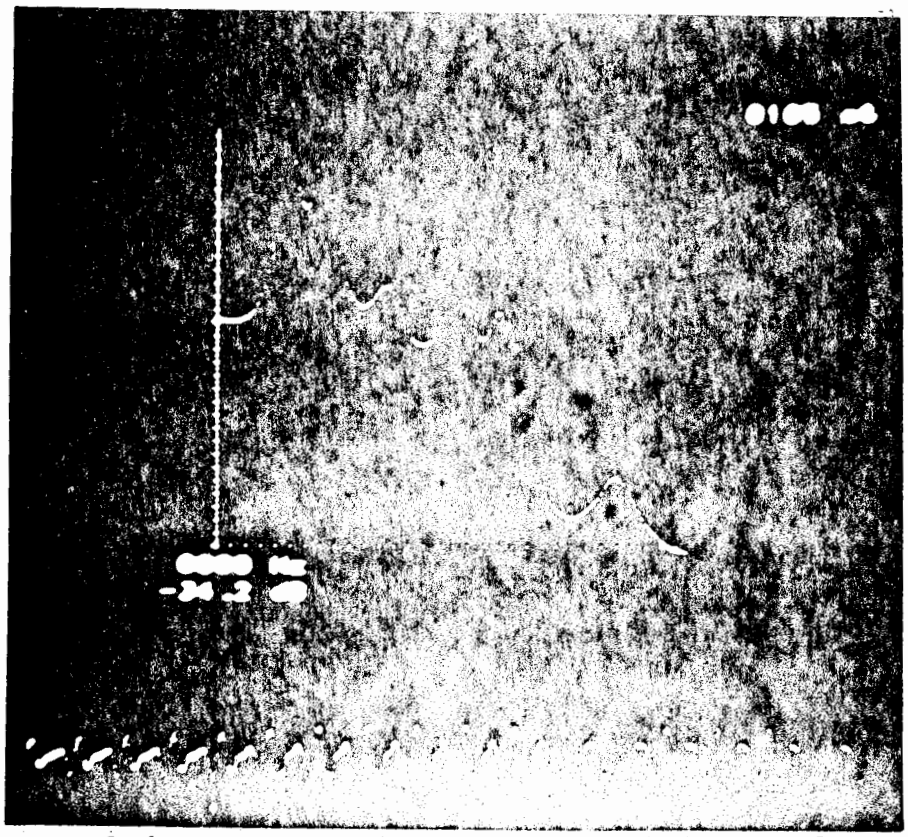


(a)

Figure 3.5 Spectra of a synthetic /i/ with different fundamental frequencies ((a) FØ=200 Hz, (b) FØ=80 Hz)

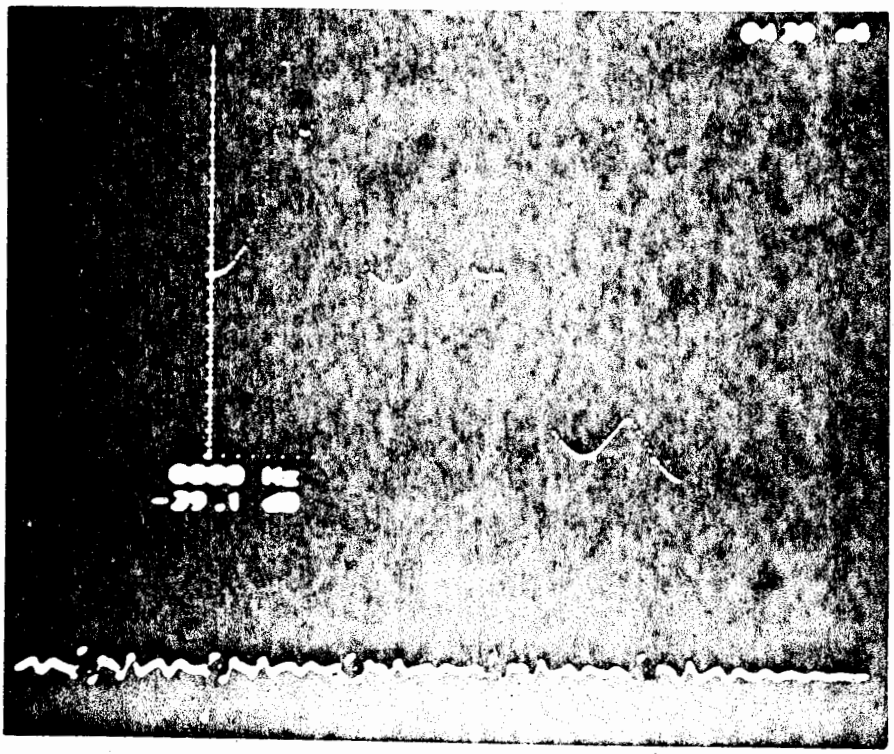


(b)

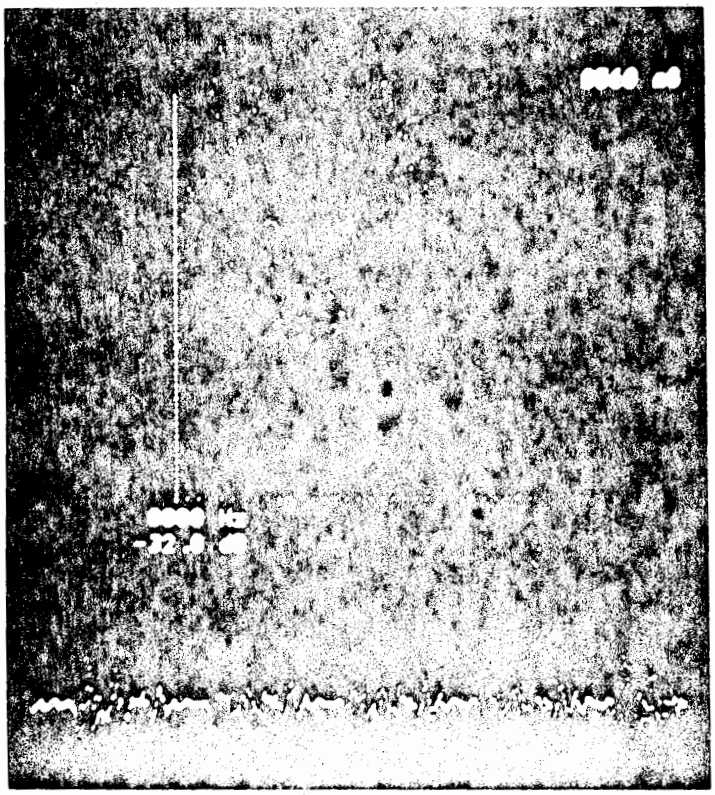


(a)

Figure 3.6 Spectra of a synthetic /ɑ/ with different fundamental frequency contours (linear FØ ramps (a) slope=800 Hz/sec, (b) slope=400 Hz/sec; note extra peak in (a))

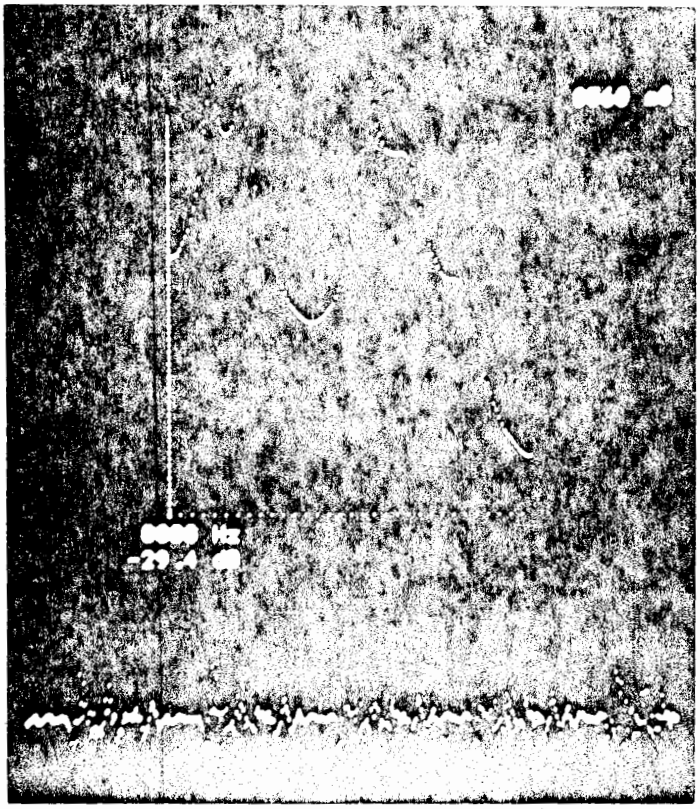


(b)

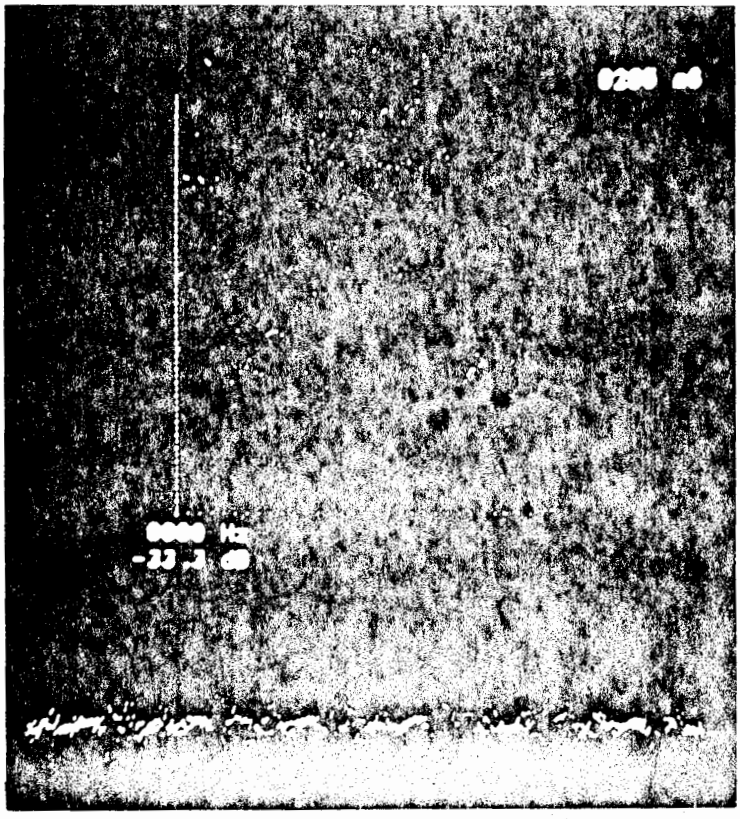


(a)

Figure 3.7 Spectra of an /a/ (obtained by (a) DFT, and (b) linear prediction)

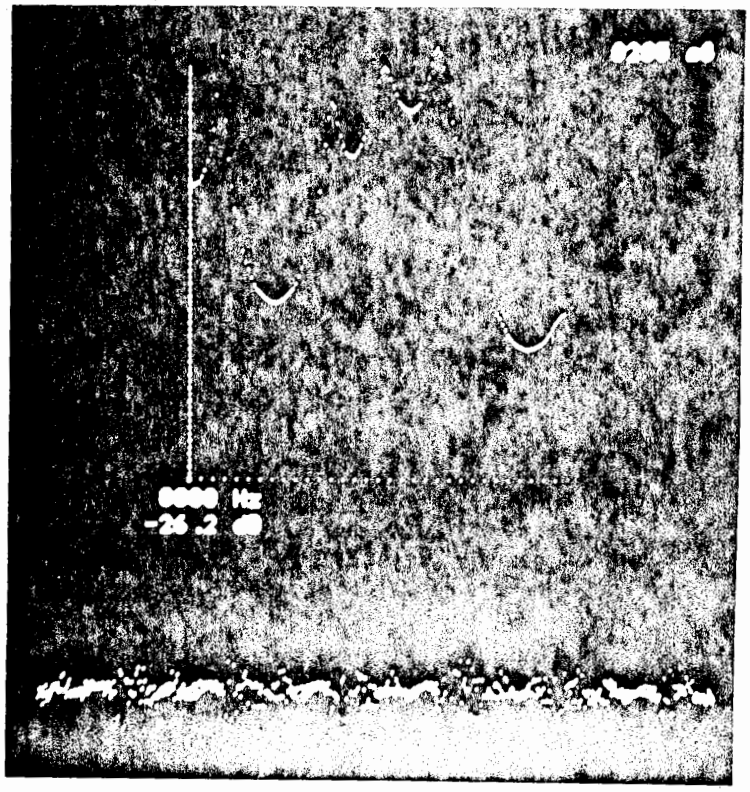


(b)

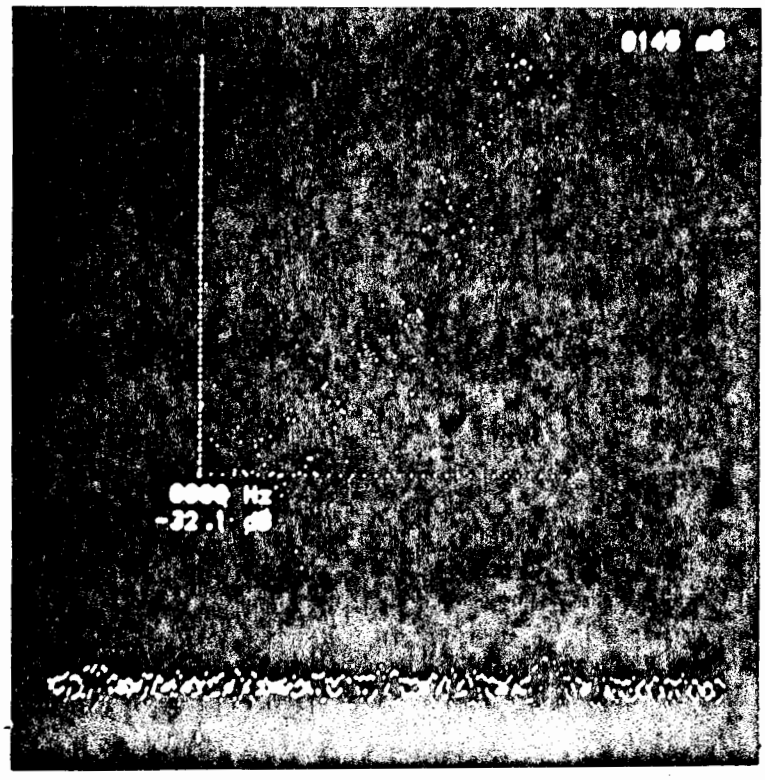


(a)

Figure 3.8 Spectra of an /i/
(obtained by (a) DFT, and
(b) linear prediction)

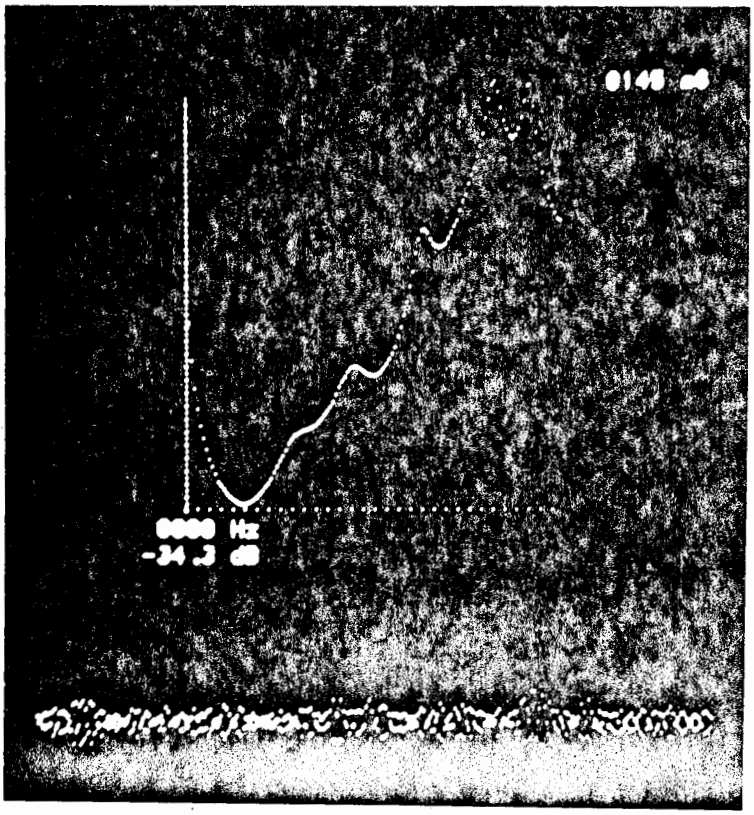


(b)



(a)

Figure 3.9 Spectra of an /s/
(obtained by (a) DFT, and
(b) Linear Prediction)



(b)

quite clearly in these figures. Figure 3.10 contains spectrographic dislays of an utterance where the spectra are obtained by linear prediction or DFT. Properties of both techniques are self apparent.

3.2 Computer Systems

Because of the enormous size of the data-base, the processing and storage of data constitutes an integral part of the research problem. To perform such a complicated analysis as linear prediction on thousands of utterances in a reasonable amount of time requires a fast signal processor. To allow storage and easy access to a fair number of utterances needs a computer with good size storage capacity. Since these two requirements seems to be orthogonal in the sense that no single computer available can accommodate both of these ends, two different computer facilities were eventually used. The Univac-FDP system was used for signal processing and the TX-2 system was used for the storage and on-line data analysis.

3.2.1 The Univac-FDP System

The signal processing part of the data-base utilizes the computer facility of the Digital Processor Group at MIT Lincoln Laboratory. The facility includes a Univac 1219,

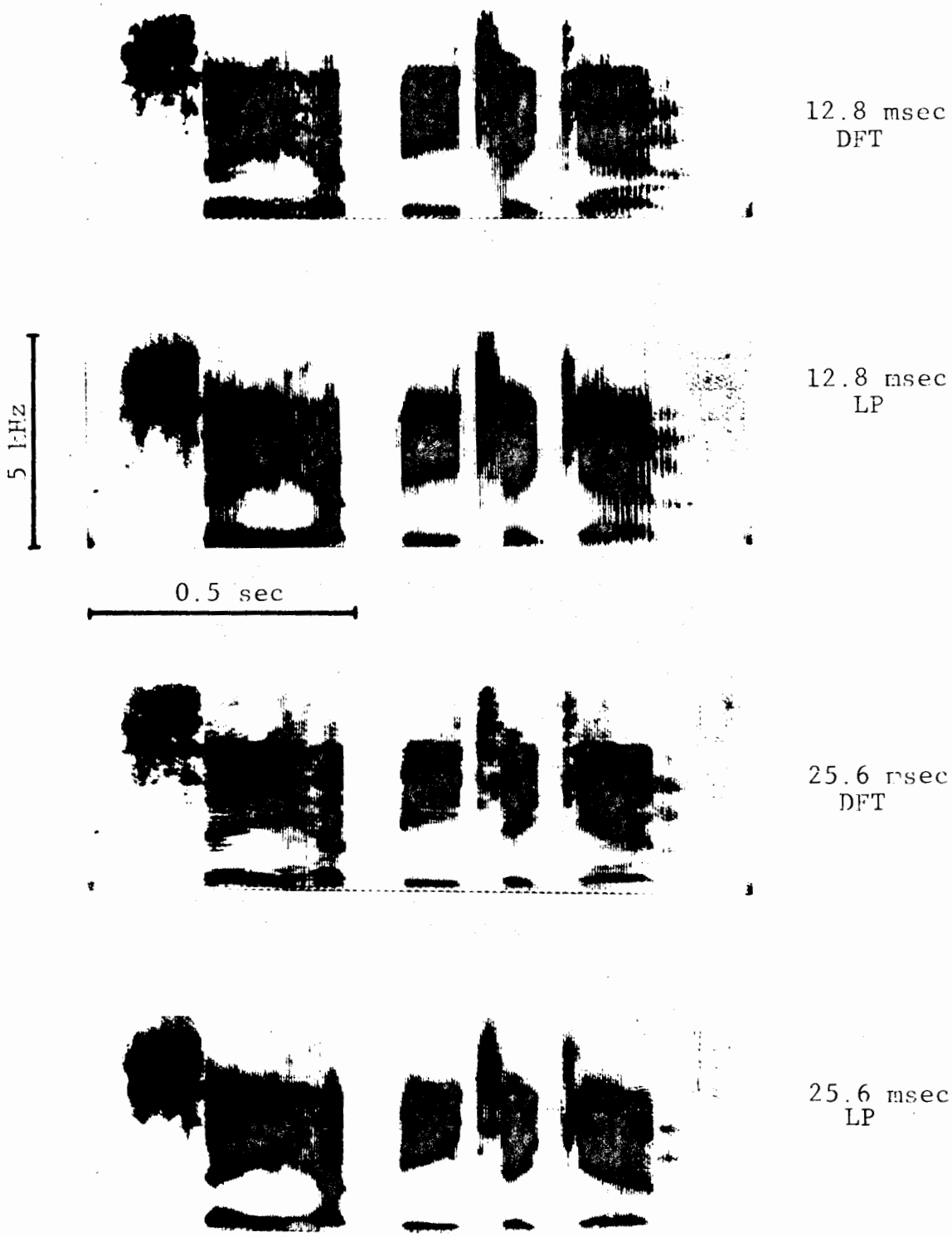


Figure 3.10 Digital spectrograms obtained by various analysis techniques and parameters (serves to illustrate the differences between DFT and linear prediction, and the trade-off between time and frequency resolutions)

which is a medium-sized general-purpose computer, the fast digital processor (FDP), which is a very high speed programmable signal processor, and a number of peripheral devices .

The Univac-1219 is an 18 bit, one's complement digital computer with 32K words of core memory and a cycle time of 2 microseconds. Peripheral devices include a paper tape reader and punch, a drum, two tape drives for 7-track magnetic tapes, A/D and D/A converters, and two CRT displays. The drum has a capacity of 917K words, and a maximum data rate of 103K words per second. One of the CRT's is a point plotting scope where time waveform and spectrum can be displayed. The other display system generates a 256 by 256 point raster and controls the brightness of each point by varying the intensity of the electron beam of the CRT. The refresh rate is approximately 10 frames per second. For speech research, this display is well suited for generating spectrograms.

The FDP, an 18 bit two's complement programmable signal processor, was built by the Group and the architecture of the machine was designed such that it is well suited for signal processing applications. The FDP has two independent 4096 word data memories, a separate 480 word control memory and four arithmetic elements each with its own multiplier

and adder. The machine executes two 18 bit program instructions simultaneously utilizing instruction overlap so that the effective double instruction cycle time is 150 nanoseconds. Effective multiplication time is 450 nanoseconds. Maximum data transfer between the FDP and the Univac-1219 is 166K words per second. The FDP also has a set of peripheral devices, including A/D and D/A converters and a 166K word peripheral core memory with a cycle time of a little over three microseconds.

3.2.2 The TX-2 System

The TX-2 computer was designed and built at Lincoln Laboratory during 1956-1959. It is a 36 bit, one's complement , single address machine with indexing and multilevel indexable indirect addressing. The TX-2 has 164K words of memory and programs are run under the APEX time-sharing system that provides each user with up to 128K of virtual address space. Major peripheral devices include a 800M bit drum, magnetic tape transports, A/D and D/A converters, and various display facilities. Each time shared terminal has a set of displays and a tablet that provide a highly interactive environment of man-machine communication. In addition, hard copy of the display can be obtained from a LDX printer.

3.3 Procedures for Data Processing and Data-Base Facility

Recorded utterances are first digitized and processed on the Univac-FDP. The processing also includes the relatively time-consuming task of manually marking the phonetic categories and durations of all segments. Results of the processing are then stored on digital magnetic tapes and entered into the data-base on the TX-2.

3.3.1 Processing on the Univac-FDP

Analog speech is band-limited to 5 kHz and sampled to 12 bits at 10 kHz. The sampling rate was chosen because, although there exists interesting acoustic information above 5 kHz for some speech sounds, most of the essential acoustic features are below 5 kHz. Increasing the sampling rate proportionally increases storage space and computation time. Increasing the sampling rate also requires a higher-order linear predictor, which is quite an undesirable feature for many reasons.

Short-time spectra can be obtained either through DFT or linear prediction analysis. Both methods operate on 12.8 milliseconds of Hanning-windowed speech, and a new spectrum is computed every 5 milliseconds. For the linear prediction analysis, we have experimented with the different formulations of the technique and have found minimal

differences in the results. It was then decided to use the autocorrelation method for the reasons discussed earlier.

With the help of the various display programs, the utterance is hand-marked with segmental boundaries. This labeling procedure is illustrated in Figure 3.11. The user makes use of the simultaneous displays of the digital spectrogram, RMS amplitude function, time waveform and spectral cross sections. All displays are time-synchronized in that the short vertical bar immediately above the spectrogram marks the position corresponding to the spectral slice and the waveform. Selected portions of the utterance can be played back through the D/A converter to resolve ambiguity. The level-coded waveform above the spectrogram is the result of a typical labeling session. The utterance is: "Say h ə ˊdɑt again", and the hand-label marks the schwa before the stop, the silence and release of the prestressed /d/, the vowel, the silence, release and aspiration of the poststressed /t/, and the following schwa. Only the schwa-CVC-schwa portion of the utterance is actually retained on the computer.

Hand-labeling utterances is a tedious task that also involves a lot of subjective decisions. It is a well known fact to people who have had experience in phonetic labeling that a given utterance can be labeled quite differently,

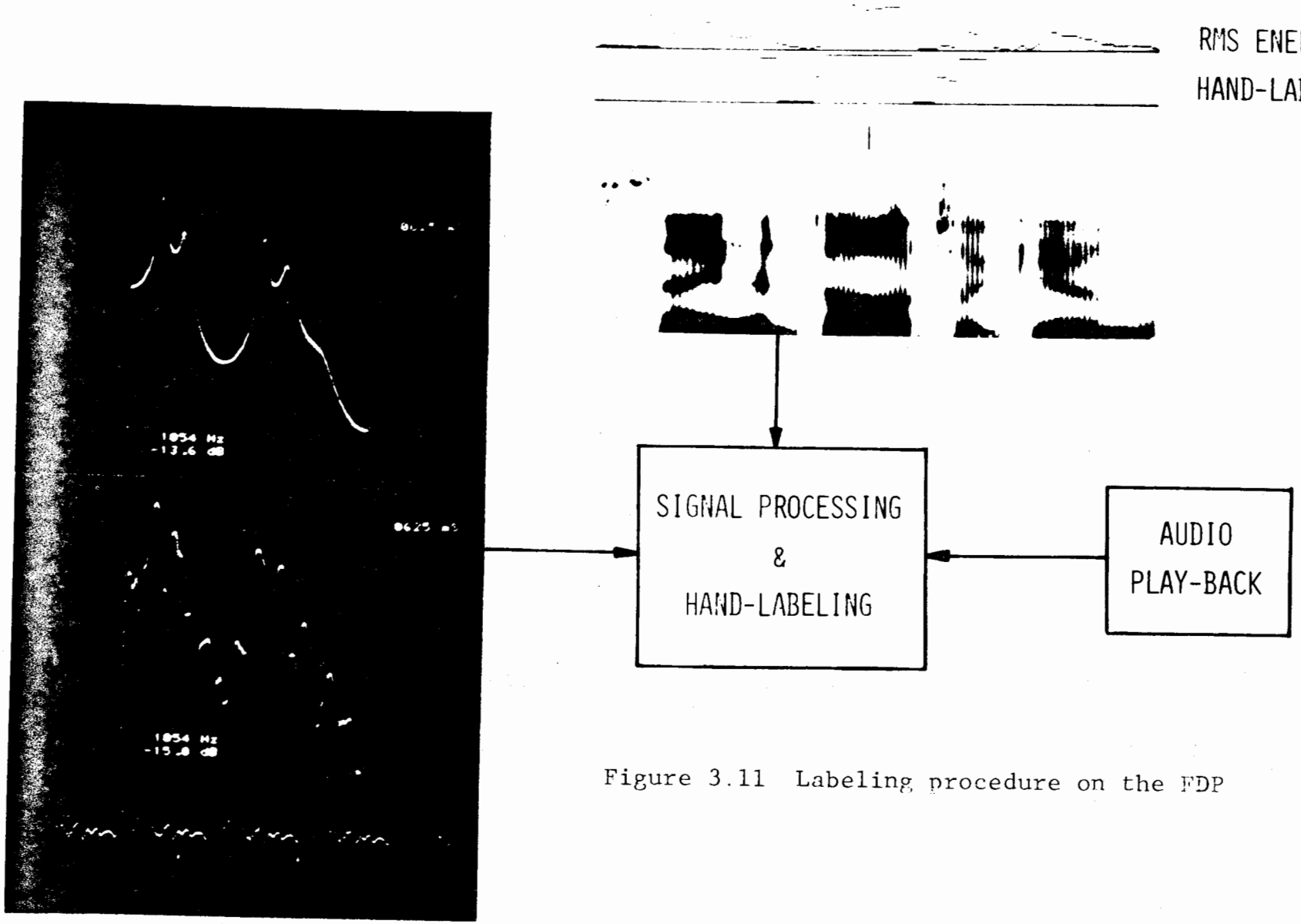


Figure 3.11 Labeling procedure on the FDP

both in phonetic content and acoustic boundaries, from one person to another. For example, the boundary between the burst and the aspiration of a voiceless aspirated stop is very difficult to determine, since acoustically there is a considerable amount of overlap between these two phases of the stop release. The only way to overcome a problem of this nature is for one to lay down, in advance, a set of rules to deal with ambiguous situations, and to adhere to these rules as consistently as possible when labeling.

After an utterance is hand-labeled, the digitized waveform, the linear prediction spectra, the RMS amplitude function, and the level-coded label function are all written into the digital magnetic tape. These unformatted tapes are then read into the TX-2, formatted, and written onto data-base tapes for later use.

3.3.2 Data-Base Facility

The speech data-base operates under the Speech Processing Controller (SPC) [Stowe 1972]. SPC is specifically developed for the storage and manipulation of a large amount of speech data. The SPC and the essential structure of the data-base, incidentally, were developed in connection with the speech understanding research at LIncoln Laboratory. Although the project terminated in 1973, the

structure and programs remained on the TX-2 and have proved to be quite useful. Each utterance in this study is identified in the TX-2 data-base by an 11 character string, and a field-type (FT) number. The characters in the string designate the recording session, speaker initials, prestressed consonant (or consonant cluster), and poststressed consonant. The exact format of the string is illustrated in Figure 3.12. The field type (FT) number, ranging from one to 15, specifies the stressed vowel in the utterance. Table 3.I lists the consonants and vowels with the assigned identification number. For example, the string "2KNS9021805" and FT=13 uniquely designates the utterance " ə´straytə " spoken by speaker KNS during the second recording session.

Raw data in the data-base can be examined in two basic forms. Figure 3.13 gives an example of the spectrographic display on the TX-2. The spectrogram, admittedly inferior in quality to its analog counterpart in that acoustic details are not as apparent, provides an excellent means to examine the global characteristics of the utterance. Simultaneous displays of time, synchronized hand labels, the RMS amplitude function and phonetic content with the spectrogram makes this display a powerful tool with which to correlate important acoustic features.

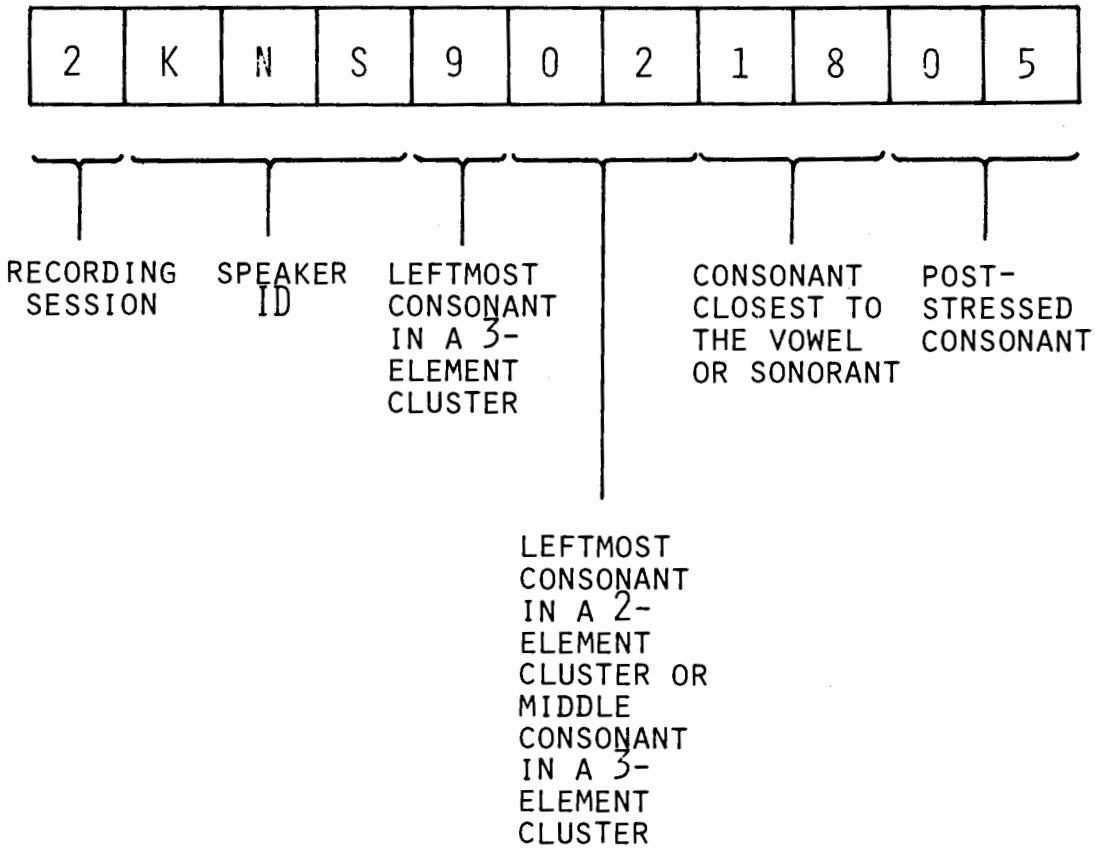


Figure 3.12 Format of the ID for an utterance in the data-base

CONSONANT	INTERNAL NUMERICAL REPRESENTATION	VOWEL
p	1	i
t	2	I
k	3	e
b	4	ɛ
d	5	æ
g	6	ɑ
m	7	ʌ
n	8	o
s	9	ɔ
š	10	u
f	11	U
θ	12	3
z	13	ɑy
ž	14	ɔy
v	15	ɑw
ð	16	
l	17	
r	18	
w	19	
y	20	

Table 3.I Correspondence between the vowels and and consonants and the internal representations on the TX-2

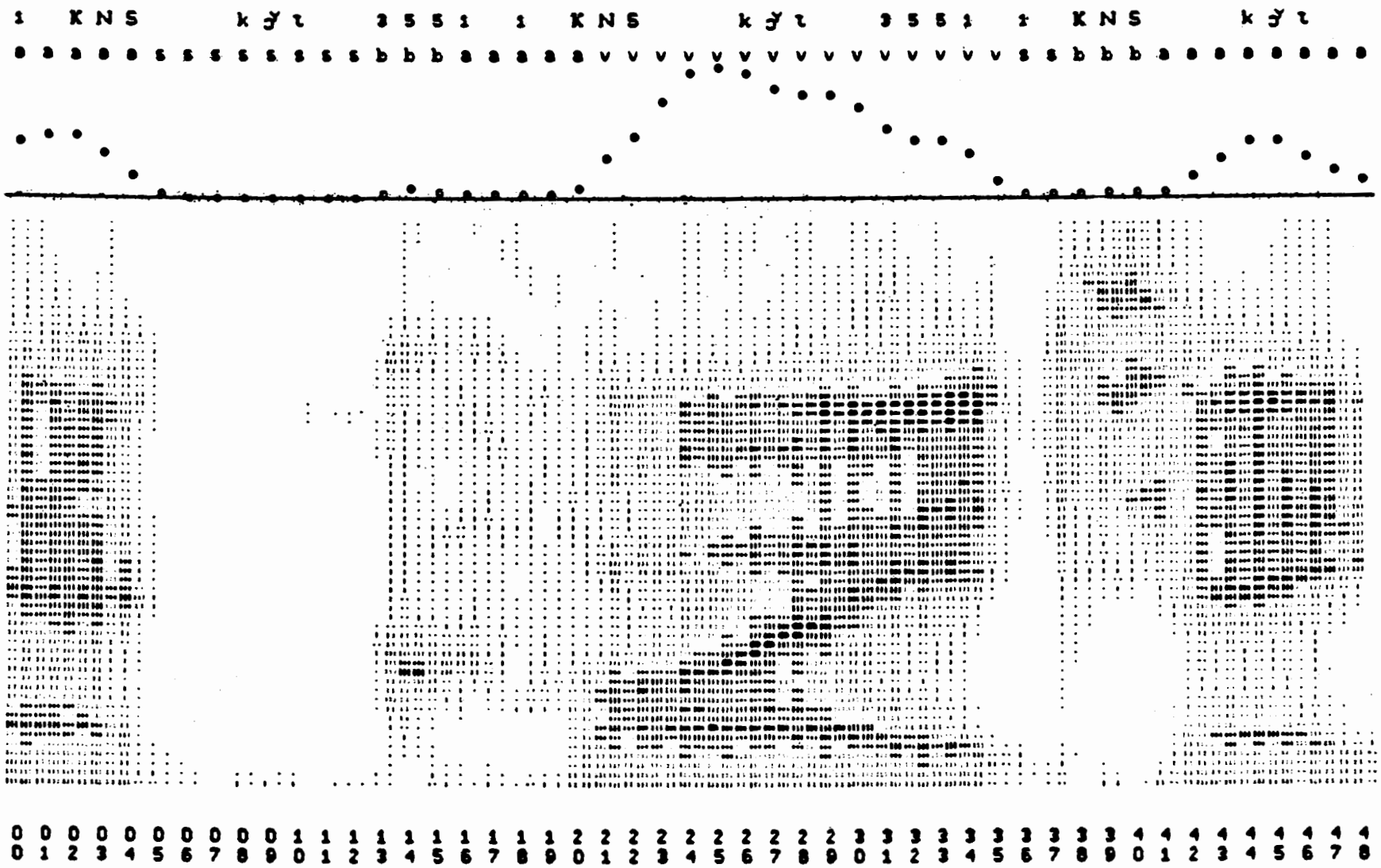


Figure 3.13 Hard copy spectrogram display from TX-2 (the utterance is "ə'kɔytə"; recording session, speaker ID and maximum RMS amplitude shown on the top line, followed by hand label and RMS amplitude function; time in msec)

Figure 3.14 illustrates the second form of data display where eight consecutive spectral cross sections are shown with the corresponding time waveform. The vertical bars on the waveform, spaced 5 milliseconds apart, mark the center of each analysis window. The numbers under the vertical bars correspond to the numbers next to the spectrum. The recording session, speaker initials, and phonetic content are indicated automatically at the top of the figure. This display is useful in studying rapid spectral changes at the release of stops, and measuring various durations of speech sounds.

It should be emphasized that the most prominent feature of the data-base is not its size as much as the accessibility of its content. By simply specifying the ID of the desired utterance (i.e., ID string and FT number) and typing a few commands, waveform, spectra etc. of the utterance become available almost immediately. Partial specification of the ID will result in an exhaustive search of the entire data-base for utterances fitting the description. For example, if the recording session and speaker initials are not specified, the computer will return all utterances fitting the phonetic context specified. Table 3.II gives an example of the tabulated results of an experimental session where the vowel identity is not specified, resulting in the tabulation of measurements for

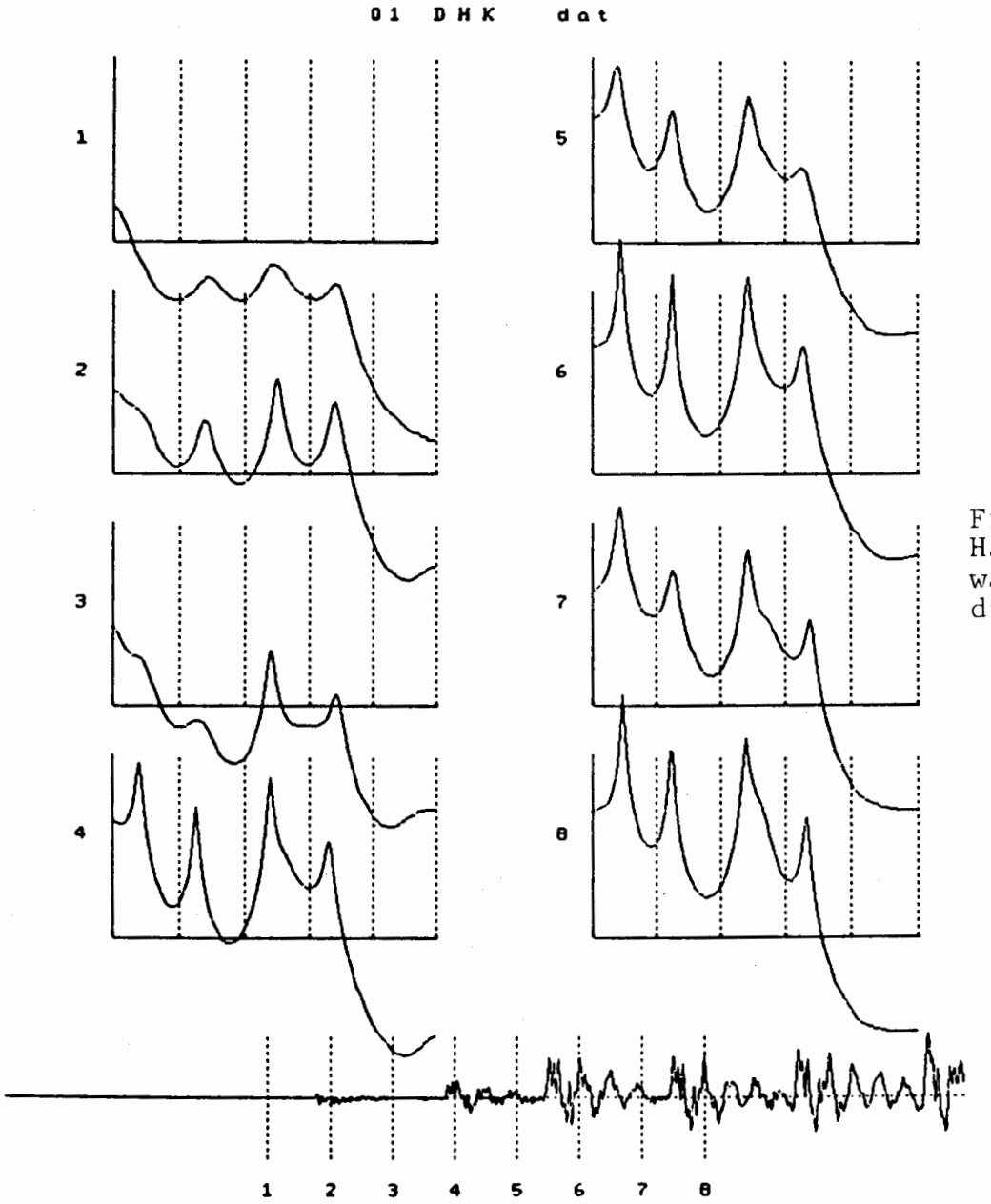


Figure 3.14
Hard copy consecutive waveform and spectra display from TX-2

1	2	3	4	5	6	7	8	9	10	11	12	13	14
70	90	'g'	'ɪ'	't'	75	15	15	0	797	2734	2128	135	2
80	75	'g'	'ɪ'	't'	55	20	20	0	163	2578	1300	120	11
65	80	'g'	'e'	't'	45	35	35	0	177	2578	990	150	15
75	70	'g'	'ɛ'	't'	55	15	15	0	107	2578	752	160	5
80	80	'g'	'æ'	't'	50	30	30	0	203	2421	1391	200	11
70	75	'g'	'ɑ'	't'	50	25	25	0	111	1718	736	200	10
75	100	'g'	'ʌ'	't'	60	40	40	0	165	1796	974	115	4
75	90	'g'	'o'	't'	50	40	40	0	121	1484	953	160	-6
75	80	'g'	'ɔ'	't'	35	45	45	0	90	1562	677	200	2
80	100	'g'	'u'	't'	45	55	55	0	87	1484	639	130	-2
70	105	'g'	'ʊ'	't'	60	45	45	0	163	1406	1837	100	6
65	80	'g'	'ɝ'	't'	45	35	35	0	56	1875	280	155	9
60	75	'g'	'ɑʸ'	't'	40	35	35	0	171	1796	1393	190	-7
80	80	'g'	'ɔʸ'	't'	40	40	40	0	109	1484	1021	195	8
65	75	'g'	'ɑʷ'	't'	45	30	30	0	83	1640	542	190	9

Table 3.II A sample measurement list
(the columns are: 1. start time;2. total duration;
3. prestressed consonant;4. vowel;5. poststressed
consonant;6. silence duration;7 VOT·8. Burst;
9. aspiration;10. overall burst intensity; 11. burst
frequency;12. burst amplitude;13. vowel duration;
14. relative burst amplitude)

all vowels.

Table 3.II also illustrates the types of measurements that can be obtained conveniently through search of the data-base. Each row in the table corresponds to a new entry, with the phonetic environment specified in the third column. Each column represents a different measurement. For example, columns 7 and 13 correspond to durational measurements of voice-onset time (VOT) and of the following stressed vowel, respectively. Column 14 measures the intensity of the release for the stressed plosives.

Facilities also exist for averaging and displaying the results of an experiment. Figure 3.15 shows a plot of average vowel durations in the context of ə´t__tə for the fifteen vowels and diphthongs. The horizontal axis represents the 15 vowels and diphthongs, and the vertical axis measures vowel durations in milliseconds. The results are in congruence with those reported in the literature [for example, House 1961, Peterson and Lehiste 1960].

Figure 3.16 represents a scatter diagram of voice on-set time versus burst frequency for all singleton stops of speaker KNS. The terminology and results will be discussed in later chapters. The purpose of introducing this figure now is to demonstrate that a large amount of data can be analyzed and displayed conveniently. Close

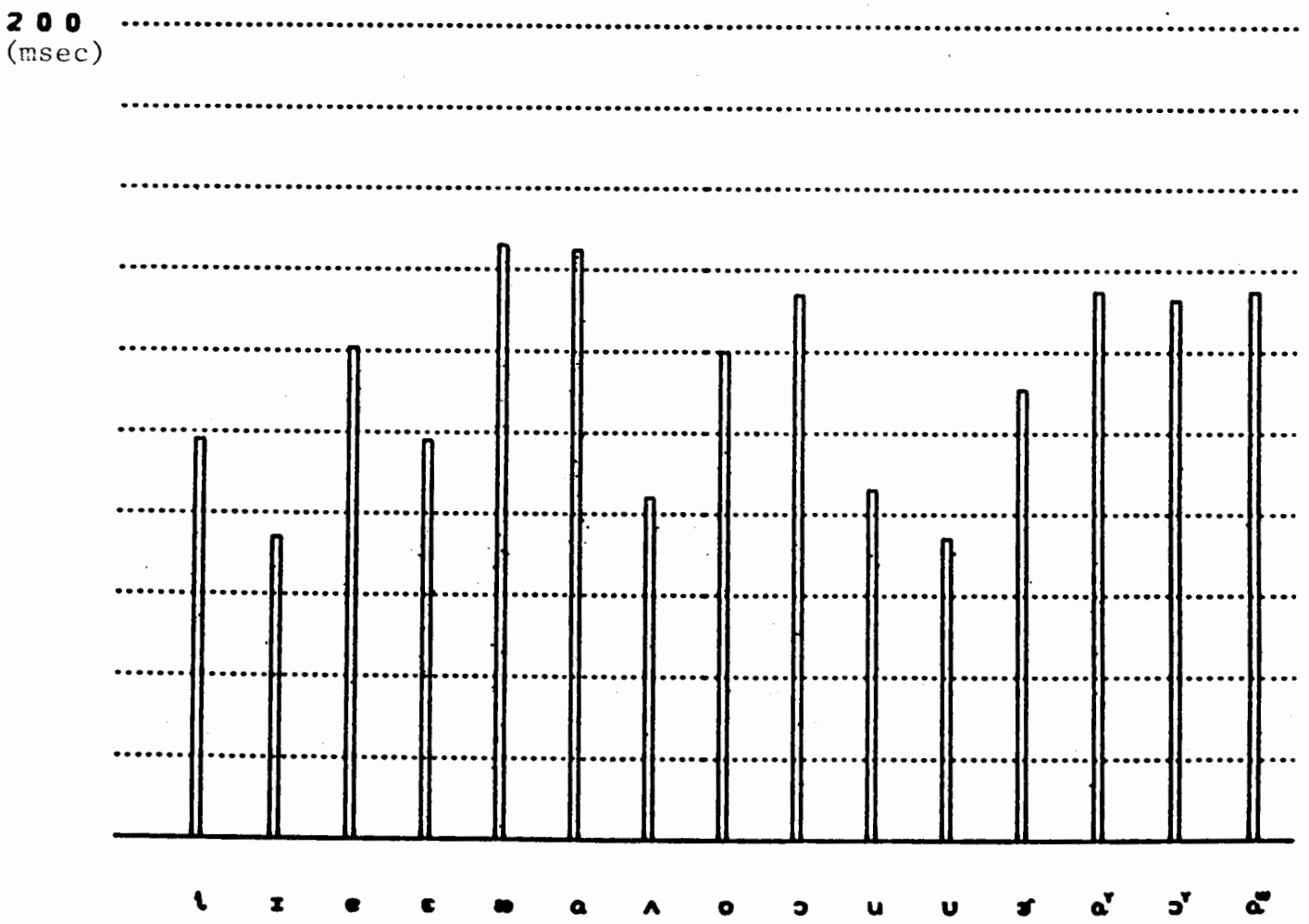


Figure 3.15

Average Vowel Duration in the syllable tVt. Speaker KNS

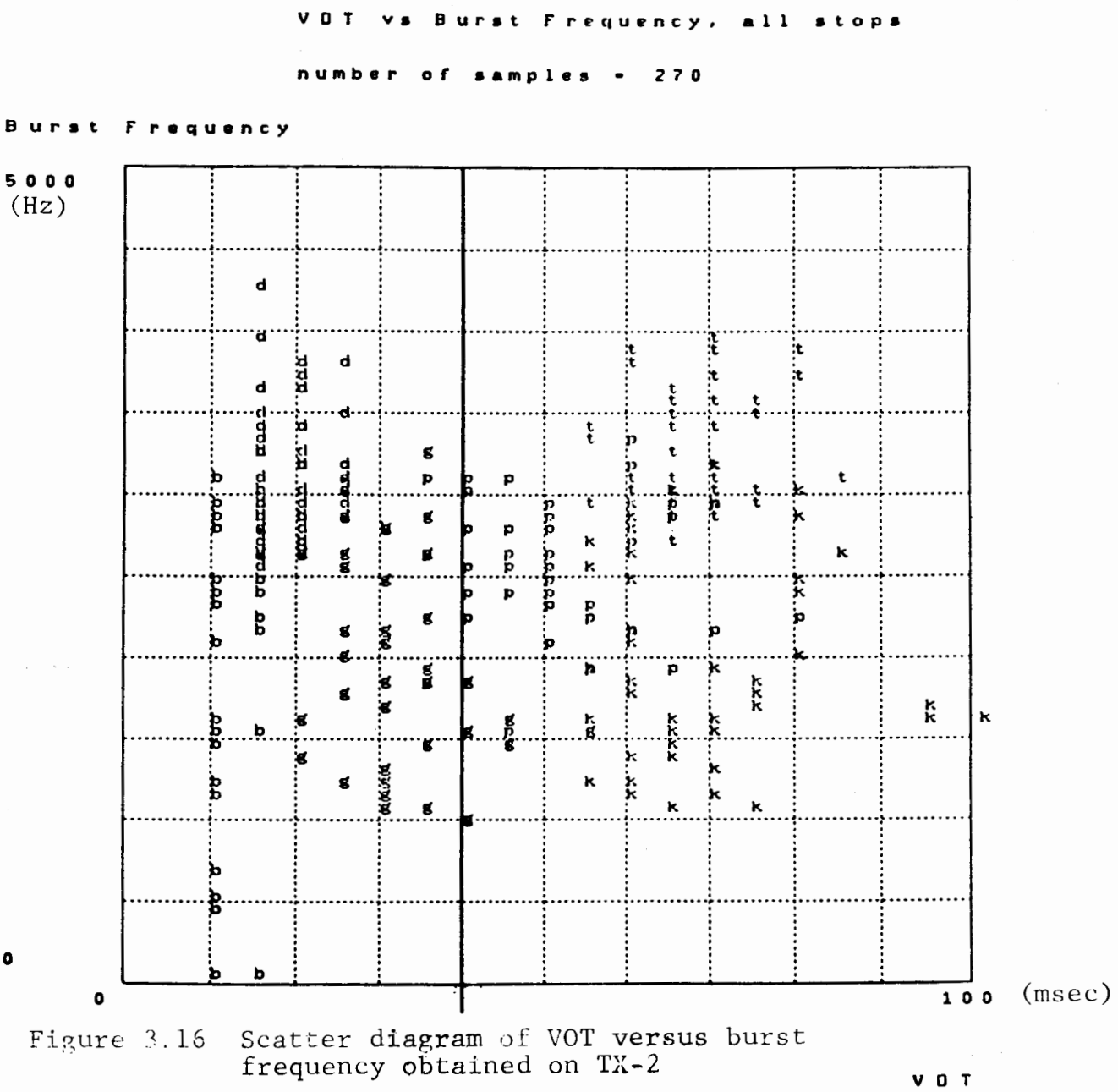


Figure 3.16 Scatter diagram of VOT versus burst frequency obtained on TX-2

examination of this figure reveals that over 95% of the voiceless stops have a VOT greater than 40 milliseconds, and that coronal stops in general have a burst frequency greater than 3,000 Hz. The time required to generate such a plot is of the order of 2 to 5 minutes, depending on the size of the data set.

CHAPTER 4

TEMPORAL CHARACTERISTICS OF ENGLISH STOPS

The effort described in Chapters 2 and 3 is directed towards the development of a general facility where controlled studies of the acoustic characteristics of selected consonants, consonant clusters, and vowels in a prescribed phonetic environment can be carried out. The type of acoustic data we have collected, and the process through which they were collected, have been described in detail in Chapter 2. The structure of the data-base and the facility were illustrated in Chapter 3.

The next two chapters deal with the results of a study of the English stops, /p,t,k,b,d,g/, using the collected data and the data-base facility. Various aspects of the temporal characteristics of stops in prestressed position will be investigated in this chapter, with Chapter 5 concentrating on the spectral characteristics.

While certain aspects of out results might have been reported in the past by others, it is felt that the careful control of the phonetic context, the improved measuring techniques, and the analysis of a large corpus of data of

the present study will provide us with statistically more reliable results, which in turn will make the interpretation of finding easier.

In this chapter we first define the different terminologies that we will use in describing the temporal characteristics of stops. Results on the durations of these stops, both in singleton and in clusters, as a function of their underlying features and the phonetic environment will then be presented.

4.1 Measurements and Techniques

The production of prestressed plosives in the ə´CVCə context is marked acoustically by several stages. The closure phase occurs after the initial schwa (or the initial /s/ in certain clusters) and is produced articulatorily by forming a complete constriction somewhere in the vocal tract. Acoustically this phase is characterized by the absence of sound energy in the resulting speech signal for unvoiced stops. For the prevoiced stops, however, there exists a small amount of acoustic energy at very low frequencies during the closure phase, due to the spontaneous vibration of the vocal cords and the subsequent radiation of sound energy through the walls of the vocal tract.

During the closure, air pressure continues to build up behind the constriction and is released finally by a sudden opening of the consriction. The release is usually accompanied by a burst of frication noise as air rushes through the narrow constriction, creating turbulence noise. The opening at the constriction continues to increase, until finally it becomes large enough such that turbulence noise can no longer be generated. Acoustically the release is marked usually by a sudden increase of sound energy at all frequencies, and the burst at the release usually has energy with frequency characteristics depending upon, among other things, the position of the constriction in the vocal tract.

For the voiced (unaspirated) stops in English, i.e. /b,d,g/, voicing may continue throughout the release, and after the burst of frication noise, normal voicing of the following vowel (or sonorant) commences. The release of the aspirated English stops, i.e. /p,t,k/, however, is quite different. Since there is no spontaneous voicing before the release, the vocal cords are not adjusted to a configuration or state appropriate for vibration. As the constriction opens, the vocal cords are brought closer to each other, and the tension of the cords is also adjusted so that voicing for the following vowel (or sonorant) can start. As the size of the glottis decreases, noise is generated at the glottis. This noise source in turn excites the vocal tract

and creates an aspirated sound.

The aspiration has several acoustic characteristics that distinguish it from the frication noise of the burst release. The noise source during the burst is directly in front of the constriction and, for a small enough constriction, excites only those resonances that are associated with the front cavity. The noise source for aspiration is situated close to the glottis, and therefore all except the lowest resonance of the vocal tract are being excited. Movements of the articulators and the resulting formant motions can, for the most part, be observed on the spectrogram during aspiration. Another characteristic of aspiration is that there is a certain degree of coupling between the supra- and sub-glottal cavities, depending on the opening at the glottis. Subglottal formants [Fant et al. 1972] frequently appear during aspiration.

Figure 4.1 illustrates the various durational measurements that are made on the stop consonants. The closure interval is measured from the end of the preceding schwa (or /s/ in certain clusters) to the release which in general is marked by a sharp increase in acoustic energy. Both the beginning and the end of the closure interval can usually be identified quite easily, except in the case of /p,b/ whose releases are often weak, partly due the the fact

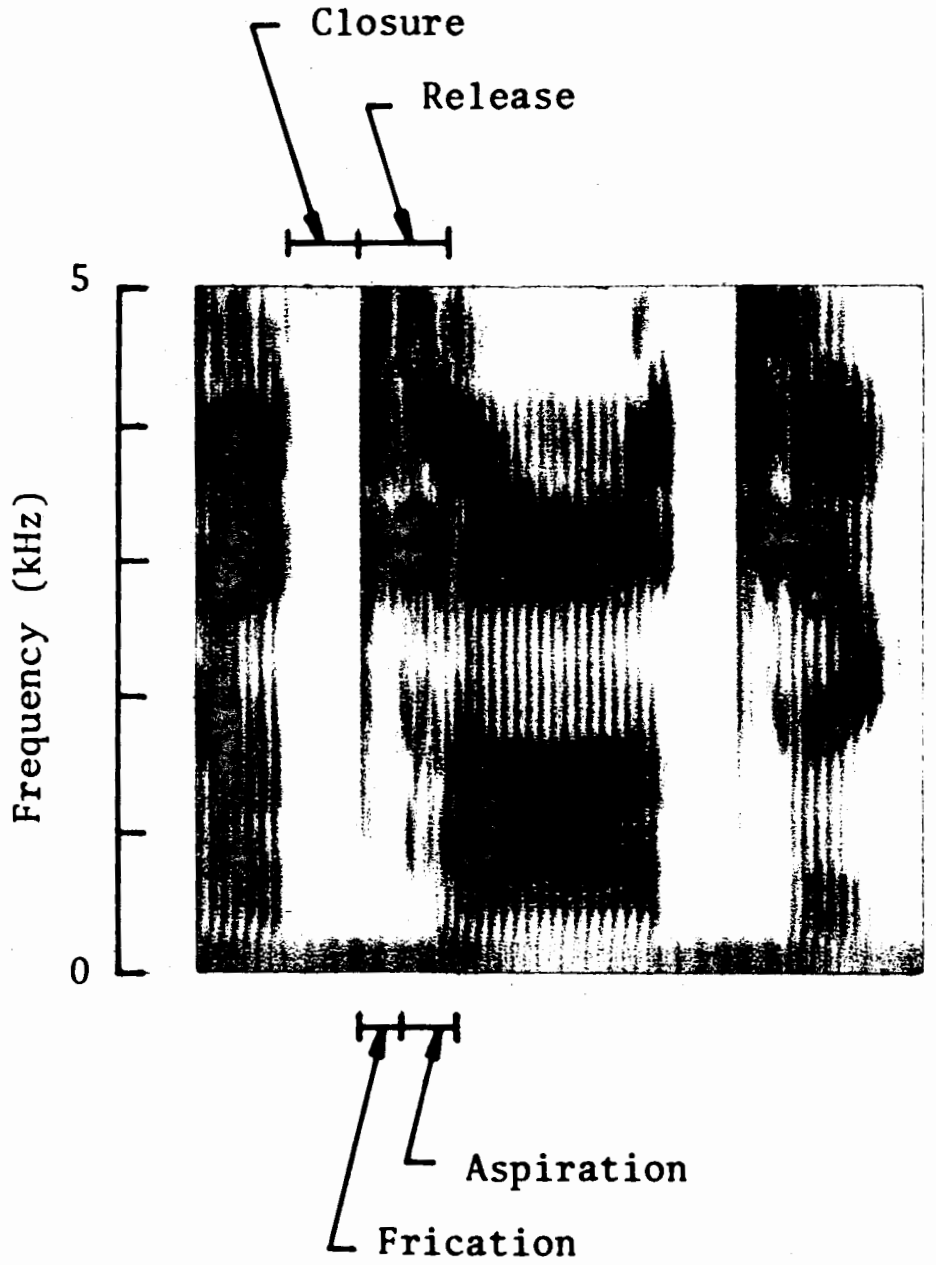


Figure 4.1
Illustration of the various durational measurements made in this study

that the constriction is formed at the lips with no excitable cavity in front of the noise source. The release interval is measured from the onset of burst release to where the time waveform first shows signs of periodicity following the release. For the voiceless aspirated stops, the release interval is further divided into frication and aspiration intervals.

As noted by other investigators [Lisker and Abramson 1964, Klatt 1975], voiced stops in such intervocalic positions often maintain voicing during the closure interval. Voicing may continue through the plosive release, or the vocal cords may cease to vibrate when the supraglottal pressure buildup becomes too great. Since prevoicing is not a phonemic determinant in English, all prevoicing is ignored in our study. For the remainder of this chapter, the term voice-onset time (VOT) will be used to designate the duration of the release interval for voiced and voiceless stops alike. For the unvoiced stops, voice-onset time is really a measure of the burst duration, and should be interpreted with caution.

As mentioned earlier, although the articulatory gesture and the positions of the noise sources are quite different for burst and aspiration, the acoustic manifestation of these two phases of the stop release are not always so

unambiguous. In fact, there is invariably a certain degree of overlap in the acoustic cues, making the effort to distinguish them extremely difficult. The only way to deal with such a problem is to lay down a set of specific rules and then adhere to them as consistently as possible. In making the boundary between frication and aspiration, we primarily use the knowledge that the burst usually has acoustic energy concentrated within a certain region; subglottal formants frequently appear in aspiration and second and higher formants can usually be traced from the following vowel (or sonorant) back into the aspiration of the prestressed stop.

4.2 Summary of Data

Table 4.I summarizes the nature of the data used in this study. It contains 1,728 utterances spoken by three male speakers. Fifteen vowels were used to form the syllable nuclei as shown in Table 4.I. Besides the singleton stops /p,t,k,b,d,g/, certain word initial clusters containing stops are also included. For the singleton stops, each CV combination contains 9 separate tokens from 3 speakers (5 tokens from speaker KNS, 3 from DHK, and 1 from JSP). For stops in clusters, each CV combination contains 3 separate tokens from only 2 speakers (2 from KNS and 1 from DHK). The post-stressed consonant is either /t/ or /d/.

Number of Speakers:	3	
Number of Utterances:	1,728	
Number of Vowels and Diphthongs:	15	[i,I,e,ɛ,æ,ɑ,ʌ,o,ɔ,u,U,ɜ̂,ɑy,ɔy,ɑw]
Number of Stops:	6	[p,t,k,b,d,g]
Number of Clusters:	21	[pl,pr,tr,tw,kl,kr,kw]
		[bl,br,dr,dw,gl,gr,gw]
		[sp,st,sk]
		[spl,spr,str,skr]

Table 4.I Summary of data

4.3 Results

Since variability among speakers and recording sessions has not been appreciable, the results presented in the remainder of this chapter have been pooled within and across speakers, unless otherwise specified.

4.3.1 Singleton Stops

Results on the voice-onset time for the voiced stops /b,d,g/ are shown in Figure 4.2. Averaging across place of articulation, the mean VOT was found to be 20.6 msec. This value is smaller than that reported for monosyllabic words spoken in isolation [Lisker and Abramson 1964] and somewhat larger than that reported for words excised from spoken sentences [Lisker and Abramson 1967, Klatt 1975]. Standard deviations of the measurements were 3 msec for /b,d/ and 5 msec for /g/. These variabilities are again consistent with those reported in the literature.

Results in Figure 4.2 also show that the VOT varies as a function of the place of articulation of the stop. Mean VOT is 13 msec for /b/, 19 msec for /d/, and 30 msec for /g/. Release time is consistently smaller than the mean for labials and larger than the mean for velars. The dependency of VOT on the place of articulation of the stop

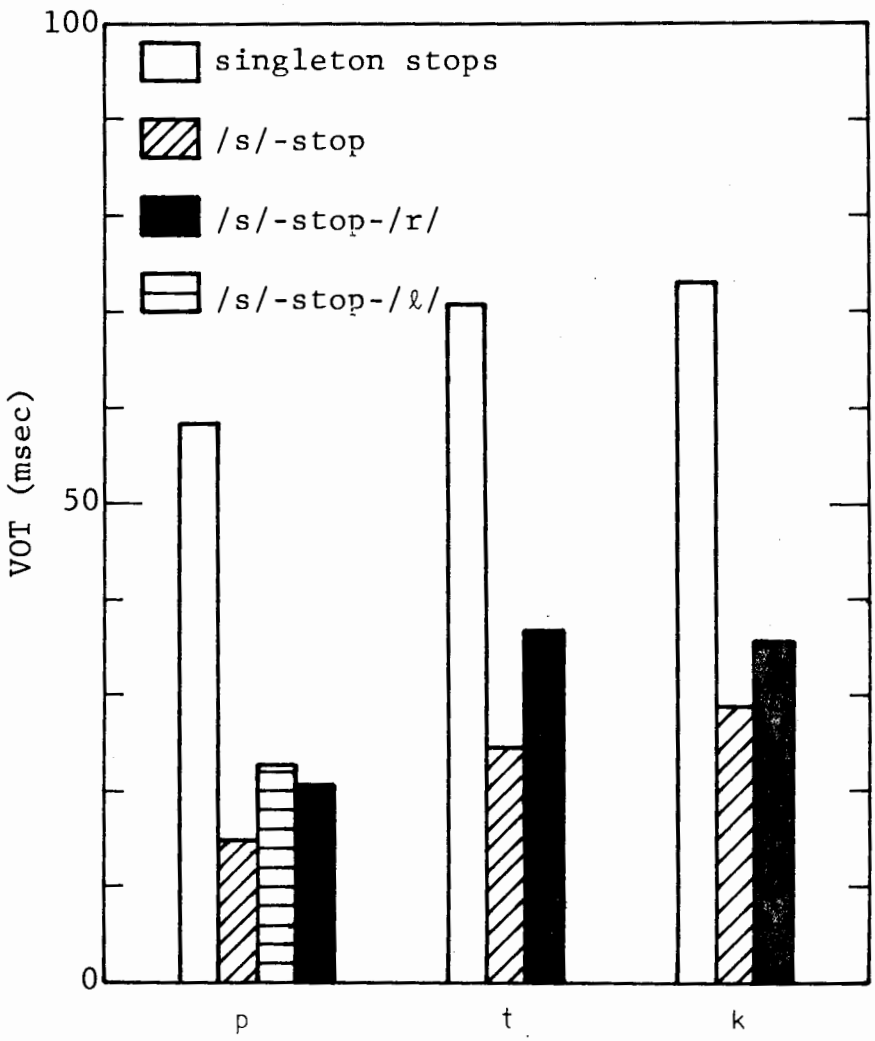


Figure 4.10 Average VOT for the voiceless stops in /s/-clusters

13% of singleton /g/'s have VOT greater than or equal to 40 msec. Our data reaffirm the finding by Klatt [Klatt 1975] that the perceptual decision on the voicing feature for English stops can not be made on the basis of VOT alone. Some derivative of the VOT, such as the presence or absence of rapid spectral changes at the onset of voicing suggested by Stevens and Klatt [Stevens and Klatt 1974], or some a priori knowledge of the phonetic environment and the place of articulation of the stops may play a role in the distinction between voiced and voiceless stops.

The increase of VOT from /b/ to /d/ to /g/ can be explained by the position and shape of the oral constriction as well as the articulators involved in forming the constriction [Klatt, 1975]. The constriction for /b/ is formed at the lips, which can move away quite rapidly following the release. As mentioned earlier, the labial release is usually weak in intensity, which also contributes to the appearance of a short burst.

The constriction for /g/ is formed by the tongue body, which is rather massive and can not move away from the palate too rapidly following the release. It has also been observed [Houde 1967, Perkell 1969] that the motion of the tongue body after the release is in such a way that a tapered, narrow opening is maintained for a longer period of

time. Therefore the constriction for /g/ opens slowly, allowing turbulence noise to be generated for a longer period of time.

The burst durations for /p,t,k/ have the same relationship as that of the voiced stops. This inherent difference can presumably be explained in the same way as outlined above.

That the voiceless aspirated stops all have identical total duration is somewhat surprising. This result cannot be attributed simply to the rhythmic pattern and constant inter-stress interval of the utterance. A possible explanation can be proposed based on the laryngeal behavior during the production of these stops and its timing relative to the release of the supraglottal constriction [Stevens 1975].

The time course of events in the production of voiceless aspirated stops is schematized in Figure 4.11. The supraglottal movement from the preceding schwa to the following vowel is shown as a simple closing and opening at the place of articulation of the stop. The vocal cords, however, will be stiffened and the glottis must be spread apart to prevent voicing and allow the generation of aspiration after oral release [Halle and Stevens 1971]. If we propose that the timing of the stop release is controlled

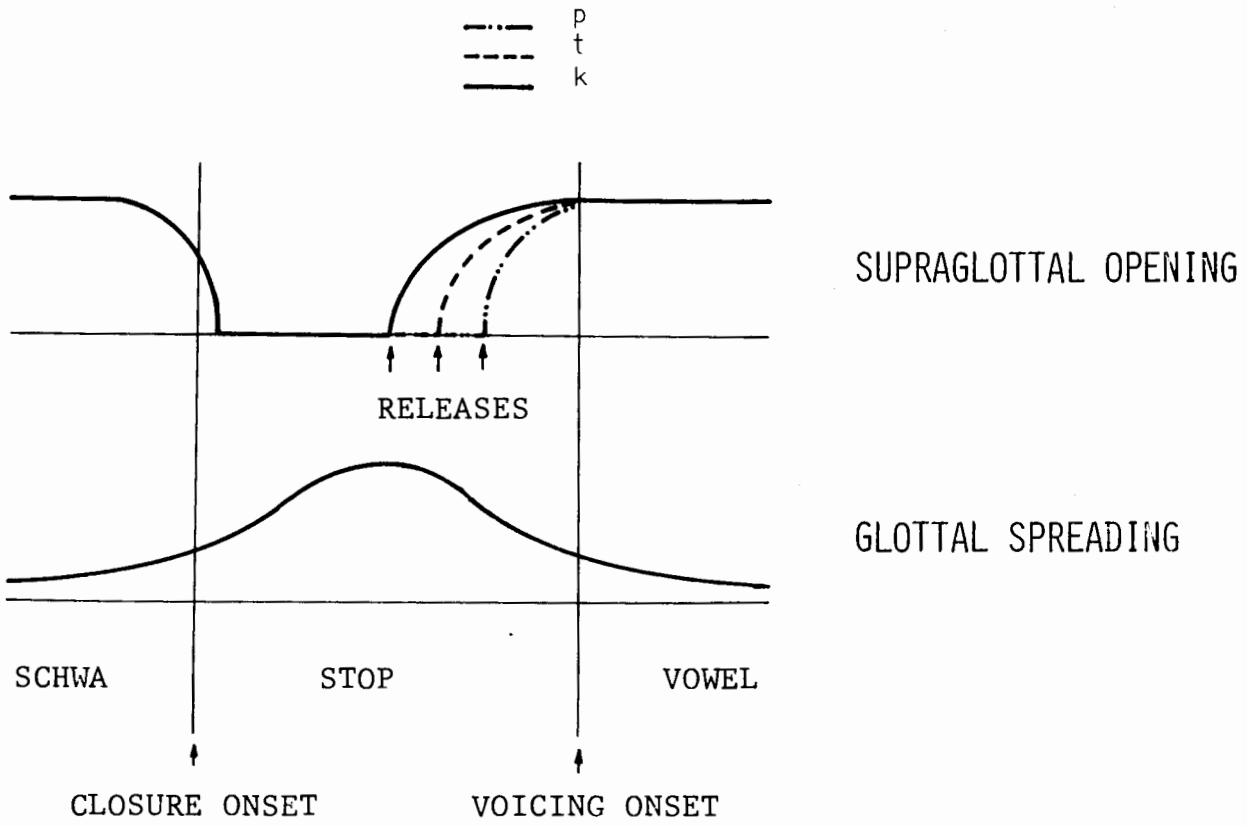


Figure 4.11 Schematized relationship between glottal spreading and supraglottal opening

by two more or less separate mechanisms, one for the supraglottal release and one for glottal abduction, then the phenomena observed above can be explained in the following way. The constant total duration of the stops could be a direct consequence of the fact that the amount of time required to spread and close the glottis is independent of the timing of the supraglottal movement. Given this constant time interval, the release will then have to be adjusted, according to the articulators involved, such that by the time of voicing onset, essentially all of the transition is completed. This lack of rapid spectral change at the onset of voicing is necessary for the proper perception of the voiceless stops, as suggested by Stevens and Klatt [Stevens and Klatt 1974].

The increase in VOT for stops in stop-sonorant clusters has been reported in the past in phonetic literature. Our data indicate that the following voiced segment is lengthened on the average by 26 msec for voiceless stops and by 13 msec for voiced stops. These data support the claim by Klatt [Klatt 1975] that only the initial part of the sonorant is devoiced. In fact, as suggested by Klatt, the greater increase for voiceless stops could be a phonological rule where the sonorant is lengthened following voiceless stops so that a substantial amount of the transition still remains after voicing onset.

The greater increase of VOT for dental-sonorant clusters is probably a direct consequence of the coarticulatory effect. /w/ in stop-sonorant clusters forms a secondary constriction due to lip rounding. A brief interval of silence can often be observed between the aspiration and voicing onset, due to a lack of acoustic output. Such a silent interval may account for the increase in VOT.

When dentals appear in dental-/r/ clusters, the larger increase in VOT may again be a consequence of the articulatory constraints. Since the stop and the sonorant both utilize the tongue tip and /r/ tends to curl the tongue blade towards the back, the release is such that the constriction is opened slowly. Klatt [Klatt 1975] also has observed a longer burst duration for the dental-/r/ clusters.

Many aspects of our results are in good agreement with the data reported recently by Klatt [Klatt 1975] on a substantially smaller data-base. We are, however, unable to find any dependency of VOT on the vowel context to support his claim that VOT is longer following high vowels. This discrepancy can probably be attributed to a difference in measurement technique. Most of the past studies [for example Lisker and Abramson 1964, 1967, Klatt 1975] utilize

spectrograms, and VOT is measured from the release to the onset of voicing, defined as the time where second and higher formants are visibly excited by voicing striations. In the present study, VOT is measured from the release to the time where the waveform first shows signs of periodicity. At the onset of voicing, the initial glottal pulses may be produced with the glottis still relatively spread. This "edge vibration" of the vocal cords will lend itself to a vibration pattern that is initially weak in high frequency energy. Therefore, depending on the amount of second and higher formant cut-back, the two measuring techniques can result in different VOT values. Measuring VOT from visible excitation of higher formants will perhaps result in amore perceptually based interpretation of data, whereas measuring VOT directly from the waveform will result in data which can possibly shed more light on the time course of articulatory events in the production of these stops.

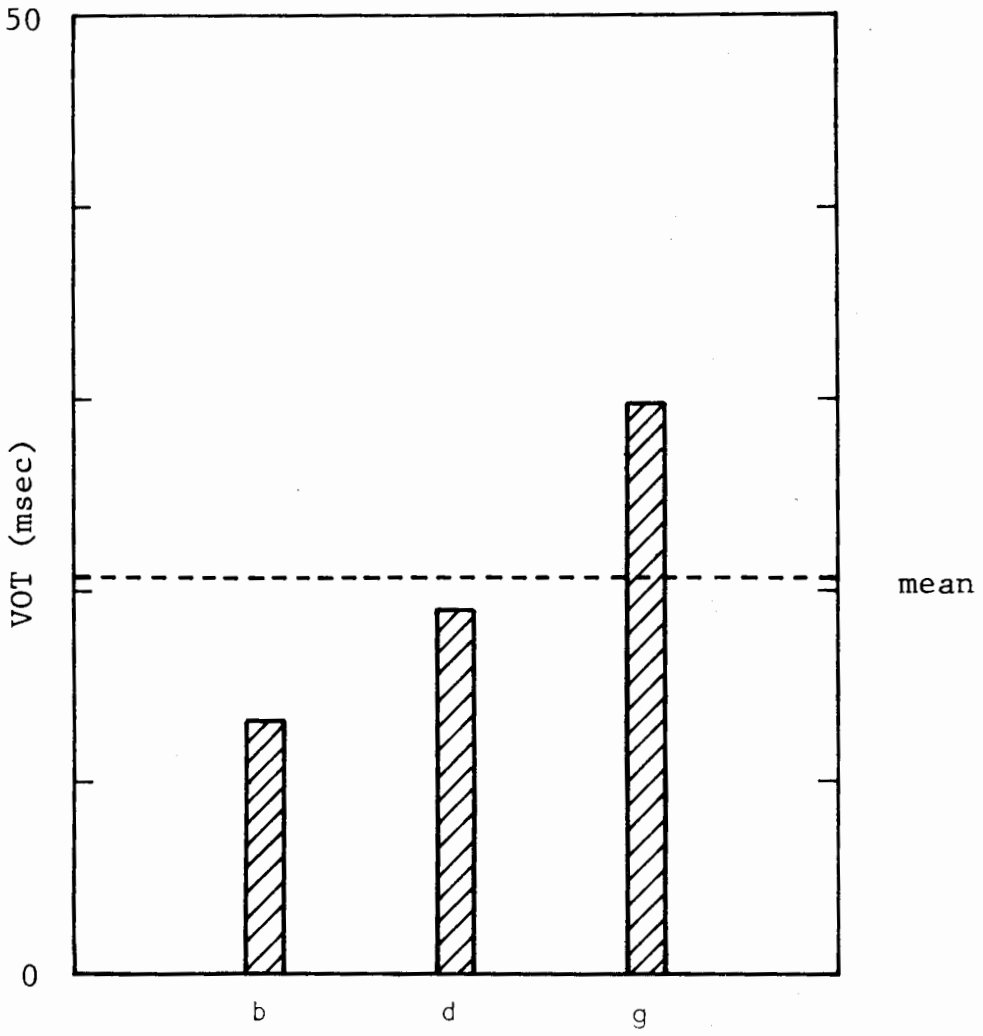


Figure 4.2 Average VOT for the singleton voiced stops

has been observed by other investigators in English [Klatt 1975] as well as across several languages [Lisker and Abramson 1967].

The voice-onset time for the voiced stops are shown in Figure 4.3 as a function of the vowel environment. As in the case of the combined results shown in the previous figure, the voice-onset time for velars is consistently longer than for dentals, which in turns is longer than for labials. No consistent dependency of VOT as a function of the following vowel was found across all three stops, although some individual variations appear to be sensitive to the vowel context. For example, VOT for /g/'s preceding back vowels are among the longest, whereas VOT for /b/'s preceding low vowels are the shortest.

Figure 4.4 summarizes results on the voice-onset time for the voiceless aspirated stops /p,t,k/. Averaging across place of articulation the mean VOT was found to be 67.5 msec. As in the case of the VOT for voiced stops, the VOT is longest for velars (7 msec longer than the mean value) and shortest for labials (9 msec shorter than the mean value). Standard deviations were 14 msec for /p/, 7 msec for /t/ and 11 msec for /k/. The larger standard deviation for /p/ can presumably be attributed to measurement error. Since the release of /p/, with virtually no excitable cavity

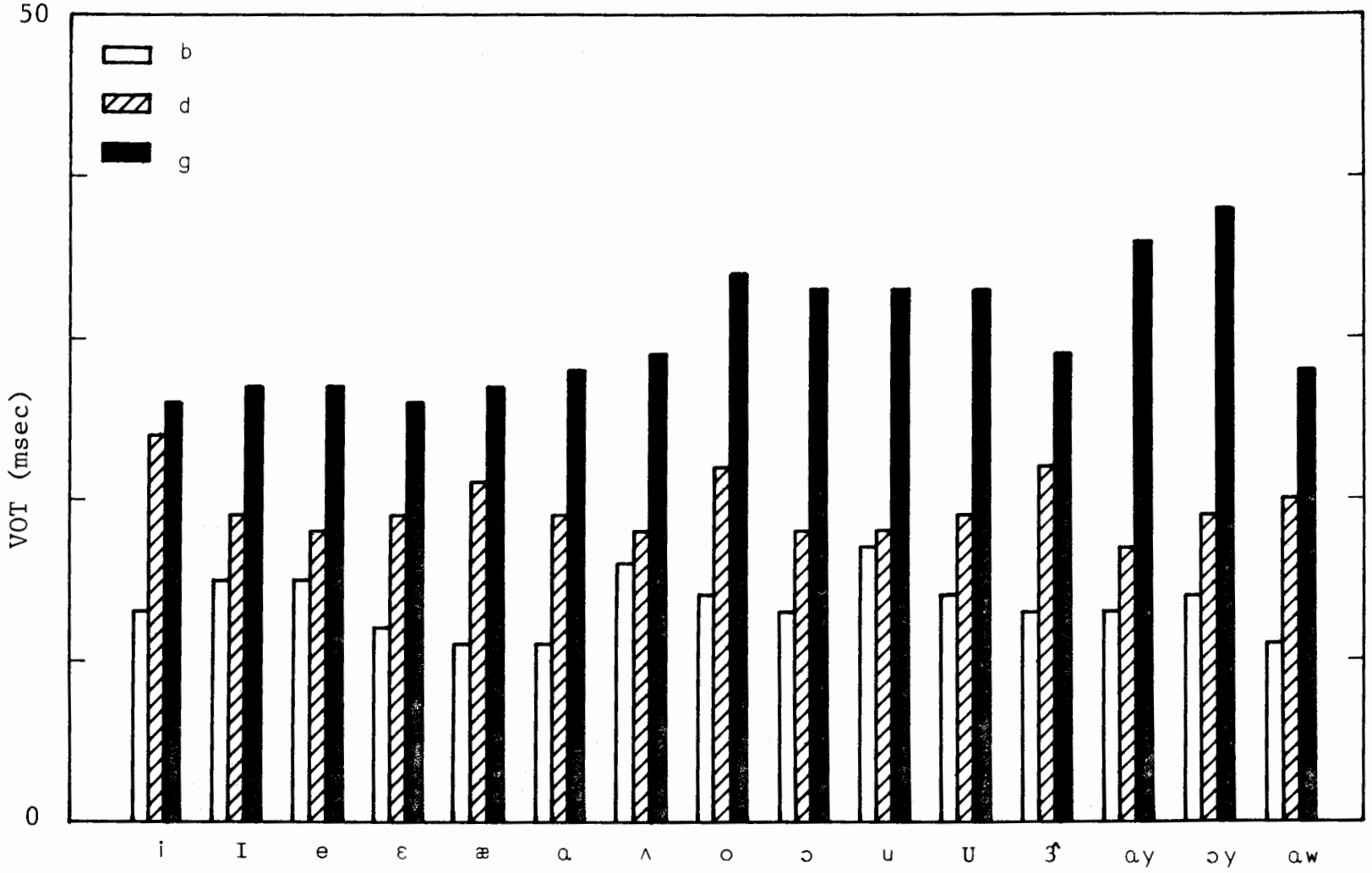


Figure 4.3 VOT for the singleton voiced stops as a function of vowel context

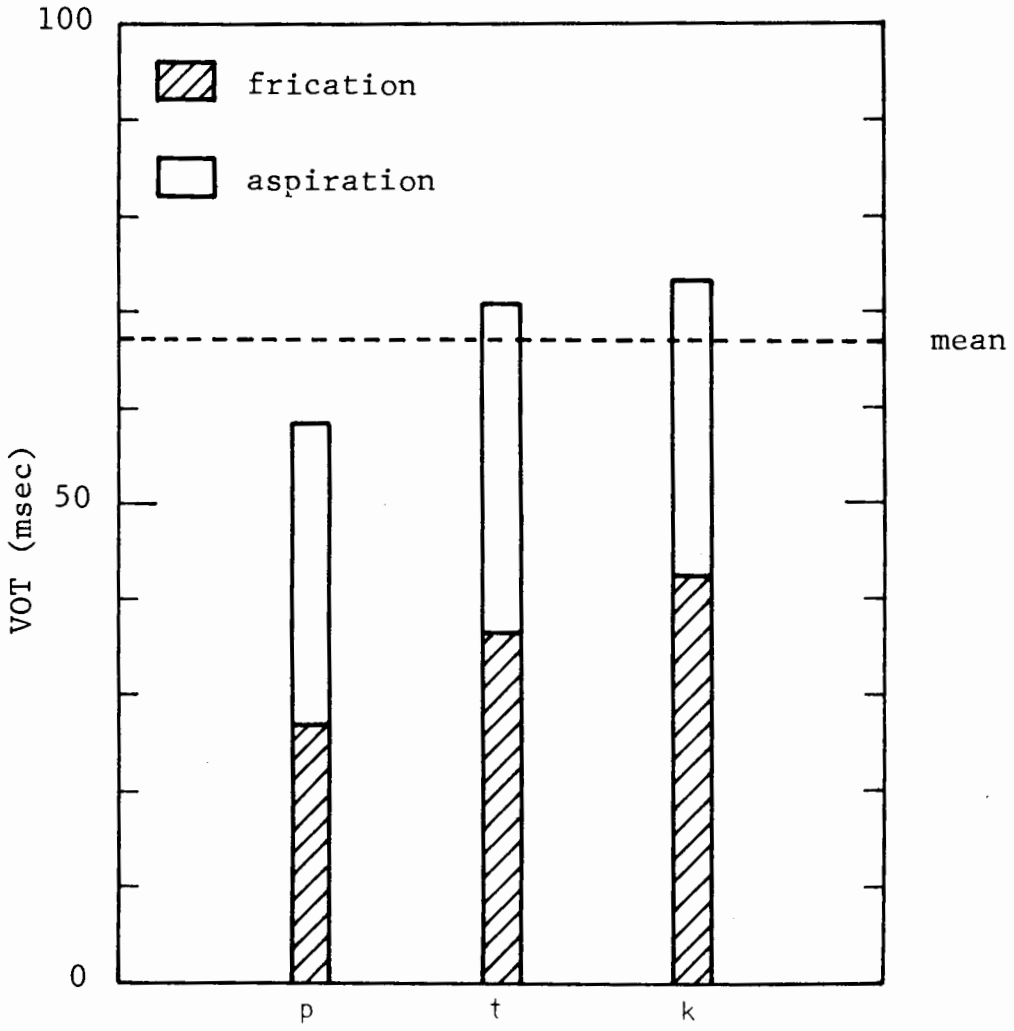


Figure 4.4 Average VOT for the singleton voiceless stops

in front of the noise source, is generally weak and difficult to locate.

Also plotted in Figure 4.4 are the durations of the frication and aspiration intervals. The burst duration for the voiceless aspirated stops is, on the average, 15 msec longer than the voiced stops. Aspiration durations for the three stops are approximately equal, thus leaving the burst durations for the three voiceless stops with the same relationship as described previously.

In measuring the durations of the closure interval for the voiceless aspirated stops, it was found that the inverse relationship holds. The closure interval is longest for /p/ and shortest for /k/. In fact, as shown in Figure 4.5, where closure as well as voicing onset durations are included, the total duration of these stops, measured from the beginning of silence to the first onset of voicing, was found to be practically identical. The total durations for /p/ and /t/ are 150 msec, whereas the total duration for /k/ was found to be 148 msec.

VOT for /p,t,k/ as a function of the vowel context is shown in Figure 4.6. Contrary to the findings by Klatt [Klatt 1975], there appears to be no systematic variation of VOT as a function of whether or not the following vowel is high. VOT for /t/ shows no appreciable variation from one

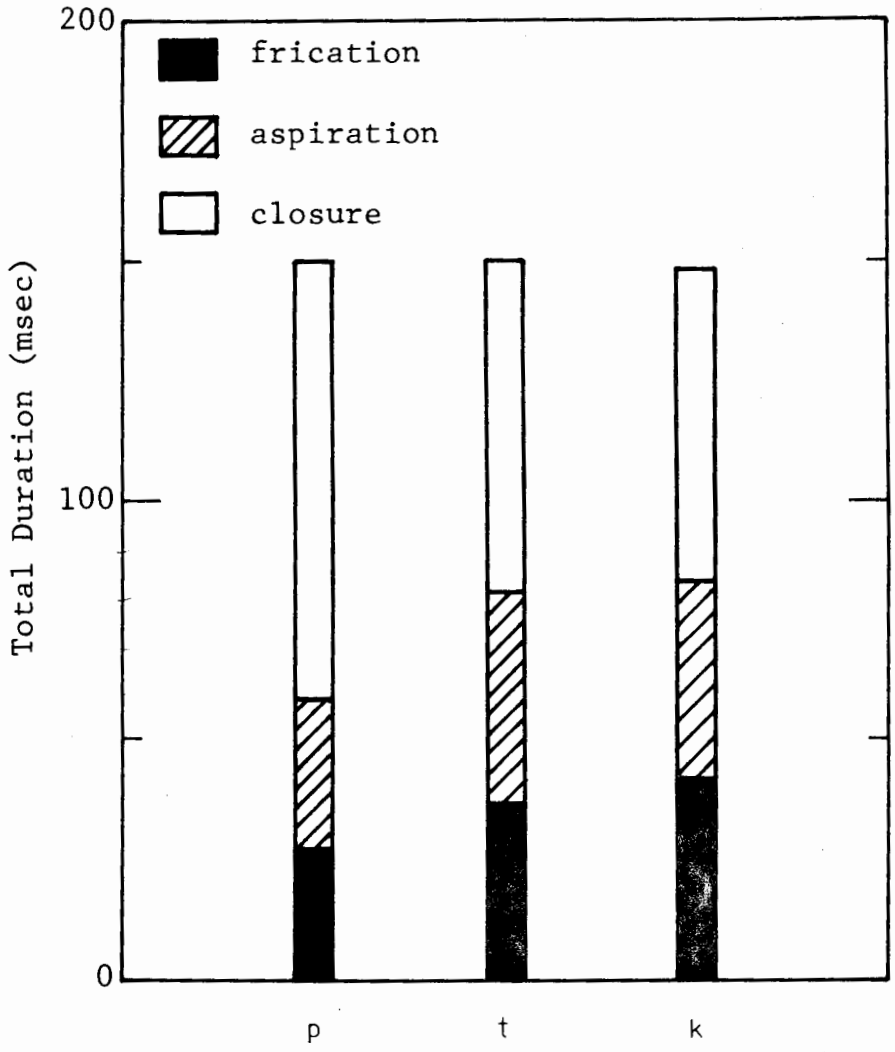


Figure 4.5 Average total duration for the singleton voiceless stops

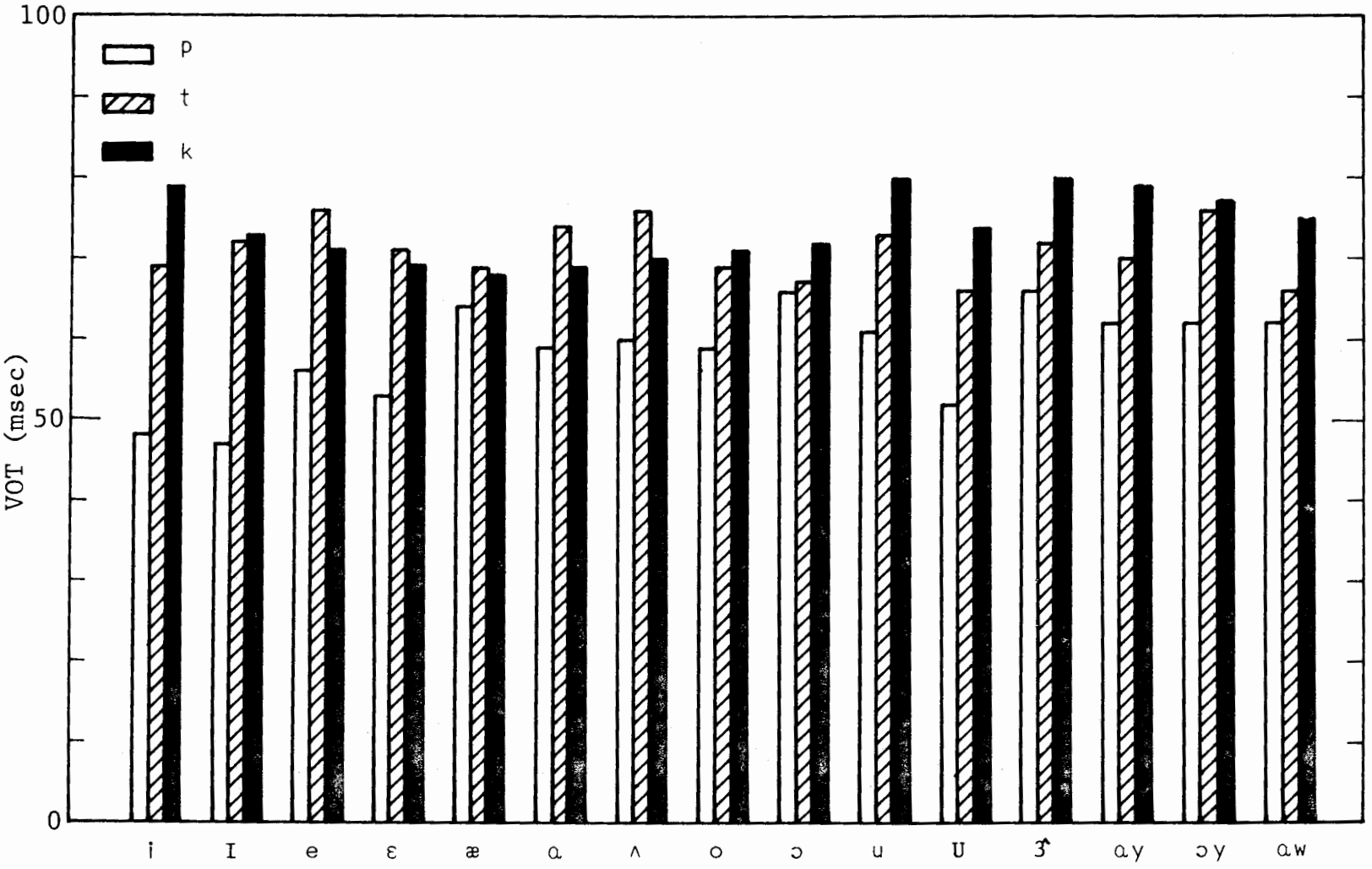


Figure 4.6 VOT for the singleton voiceless stops as a function of vowel context

vowel to another. VOT for /k/ is longer preceding /i,u/, both having the feature [+high]. However, VOT for /k/ is also long preceding /ɑy,ɔy/.

Table 4.II summarizes results on the dependency of VOT on the underlying features of the following vowel. For example, the first row in Table 4.II lists averaged VOT's for /p,t,k/ preceding high vowels and the second row lists averaged VOT's for /p,t,k/ preceding vowels which are not high. The directions of the feature-dependency from one stop to another have not been consistent for all the features tested. Neither did the averaged VOT's in the last column show any significant difference for all the features tested.

4.3.2 Stops in Clusters

A total of 21 word-initial consonant clusters containing stops have been investigated in this study as shown in Table 4.I. Fourteen of the 2-element clusters involve stop-sonorant combinations, with the remaining clusters involving /s/-stop or /s/-stop-sonorant combinations. It should be reiterated that in the case of stops preceding a sonorant, the durational measurements were made such that the end of the stop is marked by the first signs of periodicity in the time waveform following the

FEATURE OF THE FOLLOWING VOWEL	p	t	k	all
+HIGH	52.0	70.0	76.3	66.1
-HIGH	60.9	71.5	72.5	68.3
+ROUNDED	61.2	70.5	75.0	68.9
-ROUNDED	56.8	71.4	72.4	66.8
+ATR	56.0	71.7	75.0	67.6
-ATR	58.5	70.9	71.4	66.9
+BACK	61.0	70.9	74.3	68.7
-BACK	53.6	71.4	71.8	65.6
ALL VOWELS	58.4	70.8	73.2	67.5

Table 4.II Average VOT for the singleton voiceless stops as a function of the vowel features (each entry represent the average of all tokens peceding vowels with the given feature)

release.

Results on the voice-onset time for the voiced stops /b,d,g/ in clusters are shown in Figure 4.7. For the sake of comparison, values for the singleton stops were also plotted alongside. Averaging across place of articulation, the mean VOT was found to be 26.3 msec, a 5.7 msec (or 28%) increase from stops in isolation. The mean voice-onset time, as a function of the place of articulation of the stops, varies in the same fashion as stops in isolation. Figure 4.7 also shows that the increase in VOT from singleton to clusters is the greatest, 7 msec or 37%, for dentals.

Figure 4.8 summarizes results on the voice-onset time for the voiceless stops in clusters. Averaging across place of articulation, the mean VOT was found to be 85.6 msec, an 18.1 msec (or 27%), increase over stops in isolation. As in the case of voiced stops in clusters, the increase in VOT is the greatest, 28.4 msec or 40%, for dentals. In fact, the results in Figure 4.8 indicate that dentals in clusters have a mean VOT greater than that of velars.

The results in Figures 4.7-8 can be broken down further with respect to the following sonorant, as is shown in Figure 4.9. The mean VOT for stops in clusters, averaged over all vowel context, varies from 15 msec. for /bl/ to 100

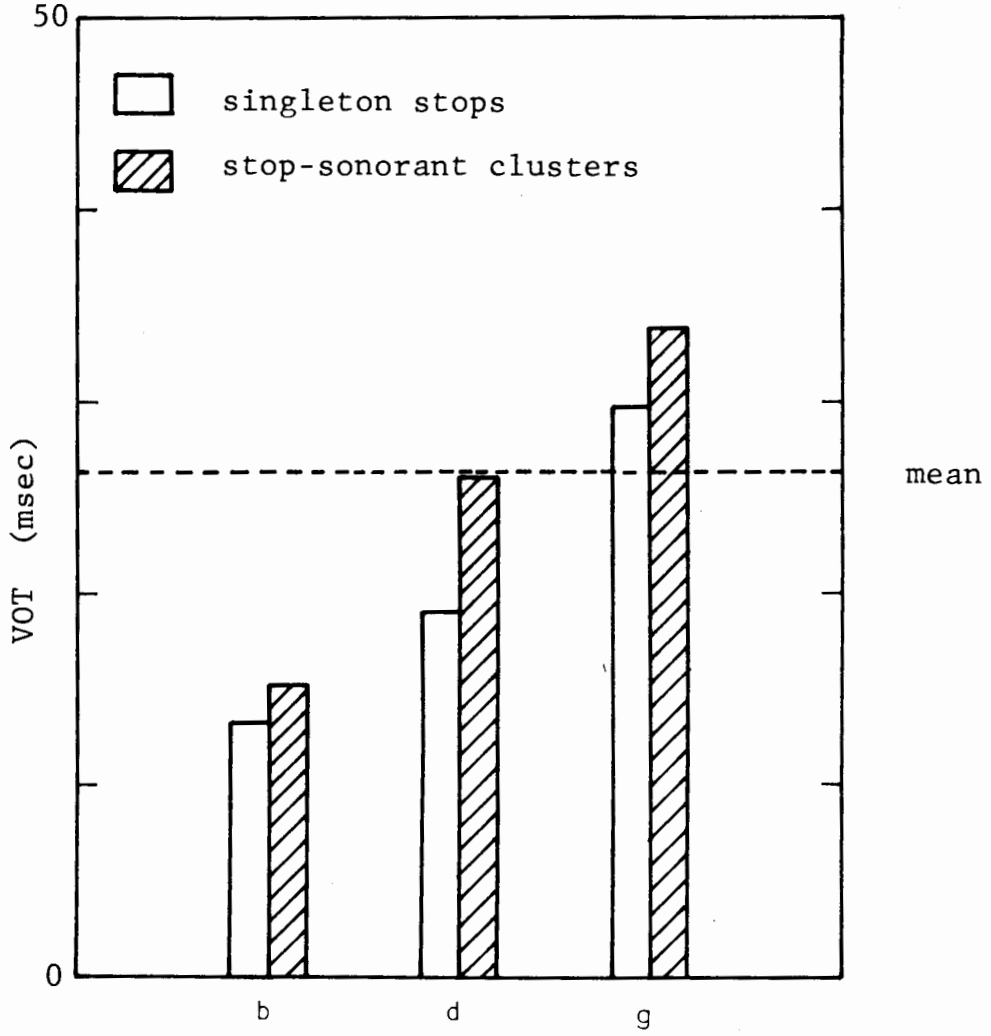


Figure 4.7 Average VOT for the voiced stops in stop-sonorant clusters

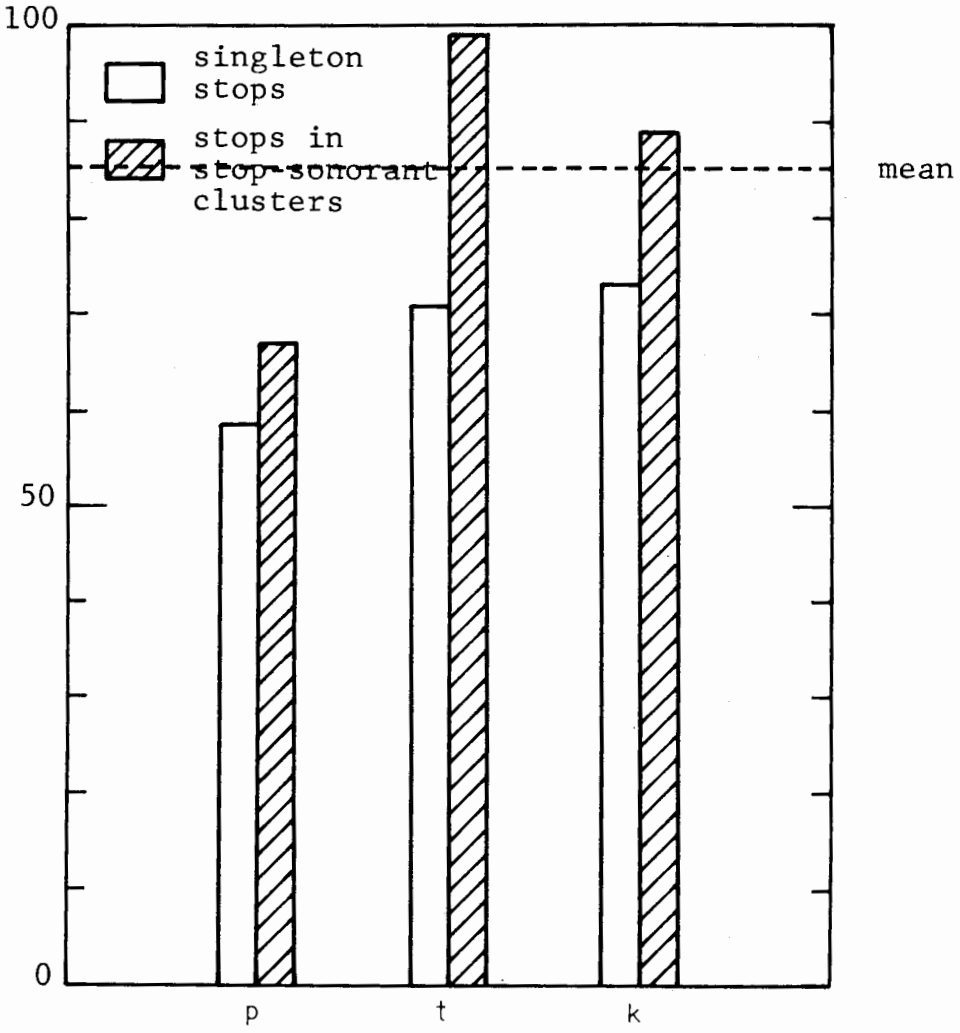


Figure 4.8 Average VOT for the voiceless stops in stop-sonorant clusters

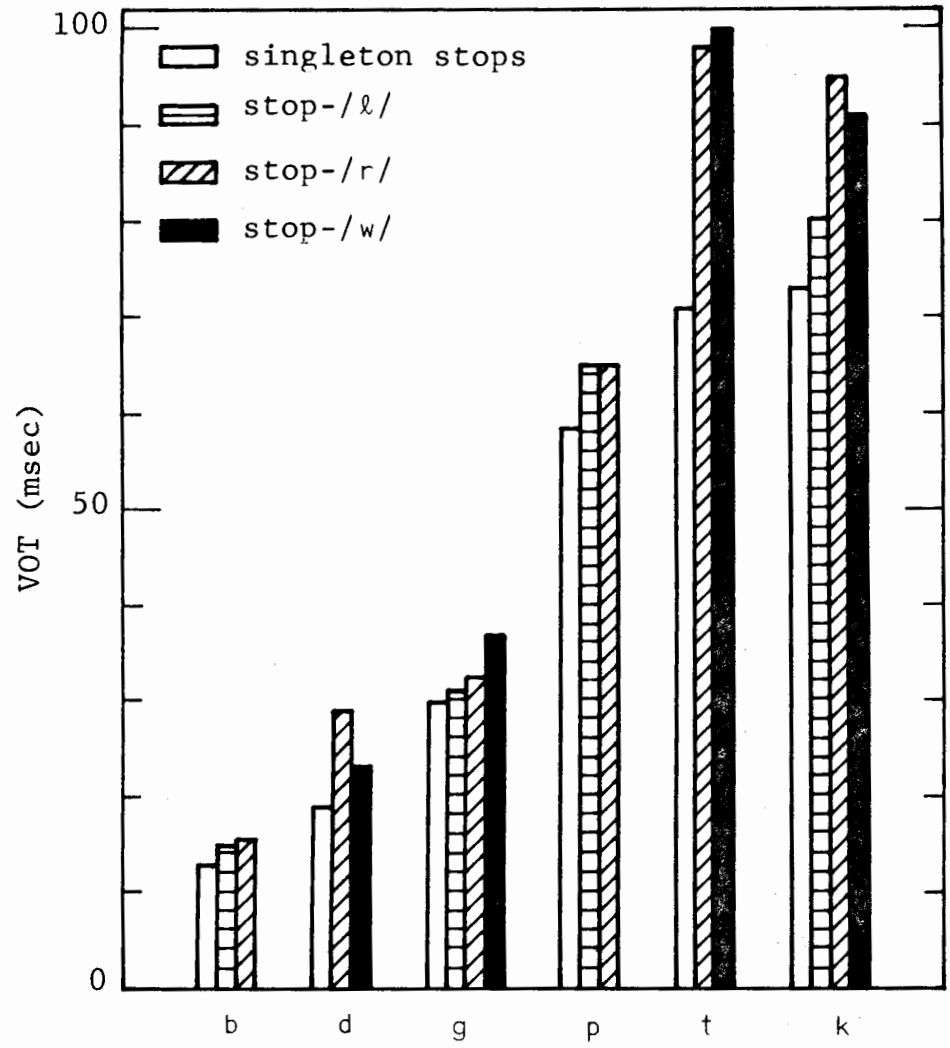


Figure 4.9 Average VOT for the stops in clusters as a function of the following sonorant

msec for /tw/. The increase in VOT over singleton stops ranges from 1.5 msec for /gl/ to 29.3 msec for /tw/.

Results of VOT for stops in s-clusters are shown in Figure 4.10. The voice-onset time decreases sharply when a voiceless stop appears in s-clusters. Averaging across place of articulation, the mean VOT for stops in s-clusters preceding a vowel was found to be 22.7 msec. This value is just slightly higher than the mean VOT found for voiced stops in isolation. When the stops appear in s-cluster preceding a sonorant, the mean VOT was found to be 30 msec.

4.4 Discussion

The results reported in the previous section show that even in a controlled environment, i.e. the prestressed position, the voice-onset time data span a wide range of variations. The averaged VOT for voiced stops ranges from 13 msec for /b/ to 38 msec for /gw/. The averaged VOT for voiceless unaspirated stops ranges from 15 msec for /sp/ to 36.7 msec for /str/ and the averaged VOT for voiceless aspirated stops ranges from 58.4 msec for /p/ to 100 msec for /tw/. Even with identical phonetic environment, a certain degree of overlap in VOT can be found between voiced and voiceless stops. For example, 22 or 16% of singleton /p/'s have VOT less than or equal to 40 msec, whereas 18 or

CHAPTER 5

SPECTRAL CHARACTERISTICS OF ENGLISH STOPS

This chapter reports our findings on the spectral characteristics of English stops, in prestressed position, using the collected data and the developed facitility as described earlier. The data set is the same as the one used in the previous chapter.

Whereas the temporal characteristics of stops, such as closure, burst, and voicing onset durations, have been studied extensively both through perceptual experiments using synthetic speech and analysis of real speech data, the spectral characteristics of stops have been studied primarily in the context of experiments looking for perceptually important acoustic cues. These perceptual experiments [for example, Cooper et al. 1952, Stevens and Klatt 1974] use almost exclusively synthetic speech material. While it is true that the synthesis features are determined through our acoustic phonetic knowledge, synthetic speech might sometimes be an overly-simplified, or even erroneous, representation of the real data. The problem lies both in the limited scope of the data examined and the methods by which they were analyzed. Most of the

past observations and measurements were made on spectrograms, which have a very limited dynamic range to represent spectral intensity. Locating the exact frequency of a stop burst from the spectrogram is extremely difficult, and deducing the spectral shape at the release is next to impossible. In addition, the AGC circuitry in a spectrograph machine has a tendency to increase the burst intensity following the silence interval, thereby making the burst appear to be more intense than it really is. Quantitative data collected from real speech data, such as these reported here, are undoubtedly necessary to substantiate and corroborate the findings of the perceptual experiments.

This chapter concentrates on two aspects of the stop release, namely the peak intensity of the stop burst and the spectral characteristics of the stop 10 or 15 msec following the release. We shall begin by first defining the measurements that were made, and then we shall present the results and discuss their implications.

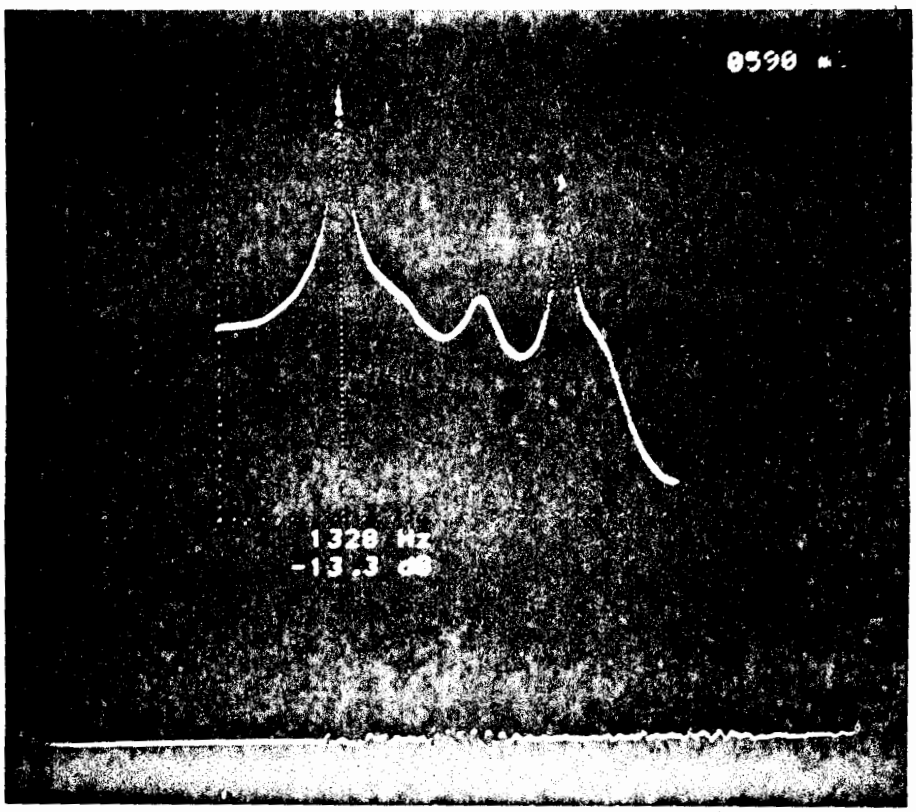
5.1 Measurements and Techniques

For the remainder of this chapter, we shall use the term burst frequency to designate the frequency of maximum spectral amplitude in the spectrum computed from the first

10 to 15 msec of the waveform following the release. For display and measurement purposes, the computed spectrum is post-emphasized with a frequency response similar to that of a conventional spectrograph.

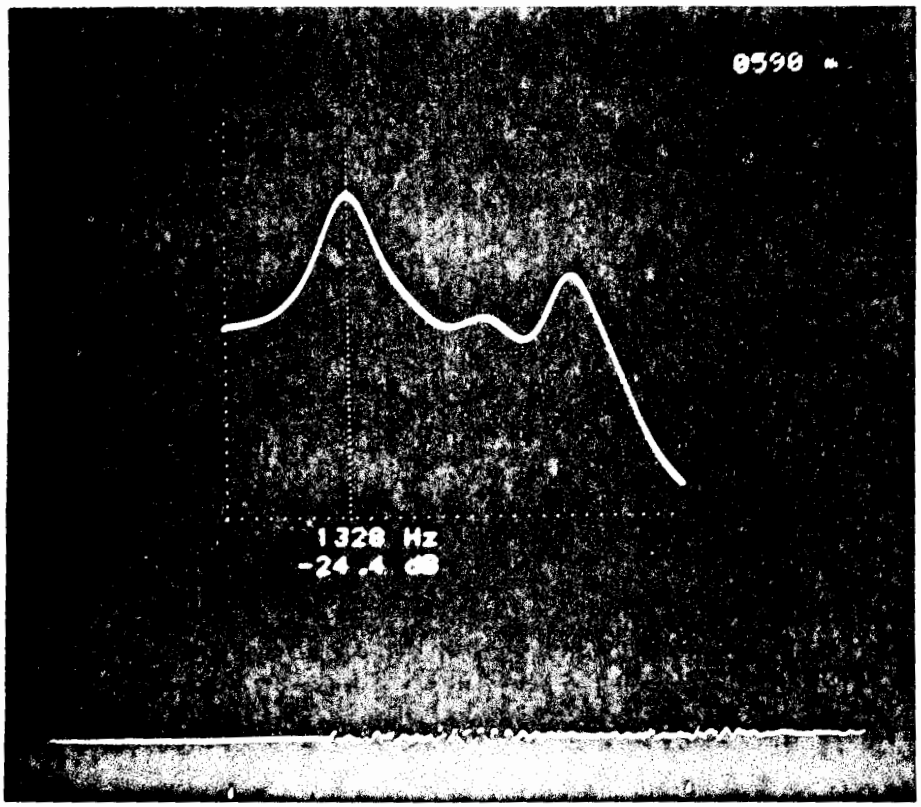
Burst spectra computed from linear prediction often show sharp spectral peaks. The relative sharpness of these spectral peaks is a direct consequence of using linear prediction as a spectral smoothing technique [Makhoul and Wolf 1972]. In any case, we are interested in the frequency locations of major energy concentrations, not in the effect of the local spectral maxima. Therefore, the burst spectra computed from linear predicion are further smoothed, using a 3-sample digital filter, to minimize the effect of the high-Q poles. This final smoothing enables us to locate the burst frequency reliably through a semi-automatic process where the program automatically locates and displays the burst frequency and the user can interactively modify the result before filing. Our experience has been that the occasions where user intervention is needed have been rare, on the order of 1% of our data set.

Figures 5.1-2 give examples of the procedure outlined above. In Figure 5.1 the burst frequency corresponds exactly to the peak in the original spectrum. This correspondence is invariably the case for velars. For



(a)

Figure 5.1 Spectra of a /k/ burst (a) before, and (b) after further smoothing (pointer indicates burst frequency)



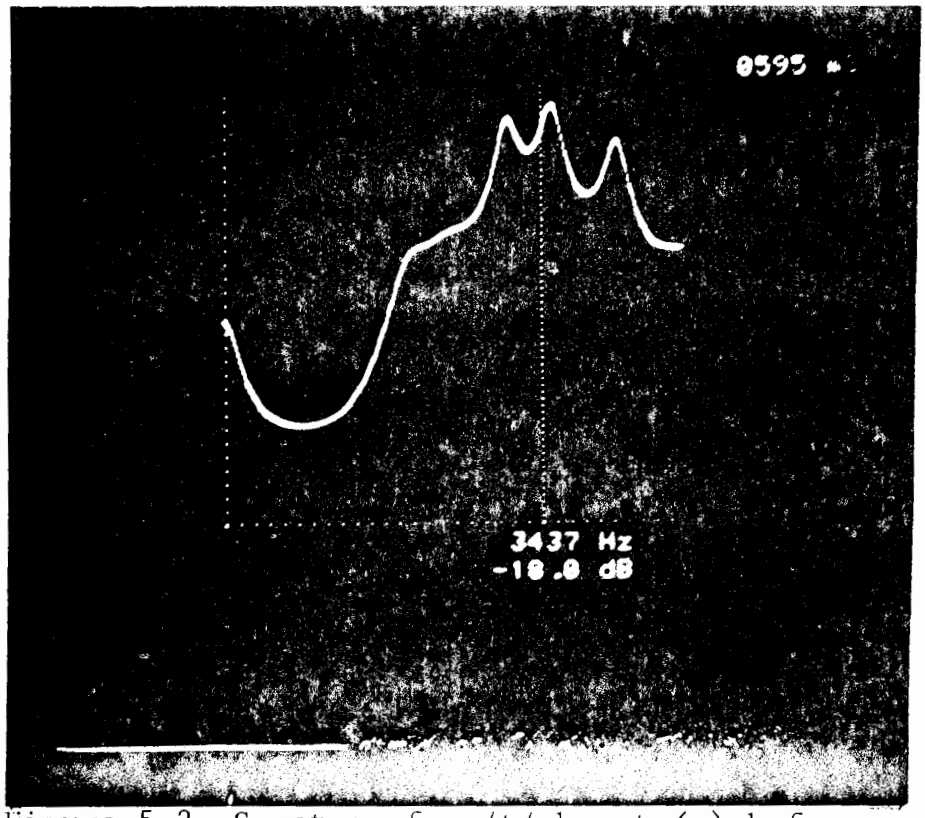
(b)

dentals, as shown in Figure 5.2, the burst frequency, as defined here, often does not coincide with spectral peaks in the original spectrum.

Once the burst frequency is located, the amplitude of the burst is also determined, as shown in Figures 5.1-2. For the sake of comparison, measurement is also made on the amplitude of the spectral peak of the following vowel in the same frequency region. Spectral amplitude for the vowel is made on the smoothed spectrum computed at the mid-point of the vowel. For the remainder of this chapter, the term burst amplitude is used in a relative sense, defined as the difference between the burst amplitude and the vowel amplitude. A positive value would indicate that the burst is higher in amplitude than the following vowel.

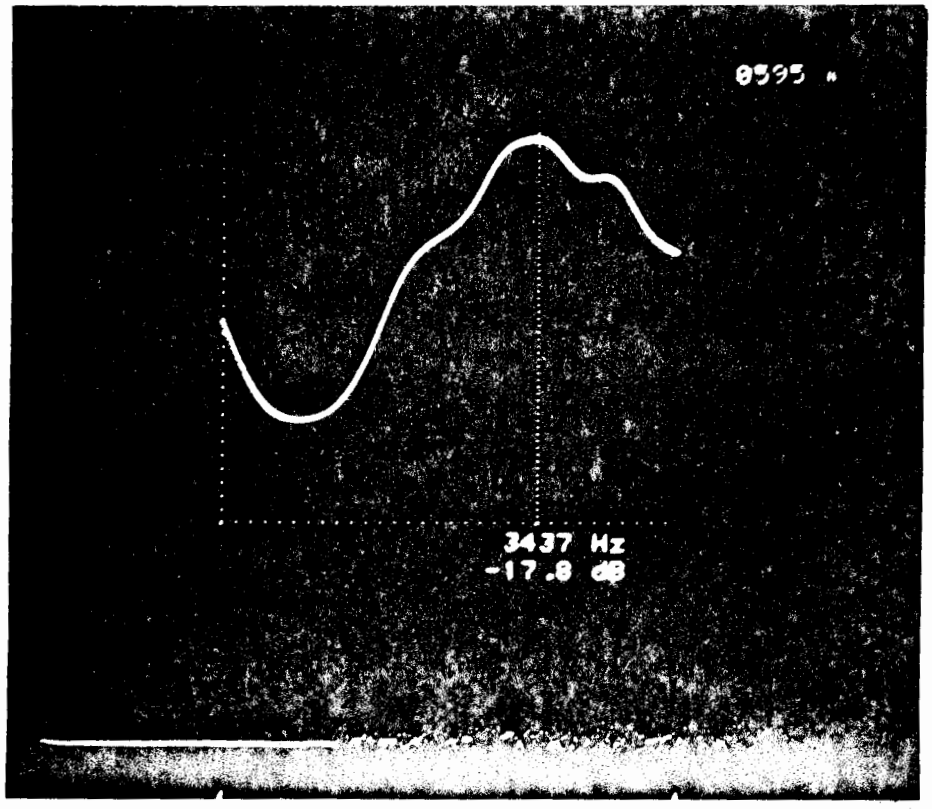
Figure 5.3 shows the burst spectrum of a /d/ in the sylllable /dɑt/. The spectrum at the mid-point of the vowel is superimposed. For this example, the burst amplitude is on the order of -2 dB.

The overall RMS amplitude of the burst is also measured in this study. The RMS amplitude is computed on the first 10 to 15 msec of the stop release, and is normalized by the maximum RMS amplitude for each utterance. Although the maximum RMS value is always within the stressed vowel, the location of the maximum rarely occurs at the mid-point of



(a)

Figure 5.2 Spectra of a /t/ burst (a) before, and (b) after further smoothing (pointer indicates burst frequency)



(b)

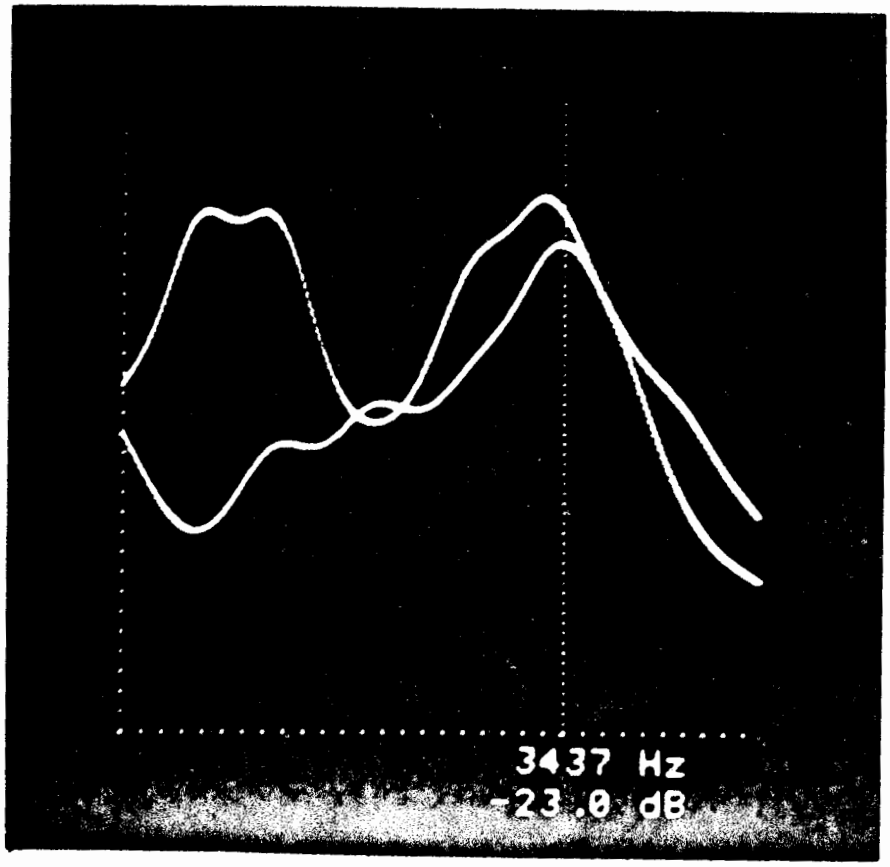


Figure 5.3 Composite display of the spectra of a /d/ burst and the following vowel /ɑ/ (both spectra have been smoothed: the difference in amplitude is about 2 dB)

the vowel.

5.2 Results

Results presented in this chapter are pooled across speakers and recording sessions. Characteristics attributable to inter-speaker variations will be discussed whenever appropriate. It should also be mentioned that the results on stops in clusters are obtained from a smaller sample than those on singleton stops.

5.2.1 Singleton Stops

Before presenting any quantitative data , it is appropriate to examine first some qualitative characteristics of the stop releases. Figure 5.4 shows the general spectral characteristics at the release of /t/ and /k/ for two different speakers. The pictures in Figure 5.4 are multiple exposure shots obtained from some forty stops. Since the burst intensity varies from one utterance to another, all spectra in Figure 5.4 have been normalized by their respective geometric means. This normalization procedure, which tends to reduce the amplitude differences among spectra and produce a cleaner clustering, is used strictly for display purposes.

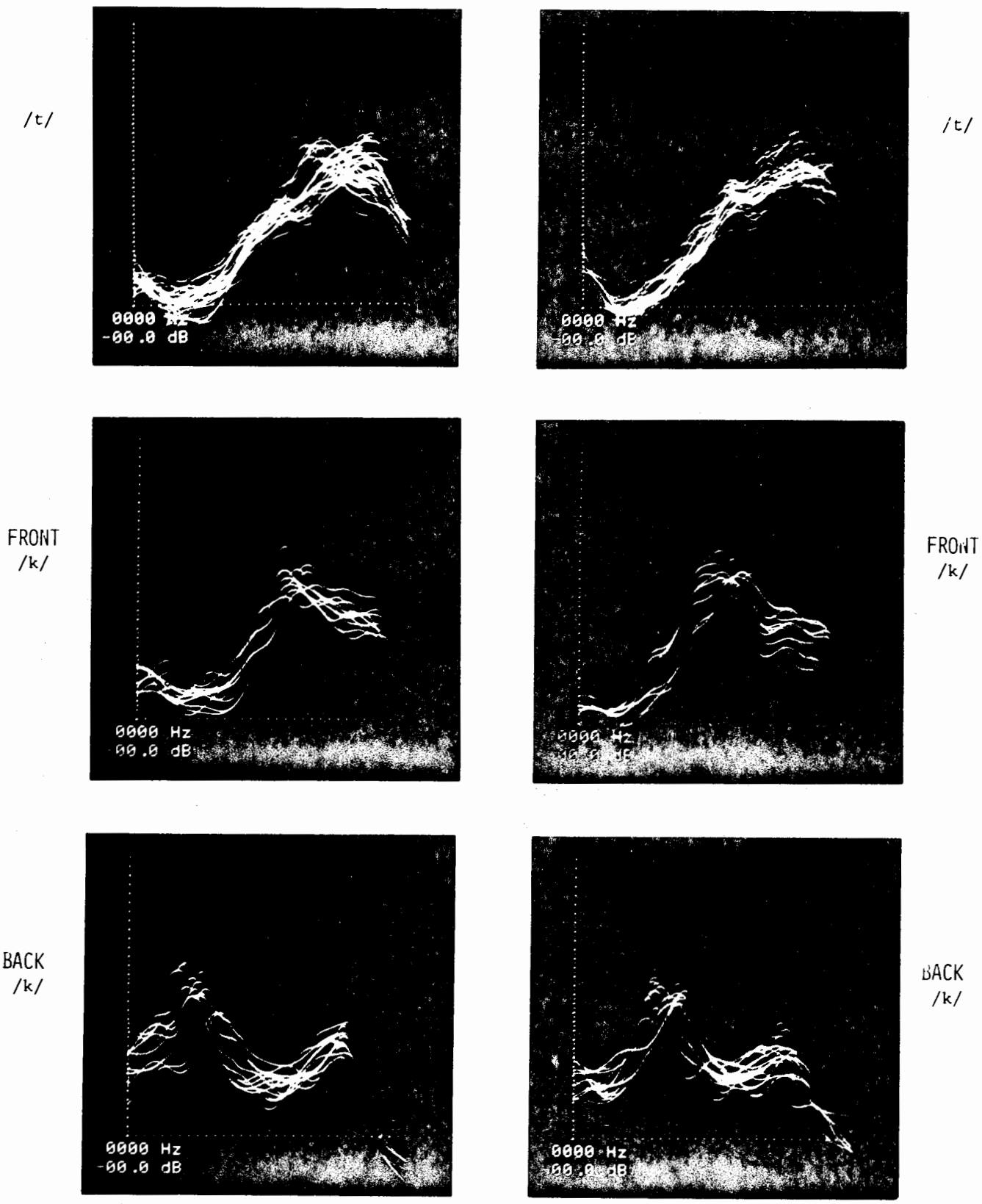


Figure 5.4 Composite display of the bursts of /t/ and /k/

Although the frequency locations of spectral concentrations vary from one speaker to another, the bursts of /t/ and /k/ exhibit some general spectral characteristics independent of speaker and, to some extent, context. The /t/ bursts are, in general, rather broad-band and occupy primarily the high-frequency region of the spectrum, say, above 2,000 Hz. The /k/ bursts show remarkably dissimilar shapes depending on the nature of the following vowel. The burst of a /k/ preceding a front vowel has a more compact spectral shape than the /t/ burst, and the location of the burst is in the mid-frequency region. The burst of a /k/ preceding a back vowel has a predominant sharp peak in the low-frequency region. In addition, a secondary spectral peak at hgh frequency can always be observed for the back /k/'s. This secondary peak, presumably associated with the 3/4-wavelength resonance of the front cavity [Stevens 1973], has generally been ignored in describing the /k/ burst. Inclusion of this additional peak has been shown to improve the quality of synthetic /k/'s [Klatt, personal communication].

Closer examination of Figure 5.4 also reveals that the burst location for back /k/ is at one of two frequency regions, depending on the vowel context. Similar results can be observed for the /t/ bursts.

The average RMS amplitude of the burst for /p,t,k,b,d,g/ are shown in Table 5.I. The data in Table 5.I indicate that there is no significant difference in RMS amplitude between voiced and voiceless stops. Dentals and velars have about the same RMS amplitude, whereas the labial bursts are weaker, by about 12 dB.

In measuring the burst frequency for the labials, it was found that there is a wide range of variation in the values found. Since the spectra of /p,b/ show no distinct burst frequency and the RMS amplitudes of these stops are weak, we have decided not to present results on the burst spectrum for labials.

The distribution of burst frequency for /t/ is shown in Figure 5.5. Averaging across all vowels, the mean burst frequency was found to be 3,660 Hz. As observed earlier, the distribution is skewed towards the high-frequency region, with essentially all of the measured values above 3,000 Hz. The distribution appears to be bi-modal. Closer examination of the data in fact shows that the lower peak in the distribution is directly related to the underlying features of the following vowel. Excluding the rounded and retroflexed vowels, the mean burst frequency for /t/ was found to be 3,900 Hz, whereas the mean burst frequency for /t/ preceding rounded or retroflexed vowels was found to be

STOP	OVERALL RMS AMPLITUDE (in dB)
/p/	-27.6
/t/	-16.6
/k/	-17.2
/b/	-28.0
/d/	-15.8
/g/	-16.6

Table 5.I Overall RMS amplitude of the burst for the singleton stops

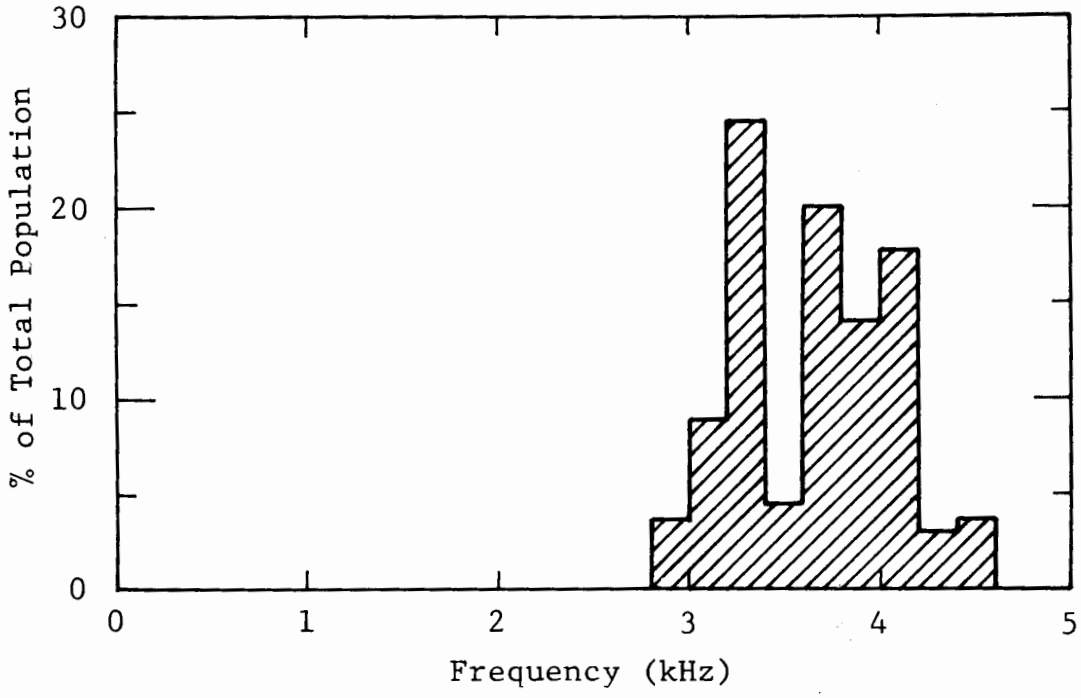


Figure 5.5 Distribution of the burst frequency for /t/

approximately 3,300 Hz. These two mean values correspond well with the two peaks in the distribution.

The averaged burst frequencies for /t/ are plotted as a function of the following vowel, and are shown in Figure 5.6. It can be seen that burst frequencies for /t/ preceding rounded or retroflexed vowels are consistently less than the overall mean value, whereas the burst frequencies preceding all other vowels are always greater than the mean value.

The distribution of burst frequency for /d/ is shown in Figure 5.7. Compared to the distribution of /t/ bursts, the burst frequency distribution for /d/ appears to have shifted down in frequency. Averaging across vowel context, the mean value was found to be about 3,300 Hz. As is the case for /t/, the distribution appears to be bi-modal in nature. The mean burst frequency for /d/ preceding rounded or retroflexed vowels was found to be 2,950 Hz, while the mean value preceding all other vowels was found to be 3,530 Hz.

In Figure 5.8 the /d/ burst frequency is plotted as a function of the following vowel. Again, it can be seen that the averaged burst frequencies for /d/ preceding rounded and retroflexed vowels are consistently less than the overall mean value, whereas the averaged burst frequencies preceding all other vowels are consistently greater than the mean.

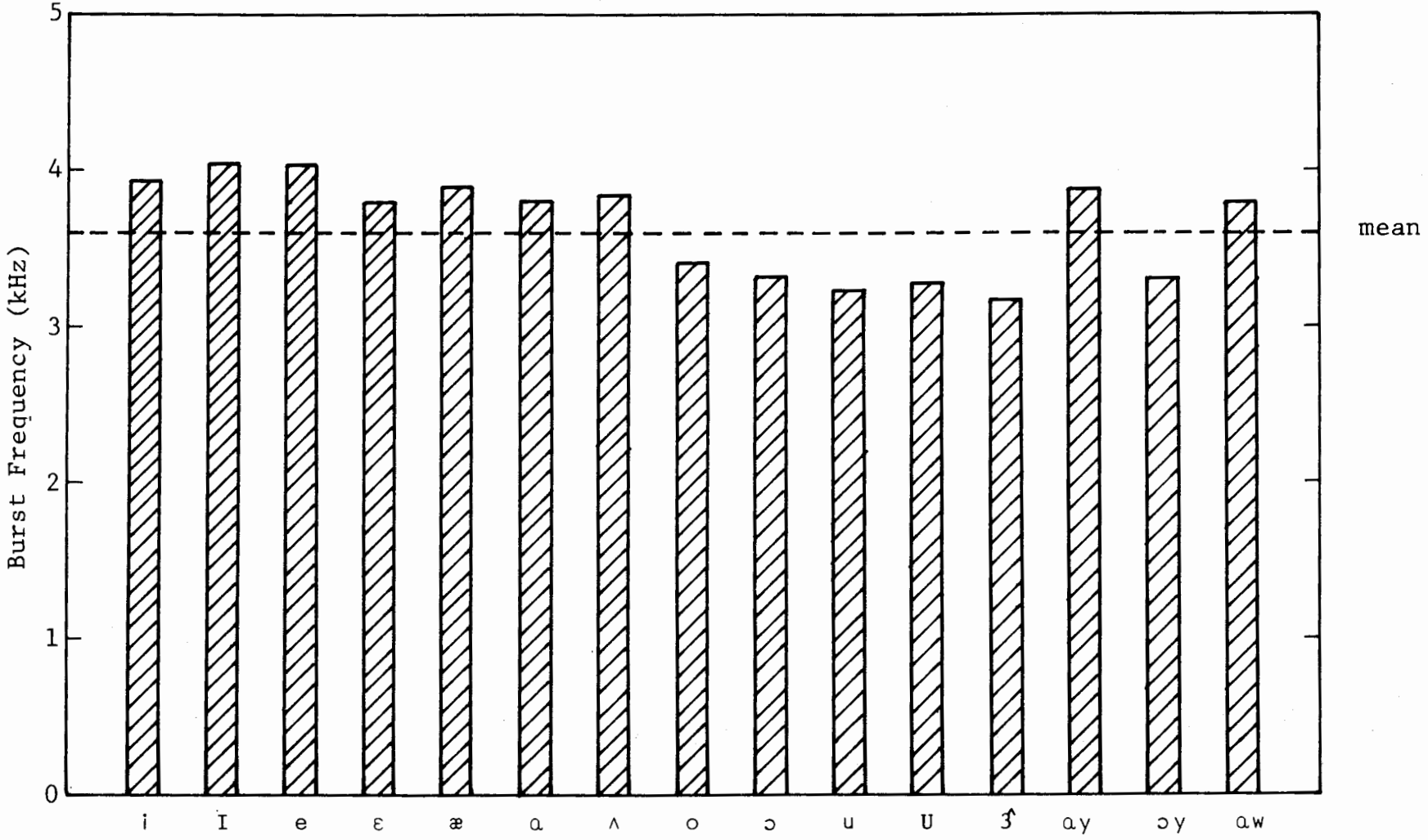


Figure 5.6 Average burst frequency for singleton /t/ as a function of vowel context

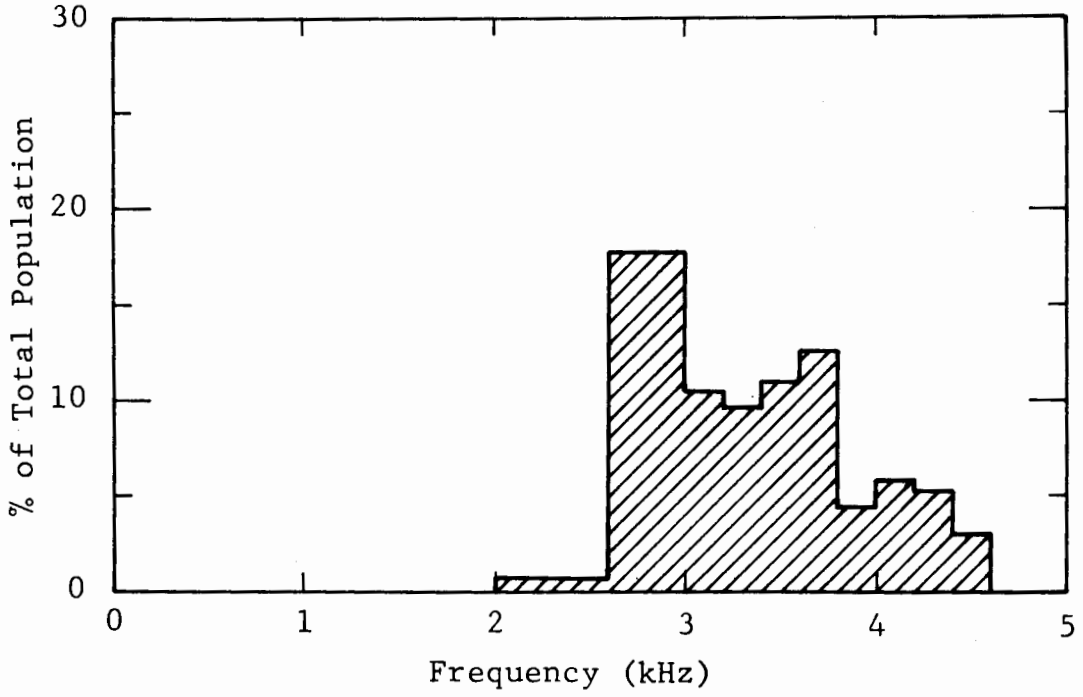


Figure 5.7 Distribution of the burst frequency for /d/

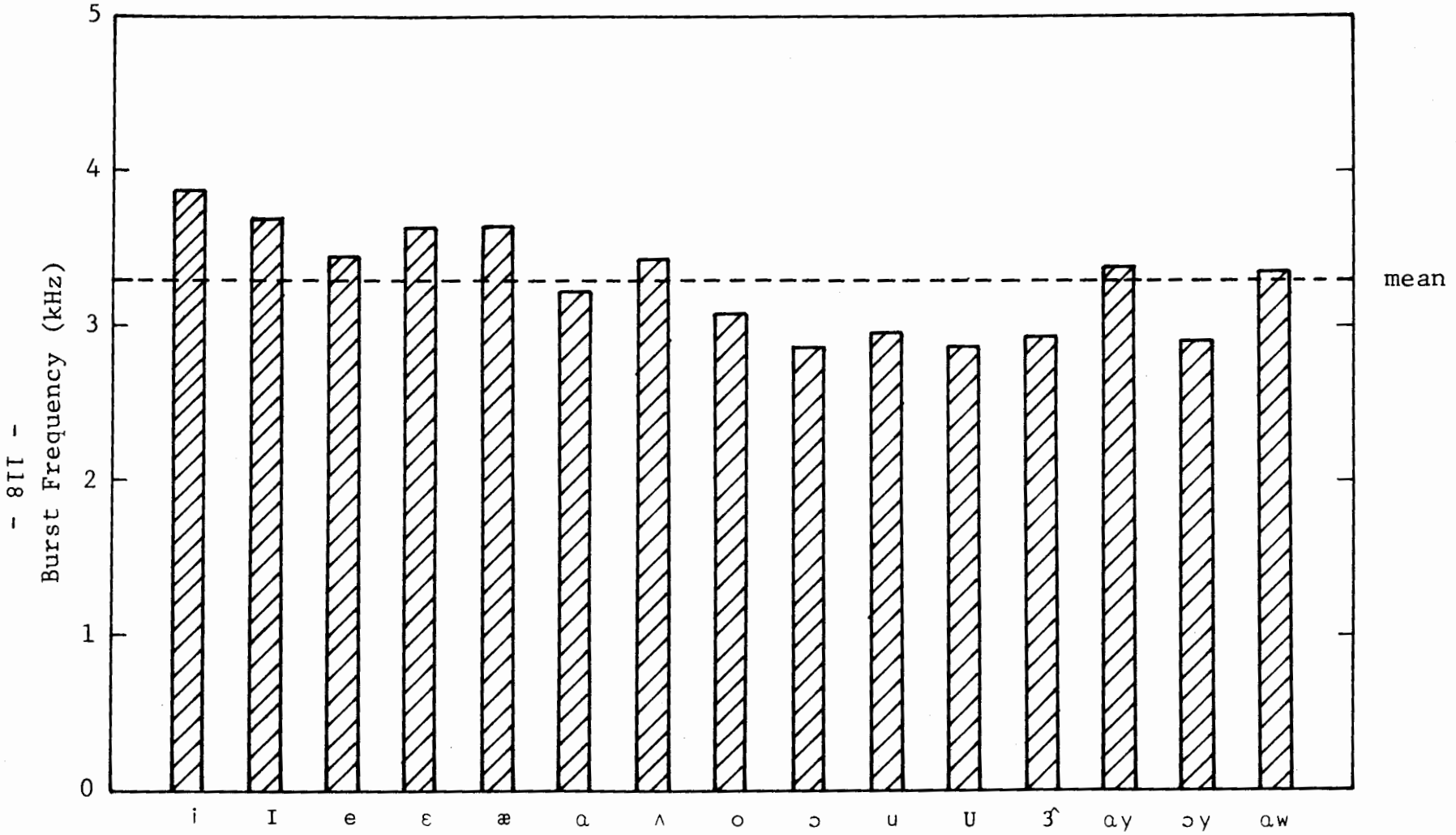


Figure 5.8 Average burst frequency for singleton /d/ as a function of vowel context

The distribution of burst frequency for /k/ is shown in Figure 5.9. The burst frequencies for /k/ tend to be distributed in the low- to mid-frequency region, with less than 5% of the values above 3,000 Hz. Averaging across all vowels, the mean burst frequency was found to be 1,910 Hz. However, there appear to be three distinct peaks in the distribution. Closer examination of data reveals that these peaks are again attributable to the underlying features of the following vowel. The mean burst frequency for /k/ preceding front vowels was found to be 2,720 Hz. Preceding back and unrounded vowels, the mean burst frequency was found to be 1,770 Hz. The mean burst frequency for /k/ preceding back and rounded vowels was found to be 1,250 Hz. These three values correspond well with the locations of the three peaks in the distribution shown in Figure 5.9.

The burst frequency for /k/ is plotted in Figure 5.10 as a function of the following vowel. Those values preceding front vowels are always greater than the mean. The frequencies of bursts preceding back vowels are consistently less than the mean value, with those preceding the rounded vowels having the smallest values.

The burst frequency distribution for /g/, as shown in Figure 5.11, is very similar to that of /k/. Averaged over all vowel context, the mean value is 1,940 Hz. The mean

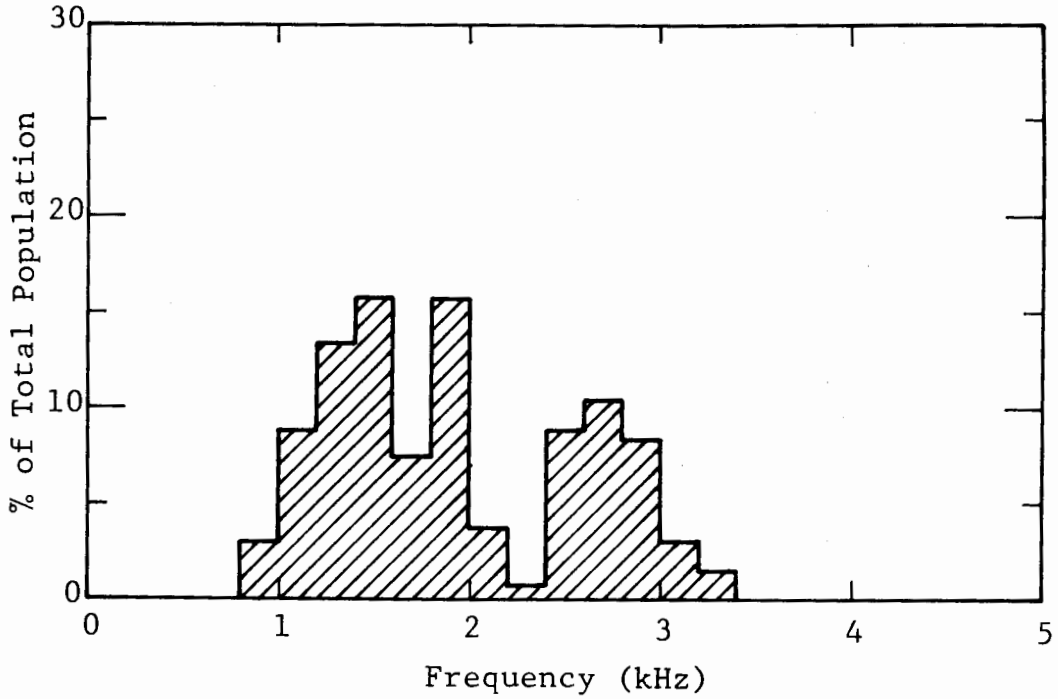


Figure 5.9 Distribution of the burst frequency for /k/

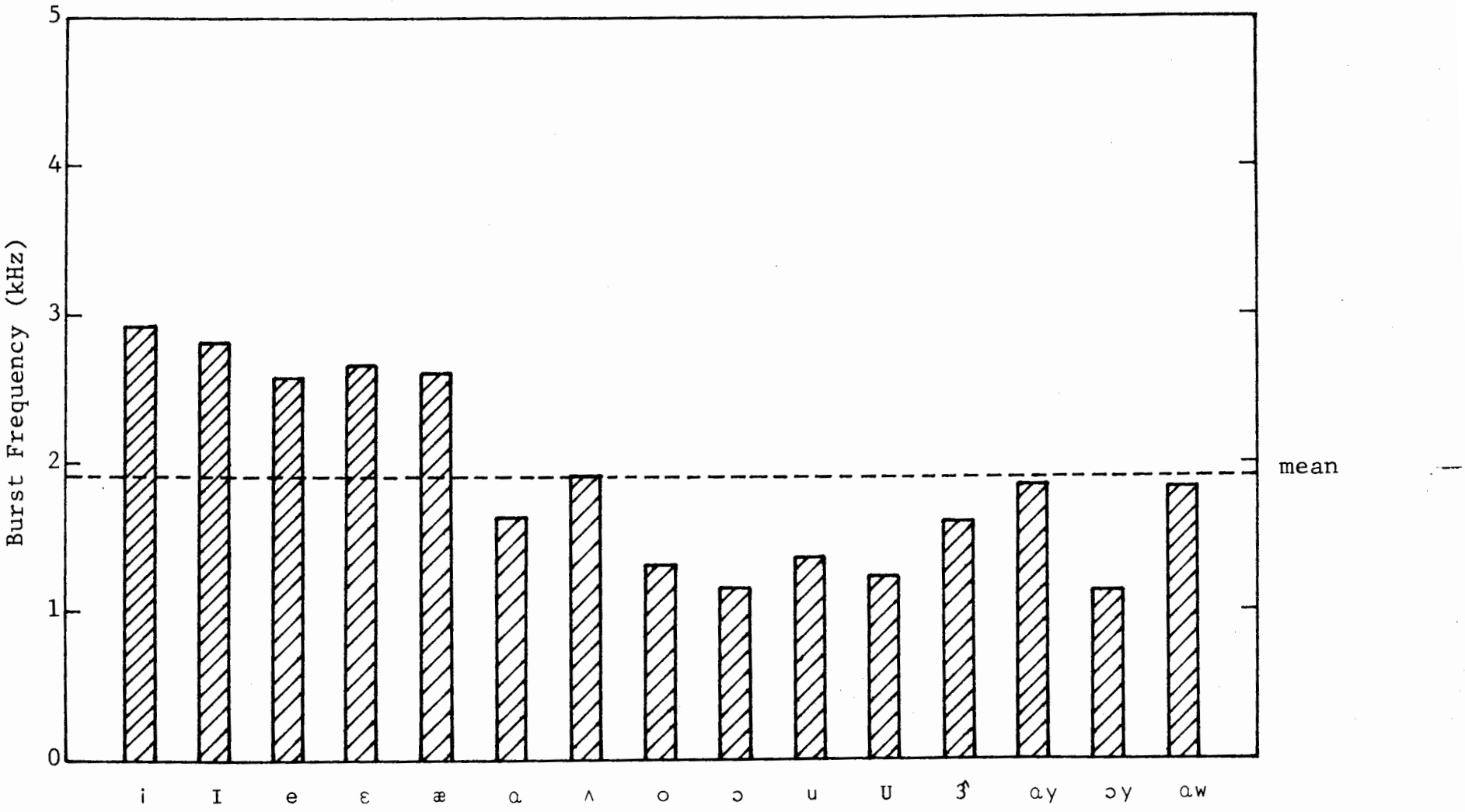


Figure 5.10 Average burst frequency for singleton /k/ as a function of vowel context

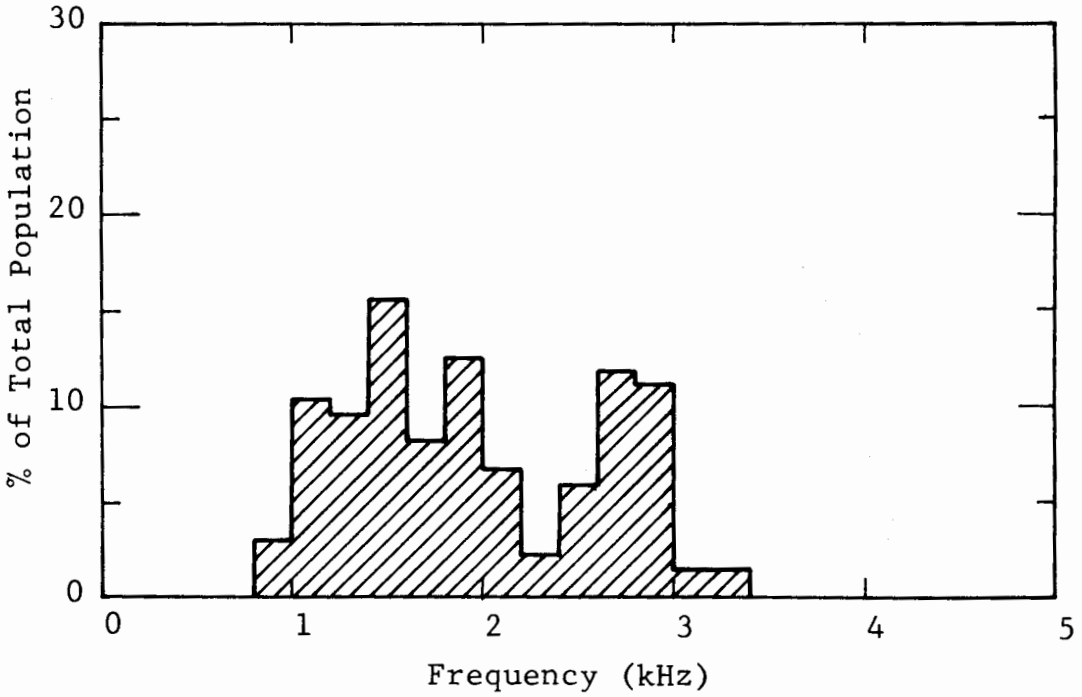


Figure 5.11 Distribution of the burst frequency for /g/

values of burst frequency for /g/ preceding front vowels, preceding back and unrounded vowels, and preceding back and rounded vowels are 2,720 Hz, 1,770 Hz, and 1,250 Hz, respectively. The average burst frequencies for /g/ are plotted in Figure 5.12 as a function of the vowel context. Comparison of Figures 5.10 and 5.12, shows that the vowel dependency of /g/ burst is the same as that of /k/ burst.

The average burst amplitudes for /t,k,d,g/ are summarized in Table 5.II. From Table 5.II, we see that the average burst amplitude of the voiceless stops is consistently greater, by about 2 dB, than that of the voiced stops. The burst amplitude is greater for dentals than for velars. The averaged burst amplitude for the dentals and the velars is about the same as that of the following vowel.

5.3.2 Stops in Clusters

The average RMS amplitude of the burst for stops in clusters is compared with that for singleton stops in Table 5.III.

The values in Table 5.III represent the difference between the RMS amplitude of the singleton stops and that of stops in clusters. For the stop-sonorant clusters, there is a marked decrease, by some 8 dB, in RMs amplitude for the velars whereas there is very little change in RMs amplitude

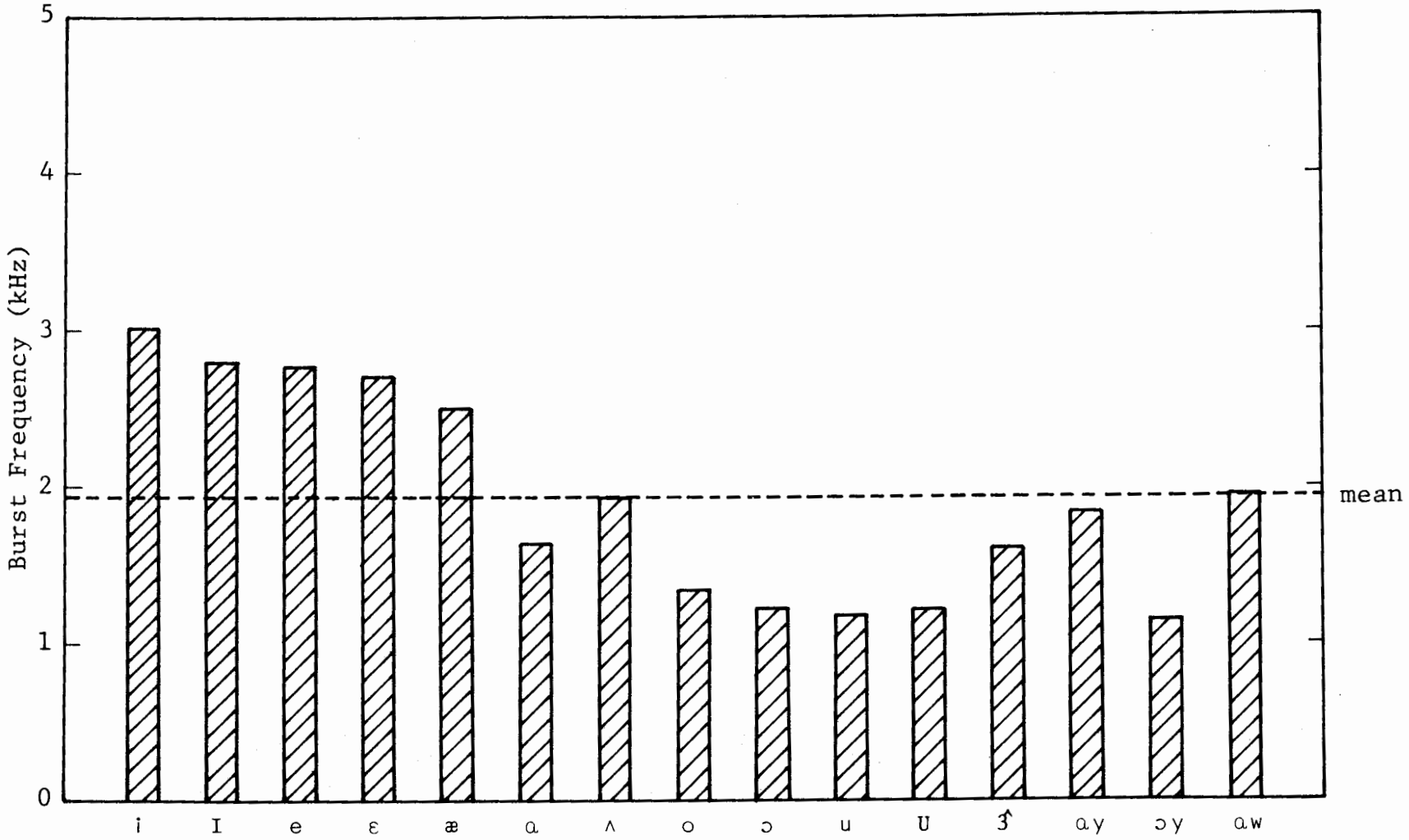


Figure 5.12 Average burst frequency for singleton /g/ as a function of vowel context

STOP	AVERAGE BURST AMPLITUDE (in dB)
/t/	2.4
/k/	-1.8
/d/	-0.4
/g/	-3.4

Table 5.II Average relative amplitude of the burst for the singleton stops

STOP	STOP-SONORANT CLUSTERS	/s/-STOP CLUSTERS
/p/	-0.2	-2.0
/t/	2.3	5.0
/k/	8.2	3.0
/b/	-0.6	
/d/	4.0	
/g/	8.7	

Table 5.III Decrease in overall RMS amplitude from singleton to clusters

for the labials. The average decrease in RMS amplitude is about 3 dB for the dentals. For the s-stop clusters, both dentals and velars show a decrease in RMS amplitude, with dentals having a larger change. The labials, however, show an increase in RMS value, by 2 dB.

Figures 5.13-14 plot the average burst frequencies for /t/ in /tr/ cluster and /k/ in /kw/ cluster, respectively, as a function of the vowel context. Also plotted in these figures are the corresponding values for the singleton stops. Figures 5.13-14 serve to illustrate the fact that, while there is a marked change in burst frequency from singleton stops to stops in clusters, this change is more related to the nature of the sonorant involved rather than the vowel context. Since this phenomenon was observed consistently for all the clusters, we shall pool the results across vowel context for the stop-sonorant clusters.

The averaged burst frequencies for stop-sonorant clusters are shown in Figure 5.15, along with values for singleton stops. The average burst frequency for /t/ in /tr/ clusters was found to be 2,460 Hz, a decrease of more than 1000 Hz. The burst frequency for /t/ in /tw/ cluster is 2,570 Hz, slightly greater than the /t/ in /tr/ cluster. As is the case for singleton stops the burst frequencies for /d/ in clusters were found to be some 200 Hz less than their

BIOGRAPHICAL NOTE

Victor Waito Zue was born in Szechuan, China, on February 19, 1946. After graduating from Tak Ming High School in Hong Kong, he came to the United States in 1965 to attend college. He received his B.S. degree in Electrical Engineering with High Honors from the University of Florida in 1968, and received his M.S.E. degree from the same university in 1969. He first attended the Massachusetts Institute of Technology in the spring of 1970, and is expected to receive his Doctor of Science degree in May, 1976.

From 1969 to 1970, he engaged in full time research at the University of Florida, making chronic brain tissue impedance measurements. He was a summer staff at the MIT Lincoln Laboratory in 1972 and 1973, working in the area of speech compression and speech understanding. Since 1973, he has been, in various capacities, actively involved in the speech understanding projects at the MIT Lincoln Laboratory and Bolt, Beranek and Newman, Inc. His research interests are in the area of acoustic phonetics, phonology, and the digital analysis of speech.

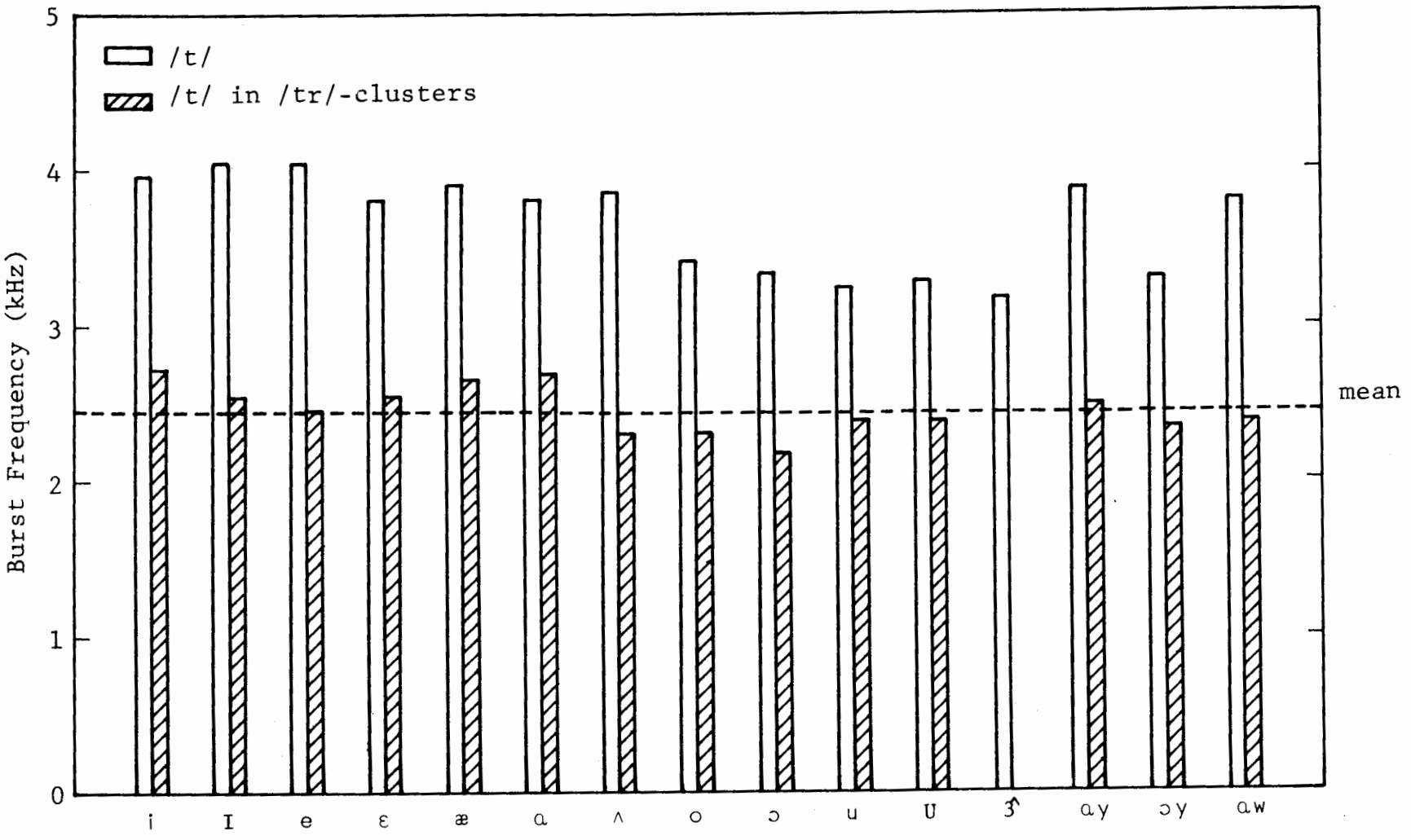


Figure 5.13 Average burst frequency for /t/ in /tr/-clusters as a function of vowel context

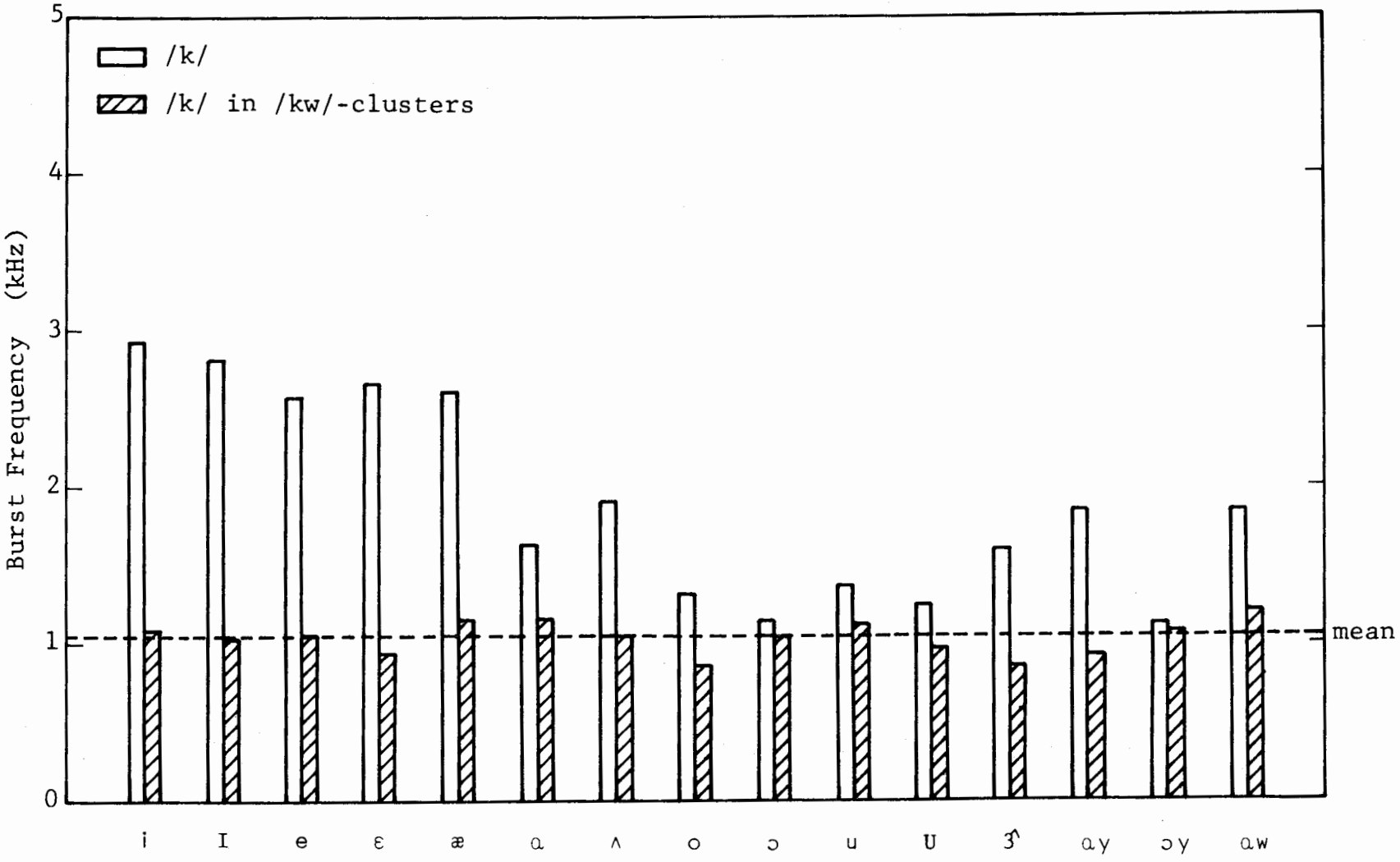


Figure 5.14 Average burst frequency for /k/ in /kw/-clusters as a function of vowel context

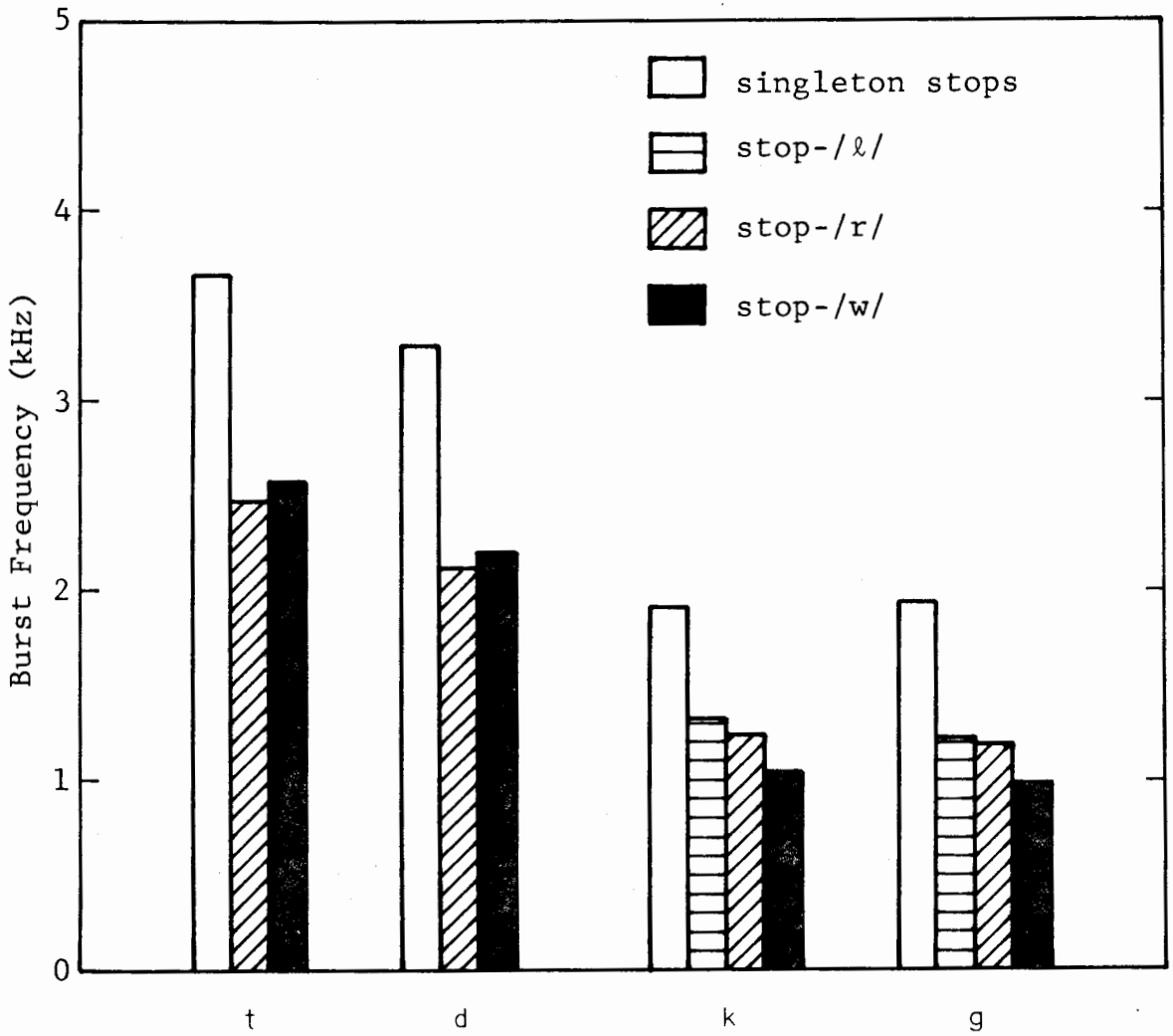


Figure 5.15 Average burst frequency for /t,d,k,g/ in stop-sonorant clusters

voiceless counterparts; 2,120 Hz for /d/ in /dr/ cluster and 2,210 Hz for /d/ in /dw/ clusters, respectively.

The burst frequencies for the velars in stop-sonorant clusters show no difference across the voicing distinction. Preceding /l/, the burst frequency was found to be 1,320 Hz for /k/ and 1,230 Hz for /g/, respectively. Preceding /r/, the values are 1,240 Hz for /k/ and 1,190 Hz for /g/, respectively. The burst frequencies were found to be the lowest when preceding /w/, 1,050 Hz for /k/ and 980 Hz for /g/, respectively.

Averaged across all vowels, the mean burst freuency for /t/ in /st/ cluster was found to be 3,240 Hz. This value is closer to the mean value of singleton /d/ than that of /t/. As shown in Figure 5.16, the burst frequencies for /t/ in /st/ cluster vary with vowel context in much the same way as singleton /d/.

The burst frequency for /k/ in /sk/ cluster is plotted in Figure 5.17 as a function of the following vowel, along with that of singleton /g/. Averaged across all vowels, the mean burst frequency was found to be 2,010 Hz, about the same as singleton /g/.

Table 5.IV compares the averaged burst amplitude for stops in clusters with that of singleton stops. For the

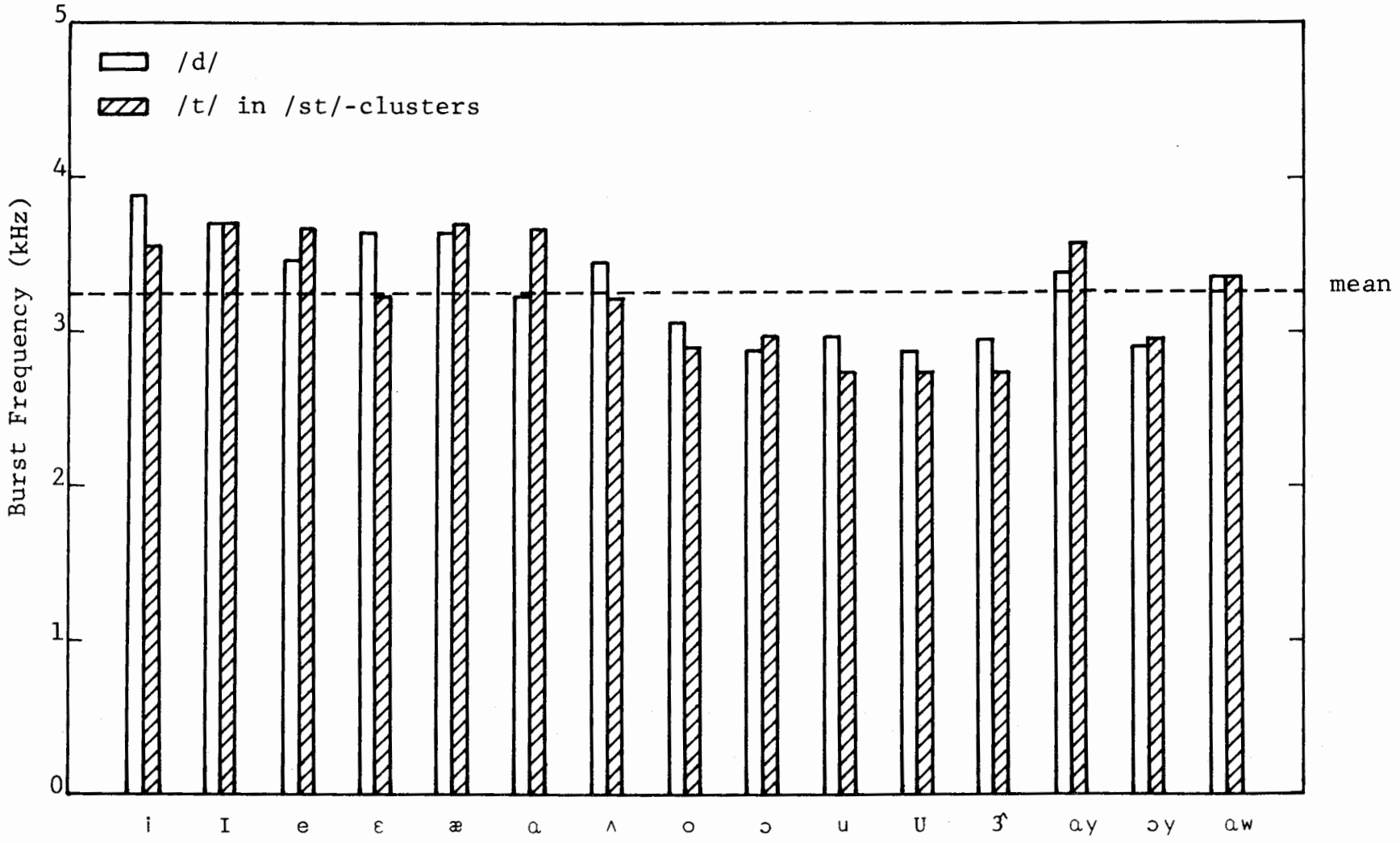


Figure 5.16 Average burst frequency for /t/ in /st/-clusters as a function of vowel context

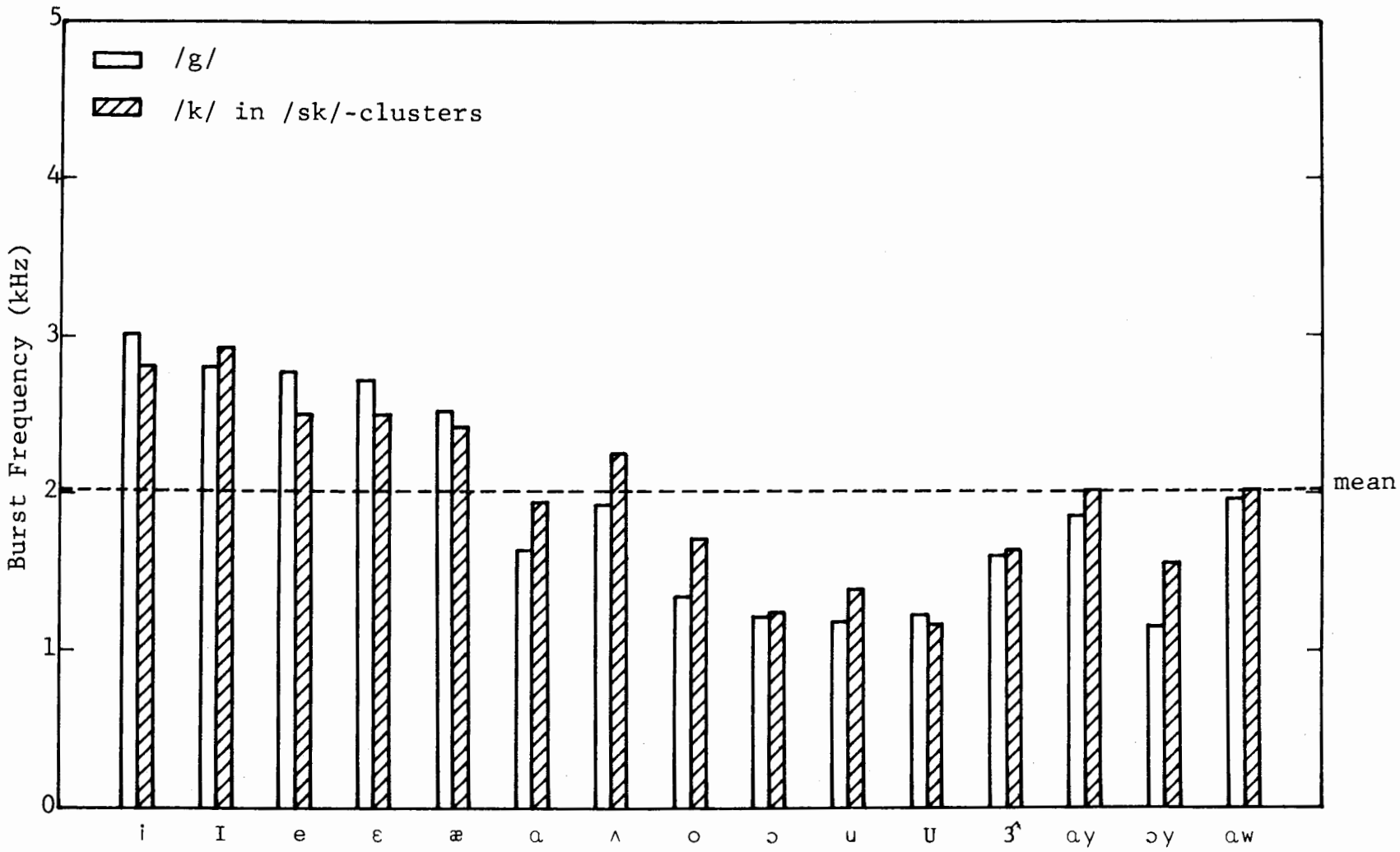


Figure 5.17 Average burst Frequency for /k/ in /sk/-clusters as a function of vowel context

STOP	STOP-SONORANT CLUSTERS	/s/-STOP CLUSTERS
/t/	3.5	-2.0
/k/	-4.7	-3.0
/d/	1.5	
/g/	-3.7	

Table 5. IV Increase in relative burst amplitude for stops from singleton to clusters

stop-sonorant clusters, there is a decrease of the average burst amplitude of about 4 dB for the velars, whereas the average burst amplitude for the dentals increases slightly. The averaged burst amplitudes for /t/ and /k/ in s-stop clusters decrease by 2 to 3 dB, resulting in values approximately the same as the singleton /d/ and /g/, respectively.

5.3 Discussion

Most of the results presented in the previous section concern only the dental and velar stops. We have excluded results on the labials because their releases are, in general, very weak. The average RMS amplitudes for the labials were some 12 dB less than the dentals and the velars. This weak release makes it extremely difficult to locate the burst frequency. The fact that the labials lack a clear, distinct burst frequency raises the question of whether the burst frequency is a perceptually important cue in the identification of the labials. Although the presence of a burst has been demonstrated to contribute to the perception of /p,b/, it may be postulated that the spectral concentration of the burst perhaps is not as important as the fact that its relative amplitude is much weaker than that of a stop with another place of articulation.

The burst frequency for the dentals was found to have a distribution that is bi-modal in nature. Bursts for /t,d/ preceding rounded vowels were found to be 600 Hz lower than those preceding all other vowels. This lowering of burst frequencies is presumably the consequence of anticipatory rounding of the lips during the stop release. Rounding reduces the opening of the vocal tract, and the protruding lips also increase the length of the front cavity. Both of these factors have the effect of lowering the resonant frequency of the cavity. It is also possible that the measured burst frequencies for the rounded and the unrounded dentals actually correspond to the natural frequencies of different cavities entirely. With no rounding, the burst frequency may be a consequence of the constriction itself, rather than the front cavity.

The distribution of burst frequency for the velars has three distinct peaks. The average burst frequencies for /k,g/ preceding front and back vowels differ by 1,200 Hz. For the back vowels, the bursts for /k,g/ are 550 Hz lower preceding rounded vowels than preceding unrounded vowels. The front and back velar distinction is quite well known in phonetics literature [for example Heffner 1950]. X-ray studies of articulatory movements [for example, Perkell 1969] had also shown that the velars preceding front and back vowels have different places of articulation. As shown

in Figure 5.4, this articulatory difference is transformed into acoustic characteristics that are quite dissimilar for the front and back /k,g/'s. Although the places of articulation might change from front to back along a continuum, the measured burst frequencies clearly show only two or three discrete values. Our measurement of burst frequency, therefore, lends further support to the theory of the quantal nature of speech production [Stevens 1972].

Due to the differences in the sizes and shapes of the vocal tract, the burst frequency, in general, does vary from one speaker to another. However, we have found that the distribution for a particular stop maintains a certain characteristic independent of speaker variation. The effect of pooling the results across speakers and sessions only tends to be a broadening of the skirts of the distribution. Figure 5.18 shows the distribution of /t/ for a single speaker. Comparing Figures 5.5 and 5.18, one can clearly observe the same characteristics as discussed earlier. Speaker KNS, incidentally, has a tendency to centralize the vowel /ʌ/ and diphthongs /ɑy,ɑw/ due to his dialect background [McDavid 1967]. This phenomenon accounts for the higher values of burst frequency or the velars preceding /ʌ,ɑy,ɑw/.

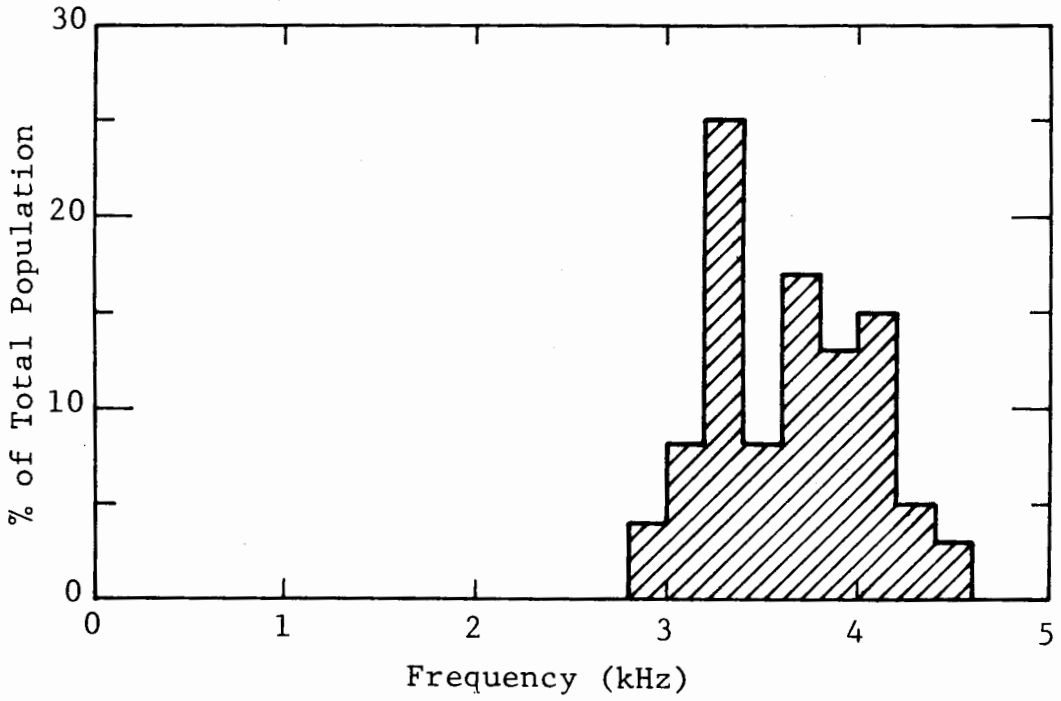


Figure 5.18 Distribution of the burst frequency for /t/ for a single speaker KNS (75 samples)

We have found essentially no difference in burst frequency values for the voiceless and voiced velars. There is, however, a consistent tendency for the voiced dental to have a burst frequency that is 200 to 300 Hz less than its voiceless counterpart under the same phonetic environment. In fact, the same trend can also be observed in clusters. In addition, burst frequency for the voiceless unaspirated /t/ was found to have values that correspond closer to /d/ than to the aspirated /t/. It is possible that /t/ and /d/ articulations actually involve different tongue positions, thus resulting in the burst frequency shift. It is also possible that the shift is due to the different positions of the larynx for the voiced and voiceless stops. Changing the length of the back cavity, assuming finite coupling, can shift the burst frequency. Another possible explanation lies in the differences in the source spectra between the voiced and voiceless stops. The frequency location of the peak in the source spectrum can shift as a result of the differences in volume flow between the voiced and voiceless stops [Stevens, personal communication]. The voiceless stops, with a higher volume flow across the constriction, will have a source spectrum that peaks at higher frequency.

Our results indicate that the average burst amplitude for /t,k,d,g/ is approximately the same as that of the following vowel in the same frequency region. In a recently

reported perceptual experiment, Bush, Stevens, and Blumstein [Bush et al. 1976] also obtained a similar finding. The burst amplitude was varied relative to the spectral peak of the following vowel in the same frequency region, and the subjects were asked to judge the quality of the stops. Bush et al. reported that the subjects consistently judged the best /d,g/ as having a burst amplitude comparable to, or even less than, that of the following vowel. These results contradict the common notion that burst amplitude should be greater. It is possible that the auditory system in doing the frequency analysis utilizes a different window that might be more sensitive to onset or transient phenomena [Stevens, person communication], thus resulting in a greater burst amplitude. It is also possible that the notion of greater burst amplitude was concluded from observations made on spectrograms, where the AGC circuitry tends to boost the burst intensity.

Results in Table 5.II also show that the burst amplitude is on the order of 2 dB higher for the voiceless stops. Stevens [Stevens 1971] has shown that in the generation of turbulence noise, the radiated sound pressure of the noise is proportional to the three-halves power of the pressure drop across the constriction, all else being equal. The measured 2 dB difference in burst amplitude, according to the above relationship, will result ina

pressure drop ratio of approximately 1.35. This calculated ratio in pressure drop is in good agreement with the measured pressure drops reported by several investigators [Lubker and Parris 1970, Lisker 1970].

Based on the measurements on the burst durations of the voiced and voiceless stops, Klatt [Klatt 1975] has estimated that the difference in duration alone can make the burst of a voiceless stop be perceived as at least 4 dB louder. Our measured difference in burst amplitudes will further contribute to the difference in perceived loudness between voiced and voiceless stops.

Results on the RMS amplitude of stops in clusters indicate a marked decrease in the values for velars. This could be a direct consequence of the fact that /kl,kr,kw/ all have a secondary constriction in front of the noise source, thus reducing the pressure drop across the primary constriction and the resulting radiated sound pressure.

When stops appear in stop-sonorant clusters, the burst frequency is dependent on the nature of the following sonorant, with the following vowel having little or no influence. The lowering of burst frequency for stops in clusters is accountable by considering the coarticulatory effect of the following sonorant [Fant 1973]. Rounding and retroflexing cause about the same amount of lowering for

dental burst, whereas rounding has the predominant effect of lowering the burst frequency for velars. It is of interest to note that the average burst frequency for /t/ in /tr/-clusters was found to be approximately 2,500 Hz. This measured value is slightly less than the most preferred burst frequency for retroflexed /t/'s found in a perceptual experiment recently reported by Stevens and Blumstein [Stevens and Blumstein 1975]. This difference is possibly due to the fact that the position of the constriction for retroflexed /t/ is more anterior than that of /t/ in /tr/-clusters.

CHAPTER 6

CONCLUDING REMARKS

This thesis has two distinct and integral parts. The first part is directed towards the development of a general facility where controlled studies of the acoustic characteristics of selected consonants, consonant clusters, and vowels in a prescribed phonetic environment can be carried out. The type of data we have collected, and the process through which they were collected have been described in Chapter 2. Various aspects of the analysis system and the data-base facility have been described in Chapter 3.

The second half of the thesis utilizes the collected data and the developed facility to study the acoustic characteristics of the English stops. The temporal characteristics of these stops were presented in Chapter 4 and the spectral characteristics in Chapter 5.

We approached the thesis with the belief that in order to study the acoustic characteristics of speech sounds and provide a strengthened basis for linguistic and phonetic theories, one must examine a large body of data, under a

controlled enviroment, and with the help of an interactive computer facility. Our experience and the results presented in this thesis clearly substantiated our claim. Although the data-base facility is no longer available, due to the termination of service of the TX-2 computer, this thesis hopefully has so demonstrated the importance and necessity of such a facility that similar ones will be developed on other computer systems.

The results we found on the temporal characteristics of the English stops are, with minor exceptions, in good agreement with those recently reported by Klatt [Klatt 1975] on a substantially smaller data-base. However, the results on the spectral characteristics represent, to our knowledge, a first attempt to quantify the release, both in frequency and in amplitude. These data will hopefully be valuable for immediate applications such as speech synthesis and speech recognition. They will also serve to help us gain a better understanding of the production and perception of speech.

The question of the presence of acoustic invariance of phonetic features has constantly been raised over the past two decades and the answer has thus far eluded us. Even under a controlled environment, such as the one in this thesis, we found that a very complex set of relationships must operate and mediate among the various acoustic

realizations of a given feature. The feature voicing for example, manifests itself in a difference in fundamental frequency, voice-onset time, burst amplitude, and a multitude of others. All things being equal, each of these dimensions might be sufficient in making the voice-voiceless distinction. When the condition is less ideal, however, the interactions among features will have to be considered. For example, some decision of the place of articulation may have to be made before VOT can be used to distinguish voiced and voiceless stops.

In the course of this thesis research, we have been able to isolate and quantify certain acoustic attributes of the underlying features, such as voicing and place of articulation. Although the effect of context and the interaction with other features are often observed, we nevertheless found many aspects of the results to be context independent. While the exact nature of the interaction among features and consequently their acoustic correlates are not yet understood, we feel that a necessary step has been taken towards a better understanding of the problem.

REFERENCES

Atal B. S., Hanauer S. L. (1971)
"Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," JASA, Vol. 50, pp. 637-665, August

Atal B. S. (1971)
"Sound Transmission in the Vocal Tract with Applications to Speech Analysis and Synthesis," Proc. 7th Int. Congr. Acoust., Budapest, Hungary, August

Atal B. S., Schroeder M. R. (1974)
"Recent Advances in Predictive Coding -- Application to Speech Synthesis," Preprints of the Speech Communication Seminar, Stockholm, Sweden, Vol. 1, pp. 27-31, August

Cooper F. S., Delattre P. C., Liberman A. M., Borst J. M., Gerstman L. J. (1952)
"Some Experiments on the Perception of Synthetic Speech Sounds," JASA, Vol. 24, pp. 597-606

Delattre P. C., Liberman A. M., Cooper F. S. (1954)
"Acoustic Loci and Transitional Cues for Consonanst," JASA, Vol. 27, pp. 769-773

Faddeeva V. N. (1959)
"Computational Methods of Linear Algebra," Dover Publications

Fant G., Ishizaka K., Lindqvist J., Sundberg J. (1972)
"Subglottal Formants," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden, Quarterly Progress and Status Report, pp. 1-15, April

Fant G. (1973)
"Speech Sounds and Features," MIT Press

Fischer-Jorgensen E. (1954)
"Acoustic Analysis of Stop Consonants," Miscellanea Phonetica, Vol. II, pp. 42-59

Halle M., Hughes G. W., Radley J-P A. (1957)
"Acoustic Properties of Stop Consonants," JASA, Vol. 29, pp.107-116

Halle M., Stevens K. N. (1971)

"A Note on Laryngeal Features," Resaerch Laboratory of Electronics, Massachusetts Institute of Technology, Quarterly Progress Report #101, pp. 198-213, April

Heffner R-M. S. (1950)
"General Phonetics," Univ. of Wisconsin Press, Madison, Wis.

Houde R. A. (1967)
"A Study of Tongue Body Motion during Selected Speech Sounds," Doctoral dissertation, Univ. of Michigan

House A. S., Fairbanks G. (1953)
"The Influence of Consonant Environment upon the Secondary Acoustic Characyeristics of Vowels," JASA, Vol. 25, No. 1, pp.105-113, January

House A. S. (1961)
"On Vowel Duration in English," JASA, Vol. 33, pp. 1174-1178

Itakura F., Saito S. (1968)
"Analysis Synthesis Telephony Based on the Maximum Likelihood Method," Report of the 6th Int. Congr. Acoust., Tokyo, Japan, Vol. II, Paper C-5-5

Jakobson R., Fant G., Halle M. (1951)
"Preliminaries to Speech Analysis," MIT Press

Klatt D. H. (1975)
"Voice Onset Time, Frication, and Aspiration in Word-Initial Consonant Clusters," J. of Speech and Hearing Research, Vol. 18, #4, pp. 686-706, December

Levinson N. (1969)
Appendix B of "Extrapolation and Smoothing of Stationary Time Series," by N. Wiener, MIT Press

Lisker L., Abramson A. S. (1964)
"A Cross-Language Study of Voicing in Initial Stops: Acoustic Measurements", Word, Vol. 20, No. 3, pp. 384-422, Deccember

Lisker L., Abramson A. S. (1967)
"Some Effect of Context on Voice Onset Time in English Stops," Lang. Speech, 10, pp. 1-28

Lisker L. (1970)
"Supraglottal Air Pressure in the Production of English Stops,"Lang. Speech, Vol. 13, pp. 215-230

Lubker J. F., Parris P. J. (1970)

"Simultaneous Measurements of Intraoral Pressure, Force of Labial Contact and Labial Electromyorgraphic Activity during the Production of Stop Cognates /p/ and /b/," JASA, Vol. 47, pp. 625-633

McDavid R. I. Jr (1967)
"Some Social Differences in Pronunciation," in "Readings in Applied English Linguistics," ed. by H. B. Allen, Appleton-Century-Crofts N. Y., 2nd edition

Makhoul J. I., Wolf J. J. (1972)
"Linear Prediction and the Spectrum Analysis of Speech," BBN Report #2304

Markel J. D. (1971)
"Formant Trajectory Estimation from a Linear Least Squares Inverse Filter Formulation," SCRL Monograph #7

Oppenheim A. V., Schafer R. M. (1968)
"Homomorphic Analysis of Speech," IEEE Trans. Audio and Electroacoust., Vol. AU-16, No. 2, pp. 221-226, June

Perkell J. S. (1969)
"Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study," Research Monograph #53, MIT Press

Peterson G., Lehiste I. (1960)
"Duration of Syllable Nuclei in English," JASA, Vol. 32, pp. 693-703

Portnoff M. R., Zue V. W., Oppenheim A. V. (1972)
"Some Considerations in the Use of Linear Prediction for Speech Analysis," Research Laboratory of Electronics, Massachusetts Institute of Technology, Quarterly Progress Report #106, pp. 141-150, July

Stevens K. N., Blumstein S. E. (1975)
"Quantal Aspects of Consonant Production and Perception: A Study of Retroflex Stop Consonants," Journal of Phonetics, Vol. 3, pp. 215-233

Stevens K. N., House A. S. (1963)
"Perturbation of Vowel Articulations by Consonant Context: An Acoustic Study," J. of Speech and Hearing Research, #6, pp. 111-128

Stevens K. N., House A. S., Paul A. P. (1966)
"Acoustic Description of Syllable Nuclii: An Interpretation in Terms of a Dynamic Model of Articulation," JASA, Vol. 40,

No. 1, pp. 123-132, July

Stevens K. N., Klatt D. H. (1974)
"The Role of Formant Transitions in the Voice-Voiceless Distinction for Stops," JASA, Vol. 55, pp. 653-659

Stevens K. N., Klatt M. (1968)
"Study of Acoustic Properties of Speech Sounds," BBN Report #1669

Stevens K. N. (1969)
"Study of Acoustic Properties of Speech Sounds II, and Some Remarks on the Use of Acoustic Data in Schemes for Machine Recognition of Speech,", BBN Report #1871

Stevens K. N. (1971)
"Airflow and Turbulence Noise for Fricative and Stop Consonants: Static Considerations," JASA, Vol. 50, pp. 1180-1192

Stevens K. N. (1972)
"The Quantal Nature of Speech: Evidence from Articulatory-acoustic Data," in "Human Communications: A Unified View," ed. by E. E. David Jr. and P. B. Denes, McGraw-Hill N. Y.

Stevens K. N. (1973)
"Further Theoretical and Experimental Bases for the Quantal Places of Articulation," Research Laboratory of Electronics, Massachusetts Institute of Technology, Quarterly Progress Report #108, pp. 247-252

Stevens K. N. (1975)
"Modes of Conversion of Airflow to Sound, and Their Utilization in Speech," Paper Presented at the 8th. International Congress of Phonetic Sciences, Leeds, England, August

Stowe A. (1972)
"SPC on TX-2" Internal Memo., MIT Lincoln Laboratory

Zue V. W. (1972)
"Speech Analysis by Linear Prediction," Research Laboratory of Electronics, Massachusetts Institute of Technology, Quarterly Progress Report #105, pp. 133-142, April